LangTech 2008

Rome, 28-29 February 2008

Proceedings

Under the high patronage

of the

President of the Italian Republic

Langtech 2008 is supported by

Fondazione Ugo Bordoni

MINISTERO DELLE COMUNICAZIONI

Istituto di Linguistica Computazionale - CNR

CNIPA Centro Nazionale per l'Informatica nella Pubblica Amministrazione

CILTA fAD

elda

ELRA European Language Resources Association

Gold sponsor

Loquendo VOCAL TECHNOLOGY AND SERVICES

Index

ORAL PRESENTATIONS

Committees

Conference Chair
Giordano Bruno Guerri, *Fondazione Ugo Bordoni*

Co-Chair
Andrea Paoloni, *Fondazione Ugo Bordoni*
Nicoletta Calzolari, *ILC-CNR*

Organising Committee
Nicoletta Calzolari, *ILC-CNR*
Khalid Choukri, *ELRA-ELDA*
Paolo Coppo, *Loquendo*
Mauro Falcone, *Fondazione Ugo Bordoni*
Giordano Bruno Guerri, *Fondazione Ugo Bordoni*
Dorota Iskra, *Logica CMG*
Gianni Lazzari, *FBK-IRST*
Bente Maegaard, *Copenhagen University*
Joseph Mariani, *LIMSI-CNRS*
Andrea Melegari, *Expert System*
Makoto Nagao, *Kyoto University*
Gianni Orlandi, *AURIS*
Andrea Paoloni, *Fondazione Ugo Bordoni*
Roberto Pieraccini, *SpeechCycle*
Fausto Ramondelli, *Intersteno*
Floretta Rolleri, *CNIPA*
Pasquale Santoli, *RAI*
Hans Uszkoreit, *DFKI*
Carlo Viola, *CONSIP*

Scientific Committee

Aladdin Ariyaeeinia, *Hertfordshire University*
Paolo Baggia, *Loquendo*
Nicoletta Calzolari, *ILC-CNR*
Amedeo Cappelli, *CELCT*
Loredana Cerrato, *Acapela Group*
Piero Cosi, *ISCT-CNR*
Franco Cutugno, *Naples University*
Amedeo De Dominicis, *Viterbo University*
Renato De Mori, *Avignon University*
Mauro Draoli, *CNIPA*
Andrzej Drygajlo, *EPFL*
Mauro Falcone, *Fondazione Ugo Bordoni*
Carmen Garcia-Mateo, *Vigo University*
Marco Gori, *Siena University*
Steven Krauwer, *Utrecht University*
Leonardo Lesmo, *Torino University*
Claudia Manfredi, *Florence University*
Giuseppe Mastronardi, *Bari Polytechnic*
Jan Odijk, *Utrecht University*
Maurizio Omologo, *FBK-IRST*
Javier Ortega-Garcia, *UAM*
Andrea Paoloni, *Fondazione Ugo Bordoni*
Domenico Parisi, *CNR*
Maria Teresa Pazienza, *Rome University*
Giuseppe Riccardi, *Trento University*
Pierluigi Ridolfi, *CNIPA*
Luciano Romito, *Calabria University*
Fabio Tamburini, *Bologna University*
Salvatore Tucci, *Rome University*
Guido Vetere, *IBM Italia*

Local Committee

Cristina Delogu, *Fondazione Ugo Bordoni*
Mauro Falcone, *Fondazione Ugo Bordoni*
Annalisa Filardo, *Fondazione Ugo Bordoni*
Andrea Paoloni, *Fondazione Ugo Bordoni*
Consuelo Tuveri, *Fondazione Ugo Bordoni*
Stefania Vinci, *Fondazione Ugo Bordoni*
Paola Baroni, *ILC-CNR*

POSTER PRESENTATIONS

A description language at the accentual unit level for Romanian intonation

Doina Jitcă [1], Vasile Apopei[1], Magdalena Jitcă [2]

[1] Institute for Computer Science, Romanian Academy, Iasi Branch, Romania
[2] University "Alexandru Ioan Cuza", Iasi, Romania
jdoina@iit.tuiasi.ro, vapopei@iit.tuiasi.ro, magdalena.jitca@infoiasi.ro

Abstract

The paper presents a classification of accentual unit patterns (AU patterns) and a corresponding label set used for generating an AU label based description language of intonation. Our AU patterns classification performed over a Romanian speech corpus, is based on the consideration that each intonational phrase corresponds to a basic discourse unit (BDU) or a subunit of BDUs. Therefore, we assign to each AU category a function in the spoken discourse. The description of the F0 contour by AU labels is suited for a text-to-speech system to create a description language of intonation for building the output of the linguistic module. It structures the input text into intonational units including the F0 contour characterizations as attributes. The structured text will be used as input for the phonetic module that generates the F0 contour for the synthesizer.
Index Terms: F0 contour, AU patterns, discourse events, speech synthesis

1. Introduction

In paper [1] was presented a solution for an F0 contour generating module of a Romanian TtS. The module has used an XML input text file, manually generated, with the text structured by intonational information using ToBI annotation labels. The intonational description is loosely related to the F0 contour patterns and it generates the following difficulties: at the linguistic module, in automatically generating correlations between the prosodic information and the syntactic structures, and at the phonetic modules, in translating the prosodic description into pattern segments within the F0 scale.
The present paper proposes an intonational description language based on labels assigned to accentual units (stress units). In order to achieve this goal the accentual unit was considered to be the basic pattern unit (BPU) within F0 contour and we have grouped the AU patterns over a Romanian speech corpus taking into consideration their different functions in spoken discourse. A solution for the F0 contour synthesis by concatenating AU natural patterns is presented in [2]. The model defines three functional types of AUs in different acoustic and phonetic contexts. In our model we increase the number of functional types of AU and the final result is a language for describing the melodic contours.
The intonational hierarchy from figure 1 illustrates our perspective over the melodic contours. An intonational phrase/intermediate phrase (IP/ip) consists of several AUs, depending on the number of stressed words. The AU sequence is nonlinear, as AUs alternate with AU groups on the same level of hierarchy. There are objective reasons for intermediate grouping between AUs and IPs/ips levels [3]. We called these groups accentual unit groups (AUGs). On the lower level of this hierarchy, an AU sequence results by splitting the AUGs.

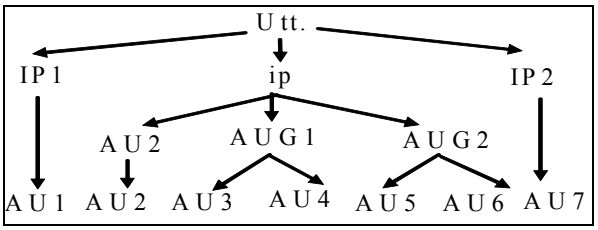


Figure 1: *The intonational hierarchy*

For classifying the BPUs we have used the F0 contour (AU pattern types) as a symbolic correlate between the spoken discourse events and the acoustic events [4]. As, referring to this idea we found that, generally, an IP should contain an AU that contributes to bring into attention the new basic discourse unit (BDU) and also, another AU to end it and eventually to prepare the audience for the next BDU. We called them AUs of PUSH/POP type. We found that the AUs having these functions usually manifest large F0 frequency variations during their accented and eventually, the following unaccented syllables. Within the first ones the F0 contour reaches highest tonal level (Top level) and within the last ones it reaches the lowest tonal level (Bottom level) of the respective IP.
There are other types of AUs that have a role in word focusing by highlighting certain tonal levels within an IP/ip. The highlighted tonal levels can be implied in generating the semantic focus by pitch accents widen around them or by contrasts with other tonal levels/targets from the adjacent AUs. We distinguished non-neutral tonal focus patterns, that manifest significant pitch accents, and weak tonal focus patterns generated by small pitch movements around a highlighted tonal level. The AU patterns having PUSH/POP function within an IP can manifest a tonal focus too.
The IP structure can contain AUs that have the F0 contour widen along the interpolate line between two tonal target/levels highlighted by AUs of previously mentioned types.
The AUG expresses either a semantic relation between the corresponding words, or just a rhythmical one. Generally, the AUG structure can be viewed as an IP structure at a small frequency scale that have a local top and bottom coordinate. Therefore, an AUG consists of two or three AUs equivalent to the PUSH/POP ones from IPs, or to the tonal focusing AUs.
We build a set of annotation labels consisting in mnemonics that suggest the function of AU in spoken discourse. Each label has a set of attributes, to indicate a particular F0 contour pattern for the corresponding AU. For example, the pitch accent type on accented syllables (as in ToBI annotation system) is explicitly descripted as an attribute of the AU label, only in case a certain pattern has no default pitch type defined for the respective label.
A Romanian speech corpus analysis has led us to define F0 contour patterns as prototypes for each category. The variability within the defined functional categories is

generated by the position of the accented syllable, the type of pitch accent, the prominence of the AU, the number of syllables etc. The labels give information about the position of the AUs in the F0 frequency scale, about types of pitch movements within it, about the amplitude of F0 frequency variations.
In conclusion, the perspective of this model gives a meaning to the melodic contour of an IP, close related to the F0 contour pattern for AUs. The description based on the corresponding labels will be more easily converted by a phonetic module of a TtS system into a sequence of patterns for building the F0 contour.

2. The label set for intonation description

The analysis of Romanian intonation over a speech corpus containing neutral text reading led us to identify some AU pattern types and to develop a set of corresponding labels for intonation annotation, divided into five categories: labels for AUs of PUSH/POP type at IP/ip level, labels of push/pop type at AUGs level, labels for tonal focusing, derived labels for AUs that have both preceding functionality and a category with AUs that link two tonal levels (tonal link labels). The labels are presented in the following paragraphs using a symbolic description of intonation consisting in AU labels separated by slash "/" and grouped by round parenthesis "()" into AUGs and by squared parenthesis into IPs/ips. In order to use them for an XML description, the labels have been transformed into values of the functional attribute of the AU tag that marks the text of an accentual unit.

2.1 The PUSH/POP labels

Introducing into attention a new BDU is performed usually by an AU that manifests large F0 variation during its accented syllable and/or the next unaccented one, up to the top level of the IP. We labeled these AU by „PH" label (PUSH). In the corresponding manner, the IP contains an AU that marks the end of the BDUs and we annotate it by „PO%" label (POP) and „PO" label in the IP and ip ending, respectively (the *ip* having "L-" type accent phrase).
In neutral case, an AU of "PH" type has a pitch accent of H* type and corresponds to the so-called "accent without focus" defined in [5] or to an initial accent (AI) defined in [6].
The AUs of PUSH type usually have an initial position within the IP (except yes-no question cases), but in case the first word/words must be focused at low tonal level, in neutral manner („f" labeled AU), a delay occurs to the PUSH event with the corresponding length of the focused word/word group.
The end of an IP is generated within an AU labeled with „PO%", in case an end point is present in the corresponding text. The „PO%" pattern is characterized by a decreasing F0 variation until the bottom level of IP is reached.
Another kind of AU pattern for BDU ending must be defined when both the end of the current BDU and the beginning of the next BDU are marked in spoken discourse. In this case, after the F0 contour reaches the lowest level during the last accented syllable, a rising F0 contour segment corresponding to a high boundary tone begins. We labeled this type of AU by „PU%" (POP-UP) and „PU" label in the IP and ip ending, respectively (the *ip* having "H-" type accent phrase).
Figure 3 illustrates the F0 contour of the utterance of the Romanian text *Avem de discutat lucruri serioase*. The "PH" label is assigned to the word *Avem* and the POP event corresponds to the word *serioase* labeled by a derived label "PO%+F" that suggest the occurrence of both POP event and focus event.

2.2 Tonal focusing labels

In an IP, between an AU of type "PH" and an AU of type "POP/POP-UP" there are one or several AUs within which certain tonal levels are highlighted. In figure 3, two types of tonal focus are present, corresponding to the words *lucruri* and *discutat.*

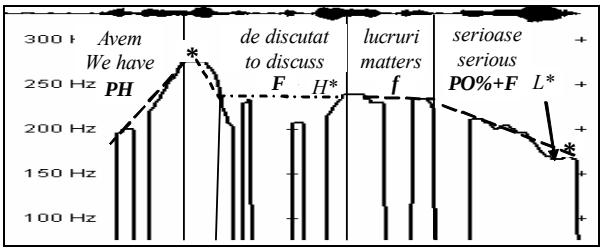


Figure 3: *The natural and the stylized F0 contour of the utterance of the text „Avem de discutat lucruri serioase" ("We have to discuss about serious matters")*

The first pattern, illustrated by the word *lucruri*, refers to the case of small variations around a tonal level, after having reached the focusing tonal level. We annotate this kind of AU pattern by „f" label, considering it a weak focus, usually occurring on the topic word.
The second pattern illustrated by the word *discutat* refers to the strong focus generated by a pitch accent of H* type. For this type of AU the tonal level on the last syllable equals the one from the beginning of the stressed word. Keeping the beginning and ending tones within a word on the same level, in the presence of a major pitch accent (H* or L*), one generates a strong focus on a certain word. We annotate these AU patterns by „F" label. In case the accented syllable is in the initial position of a focused word, the well-known shape of "peak" is generated within the corresponding AU and then the tone on the last unaccented syllables falls until the initial level is reached.
The AU of the last word *serioase* performs BDU ending and highlights the final target tone by a pitch accent of L* type generated by a slope during last accented syllable. Therefore, the AU was annotated by "PO%+F" derived label. The description of the melodic contour is the following:
PH / F / f / PO%+F
An AU having a F0 pattern of type "f", positioned at low level before or after an AU with high target tone, carries a semantic focus generated by the tonal contrast between the target tones of two adjacent AUs (figure 4). The semantic focus based on tonal contrast (corresponding to the metrical view of sentence stress defined in [5]) is frequently used in Romanian neutral rendering. We have no special label to annotate the semantic focus generated by the tonal contrast between the target tones of two adjacent AUs, because a new label can't add other information about the F0 pattern. The attribute for tonal levels ("-l") should be used with the labels of the implied AUs in order to change the size of tonal contrast.
Figure 4 illustrates the F0 contour of the utterance of the Romanian text *Era genul de om… (He was the type of man…).* The verb *era* is rendered at a low level with a "f" pattern and becomes focused by the following high target tone reached during the next AU of PUSH type.
The label sequence f / PH / PU% -l:m describing the F0 contour from figure 4 specifies a medium level for the boundary target tone of the "PU%" label, using the attributes "-l:", specifying that the last rising segment hasn't got a large amplitude.

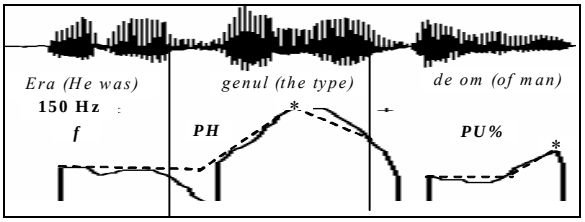


Figure 4*: The natural and the stylized F0 contour of the utterance of the text „Era genul de om…" ("He was the type of man…")*

The information about contrast focus is implicitly presented in the description. Before an AU of "PH" type, the "f" pattern corresponding to the first AU has the level attribute, "-l:", at the implicit low value.

2.3 Derived labels

The beginning of a BDU can be conveyed by an AU that reaches a high tonal level by stepping upward before its first syllable. During the AU the F0 frequency keeps this high value manifesting small variations in amplitude. We annotated such an AU pattern by „PH+f" label.
Figure 5 illustrates the F0 contour of the utterance of the Romanian text *Bine, atunci luaţi loc şi să stăm de vorbă mai comod (OK, then sit down and let's talk more comfortable).* The „PH+f" label corresponds to a beginning IP at a high focused tonal level.

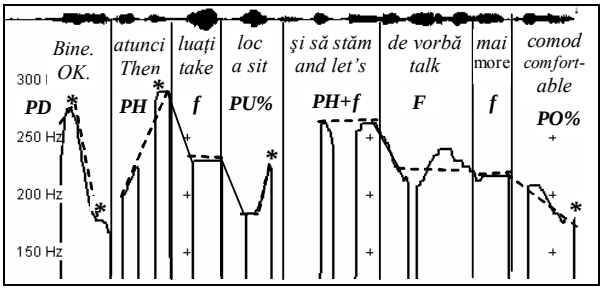


Figure 5: *The natural and the stylized F0 contour of the utterance of the text „Bine, atunci luaţi loc şi să stăm de vorbă mai comod" ("OK, then sit down and let's talk more comfortable")*

The utterance contains three IPs, the third one begins at high level and its first AU (corresponding to the verb *să stăm*) keeps this high level with small variations.
Sometimes, after a prominent H* pitch accent during a PUSH event, a tonal return to the value from the beginning of the word occurs in the end of the AU. Therefore, a focus is being generated and we label the corresponding AU with „PH+F". A focus can also be generated by a decreasing L* pitch accent after the high tonal level is reached and, in this case, we use the „PH+F" label and the attribute "–p:" (pitch) appears explicitly in the description with the "l" value.
In a similar manner, we can generate derived label for POP/ POP-UP events in case a prominent pitch accents occurs by a tonal decrease on the accented syllable, down to a minimum value. We label these Aus with "PO%+F"/ "PU%+F".
Another type of L* pitch accent within a POP-UP event is generated by just mantaining the low level during the accented syllable, reached before, and then by the rise of the F0 contour, in order to generate the boundary tones. This pitch accent characterization corresponds to the pattern of underived labels "PU%", illustrated in figure 5 by the AU corresponding to the word *loc*. The intonational description of the three IPs from figure 5 is the following:

[PD] [PH / f / PU%] [PH+f / F / f / PO%]

The label PD corresponds to a PUSH-DOWN event. We consider that in yes-no questions the PUSH event occurs in the end of the BDU in the case of nonfinal emphasis. The final rising on accented syllable by a H* pitch accent generates an PUSH event (oxitone final word) or a PUSH-DOWN event (nonoxitone final word). For yes-no questions an attribute for F0 range ("-r:") is used with large "l" value in order to characterize their final rising up to high levels.
An example of using "PD" labels in affirmative intonation is illustrated by the word *bine* in figure 5 where two target tones can be distinguished within the AU, both being significantly highlighted. The first one corresponds to a high target tone of the pitch accent and the second to a lower boundary tone. The F0 contour rises on the first syllable and falls down on the last syllable.
Another type of PUSH-DOWN event in affirmative intonations is generated within an AU during the initial unaccented syllables, followed by the fall of the F0 contour to the level where a focus occurs, starting from the accented syllable. We labeled this type of AU with "PD+f", in case of neutral focus and with "PD+F" in case of strong focus (L* pitch accent).

2.4 The push/pop labels

The annotation of AUGs must consist of a label sequence corresponding to its AU components and it may be assigned a label that characterizes its function within the IP/ip.
To annotate the AU component we introduce a set of mnemonics equivalent to those used at IP/ip level, based on their functional resemblance, as follows: „ph" labels for the first AU, "po/pu" labels for the last AU and F/f labels for the focused AUs. The AUs that have functions at IP/ip level too, keep the label at this level. The AU patterns of „ph" or „po" type contain a H* pitch accent and the patterns of „pu" type contain a L* pitch accent. In figure 6 the arrows mark the target tones of all pitch accents of H* type within the four AUGs. The F0 patterns at AUG level are also characterized by the difference between the AU target tone levels. In figure 6, it has an implicit positive value for the first three AUGs and an implicit zero value for the last one. The last AUG having the second target tone at a level equal to its first target tone, generates the *ip* ending with an accent phrase of "H-" type(the AU of "PU" type).

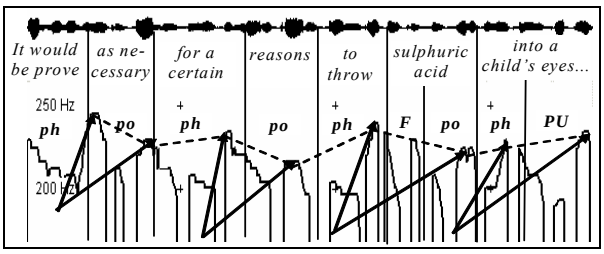


Figure 6: *The natural and the stylized F0 contour of the utterance of the text „...s-ar dovedi necesar dintr-un motiv oarecare să aruncaţi acid sulfuric în ochii unui copil..."*

The description of the F0 contour from figure 6 at AU level, corresponding to the middle and the end of an intermediate phrase, is the following (the PUSH AU is not presented in figure 6):

...(ph /po) /(ph / po) / (ph / F /po) / (ph / PU)

The labels for AUGs are the same as those used for ungrouped AUs at IP/ip level annotation, conveying they have equivalent functions. For example, an AUG in the beginning of the phrase described by (PH/po) is labeled by "PH", or an AUG in the ending of the phrase described by (ph/PO%) is labeled by PO% and a focused AUG (ph/po) can be described by the "F" label.
The following compact description results for the F0 contour by using only AUG labels:
...(F) / (F)/ (F) / (PU)
The compact description is useful for searching through a lexicon of stylized melodic contours at IP/ip level.

2.5 Tonal linking labels

The labels of type "L" are used for annotating the AUs that link two tonal levels. Their F0 contour patterns manifest a downsteping or upsteping trend. If a pitch accent occurs during the accented syllable of an AU of this type, then its label becomes "L+f". Using the attribute "t" (trend) and one of the values "u" or "d" helps to specify the upstep direction ("u" value) or downstep direction ("d" value). Figure 7 illustrates the F0 contour of the utterance of the Romanian text *Sunt destule scaune?(Are there enough chairs?*). The linebase has a decreasing tendency until the lowest tone is reached, before the final rising of the yes-no question.

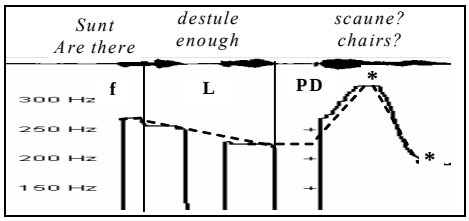


Figure 7: *The natural and the stylized F0 contour of the utterance of the text „Sunt destule scaune" ("Are there enough chairs?")*

The AU pattern of the word *destule* fits the implicit downstep trend beginning with the medium level at which the first AU of auxiliar verb "*sunt*" performs a weak tonal focus. The description of the melodic contour from figure 7 is the following:
f -l : m/ L / PD

3. The intonational melodic contour description using AU labels

In the perspective of this intonational model, the melodic contours of IPs/ ips can be stylized to a sequence of tonal floors and target tones. The target tones are reached during PH / PO/ ph / po events. Along each tonal floor the variations of an "f or F" AU pattern are spread. The "L" AU patterns link two consecutive tonal floors/targets. The size of the difference between adjacent tonal coordinates of the stylized contour can be used to control the semantic focuses prominence based on tonal contrast. The prominence of PUSH or POP/POP-UP event influences the whole F0 frequency range of the IP, respective top and bottom level. An IP has a number of tonal floors at intermediate levels. We view the IP as a temporal sequence of tonal floors and target points that generate a tonal skeleton to which the contained AU patterns are anchored. The F0 stylized contours from figures 3-7 illustrate the tonal skeletons of the corresponding natural curves.
Based on the speech corpus analysis we built two lexicons, one containing tonal skeleton prototypes described by different label sequences and another containing the AU patterns corresponding to each label in different lexical contexts. By using the two lexicons a phonetic module can be designed in order to generate the F0 contour in Romanian speech synthesis.

4. Conclusions

We consider that this description language for Romanian intonation can be understood both by NLP researchers and speech technology researchers that are interested in spoken language. The description using AU labels characterizes better what happens within F0 contour between pitch accents.
Our future task consists in building a Romanian speech corpus annotated at intonational level using the AU label set and at morfo-syntactic level, and then training a module to predict an intonational structure for an input text in Romanian text-to-speech systems. We think the syntactic units can be more easily assigned to a structured sequence of AU labels. Furthermore, the F0 contour generating module must be modified in order to use the pattern lexicons at AU level and IP/ip level.
The results will be useful in connecting a TtS system to the eDTLR (electronic Romanian dictionary), performing read aloud sense definitions of the words contained in the dictionary, in order to be used by the persons with sight disabilities.

5. Acknowledgements

This research was performed in the Romanian Academy and is partially supported by the grant PNCDI2 nr.910013/18.09.2007 entitled "eDTLR – Dicționarul Tezaur al Limbii Române în format electronic".

6. References

[1] Apopei V. and Jitcă D., "Module for generating the F0 Contour using as input a Text structured by prosodic information", Advances in Spoken Language Technology, Romanian Academy, 119-126, Bucharest, 2007.
[2] Heggtveit, P. O. and Natvig, J. E., "Intonation Modelling with a Lexicon Natural F0 Contours", Eurospeech, 2001.
[3] Sun-Ah Jun, "Intonational phonology of Seul Korean" Revisted, Japanese /Korean Linguistics Conference, Tucson, Arizona, nov.5-7, 2004
[4] Morton, K., Tatham, M. and Lewis, E., "A new Intonation model for text-to-speech Synthesis"
[5] Ladd, D. R. , "Intonational Phonology", Cambridge University Press, 1996
[6] Mertens, P., "Synthesizing elaborate Intonation Contour in Text-to-speech For French"

Blind Dereverberation Based on Spectral Subtraction by Multi-channel LMS Algorithm for Distant-talking Speech Recognition

L. Wang[1], *S. Nakagawa*[1], *N. Kitaoka*[2]

[1]Department of Information and Computer Sciences, Toyohashi University of Technology, Japan
[2]Department of Media Science, Nagoya University, Japan
{wang,nakagawa}@slp.ics.tut.ac.jp, kitaoka@nagoya-u.jp

Abstract

In this paper, we propose a blind dereverberation method based on spectral subtraction by Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Therefore, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional, and conventional Cepstral Mean Normalization (CMN) is not effective to compensate for the late reverberation under these conditions. By treating the late reverberation as additive noise, a noise reduction technique based on spectral subtraction is proposed to estimate power spectrum of the clean speech using power spectra of the distorted speech and the unknown impulse responses. To estimate the power spectra of the impulse responses, a Variable Step-Size Unconstrained MCLMS (VSS-UMCLMS) algorithm for identifying the impulse responses in a time domain is extended to the spectral domain. We conducted the experiments on distorted speech signal simulated by convolving multi-channel impulse responses with clean speech. An average relative recognition error reduction of 17.8% over conventional CMN under various severe reverberant conditions was achieved using only 0.6 second speech data to estimate the spectrum of the impulse response.

Index Terms: distant-talking speech recognition, blind dereverberation, Multi-channel LMS, spectral subtraction.

1. Introduction

Hands-free speech recognition has been more and more popular in some special environments such as an office or a cabin of a car. Unfortunately, in a distant-talking environment, channel distortion may drastically degrade speech recognition performance.

Compensating an input feature is the main way to reduce a mismatch between the pratical environment and the training environment. Cepstral Mean Normalization (CMN) has been used to reduce channel distortion as a simple and effective way of normalizing the feature space [1]. In order to be effective for CMN, the length of the channel impulse response needs to be shorter than the short-term spectral analysis window. However, the duration of the impulse response of reverberation usually has a much longer tail in a distant-talking environment. Therefore, the conventional CMN is not effective under these conditions. Several studies have focused on decreasing the above problem. Raut et al. [2] used preceding states as units of preceding speech segments, and by estimating their contributions to the current state using a maximum likelihood function, they adapted the models accordingly. However, model adaptation from *a priori* training data make it less practice to use. A reverberation compensation method for speaker recognition using spectral subtraction in which the late reverberation was treated as additive noise was proposed in [3]. However, the drawback of this approach is that the optimum parameters for spectrum subtraction are empirically estimated on a development dataset and the late reverberation cannot be subtracted well since the late reverberation is not modelled precisely. In [4, 5], a novel dereverberation method utilizing multi-step forward linear prediction were proposed. They estimated the linear prediction coefficients in a time domain and suppress amplitude of late reflections using spectral subtraction in a spectral domain.

In this paper, we propose a blind reverberation reduction method based on spectral subtraction by adaptive Multi-Channel Least Mean Square (MCLMS) algorithm for distant-talking speech recognition. Speech captured by distant-talking microphones is distorted by the reverberation. With long impulse response, the spectrum of the distorted speech is approximated by convolving the spectrum of clean speech with the spectrum of impulse response. We treat the late reverberation as additive noise, and a noise reduction technique based on spectral subtraction can be easily applied to compensate for the late reverberation. By excluding the phase information from the dereverberation operation as in [6, 5], the dereveration reduction on a power spectrum domain provided a robustness to certain errors that conventional sensitive inverse filtering method could not achieve. The compensation parameter (that is, the spectrum of the impulse response) for spectral subtraction is required. In [7, 8], an adaptive MCLMS algorithm was proposed to blindly identify the channel impulse response in a time domain. In this paper, we extend this method to blindly estimate the spectrum of impulse response for the spectral subtraction in a frequency domain.

2. Dereverberation Based on Spectral Subtraction

When speech $s[t]$ is corrupted by convolutional noise $h[t]$ and additive noise $n[t]$, the observed speech $x[t]$ becomes

$$x[t] = h[t] \otimes s[t] + n[t]. \quad (1)$$

In this paper, additive noise is ignored for simplification, so Eq. (1) becomes $x[t] = h[t] \otimes s[t]$.

To analyze the effect of impulse response, the impulse response $h[t]$ can be separated into two parts $h_{early}[t]$ and $h_{late}[t]$ as [3]

$$h_{early}[t] = \begin{cases} h[t] & \text{t < T} \\ 0 & \text{otherwise} \end{cases}, h_{late}[t] = \begin{cases} h[t+T] & \text{t} \geq 0 \\ 0 & \text{otherwise} \end{cases}, \quad (2)$$

where T is the length of the spectral analysis window, and $h[t] = h_{early}[t] + \delta(t-T) \otimes h_{late}[t]$. $\delta()$ is a dirac delta function (that is, a unit impulse function). The formula (1) can be rewritten as

$$x[t] = s[t] \otimes h_{early}[t] + s[t-T] \otimes h_{late}[t], \quad (3)$$

where the early effect is within a frame (analysis window), and the late effect is over multiple frames.

When the length of impulse response is much shorter than analysis window size T used for short-time Fourier transform (STFT), STFT of distorted speech equals STFT of clean speech multiply by STFT of impulse response $h[t]$ (in this case, $h[t] = h_{early}[t]$). However, when the length of impulse response is much longer than an analysis window size, STFT of distorted speech is usually approximated by

$$\begin{aligned} X(t,\omega) &\approx S(t,\omega) \otimes H(w) \\ &= S(t,\omega)H(0,\omega) + \sum_{d=1}^{D-1} S(t-d,\omega)H(d,\omega). \quad (4) \end{aligned}$$

where $H(d,\omega)$ denotes the part of $H(\omega)$ corresponding to frame delay d. That is to say, with long impulse response, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional [2].

In [3], the early term of Eq. (3) was compensated by the conventional CMN, whereas the late term of Eq. (3) was treated as additive noise, and a noise reduction technique based on spectrum subtraction was applied as

$$\hat{S}(t,\omega) = max(X(t,w) - \alpha \cdot g(\omega)X(t-T,\omega), \beta \cdot X(t,\omega)), \quad (5)$$

where α is the noise overestimation factor, and β is the spectral floor parameter to avoid negative or underflow values. However, the drawback of this approach is that the optimum parameters α, β, and $g(\omega)$ for the spectrum subtraction is empirically estimate on a development dataset and the STFT of late effect of impulse response as the second term of the right-hand side of Eq. (4) is not straightforward subtracted since the late reverberation is not modelled precisely.

In this paper, we propose a dereverberation method based on spectral subtraction to estimate the STFT of the clean speech $\hat{S}(t,\omega)$ based on Eq. (4), and the spectrum of the impulse response for the spectral subtraction is blindly estimated using the method described in Section 3. Assuming that phases of different frames is noncorrelated for simplification, the power spectrum of Eq. (4) can be approximated as

$$|X(t,\omega)|^2 \approx |S(t,\omega)|^2|H(0,\omega)|^2 + \sum_{d=1}^{D-1} |S(t-d,\omega)|^2|H(d,\omega)|^2. \quad (6)$$

The power spectrum of clean speech $|\hat{S}(t,\omega)|^2$ can be estimated as

$$|\hat{S}(t,\omega)|^2 = \frac{max(|S(t,\omega)|^2 - \alpha \cdot \sum_{d=1}^{D-1} |\hat{S}(t-d,\omega)|^2|H(d,\omega)|^2, \beta \cdot |X(t,\omega)|^2)}{|H(0,\omega)|^2}, \quad (7)$$

where $H(d,\omega), d = 0, 1...D-1$ is the STFT of impulse response which can be calculated from known impulse response or can be blindly estimated.

3. Compensation Parameter Estimation for Spectral Subtraction by Multi-channel LMS Algorithm

3.1. Adaptive Multi-channel LMS Algorithm for Blind Channel Identification in Time Domain

In [7, 8], an adaptive multi-channel LMS algorithm for blind Single-Input Multiple-Output (SIMO) system identification was proposed.

Before introducing the MCLMS algorithm for the blind channel identification, we express what SIMO systems are *blind identifiable*. According to [9], the following two assumptions are made to guarantee an identifiable system:

1. The polynomials formed from $h_n, n = 1, 2, ..., N$ where h_n is *n-th* impulse response and N is the channel number, are co-prime, i.e., the channel transfer functions $H_n(z)$ do not share any common zeros;
2. The autocorrelation matrix $\mathbf{R}_{ss} = E\{s(k)s^T(k)\}$ of input signal is of full rank (such that the single-input multiple-output (SIMO) system can be fully excited).

In the absence of additive noise, we can take advantage of the fact that

$$x_i * h_j = s * h_i * h_j = x_j * h_i, i, j = 1, 2, ..., N, i \neq j, \quad (8)$$

and have the following relation at time k:

$$\mathbf{x}_i^T(t)\mathbf{h}_j(t) = \mathbf{x}_j^T(t)\mathbf{h}_i(t), i, j = 1, 2, ..., N, i \neq j, \quad (9)$$

where $h_i(t)$ is *i-th* impulse response at time t and

$$\mathbf{x}_n(t) = [x_n(t) \ x_n(t-1) \ ... \ x_n(t-L+1)]^T, n = 1, 2, .., N, \quad (10)$$

where $x_n(t)$ is speech signal received from *n-th* channel at time t and L is the number of taps of the impulse response.

When the estimation of channel impulse responses deviates from the true value, an error vector is produced:

$$e_{ij}(t+1) = \mathbf{x}_i^T(t+1)\mathbf{h}_j(t) - \mathbf{x}_j^T(t)\mathbf{h}_i(t), i, j = 1, 2, ..., N, i \neq j. \quad (11)$$

This error can be used to define a cost function as

$$\mathbf{J}(t+1) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \epsilon_{ij}^2(t+1) \quad (12)$$

$$\epsilon_{ij}(t+1) = \frac{e_{ij}(t+1)}{||\mathbf{h}(t)||} \quad (13)$$

$$\mathbf{h}(t) = [\mathbf{h}_1^T(t) \ \mathbf{h}_2^T(t) \ ... \ \mathbf{h}_N^T(t)]^T \quad (14)$$

$$\mathbf{h}_n(t) = [h_n(t,0) \ h_n(t,1) \ ... \ h_n(t,L-1)]^T \quad (15)$$

where $h_n(t,l)$ is *l-th* tap of *n-th* impulse response at time t.

By minimizing the cost function $\mathbf{J}(t+1)$ of Eq. (12), impulse response is blindly derived. There are various methods to minimize the cost function $\mathbf{J}(t+1)$, for example, constrained MCLMS algorithm, constrained Multi-Channel Newton (MCN) algorithm and Variable Step-Size Unconstrained (VSS-UMCLMS) algorithm and so forth [7, 8]. Among these methods, the VSS-UMCLMS achieves a nice balance between complexity and convergence speed [8]. Moreover, the VSS-UMCLMS is more practical and much easier to use since the step size does not have to be specified in advance. Therefore, in this paper, we apply VSS-UMCLMS algorithm to identify the multi-channel impulse responses. The details of the VSS-UMCLMS were described in [8].

Table 1: Detail record conditions for impulse responses measurement. “angle”: recorded direction between microphone and loudspeaker. “RT60 (second)”: reverberation time in room. “S”: small, “L”: large.

array no	array type	room	angle	RT60
1	linear	tatami-floored room (S)	120^o	0.47
2	circle	tatami-floored room (S)	120^o	0.47
3	circle	tatami-floored room (L)	130^o	0.60
4	circle	tatami-floored room (L)	90^o	0.60
5	linear	Conference room	50^o	0.78
6	linear	echo room (panel)	70^o	1.30

3.2. Extending MCLMS Algorithm to Compensation Parameter Estimation for Spectral Subtraction

To blindly estimate the compensation parameter (that is, the spectrum of impulse response), we extend the MCLMS algorithm mentioned in Section 3.1 in a time domain to a frequency domain in this section.

The spectrum of distorted signal is a convolution operation of the spectrum of clean speech and that of impulse response as shown in Eq. (4). The spectrum of the impulse response is dependent on frequency ω, and the varibale ω is omitted for simplification. Thus, in the absence of additive noise, the spectra of distorted signals have the following relation at frame t on the frequency domain:

$$\mathbf{X}_i^T(t)\mathbf{H}_j = \mathbf{X}_j^T(t)\mathbf{H}_i, \quad i,j = 1,2,...,N, \quad i \neq j, \qquad (16)$$

Where $\mathbf{X}_n(t) = [X_n(t) \quad X_n(t-1) \quad ... \quad X_n(t-D+1)]^T$ is a D-dimention vector of spectra of the distorted speech recieved from *n-th* channel at frame t, $X_n(t)$ is the spectrum of the distorted speech recieved from *n-th* channel at frame t for frequency ω, $\mathbf{H}_n = [H_n(0) \quad H_n(1) \quad ... \quad H_n(d) \quad ... \quad H_n(D-1)]^T$, $d = 0, 1, ..., D-1$ is a D-dimension vector of spectra of the impulse response, and $H_n(d)$ is the spectrum of the impulse response for frequency ω.

Using Eq. (16) in place of Eq. (9), the spectra of the impulse responses can be blindly estimated by the VSS-UMCLMS mentioned in Section 3.1.

4. Experiments

4.1. Experimental setup

Multi-channel distorted speech signals simulated by convolving multi-channel impulse responses with clean speech were used to evaluate our proposed algorithm. Six kinds of multi-channel impulse responses measured in various acoustical reverberant environments were selected from the RWCP sound scene database [10]. Four-channel circle type or linear type microphone array was taken from a circle + linear type microphone array (30 channels). A four-channel circle type microphone array has a diameter of 30 cm, and 4 microphones are located at equal 90^o intervals. Four microphones of a linear microphone array are located at 11.32 cm intervals. Impulse responses were measured at several positions which were 2 m distance from the microphone array. The sampling frequency was 48 kHz. Table 1 shows the detail record conditions for six kinds of 4 channels microphone array.

For clean speech, twenty male speakers each with a close-microphone uttered 100 isolated words. The 100 isolated words are phonetic balance common isolated words selected from Tohoku University and Panasonic isolated spoken word database [11]. The average time of all utterances was about 0.6 second. The sampling frequency was 12 kHz. The impulse responses sampled at 48 kHz were downsampling to 12 kHz to convolve with clean speech. The frame length was 21.3 ms, and the frame shift was 8 ms with a 256 point Hamming window. Then, 116 Japanese speaker-independent syllable-based HMMs (strictly speaking, mora-unit HMMs [12]) were trained using 27992 utterances read by 175 male speakers (JNAS corpus). Each continuous-density HMM had 5 states, 4 with pdfs of output probability. Each pdf consisted of 4 Gaussians with full-covariance matrices. The feature space comprised 10 MFCCs. First- and second-order derivatives of the cepstra plus first and second derivatives of the power component were also included.

4.2. Experimental results and discussion

In this paper, only the speech signal from the first channel of each microphone array was performed for speech recognition. For our proposed method, at first speech signals from multiple microphones (2 microphones or 4 microphones) were used to blindly identify the compensation parameters for the spectral subtraction (that is, the spectra of the channel impulse responses), and then the spectrum of the first channel impulse response was used to compensate for the reverberation of the speech signal from the first channel.

The number of reverberant window D in Eq. (4) was set to 8. The length of the hamming window for DFT was 256 (=21.3 ms), and the overlaping rate was 1/2. No special parameters such as over-subtraction parameters were used for spectral subtraction ($\alpha = 1$), except that the subtracted value was controlled so that it did not become negative ($\beta = 0.15$). The speech recognition performance for clean isolated words was 96.0%.

Table 2 shows the experimental results for speech recognition. CMN performed on distorted speech was used as baseline. For given impulse response, LSE (Least Square Error) based inverse filtering [13] using single channel impulse response was used to recover the reverberant speech, which is an ideal condition. In practice, for LSE base inverse filtering, it could not appropriately deal with a non-minimum phase impulse response [13], whose case is often in real reverberant environments. Therefore, the speech recognition performance was not very accurate even using the known impulse response. There are many other more precise inverse filtering techniques such as [13, 14] and so forth. We will use the more precise inverse filtering techniques as ideal condition in the future. For our proposed method, the dereverberant speech of the first channel was obtained by using the proposed reverberation compensation technique based on the spectral subtraction, and then CMN was also performed on the dereverberant speech. The proposed method remarkably improved the speech recognition performance. By using 4 microphones to estimate the spectrum of the impulse response, the improvement was less than that of 2 microphones. The more parameters needed to estimate may result in degrade performance and we are still investigating the other reasons. By using 2 microphones to estimate the spectrum of the impulse response, different results were obtained by individual 2-channel microphone array, and the results with bold font were the best results. The different results obtained from different arrays may be complementary, thus we combined these results by a so-called *maximum-summation-likelihood method (MSLM)* proposed in [15]. The *MSLM* is to use the maximum summation likelihood of recognition results from different 2-channel

Table 2: Speech recogniton performacne for reverberant speech. Only the speech data of first channel was evaluated. For proposed method, the first channel speech was compensated by the impulse response of the first channel.

distorted speech #	No processing	CMN	proposed method (2 or 4 microphones were used to estimate the spectrum of impulse response)					inverse filtering
			4 microphones	(mic1, mic2)	(mic1, mic3)	(mic1, mic4)	*MSLM*	
1	46.0	64.2	69.9	67.5	**67.5**	66.5	69.5	77.4
2	48.4	64.2	63.4	65.9	66.1	**69.0**	68.4	71.8
3	53.3	62.8	66.8	69.2	**70.2**	64.1	70.1	77.3
4	48.7	65.1	69.5	**70.5**	70.1	67.9	70.4	76.2
5	43.5	56.2	63.4	**65.2**	62.1	62.5	66.3	70.9
6	40.3	54.7	62.5	62.5	**62.7**	61.0	63.9	72.4
Ave.	46.7	61.2	65.9	**66.8**	66.5	65.2	68.1	74.3

arrays to obtain the final result. It significantly improved the speech recognition performance for all severe reverberant conditions. An average relative recogniton error reduction of 17.8% over the conventional CMN was achieved.

5. Conclusions and Future Work

In this paper, we proposed a blind reverberation reduction method based on spectral subtraction by MCLMS algorithm for distant-talking speech recognition. In a distant-talking environment, the length of channel impulse response is longer than the short-term spectral analysis window. Therefore, the channel distortion is no more of multiplicative nature in a linear spectral domain, rather it is convolutional. We treated the late reverberation as additive noise, and a noise reduction technique based on spectrum subtraction was proposed to estimate the clean power spectrum. Power spectrum of impulse respone was necessary to estimate the clean power spectrum. To estimate the power spectra of the impulse responses, a VSS-UMCLMS algorithm for identifying the impulse responses in a time domain was extended to the spectral domain. Our proposed algorithm was evaluated by distorted speech signals simulated by convolving multi-channel impulse responses with clean speech taken from Tohoku University and Panasonic isolated spoken word database. The experimental results showed that an average relative recognition error reduction of 17.8% over the conventional CMN under various severe reverberant conditions was achieved using only a isolated word (about 0.6 second) to estimate the spectrum of the impulse response.

The proposed method relies on the assumption that there are no zeros common to all channels. However, it is known that room impulse responses have a large number of zeros close to the unit circle on the z-plane. If the channels present numerically overlapping zeros, the dereverberation performance would perform poorly. [5] indicated that spatial information can be used to deal with the problem of overlapping zeros. We try to use the spatial information to deal with the same problem of overlapping zeros of our method in the future.

6. Acknowledgements

This work was partially supported by The Global COE Program "Frontiers of Intelligent Human Sensing", from the ministry of Education, Culture, Sports, Science and Technology.

7. References

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acous. Speech Signal Processing, Vol. 29, No. 2, pp. 254–272, 1981.

[2] C. Raut, T. Nishimoto, S. Sagayama, "Adaptation for long convolutional distortion by maximum likelihood based state filtering approach," Proc. of ICASSP-2006, Vol. 1, pp. 1133–1136, 2006.

[3] Q. Jin, T. Schultz and A. Waibel, "Far-field speaker recognition," IEEE Trans. ASLP, Vol. 15, No. 7, pp. 2023–2032, 2007.

[4] M. Delcroix, T. Hikichi, M. Miyoshi, "On a blind speech dereverberation using multi-channel linear prediction," IEICE Trans. Fundamentls, E89-A(10), pp. 2837–2846, October 2006.

[5] M. Delcroix, T. Hikichi, M. Miyoshi, "Precise dereverberation using multi-channel linear prediction," IEEE Trans. ASLP, 15(2), pp. 430–440, February 2007.

[6] I. Tashev and D. Allred, "Reverberation reduction for improved speech recognition", Proc. Hands-Free Communication and Microphone Arrays, 2005.

[7] Y. Huang and J. Benesty, "Adaptive multichannel least mean square and Newton algorithms for blind channel identification," Signal Processing, Vol. 82, pp. 1127-1138, Aug. 2002.

[8] Y. Huang, J. Benesty and J. Chen, "Acoustic MIMO Signal Processing", Springer, 2006.

[9] M. Xu, L. Tong and T. Kailath, "A least-squares approach to blind channel identification", IEEE Trans. Signal Processing, Vol. 43, pp. 2982–2993, Dec. 1995.

[10] http://www.slt.atr.co.jp/ tnishi/DB/micarray/indexe.htm.

[11] S. Makino, K. Niyada, Y. Mafune and K. Kido, "Tohoku University and Panasonic isolated spoken word database," Journal of the Acoustical Society of Japan, Vol. 48, No. 12, pp. 899–905, Dec. 1992. (in Japanese)

[12] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based HMMs and triphone-based HMMs in Japanese speech recognition," Proc. International Workshop on Automatic Speech Recognition and Understanding, pp. 393–396, 1999.

[13] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.ASSP-36, No.2, pp. 145–152, 1988.

[14] T. Hikichi, M. Delcroix, M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," EURASIP J. APS, vol.2007, Article-ID 34013, April 2007.

[15] L. Wang, N. Kitaoka and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent CMN", EURASIP J. Appl. Signal Process., Vol. 2006, Article ID 95491, pp. 1–11., 2006.

Combining Statistical Parameteric Speech Synthesis and Unit-Selection for Automatic Voice Cloning

Matthew P. Aylett[1,2], *Junichi Yamagishi*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, U.K.
[2]Cereproc Ltd., U.K.

matthewa@cereproc.com jyamagis@inf.ed.ac.uk

Abstract

The ability to use the recorded audio of a subject's voice to produce an open-domain synthesis system has generated much interest both in academic research and in commercial speech technology. The ability to produce synthetic versions of a subjects voice has potential commercial applications, such as virtual celebrity actors, or potential clinical applications, such as offering a synthetic replacement voice in the case of a laryngectomy. Recent developments in HMM-based speech synthesis have shown it is possible to produce synthetic voices from quite small amounts of speech data. However, mimicking the depth and variation of a speaker's prosody as well as synthesising natural voice quality is still a challenging research problem. In contrast, unit-selection systems have shown it is possible to strongly retain the character of the voice but only with sufficient original source material. Often this runs into hours and may require significant manual checking and labelling.

In this paper we will present two state of the art systems, an HMM based system HTS-2007, developed by CSTR and Nagoya Institute Technology, and a commercial unit-selection system CereVoice, developed by Cereproc. Both systems have been used to mimic the voice of George W. Bush (43rd president of the United States) using freely available audio from the web. In addition we will present a hybrid system which combines both technologies. We demonstrate examples of synthetic voices created from 10, 40 and 210 minutes of randomly selected speech. We will then discuss the underlying problems associated with voice cloning using found audio, and the scalability of our solution.

Index Terms: speech synthesis, unit-selection, statistical parametric synthesis, voice cloning, HMM, speaker adaptation

1. Introduction

Vocal mimicry by computers is regarded with both awe and suspicion [1]. This is partly because perfect vocal mimicry is also the mimicry of our own sense of individuality: the use of a certain voice draws with it much more than the voice itself, it also draws the associations we have with that voice. Conveying this sense of character is becoming important in a whole set of innovative applications for human-computer interfaces which use speech for input and output.

For example, the ability to produce synthetic versions of a subjects voice has potential attractive commercial applications, such as virtual celebrity actors, or potential beneficial clinical applications, such as offering a synthetic replacement voice in the case of a laryngectomy. In addition, the ability to retain the character of a speaker could be combined with translation systems, where it would help personalize *speech-to-speech* translation so that a user's speech in one language can be used to produce corresponding speech in another language while continuing to sound like the user's voice. It might eliminate the need for subtitles and onerous voice-overs acting on international broadcasts or movies in the future.

In this paper we investigate the reproduction/mimicry aspects of up-to-date speech synthesis technologies: how well can we take a well-known speaker and duplicate his acoustic feature, linguistic features, and speaking styles so that a listener immediately recognises the speaker? Furthermore, how effective is this mimicry for conveying the character of the speaker in an amusing manner? We term the process of producing a speech synthesis system that can effectively mimic a speak "voice cloning". We apply two major competing technologies to this voice cloning problem, the first is a well-established and well-studied technique called "unit-selection", which concatenates segments of speakers' source speech to create new utterances [2], the second is often termed "statistical parametric synthesis," where a statistical acoustic model is trained or adapted from speakers' source speech [3]. In the experiments, we will apply both techniques to the problem of cloning the voice of **George W. Bush** (The 43rd President of the United States) and produce a short rendition of the introduction of a well known children's story, "The Emperor's New Clothes". In addition we will explore the use of a new hybrid system which attempts to utilise the strengths of both approaches to create a more scalable means of mimicking voices.

2. Voice Cloning

2.1. Constraints

There is a genuine commercial interest in voice cloning for entertainment as well as an interest for speakers to create a virtual version of there own voice for use in cyber-realities varying from web pages to virtual environments. In addition there is a serious clinical application for the technology where it can be used to produce synthetic voices for patients who, due to illness, trauma, or surgery can no longer speak normally. However there are three constraints which have made voice cloning a rare activity in speech synthesis.

1. The resulting synthesis must sound 'natural' enough to effectively mimic the voice. Only over the last few years has speech synthesis begun to reach this level of naturalness.
2. The amount of data required from a speaker is not unlimited. Ideally we wish to mimic voices from as small amount of material as possible.
3. The type of speech styles required have a big impact on

Table 1: *Speech synthesis systems under test. These mnemonics will be used throughout this paper to refer to specific voices.*

System	Source Data		
	10 minutes	40 minutes	210 minutes
HTS-2007	HTS10	HTS40	HTS210
Cereproc CereVoice	CPCV10	CPCV40	CPCV210
Cereproc Hybrid	CPHY10	CPHY40	CPHY210

current cloning techniques. To a very large extent we can successfully mimic a voice in a single speech style with between 3-5 hours of carefully recorded speech. However to mimic a voice across many speech styles, for example mimicking different emotions is still a challenging research problem.

The unit-selection techniques can produce good mimicry of a single speech style (and recently some speech style variation [4]) given sufficient carefully collected data. The statistical parametric approaches, while not reaching the same level of naturalness as that of the unit-selection techniques, can offer the ability to mimic voices with substantially smaller amount of source speech data. In addition the statistical parametric techniques make it easier to alter vocal style due to the model based approach.

In this paper we will present three systems — HTS-2007 (statistical parametric approach) [5][6], Cereproc CereVoice (unit-selection approach) [4], and Cereproc Hybrid (hybrid approach of statistical parametric and unit-selection) — based on three different amounts of source speech material, a 10 minute database, a 40 minute database, and a 210 minute database taken from audio publicly available of Mr. George W. Bush.

2.2. Data Collection

The source data for these research voices was taken from audio freely available on the web. In order to effectively create the voices audio was carefully chosen and segmented into utterances varying from 1-261 words in length (Mean 12 SD 8.08). Note that care was taken to avoid background noise (i.e. applause, music), disfluencies (Ums and ahs), and poor audio quality caused by compression or low sampling rates. The data was manually transcribed and then verified using Cereproc's proprietary voice building system.

From 257 minutes of source speech in 4006 speech utterance files obtained via the above procedures, three sets of utterance lists were randomly selected for generating voice using approximately 10, 40 and 210 minutes of speech. For all the systems only this amount of source material was then used to generate the resulting voice. For example acoustic models were not trained on all the data and then used to segment part of it. From these three lists nine voices were created for each system, statistical parametric, unit-selection, and hybrid systems. See Table 1 for the mnemonics used for each voice.

3. Statistical Parametric Synthesis: HTS-2007

The HTS-2007 system is a high-quality speaker-independent HMM-based speech synthesis system developed by CSTR and

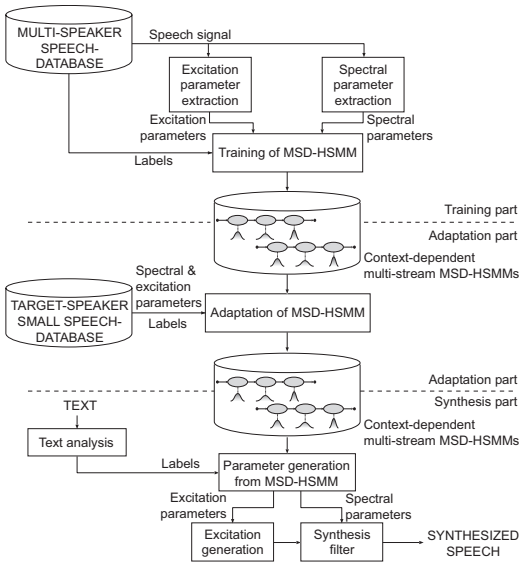


Figure 1: Overview of the HTS-2007 speech synthesis system.

Nagoya Institute Technology. In the system (Fig. 1), an average voice model using context-dependent multi-stream MSD-HSMMs is created from more than 10 hours of speech data uttered by many speakers and is adapted with speech data obtained from a target speaker. The acoustic features for the MSD-HSMMs are three kinds of parameters required for a high-quality speech vocoding method with mixed-band excitation called *STRAIGHT* [6]: the STRAIGHT mel-cepstrum, $\log F_0$, and aperiodicity measures. Using the above acoustic features, the MSD-HSMMs are trained based on the speaker-adaptive training and are adapted to the target speaker by using a combined algorithm of constrained structural maximum a posteriori linear regression (CSMAPLR) [7] and maximum a posteriori (MAP) adaptation. Speech parameters are directly generated from the adapted MSD-HSMMs using a penalised maximum likelihood method [8].

Since the average voice models can utilise the large-scale speech database and both spectral and prosodic features such as $\log F_0$ or phone duration can be statistically and simultaneously transformed from the average voice model into those of the target speaker, we can robustly create voices even from relatively small amount of speech data. However, the synthetic speech generated from the voices has a "buzzy" quality, since speech waveform is vocoded from pulse or noise excitation. Parts of the system have already been released in an open-source software toolkit called HTS (from "H Triple S," an initialism for the "HMM-based speech synthesis system") [9].

4. Unit-Selection Synthesis: Cereproc CereVoice

CereVoice is a faster-than-realtime diphone unit selection speech synthesis engine, available for academic and commercial use. The core CereVoice engine is an enhanced synthesis 'back end', written in C for portability to a variety of platforms. The engine does not fit the classical definition of a synthesis back end, as it includes lexicon lookup and letter-to-sound rule modules, see Fig. 2. An XML API defines the input to the en-

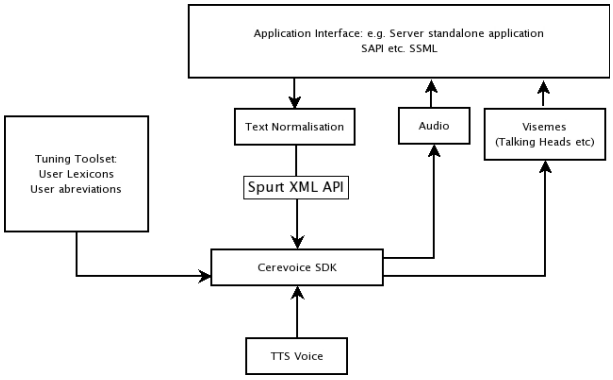


Figure 2: *Overview of the architecture of the CereVoice synthesis system. A key element in the architecture is the separation of text normalisation from the selection part of the system and the use of an XML API.*

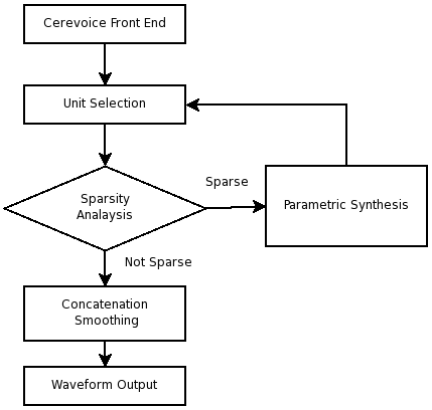


Figure 3: *Combining parametric and unit selection synthesis.*

gine. The API is based on the principle of a 'spurt' of speech. A spurt is defined as a portion of speech between two pauses. To simplify the creation of applications based on CereVoice, the core engine is wrapped in higher level languages such as Python using Swig. For example, a simple Python/Tk GUI was written to generate the test sentences for the Blizzard challenge.

The CereVoice engine is agnostic about the 'front end' used to generate spurt XML. CereProc use a modular Python system for text processing. Spurt generation is carried out using a greedy incremental text normaliser. Spurts are subsequently marked up by reduction and homograph taggers to inform the engine of the correct lexical variant dependent on the spurt context.

5. Hybrid Approach: Cereproc Hybrid

A key weakness in the unit selection approach is the issue of sparsity. In order to produce a smooth rendition of speech the database must contain appropriate units. If these are diphones and are American voice contains 40 phones this would require in 1600 different units for full coverage. In addition to the phones, you also need coverage of prosodic context, for example stress, increasing the required units to 6400 units if you include phrasing 25.6k units for full coverage. Finally the context of many units is also vital for concatenation because of co-articulation. If you also require coverage of all left and right contexts you require over 4 million different units.

Fortunately many of these contexts are very rare or do not occur. However, even with a modest requirement of 1600 different diphones, because speakers are required (in general) to produce normal connected speech for the source database, this tends to result in a database of approximately 300k diphones which can require up to 21 hours of studio recording time. Even given this size of database there will be many contexts missing and this in turn can produce concatenation errors.

Parametric approaches offer an attractive solution to this sparsity problem. Firstly, as the speech is synthesised from model parameters there are no concatenation errors. Secondly because the voice is derived from a model it is possible to use adaptation to harness information from other speakers to improve the model on only a small selection of data. The disadvantages with the parametric approach is that the generation of the speech from model parameters does not produce completely natural sounding speech. In some cases a so called 'vocoder buzz' is perceivable. However a more profound problem is that it is necessary to model **all** features in speech including prosodic variation and structure. Natural speech prosody is complex and as yet not entirely understood. Thus parametric systems can have dull or repetitive prosody.

A hybrid approach tries to use the advantages of both systems to create a more saleable and natural solution. Cereproc have developed a means for seamlessly concatenating parametric produced speech within a unit selection framework. In effect, when sparsity of concatenation errors are assessed as likely, sections of parametric speech can be used rather than standard units, see Fig. 3. The advantage of this approach is that a large proportion of the prosody can be produced within unit selection while at the same time avoiding sparsity caused by high dimensionality.

Results in this paper are for a very early prototype system. The system is combination of Cereproc CereVoice combined with HTS-2007.

6. Evaluation

All the systems were used to synthesise the opening paragraph of "The Emperor's New Clothes" by Hans Christian Anderson. This text was used for evaluation because: the material was completely different from the domain of the source material, story telling is harsh test of prosodic relevance and variation, the vocabulary was simple which meant that intelligibility was less likely to be a factor in the assessment.

The paragraph was split up into 9 utterances varying from 8 to 31 words long. Each subject heard each utterance from each of the systems and scored the naturalness on a five point scale. Finally the full paragraph from one system was played to the subject who then gave an overall score for the complete rendition from that specific system. 23 subjects took part in the experiment (of which 9 were native speakers).

Fig. 4 shows a histogram of the average mean opinion score (MOS) for each system. 95% confidence intervals are shown for each histogram.

The full audio for each system are available on the web at `http://www.cogsci.ed.ac.uk/~matthewa/LANGSPEECH2008.htm`.

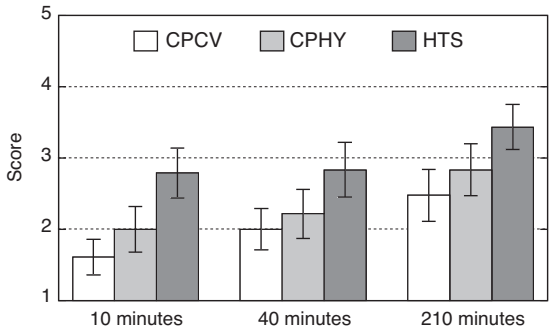


Figure 4: *Average 5 point MOS scores for all systems and databases. Error bars show 95% confident intervals.*

7. Discussion

Due to resource limitations our evaluation was small and the results should be treated with caution. In addition, MOS style evaluations can be problematic because there can be significant subject variation in what is regarded as "natural". For example non native speakers rated all systems approximately 0.5 worse on the MOS scale than native speakers.

There is also, arguably, a tendency for concatenation errors to be more heavily penalised by subjects than the problem of vocoder buzz when evaluated in single sentence MOS experiments.

However there is a clear preference for the HTS system in contrast to the past comparison of speech synthesis systems (e.g. [10]) in which hybrid approaches provide significant better quality. This in part may be related to specific issues with regards to found data (as opposed to carefully recorded data). In general one of the strengths of unit selection is that you get a lot of speech which has had very little modification made to it. When the speech is carefully recorded in a quite environment this is a great advantage. With this data, however, much of the audio was recorded in very different environments (in some cases decompressed from MP3). One of the strengths of the parametric approach was that it was able to remove the inconsistency caused by recording environment. Another important problem caused by recording environment is the possibility of phase differences between different sections of audio. Such phase variation can make time domain concatenation extremely problematic. Phase distortion was such a problem that the Cereproc system for smoothing the vocoded speech had to be switched off.

Finally we intentionally did not choose any data on the basis of it improving unit coverage. Thus the results highlight another of the strengths of parametric synthesis, where, for the 10 minutes database, the unit selection system was completely unable to function effectively in complete contrast to the Hybrid and HTS systems.

8. Conclusions

Given the data available we feel that the results for relatively small databases were excellent for the HTS system, which maintained impressive consistency. The Hybrid system exhibited teething problem in terms of effectively merging different recording environments but again showed that a dual approach is a serious research direction in speech synthesis. For the large database, all systems performed well, with HTS doing best in terms of a sentence by sentence 5 point MOS evaluation.

However we strongly suggest interested readers listen to the 9 short versions of the audio themselves to gain an insight into the differences between these systems and pros and cons of them. Since the artificial and buzzy quality of the synthetic speech generated from the HTS system remains, we need to explore a better hybrid algorithm which can work robustly and effectivley, even for found data, in order to reproduce the depth and variation of a speaker's prosody.

9. Acknowledgements

The authors would like to thank the HTS working staffs. Support for this research was provided by EPSRC (award number EP/D058139/1).

10. References

[1] L. Seward, "Scientists warn of 'vocal terror'," BBC NEWS, Sept. 2007. http://news.bbc.co.uk/1/hi/sci/tech/6994595.stm

[2] A. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", Proc. ICASSP-96, pp.373–376, May 1996.

[3] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," Proc. ICASSP 2007, pp.1229–1232, April 2007.

[4] M.P. Aylett, C.P. Pidcock, "The CereVoice characterful speech synthesiser SDK," Proc. AISB 2007, pp.174–178, April, 2007

[5] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "Speaker-independent HMM-based speech synthesis system — HTS-2007 system for the Blizzard Challenge 2007," Proc. BLZ3-2007 (in Proc. SSW6), Aug. 2007.

[6] J. Yamagishi, T. Nose, H. Zen, T. Toda, K. Tokuda, S. King, and S. Renals, "A speaker-independent HMM-based speech synthesis system for the Blizzard Challenge 2007," IEEE Trans. Speech, Audio & Language Process., 2007 (under review).

[6] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", Speech Communication, vol. 27, pp.187–207, 1999.

[7] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR Adaptation Algorithm", IEEE Trans. Speech, Audio & Language Process., 2007 (under review).

[8] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., vol.E90-D, no.5, pp.816–824, May 2007.

[9] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.B. Black and T. Nose, "The HMM-based speech synthesis system (HTS) Version 2.0.1", 2007. http://hts.sp.nitech.ac.jp/

[10] R.A.J Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results", http://festvox.org/blizzard/bc2007/

Conceptual maps and Computational Linguistics: the Italian ALTI project

Francesco Di Maio[1], *Johanna Monti*[2]

[1] Dipartimento di Scienze della Comunicazione, Università degli Studi di Salerno - Italy
[2] Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico, Università degli Studi di Napoli "L'Orientale" - Italy

dimaio@unisa.it, jmonti@unior.it

Abstract

ALTI linguistic multifunctional databases are the result of a project started in 1998 when the research interests of different Italian universities (Università degli Studi di Napoli L'Orientale, Università di Pisa, Università degli Studi di Salerno, Università degli Studi di Roma "Tor Vergata", Università degli Studi di Perugia) came together in one research project of national interest under the coordination of prof. Domenico Silvestri of the Università degli Studi di Napoli L'Orientale.

ALTI stands for **Atlanti Linguistici Tematici Informatici (**Electronic Thematic Linguistic Atlases), which represent a new typology with respect not only to traditional lexicography, but also to computational linguistics, since they put together the characteristics of traditional dictionaries and terminological collections with a conceptual map, which highlights the conceptual relation among terms. The word *atlas* is used to underline that it is not a simple dictionary but a collection of maps, organized as multimedia hyper-textual atlases which collect linguistic data belonging to specialized linguistic areas (such as onomatology, food-terminology, numerals, linguistic activities, metalanguage of linguistics, lexical-grammar) in a multilingual and interlinguistic perspective.

The Atlases describe the phenomenology of specific language areas by linking definitions and usage as given in conventional dictionaries to specific cognitive categories which create conceptual networks, several sets of maps (one for each atlas) or cognitive ellipses. These conceptual networks allow to navigate and to explore the relations between concepts inside a specialized linguistic area.

On the one hand, they are meta-dictionaries, since they refer to other dictionaries, lexical and terminological resources already available in traditional or electronic format, but they also add new information by following original research perspectives, and, on the other hand, they offer multimedia information such as graphs, photos, films which represent very useful tools for the users.

The languages investigated are: ancient and modern Indo-European and non-Indo-European languages; ancient and modern Celtic languages; Latin and Italy's ancient languages; major modern languages.

In conclusion the Atlases are an open work since it is always possible to modify and update them with new contents and so achieve rich virtual cognitive universes.

In our contribution we will describe the main features of the project, the research methodologies, the structure of the Atlases and of the lexical entries, the results achieved until now and the future aims.

Index Terms: conceptual maps, Electronic Thematic Linguistic Atlases, computational linguistics.

1. Introduction

In this contribution we describe the general outline of the Italian ALTI project. The acronym ALTI stands for **Atlanti Linguistici Tematici Informatici (**Electronic Thematic Linguistic Atlases), which are the result of the efforts of different Italian research groups (Università degli Studi di Napoli L'Orientale, Università di Pisa, Università degli Studi di Salerno, Università degli Studi di Roma "Tor Vergata", Università degli Studi di Perugia) coordinated by prof. Domenico Silvestri of the University of Naples L'Orientale. The Atlases are a collection of data belonging to specialized linguistic areas (such as onomatology, food-terminology, numerals, linguistic activities, metalanguage of linguistics, lexicon-grammar) organized inside conceptual maps or cognitive ellipses.

The main aim of the ALTI project is to implement multimedia hyper-textual atlases, which allow to navigate and explore synchronically and diachronically inside specialized thematic lexica, built mainly on the basis of existing lexical resources. They fulfill completely the main requirements of hypertexts such as reticularity, iconicity, non-finiteness and interactivity. On the one side, they are meta-dictionaries since they refer to other dictionaries or terminological and lexical collections, to which they add new information according to original research perspectives, and, on the other side, they contain multimedia information, such as graphs, photos, movies and other useful disambiguation tools for the final user. It is an open work, since every node of the map can be linked to any other node of the same atlas or different atlases. It is always possible to increase and modify the content of the atlases, which become in this way very rich virtual cognitive universes. Interactivity is of course another feature of these hyper-textual atlases, providing the user with a personal reading of the information or with several different readings of the same concept.

2. The Atlases

The ALTI project is composed of six different Atlases:

- the DETIA (Dizionario degli Etnici e Toponimi dell'Italia Antica) has the main aim to create a repository of ethnics and toponyms of Ancient Italy according to the Augustan *discriptio* in *regiones*.
- the AGAM (Atlante Generale dell'Alimentazione Mediterranea) collects all the terms connected with food in the Mediterranean area together with information regarding their preparation and contextual specification.
- the AULIL (Atlante Universale dei Logonimi e delle Istanze Logonimiche) has the main aim of creating a database of all the terms connected with the speech acts.

- the AUNIN (Atlante Universale dei Numerali e delle Istanze di Numerazione) is focused on the numerals which are described using information concerning their linguistic and semantic features.
- the DLM (Dizionario del Lessico Metalinguistico) collects all the terms which describe and explain the phenomenology of languages together with their usage, the different definitions they were given to in different periods of time, authors, works and theoretical schools. It contains 30,000 lemmas, in 31 languages, extracted from approximately 80 texts, with 10.000 authors' quotations.
- The DICOMP (Dizionario delle parole composte) collects all the compound words of a specific terminological field. In particular for this project 30,000 lemmas concerning economy, which are the results of one of the research activities carried out at the University of Salerno under the direction of prof. Annibale Elia, are made available.

As in all computational linguistic and lexicological researches, the structure of the data and of the databases, the query modalities and the updating procedures were changed during the project.
At the beginning the design of the database started from the creation and formalization of tables, then lexicographical workplaces were designed and released in order to carry out the lexicographical work using database facilities. In this way different lexicographical workplaces were implemented for each linguistic area (numerals, food-terminology, …). Once the databases had been created, a Web interface was designed in order to put all data on the Internet and let users access them with dynamic query modalities. To sum up, the ALTI system foresees: (a) different lexicographical workplaces, used by the researchers to store the entries together with relevant information in specific databases, (b) a Web interface composed of dynamic web pages implemented using the ASP language (Active Server Pages).

3. Sources and research methodology

The main sources of information of the researches carried out are:

- dictionaries, glossaries, lexicographical and/or terminological collections, both on paper or in electronic format, off-line and on-line;
- multimedia information for the description of lemmas (graphs, photo, movies, and so on);
- knowledge of the researchers who analyze the sources in order to identify the lemmas.

The researchers on the basis of the sources at their disposal, identify, organize and store:

- lemmas extracted from the sources adding the relevant bibliographical information;
- lexicological and lexicographical information (also in multimedia format) organized according to the descriptors of the cognitive maps of the different atlases;
- possible relations between lemmas inside one or more atlases.

All this information is stored inside the databases, which are updated and managed using lexicographical workplaces for each atlas typology.
These databases represent a set of structured knowledge which allows to have at disposal:

- a set of thesauri organized on a thematic basis for specific lexical areas, with advanced query possibilities;
- a documentation about specific lexical areas which allows to understand their historical evolution and their multilingual dimension;
- a set of tools for updating, managing and querying linguistic data.

4. Basic entry features

The lemmas updated in the atlases can be composed of single words or compound words. All the lemmas are updated in their canonical form, i.e. nouns in the singular form, verbs in the infinitive tense, and so on and are always lowercase, except for the spelling conventions of a specific language.
The lemmas are updated with relevant information connected to their peculiar nature. For instance, for the numerals the following information is required: the function they can have inside the numeral system of a given language, the grammatical category, morphology, structuring procedures, internal and external syntax, etymology, possible gestural and graphic codification and cultural implications.

5. Conceptual maps

For each atlas a cognitive map or ellipsis was created together with a series of electronic data concerning the lemmas.
Conceptual maps were used since they:

- are a graphical representation of knowledge where concepts are organized according to geometrical forms known as nodes and the links between them represent their relations;
- describe the structure of knowledge of the domain under investigation
- represent the main result of the activity of knowledge analysis and modelling.

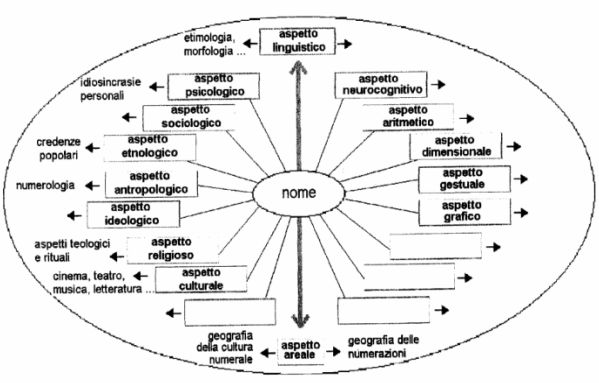


Figure 1: *Example of conceptual map or cognitive ellipsis.*

The maps allow the user to navigate inside one or more atlases.
The focus of each map is given by the lemma. Each map is divided up in two distinct zones: on the left of the map there

is the zone of subjective cultural data and on the right the one concerning objective cultural data.
The subjective cultural data are in common to all the different maps and are represented by psychological, sociological, ethnological, anthropological, ideological, religious and cultural aspects. Of course these aspects vary according to the specific linguistic area investigated. For instance if we take into account the ethnological aspect, this may concern the popular beliefs connected with a specific number (for instance 17 in Italy) in the AUNIN, whereas it may concern the raw/cooked categories in the AGAM.
The objective cultural data are specific of each atlas. For instance, in the AUNIN this zone of the map foresees the neuro-cognitive, arithmetic, dimensional, gestural and graphic aspects, whereas in the AGAM we find gastronomic, economic, technological, dietetic and medical aspects.
In each map, on the border between these two zones there are: on the top of the map the linguistic aspect (represented by etymology, morphology, …) and at the bottom the areal aspect, which is foreseen for the creation of specific geo-linguistic maps.

6. The ALTI Lexicographical workplaces

In the first phase of the project a lexicographical workplace for each atlas was created for the entry of lexical data. All these data will join in the near future in one database managed locally and on the WEB. The software was implemented in Windows, and based on MSAccess tables with SQL queries.
The databases are structured according to the requirements of the different typology of linguistic data. For instance, in the AUNIN, the data are input using a lexicographical workplace in a database composed of three different archives: bibliography, numeral systems and lemmas. Therefore the compilation of three different cards, one for each archive, is foreseen:

- System card
- Numerals card
- Bibliography card

Figure 2: *AUNIN Lexicographical workplace: the numerals card*

Whereas in the AULIL, the lexicographical workplace foresees the compilation of two different cards which correspond to two different databases: bibliography and logonyms.

7. The ALTI Web site

In the second phase of the project, once the databases had been designed and implemented, the next step was the distribution of the ALTI data on the Internet. Therefore, WEB interface pages were designed to handle data dynamically using the ASP language (Active Server Pages). In addition to pure HTML code, this language creates several scripts which generate the page code to be sent to the user browser. In this way the dynamic contents (that is, the contents extracted by the database located on the web server) can be displayed and their appearance changed according to the rules coded in the scripts, without sending the code to the final user program. Only the result is sent, with a significant saving of waiting times on the Internet. This language interfaces with different database types (like Access) with no need of converting data, that would need further programming and indexing operations with the risk of possible errors. The Web site design, which handles the contents of the different linguistic databases (AULIL, AUNIN, etc.), was performed through distinct steps. First of all, the query criteria of the database (the following figure shows the search page designed for the AULIL database) had to be decided.

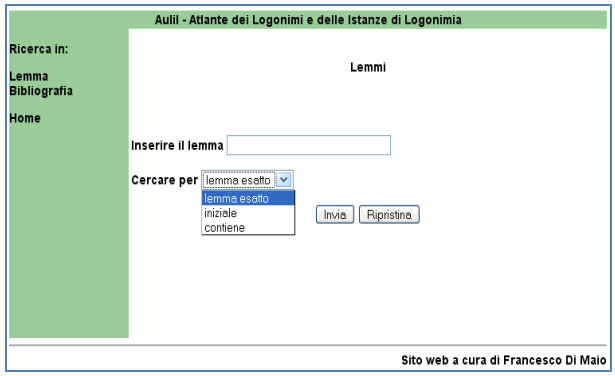


Figure 3: *The AULIL lemma query interface*

The figure shows an additional parameter which can be used in combination with the field **lemma** to perform the query. The input string of the lemma field can be combined along with the three options in the **Cercare per** (Search for) drop-down box:

- **Lemma esatto** (exact lemma): it queries the database looking for the exact string as lemma
- **Iniziale** (initial): it queries the database looking for the string as initial part of a lemma
- **Contiene** (contains): it queries the database looking for the string as part of one or more lemmas.

The input and selection/combination of these parameters in the available fields determine the transfer of parameters and the generation of results through a query to the database in SQL language. For example, searching for the **ne** string as initial string of lemmas in the database and combining the **iniziale** parameter of **Cercare per**, the parameters will be transferred to the query engine of the database when the user clicks the **Invia** (Enter) key :

http://www.alti.unisa.it/alti/aulil/risultlemma.asp?filtrolemma=ne&select=1&Submit=Invia

which for the user means “look for all the lemmas that begin with ne”. For more expert users we could split the string as follows:

- *filtrolemma* : the variable to be input in the query field **lemma**
- *select*: any of the three options of **Cercare per**, whose parameters are: 1=iniziale (initial), 2=lemma esatto (exact lemma) and 3=contiene (contains)
- *Submit*: the Invia (Enter) key, which is the transmission command of these parameters to the server hosting our database. The required data will be searched for, and, if available, they will be displayed in the risultlemma.asp page, a dynamic page specifically created. The following example shows the query used for this search:

```
<%
sel=request("select")
If (Request.QueryString("filtrolemma") <> "") and sel="1" Then
stringasql="SELECT lemma, fonte, pagine FROM logonimi GROUP
BY lemma, fonte, pagine HAVING lemma LIKE '" + filtrolemma +
"%" + "' ORDER BY lemma ASC"
end if
%>
```

If the result of the query is successful, the data will be displayed as in the figure below:

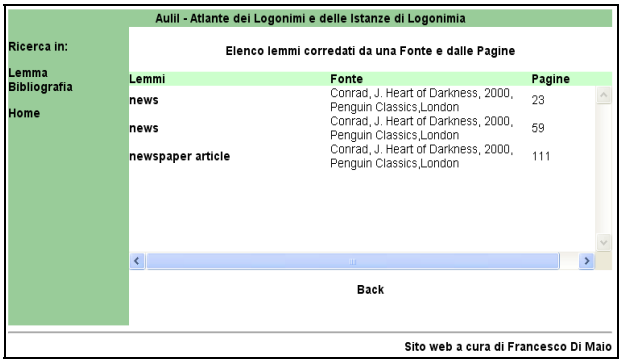


Figure 3: *The AULIL lemma query result interface*

The query result is given by the lemmas with additional information about the source and the pages of the text/dictionary which they were extracted from.
Regarding hyper-textuality, we can connect to another page containing other related information in this page, if we click on a lemma in the list, such as *newspaper article*, we can read the catalogue details of the single lemma from the table which contains them using a query, and we can display them in a different page:

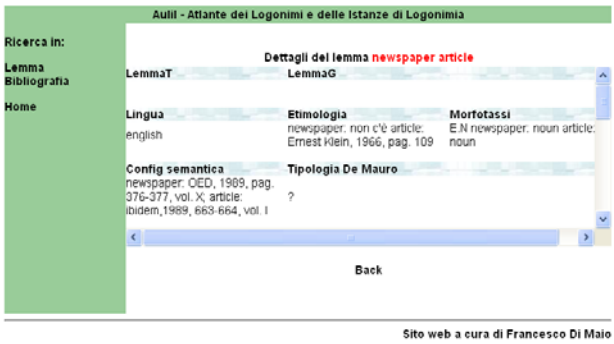


Figure 3: *The AULIL lemma catalogue details interface*

The page for this data was designed in scroll mode so that it can display the information it contains. The user can read the information using the scrollbar on the right side of the page as displayed in the above figure.
The criteria used for the bibliographical search page are similar to the ones used for the lemma search page. In this example, a combination of the search fields **autore (**author**)** and **titolo (**title**)** was used so that the user can search any of the items or both of them at the same time.

Figure 3: *The AULIL bibliography query interface*

Moreover, in this example, you can use the criteria previously described for the search page of lemmas , i.e. search by exact string, initial string and “contained in” string.

8. Conclusions

We have provided an overview of the Italian ALTI project, and in particular we have described the different components and the results so far achieved.
What we have described is only the beginning of a research project which should be considered as an “open” work, that is a work in progress since new languages and new data can be and will be added.
We believe that the framework we have designed can prove to be very useful for both theoretical research and computational applications and thus could become a model for similar lexical resources.
The results of the ALTI research project, in particular the Web site and all the linguistic data, will be available on the Internet in the very near future.

9. Note

Abstract, paragraphs 1,2,3,4,5 and conclusions are by Johanna Monti, whereas paragraphs 6 and 7 are by Francesco Di Maio.

10. References

Atti del Convegno su Numeri e istanze di numerazione tra preistoria e protostoria linguistica del mondo antico – Napoli 1-2 dicembre 1995 - AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 17, 1995.

Chiari, I., Introduzione alla linguistica computazionale, Gius. Laterza & figli, Roma-Bari, 2007.

Di Maio, F., Monti, J., Gli atlanti tematici informatici: oltre il dizionario elettronico. Esemplificazioni dalla versione tedesca dell’ Atlante Universale dei Numerali e

delle Istanze di Numerazione (AUNIN) e dell'Atlante Universale dei Logonimi e delle Istanze Logonimiche (AULIL) (in press)

Pannain, R., "Numerali ed istanze di numerazione: note per un progetto di tipologia areale dei numerali", AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 22, 63-105, 2000.

Silvestri, D. "Logos e logonimi", Vallini C. (ed.) , Le parole per le parole. I logonimi nelle lingue e nel metalinguaggio, Atti del convegno Istituto Universitario Orientale Napoli 18-20 dicembre 1997, Il Calamo, Roma, 21-37, 2000

Silvestri, D., "From the eloquence of light to the splendor of the word", Semiotica Special issue, Signs and Light: Illuminating Paths in the Semiotic Web 136 -1/4, , Guest editor : Susan Petrilli: 117-132

Silvestri, D., "I lessici tematici tra lingua standard e lessici scientifici" AIΩN (Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico – Sezione Linguistica, direttore: Domenico Silvestri) 24, 11-30, 2002

Vallini, C. (ed.), Le parole per le parole. I logonimi nelle lingue e nel metalinguaggio. Atti del convegno Istituto Universitario Orientale Napoli 18-20 dicembre 1997, "Lingue, Linguaggi, Metalinguaggi" 1, Collana diretta da C. Vallini e V. Orioles, Il Calamo, Roma, 2000.

COUPLING SPEECH RECOGNITION AND RULE-BASED MACHINE TRANSLATION WITH CHART PARSING

Selçuk Köprü, Adnan Yazıcı**, Tolga Çiloğlu***, Ayşenur Birtürk***

*Applications Technology, Inc., Turkey
**Dept. of Computer Engineering, Middle East Technical University, Turkey
***Dept. of Electrical and Electronics Engineering, Middle East Technical University, Turkey

ABSTRACT

This article presents our approach and findings in coupling statistical Speech Recognition (SR) systems with a rule-based Machine Translation (MT) system. Most of the literature about coupling focuses on how to integrate SR with statistical MT systems. We think that utilizing rule-based MT systems for Speech Translation (ST) task is important and still remains as an open research issue. In this paper we introduce the Apptek Speech Translator system, the distinctive approach used for coupling SR and rule-based MT, and the results of the experiments to justify the approach.

Index Terms— chart parsing, machine translation, coupling, speech translation, speech recognition

1. INTRODUCTION

The demand for integrating SR systems and MT systems is becoming more and more important as both technologies are advancing towards satisfactory levels. Realizing ST task requires this integration to be as much perfect as possible. Most of the research focuses on how to integrate SR with statistical MT systems. The fact that both are based on the same principles and techniques might be a reason for existing integrations. This allows both systems to interact and exchange information in a much more straightforward way. Quite few of the studies [6, 12, and 13] address coupling heterogeneous components of NLP Systems. In this respect, our research work is unique because it presents a new approach and findings in coupling statistical SR systems with a rule-based MT system. It would be very beneficial to utilize rule-based MT systems for ST task because linguistic resources are used extensively. The rules and lexicons are created in many years based on broad studies and experience. Thus, utilizing these existing linguistic resources for both coupling and analysis would be a big gain in ST.

There is a wide range of studies available in literature attacking the coupling problem. Reference [10] is the first study classifying coupling into three categories: tightly-coupled, loosely-coupled, and semi-coupled. According to [8], tightness describes how close the SR and MT units interact with each other. In a tightly-coupled system, speech and MT processing is packed into an inseparable unit. In a loosely-coupled system, processing is inside independent modules. Finally, semi-coupled systems lie between the previous two approaches in terms of isolation. Tight coupling is possible if both modules are statistically based because the unit of information interchanged between systems is meaningful for both sides. Thus, it is considered to be a difficult task to tightly couple a statistical SR with a rule-based MT.

The coupling method suggested in [2] is a tightly coupled system where the whole process is based on Bayes decision rule. The work in [3] and [4] are similar to [2] and all are applicable only for statistical MT systems. In [5], Saleem et. al. discuss another approach towards tightly coupling SR and statistical MT systems. They conclude that using word graphs as the information exchange unit does improve performance when the weighted acoustic scores are incorporated into the MT unit. An alternative for a word-graph is a confusion network which is another type of directed graph where each path from start to end includes all existing nodes. Using confusion networks as the unit of exchanged information between SR and statistical MT is explored in [7] and [9].

While integrating the SR system with the rule-based MT system, this study uses word-graphs and chart parsing with new extensions to achieve the coupling. Parsing of word lattices has been a topic of research over the past decade [1, 11, 13, 14, and 15]. The idea of chart parsing the word graph in SR systems has been used previously in different studies in order to resolve ambiguity [1, 11]. However, to the best of our knowledge, the specific method for chart parsing a word graph introduced in this paper has not been used for coupling purposes before. There are two main differences between the work presented here and the previous ones existing in literature. First, in [1] and [11], the chart is populated with the same word-graph that comes from the speech recognizer without any pruning, whereas in our approach the word-graph is shrunken to an acceptable size. Otherwise, the efficiency becomes a big challenge because the search space introduced by a chart with more than thousand initial edges can be easily beyond current practical limits. Another important difference in our approach is the extension of the chart to eliminate spurious parses. Main distinction between [13, 14, and 15] and our study is the parsing algorithm being used. In contrast to our chart parsing approach augmented by unification based feature structures, Charniak parser is used in those studies along with PCFG.

In the next section we introduce the Apptek Speech Translator and our approach on coupling. In Section 3, we present the results of the experiments that are carried out to justify the approach. The concluding remarks are given in Section 4.

2. APPTEK SPEECH TRANSLATOR

The general architecture of the Apptek Speech Translation system is depicted in Figure 1. The system is loosely coupled, i.e., there is a one directional information flow between the SR and MT. The original word-graph created by the SR is reduced according to the Viterbi search algorithm based on bigram scores inside the pruning module. The output is another word-graph representing the N-best sentence hypotheses. The pruned word graph is processed by the MT component of the speech translation system. MT task deploys a transfer based approach and processing is divided into three clear-cut phases: analysis, transfer and generation. At the end of the analysis, any ambiguity is resolved and the best sentence hypothesis is picked for the transfer stage. In this paper, our focus is on the part encircled with dashed lines in Figure 1.

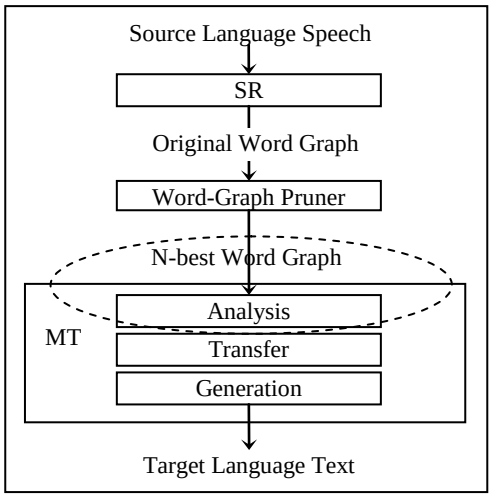


Figure 1. Apptek Speech Translator.

2.1. MT Analyzer

The analysis is accomplished in two consecutive tasks. First, morphological analysis is performed on the word level and any information carried by the word is extracted to be used in later stages. Next, syntactic analysis is performed on the sentence level. The syntactic analyzer consists of a chart parser in which the rules modeling the source language grammar are augmented with feature structures. The grammars are implemented using *Lexical Functional Grammar* (LFG) paradigm. Primary data structure to represent the features and values is a directed acyclic graph (dag). The system also includes an expressive Boolean formalism, used to represent functional equations to access, inspect or modify features or feature sets in the dag. Complex feature structures, e.g. lists, sets, strings, and conglomerate lists, can be associated with lexical entries and grammatical categories using inheritance operations. Unification is used as the fundamental mechanism to integrate information from lexical entries into larger grammatical constituents.

A sample parse tree and the feature structures in English are shown in Figure 2. For the case of simplicity, many details and feature values are not given. The dag containing the information originated from the lexicon and the information extracted from morphological analysis is shown on the leaf levels of the parse tree in Figure 2. The final dag corresponding to the root node is built during the parsing process in cascaded unification operations specified in the grammar rules.

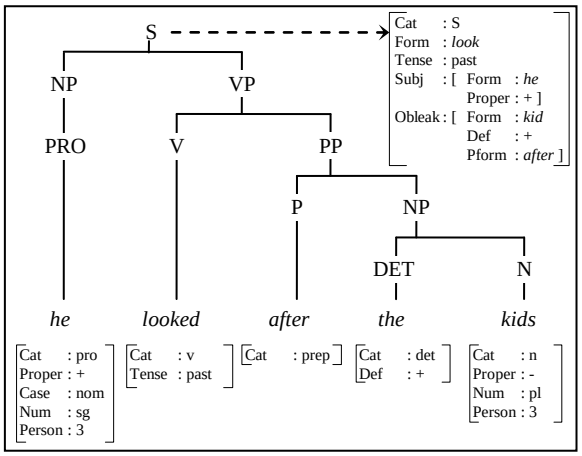


Figure 2. Unification based parsing.

2.2. Approach

In a loosely coupled system, information flow between modules can be made in different forms and quantities. In the simplest form, the SR provides only the first-best sentence. It can also provide an N-best list of sentence hypotheses to the parser to eliminate existing recognition errors [16]. In a more sophisticated form, the SR provides a word graph containing more information compared to the previous two forms.

The fundamental idea behind our solution is to initialize the chart of the MT parser using the simplified word graph coming from the Word-Graph Pruner module. Thus, all selected sentence hypotheses will be processed simultaneously. Initialized chart will be processed once until the first sentence hypothesis is picked by the parser. In its basic form, the chart models a confusion network, which might lead to spurious parse trees. We have extended the original chart representation and its processing in order to be able to model a word-graph instead. The advantage of the approach is essentially in the ability to rule out non-syntactic hypotheses in a parallel fashion.

The chart initialization algorithm assumes that a word-graph is represented by (N, S, A) where N is a set of nodes, S is a subset of N and contains the starting nodes, and A is a list of arcs as in $u \rightarrow v$, where $u, v \in N$. Algorithm 1 ensures the creation of a valid chart that can be processed by the well-known chart parser. The number of rows in the chart is equal to the number of sentence hypotheses in the input word-graph. The algorithm makes use of a stack to keep track of visited nodes and the associated cell indices. *pop* and *push* functions insert into and retrieve data from the stack data structure. The chart is represented by a table like structure. Function *split_cell(x, y, i)*, splits all the cells from position *(x, y)* until *(x, z)* into *i* vertical cells, where *z* is the total number of columns in the table. Splitting operation is a way of inserting new rows to the table. While doing this operation, all effected row numbers inside the stack are updated. Function *merge_cells (x, y, z)*, merges cells in the range from *(x, y)* to *(x, z)*. Function *out_arcs(n)*, gives the number of arcs in *A* where *n* is the source node. Finally, function *column(n)* returns the column index of node *n* in the chart.

Algorithm 1. Initialization of the chart

```
for each node n ∈ S
    row ← 1
    push (n, row, 1)
    row ← row + 1
while stack ≠ Empty
    pop (n, row, col)
    chart [row, col] ← n
    if out_arcs(n) > 1
        split_cell (row, col + 1, out_arcs(n))
    i ← 0
    for each arc: n → m
        if all nodes p exist in chart, where ∃ an arc: p → m
            push (m, row + i, col + 1)
        i ← i + 1
    for each arc: p → n
        if column (p) + 1 ≠ col
            merge_cells (row, column (p), col - 1)
```

Consider the simple word-graph, the corresponding chart and the hypotheses depicted in Figure 3. Using Algorithm 1, the chart is populated for syntactic analysis. The *for* loop at the beginning of the Algorithm 1 puts starting nodes into different rows. The *while* loop processes the remainder of the path to the final node. Words on the same column are regarded as a single lexical entry with different senses (e.g. *'boy'* and *'boycott'* at column 2). Words spanning more than one column are regarded as idiomatic entries (e.g. *'escalated'* at columns 3 to 5). Merged cells in the chart (e.g. *'the'* and *'yesterday'* at columns 1 and 6, respectively) are shared in the two sentence hypotheses.

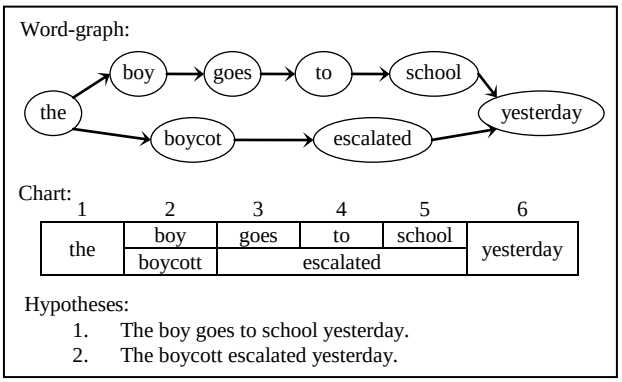


Figure 3. Sample word-graph, the corresponding chart and the hypotheses.

In an ordinary chart parser, the initial chart in Figure 3. can produce some spurious parses. For example, both of the entries at location 2, *'boy'* and *'boycott'*, can be combined with the word *'goes'*, although *'boycott goes'* is not allowed in the original word-graph. We have eliminated these kinds of spurious parses by introducing a new feature *rowid* into the lexical dags and grammar rules. This feature will contain the row number of the words. Constituents can be combined only if their *rowid* values can be unified. *rowid* for a cell spanning more than one row will include all the spanned row numbers as a set. The sample implementation of this idea is illustrated in Figure 4. *'boycott'* and *'goes'* cannot be combined in a parse tree because their *rowid* values do not unify. *'the'* can be combined with both *'boy'* and *'boycott'* because its *rowid* value contains 1 and 2. This extension to the parser makes our approach word-graph based rather than confusion network based.

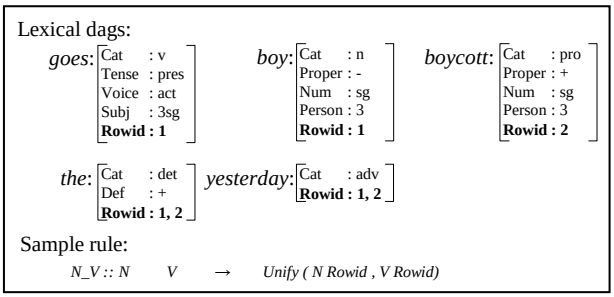


Figure 4. Sample rule enforces *rowid* labels from the constituents to be unified.

3. RESULTS OF EXPERIMENTS

In this section we present some experiments carried out on English to Arabic translation as a part of the Apptek speech translation project. The English and Arabic monolingual lexicons contain 40K and 60K entries, respectively. English to Arabic bilingual transfer lexicon contains also 40K entries. English grammar is modeled with 125 parsing rules implemented in LFG formalism.

Table 1. represents the number of complete and incomplete edges generated for the sample word-graph in Figure 3 for all three approaches. "Gain" column in the table represents the decrease in the total number of edges compared to the N-best list approach. The first-best approach performs better only if the first hypothesis is syntactically correct (first row in Table 1). When the first hypothesis is syntactically ungrammatical in the first-best approach (second row in Table 1), the N-best word-graph processing system performs better and produces more accurate output. The reason for this is because it is able to parse the first N sentence hypotheses together instead of trying to parse an ungrammatical sentence.

Table 1. Number of edges generated in the two different approaches for the sample in Figure 3.

Approach	Complete Edges	Incomplete Edges	Total	Gain
First-best (*success*)	62	236	298	75%
First-best (*fail*)	92	849	941	19%
N-best List	150	1007	1157	-
N-best Word-Graph	78	284	362	69%

Compared to the N-best list approach, the shared edges are processed only once for all hypotheses. This saves a lot on the number of complete and incomplete edges generated during parsing. Hence, the overall processing time required to analyze the hypotheses are reduced. In an N-best list approach, where each hypothesis is processed separately in the analyzer, there are different charts and different parsing instances for each sentence hypothesis. Shared words in different sentences are parsed repeatedly and same edges will be created at each instance. As it can be seen, there is a 69% gain in the number of edges if the N-best word-graph approach is used instead of the N-best list approach for the sample in Figure 3. The huge difference in the number of incomplete edges arises from the fact that the processing continues until a successful parse is found. For a syntactically wrong sentence, the search for the successful tree continues until

all possibilities are exhausted. The word-graph includes most probably a syntactically correct sentence hypothesis. Thus, the parsing process stops before exhausting all possibilities as soon as it finds the desired parse tree. The N-best list approach performs similar to the first-best approach if the first hypothesis is the correct one. As might be expected, the hypothesis that is picked after a successful parse in any of the approaches does not mean to be the desired hypothesis.

Next, we have conducted an experiment to show the relation between the word-graph size and the number of edges built during processing for each approach. For this purpose, we have picked from newspapers 20 real-life phrases for each word-graph size segment and run through the SR and word-graph pruner with N=10. On the average, the correct hypothesis was at the sixth place. As depicted in Chart 1, the ratio remains the same as the number of nodes in the word-graph increases. In the word-graph processing approach, the total number of edges remains below acceptable boundaries even for big sized word-graphs. For separate processing, however, this number is above practical limits especially in large word-graphs.

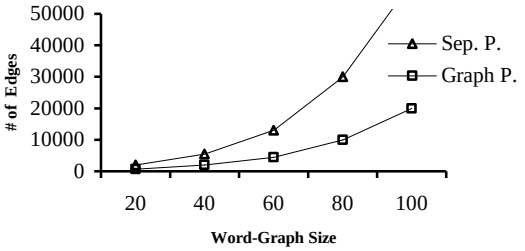


Chart 1. Relation between word-graph size and number of edges (N=10).

4. CONCLUSIONS

In this paper, we present a new approach to couple SR and a rule-based MT for speech translation purpose. This approach can be generalized to any MT system employing chart parsing in its analysis stage. Thus, the extensions proposed in this study can pertain also to statistical MT framework utilizing a chart. However, our focus is on a rule-based MT. Besides utilizing rule-based MT in ST, this study uses word-graphs and chart parsing with new extensions. The experiments described in this paper show that parsing the word-graph at one instance improves the translation performance, compared to parsing all sentence hypotheses separately. For further improvement of the ST system, our future studies include the following:

1. Evaluation of the total ST system based on standard assessment measures (word error rate, BLEU, etc.) It will be interesting to see how our coupling performs compared to other approaches.
2. Supplementing the word-graph pruning module with the language grammar rules to improve the quality of the N-best word-graph.
3. Utilizing the statistical information, coming from the SR module, inside MT towards a hybrid system. This will supplement the ambiguity resolving process in syntactic analysis.

5. ACKNOWLEDGMENTS

This work is partially financed by Applications Technology, Inc., USA. Thanks for Jude Miller and Nagendra Goel for their valuable comments and kind support.

6. REFERENCES

[1] L. Chien, K. Chen, and L. Lee, "A Best-First Language Processing Model Integrating the Unification Grammar and Markov Language Model for Speech Recognition Applications," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No.2, pp. 221-240, 1993.

[2] H. Ney, "Speech translation: Coupling of Recognition and Translation," *Proc. ICASSP*, 1999.

[3] R. Zhang, G. Kikui, "Integration of Speech Recognition and Machine Translation: Speech Recognition Word Lattice Translation," *Speech Communication*, Vol.48, Issues 3-4, pp. 321-334, 2006.

[4] E. Matusov, S. Kanthak, and H. Ney, "On the Integration of Speech Recognition and Statistical Machine Translation," *Proc. InterSpeech*, 2005.

[5] S. Saleem, S. C. Jou, S. Vogel, and T. Schultz, "Using Word Lattice Information for a Tighter Coupling in Speech Translation Systems," *Proc. ICSLP*, 2004.

[6] V. Arranz, et. al., "A Speech-to-Speech Translation System for Catalan, Spanish and English," *AMTA-2004*, Washington DC.

[7] N. Bertoldi, and M. Federico, "A New Decoder for Spoken Language Translation Based on Confusion Networks," *IEEE ASRU Workshop*, 2005.

[8] E. Ringger, "A Robust Loose Coupling for Speech Recognition and Natural Language Understanding," *Technical Report*, The University of Rochester, Computer Science Department, Rochester, New York, 1995.

[9] W. Shen, R. Zens, N. Bertoldi, and M. Frederico, "The JHU Workshop 2006 IWSLT System", *Proc. IWSLT*, 2006.

[10] M. P. Harper et. al., "Integrating Language Models with Speech Recognition," *Proc. AAAI Workshop on the Integration of Natural Language and Speech Processing*, pp. 139-146, 1994.

[11] H. Weber, "Time Synchronous Chart Parsing of Speech Integrating Unification Grammars with Statistics," *Proc. of the 8th Twente Workshop on Language Technology*, Dec. 1994.

[12] C. Boitet, and M. Scligman, "The "Whiteboard" Architecture: A Way to Integrate Heterogeneous Components of NLP Systems," *Proc. COLING*, 1994.

[13] K. Hall, "Best-First Word Lattice Parsing: Techniques for Integrated Syntax Language Modeling," *PhD Thesis*, Dept. of Computer Science; Brown University, 2005.

[14] K. Hall, and M. Johnson, "Language Modeling Using Efficient Best-First Bottom-Up Parsing," *Proc. ASRU*, 2003.

[15] K. Hall, and M. Johnson, "Attention Shifting for Parsing Speech," *Proc. ACL*, 2004.

[16] C. Yen-Lu and S. Richard, "The N-Best Algorithm: An Efficient Procedure For Finding Top N Sentence Hypotheses," *Proc. HLT*, 1989.

Human Language and Semantic Web Technologies for Business Intelligence Applications

Thierry Declerck [1], *Hans-Ulrich Krieger* [1], *Horacio Saggion* [2], *Marcus Spies* [3]

[1] Language Technology Lab, DFKI GmbH
[2] NLP Group, Department of Computer Science, Sheffield University
[3] Digital Enterprise Research Institute, Universität Innsbruck
{declerck,krieger}@dfki.de, Saggion@dcs.shef.ac.uk, Marcus.spies@deri.at

Abstract

In this LangTech poster submission, we describe the actual state of development of textual analysis and ontology-based information extraction in real world applications, as they are defined in the context of the European R&D project "MUSING" dealing with Business Intelligence. We present in some details the actual state of ontology development, including a time and domain ontologies, which are guiding information extraction onto an ontology population task.
Index Terms: language technology, semantic web, business intelligence

1. Introduction

MUSING is an R&D European project[1] dedicated to the development of Business Intelligence (BI) tools and modules founded on semantic-based knowledge and content systems. MUSING integrates Semantic Web and Human Language technologies for enhancing the technological foundations of knowledge acquisition and reasoning in BI applications. The impact of MUSING on semantic-based BI is being measured in three strategic domains:

- **Financial Risk Management (FRM),** providing services for the supply of information to build a creditworthiness profile of a subject -- from the collection and extraction of data from public and private sources up to the enrichment of these data with (semantic) indices, scores and ratings;
- **Internationalization (INT)**, providing an innovative platform, which an enterprise may use to support foreign market access and to benefit from resources originating in other markets;
- **IT Operational Risk & Business Continuity (ITOpR),** providing services to assess IT operational risks that are central for Financial Institutions -- as a consequence of the Basel-II Accord – and to asses risks arising specifically from enterprise's IT systems -- such as software, hardware, telecommunications, or utility outage/disruption.

Across those development streams of MUSING, there are some common tasks, like the one consisting in extracting relevant information from annual reports of companies and to map this information into XBRL (Extended Business Reporting Language). XBRL is a standardized way of encoding financial information of companies, but also the management structure, location, number of employees, etc. (see www.xbrl.org). This is mostly "quantitative" information, which is typically encoded in structured documents, like financial tables or company profiles etc.

But for many Business Intelligence applications, there is also a need to consider "qualitative" information, which is most of the time delivered in the form of unstructured text, which one can find in textual annexes to the balance sheets in annual reports or in news articles. The problem is here how to accurately integrate information extracted from structured sources, like the periodic reports of companies, and the day to day information provided by news agencies, mostly in unstructured text form.

So for example imagine that you have an annual report from a company, which specifies the name of the CEO (and of other members of boards). The described CEO-relationship between the named person and the company is valid in this case only for the reporting period. But very often, the final report is published 2-3 months after the end date of the covered period. In the meantime it can be that the CEO position is now in the hand of another person, and this has been announced in press articles. We need here accurate information extraction (IE) systems, that detect in the news this change of name of the CEO, and overwrites the information that might have been extracted from the annual report, for the period following the temporal coverage of the annual report until the publication date of this report. This period has to be specified, also on the base of temporal information extracted from the article (a combination of the publication date and temporal information).

As the example above shows, work on IE and ontology population in MUSING is highly depending on temporal information associated with both the publication date of documents and the content of the document itself. So for example the date of publication of an annual report doesn't coincide with the end of the reporting period, and we have to extract the values for the starting and the ending time of the reporting period from the document itself. Additionally, the temporal information associated with certain quantitative information contained in the annual reports can be of two types, whereas this is not explicitly mentioned in the report: duration or instant (for example the number of employees given is valid for a specific instant in time, whereas the value of certain financial indicators is valid for a specific period). This distinction is formulated in available background semantic resource (like XBRL taxonomies) and has to be made explicit in our semantic representation of the information extracted from relevant documents in the MUSING applications.

As a summary of our needs with respect to temporal information in the concrete tasks described above, we learned that we can not work only with synchronic relationships, but

[1] See www.musing.eu for more details.

rather that we need some means to deal with the intrinsic diachronic aspects of entities and relations.

We describe in the following the actual state of development of MUSING ontologies, including our proposal for temporal representation. We give then examples of the kind of temporal expressions we encounter in applications of MUSING, and how our IE and Ontology Population tools deal with those expressions in the light of our representation of temporal information, aiming also at supporting temporal reasoning in various applications.

2. State of MUSING ontologies

In MUSING we decided to use as the upper level ontology the PROTON ontology (http://proton.semanticweb.org), on the base of which domain-specific extensions can be easily defined.

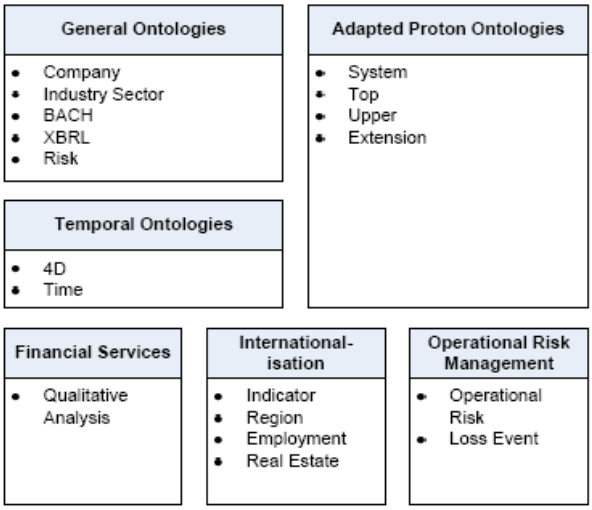


Figure 1: The various types of MUSING ontologies

The species of the model of the PROTON Upper module is OWL Full. The MUSING version available contains mostly the same information as the original one but is slightly changed to fulfill the OWL Lite criteria (see the box "Adapted Proton Ontologies" in the figure above).

"The System module of PROTON, http://proton.semanticweb.org/2005/04/protons, provides a sort of high-level system- or meta-primitives, which are likely to be accepted and even hard-coded in tools that may use PROTON. It is the only component in PROTON that is not to be changed for the purposes of ontology extension." (Terziev et al. 2005).

The contained Top-Level classes, http://proton.semanticweb.org/2005/04/protons, represent the most common definition of world knowledge concepts. These can directly be used for knowledge discovery, metadata generation and to interface intelligent knowledge access tools (Terziev et al. 2005).

The PROTON upper module, http://proton.semanticweb.org/2005/04/protonu, adds sub-classes and properties to the Top-module super classes to the concepts other than "Abstract, Happening and Object" from the original PROTON Top ontology.

The "Extension" ontology in MUSING has been designed as a single contact point between upper and MUSING application specific ontologies (the three boxes at the bottom of the figure above).

Besides the time ontology developed within MUSING, there are currently five domain ontologies, which are not assigned to any particular application (the "General Ontologies" in the figure above). They cover the following areas: Company, Industry sector, BACH (Standard for a harmonization of financial for harmonizing accounts of companies across countries), XBRL (Standard language for "Business Reporting") and Risk.

In the time ontology of MUSING, temporally-enriched facts are represented through *time slices*, four dimensional slices of what Sider (1997) calls a *space-time worm* (we only focus on the temporal dimension in MUSING). These worms, often referred to as *perdurants*, are the objects we are talking about. For instance, *Jürgen Schrempp* (JS) is a perdurant that comes up with several time slices, talking about his CEOship with Daimler Chrysler (DC), his resignation as CEO of DC (end of 2005), his membership for a certain time within the supervisory board of Allianz and Vodafone, etc. All facts are associated with a temporal dimension, even if they are instants, i.e., having an infinitely-small extension. This kind of representation is encoded in the "4D" ontology (see Figure 1). The time ontology itself contains the conceptualization of temporal objects that are relevant in MUSING. In fact, any time ontology can be combined with the "4D" ontology.

There is only one ontology module specific to the Financial Services applications: "Qualitative Analysis". This ontology describes in fact what can be the result of different kind of questionnaires used in this applications field. All the quantitative information relevant for Financial Risk management are covered by the General ontologies of MUSING, since this information plays also a role in the other application domains considered in MUSING.

The set of ontologies for the Internationalization applications in MUSING contains four ontologies. The most important ontology is the one defining the indicators used to measure properties of political regions. It is based on a list of 162 indicators grouped into 14 categories.

In the ItOperational Risk applications of MUSING, we introduce two ontologies, which deal with processes and IT infrastructure, on the one hand, and operational risk in general and IT operation risk in particular, on the other hand.

As a concluding remark about the ontologies, we would like to mention that they have been built by hand, most of them on the base of "compentency questions" (Gruninger, M., & Fox, M., 1994) addressed by domain experts. But it is also planned in MUSING to investigate the topic of (semi-)automatic ontology learning or creation, on the base of information and knowledge extracted from the analyzed data.

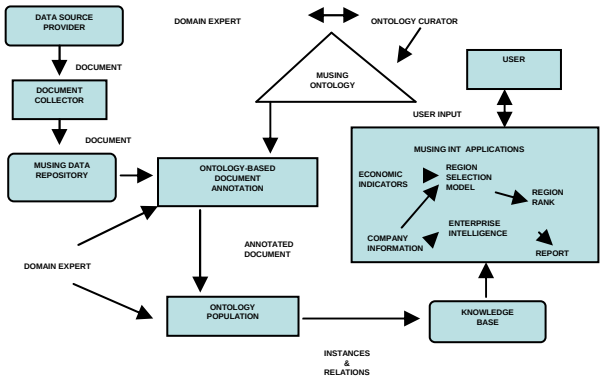
3. Ontology-based Information Extraction in MUSING

In the former chapter, we presented in some details the different types of MUSING ontologies, and the way they interact (mainly via the "Extension" ontology). This model of the relevant concepts for a set of Business Intelligence applications has to be filled (or populated) with real data, so that the applications can make use of the semantic capabilities of such an ontology infrastructure. We call this task "ontology population", which in a sense is Information Extraction (IE) guided by ontologies, the results of IE not being displayed in the form of templates, but in knowledge representation languages, e.g. OWL in the case of MUSING. The information stored in this way is considered as

"instances" of the concepts and relations introduced in the ontology. The set of instances is building the knowledge base[1] for the applications, and this knowledge base is supporting for example credit institutes on their decision-making procedures on credit issuing issues.

As we mentioned in the introduction, a substantial amount of the needed information for the development of semantic business intelligence applications is to be found in unstructured textual documents, so that the automatic ontology population task is relying on natural language processing in general and Information Extraction in particular.

In the following figure, the reader can see the interaction between textual analysis, information extraction and the semantic resources (ontologies) in MUSING, whereas the examples of classes presented in the box on the right side of the figure are taken from the Internalisation ontologies developed in MUSING:



.

3.1. (Temporal) Ontology population from News articles in the financial domain

Temporal information is given in news articles at two levels: date of publication and within the article reporting on issues related to the financial domain (we concentrate here on this application domain of MUSING).

We can make use of the typical layout of electronically available newspapers, with a title, an abstract, the names of authors and the publication date, etc., for extracting information with high accuracy from this "structured" part of the document (see for example RSS feeds). Important is here the publication date, which is offering an anchoring temporal index for the interpretation of temporal expressions present in the article itself.

In the content part of news articles, the reader can find a large variety of temporal expressions. This variety is partly motivated by the habits of journalistic style to use as much as possible variances in their formulations on facts. So writing about a CEO of a company, the journalists will use maximally one time the "dry" and precise formulation "Mr. X, Ceo of Y limited", and will in the text refer to this person mentioning other properties, like her/his nationality, age, region or city of birth, or even character properties, like "the ambitious Swiss manager …". This is similar with temporal expressions, where journalists also make use of metaphors and says. One thing being quite sure: authors will very seldom use the ISO standard for writing down a date in their article!

In the following we focus on some examples of natural language expressions in newspapers articles, and show how MUSING NLP components deal with those expressions. We take here as example news articles, from different providers, reporting on one person, Sergio Ermotti.

In one article, the title of the news and the publication date are:

Title: Ermotti statt Jentzsch (*Ermotti instead of Jentschz*)
Pubdate: 16.12.2005

The date is detected (in our case with the DFKI Information Extraction tools "SProUT), and annotated as a temporal_unit:

```
<F name="TEMPORAL-UNIT"><FS type="temporal unit"></FS></F>
<F name="YEAR"><FS type="2005"></FS></F>
<F name="MONTH"><FS type="12"></FS></F>
<F name="DAY"><FS type="16"></FS></F>
<F name="MUC-TYPE"><FS type="date"></FS></F>
```

This (simplified) feature structure representation tells us that the system found a "temporal unit" (equivalent to the MUC-TYPE="date") and the temporal expressions is being split into its components (Day, Month and Year). We associate then this temporal information with the "PubDate" feature.

Concerning the title of the article; the Information Extraction system also detects that both Ermotti and Jentzsch are person names, on the base of either the use of a gazetteer, the MUSING knowledge base or heuristics:

```
<F name="SURNAME">
      <FS type="Ermotti"></FS></F>
<F name="SURNAME">
      <FS type="Jentzsch"></FS> </F>
```

Further information about those persons, like nationality and age can be detected by SProUT in the text (if the information is present at all), or is available in the knowledge base (or in a gazetteer). But the Named Entity Detection system will not get the particular relation between both persons as this relation is expressed in the title: "Ermotti instead of Jentschz". We need here some lexical semantic information about "statt" (*instead*) in order to detect that there is a kind of competition between the two persons,

So far we got the information that at a certain date it is reported on a decision taken between two persons. The abstract of the article is giving us more information:

"Die italienische Großbank Unicredit hat Sergio Ermotti mit sofortiger Wirkung zum Chef der Sparte internationale Großunternehmen und Investmentbanking ernannt. Ursprünglich war der Posten für den ehemaligen HVB-Manager Stefan Jentzsch vorgesehen." (*The large Italian Bank Unicredit has named Sergio Ermotti with immediate effect as the head of the Branch "Large International Companies and Investment Banking". Originally was this job foreseen for the HVB Manager Stefan Jentzsch.)*

Two temporal expressions are included in this abstract: "mit sofortiger Wirkung" (*with immediate effect*) and "ursprünglich" (*originally*). And here our approach consists in first in linking those underspecified time expressions to the publication date. So that we have the relation:

[1] But to be quite correct here, we should mention that the "knowledge bases" are constituted by the ontologies together with their sets of instances.

t = 2005-12-16: <Ermotti, headOf, Unicredit Branch "Large International Companies and Investment Banking">

The actual value of *t* might not be the exact date at which Ermotti is starting (we expect that some time interval will exist between the event itself and its announcement by a press organ). But it is a very good approximation.

In the main text of the article, then more details are given, for example about Ermotti:

"Ermotti arbeitete früher kurz für den weltgrößten Finanzkonzern Citigroup und danach 17 Jahre lang bis 2004 für die Investmentbank Merrill Lynch." (*Ermotti have worked before for a short time for the world largest financial concern, Citigroup, and afterwards for 17 years, till 2004, for the investment bank Merrill Lynch.*)

This is a quite interesting sentence, since it contains a lot of temporal expressions (actually a quite normal fact in news articles). The first two expressions ("before" and "a short time") are again very vague. So here we assume that the before is actually "before the pubdate". The next temporal expressions are "for 17 years" and "till 2004". In those two expressions we get now more precise information: The relation "Ermotti works_at Merrill Lynch" is first associated with the duration of 17 years, and in a second step we can calculate the starting point of this relationship since an ending point is given: 2004 (we allow for such under-specification in the time ontology, having introduced a class called "yearDate"). In order to extract this information and to populate the ontology we need here a deeper linguistic analysis. We extract with the help of syntactic analysis (and more specially dependency analysis) that there is a working relationship between Ermotti (as the subject of the first part/clause of the sentence) and Merril Lynch. We can associate the time code to this relationship on the base of the dependency analysis of the two temporal expressions as linguistic expressions that "modify" the main verb "arbeitete" (worked).

The name of the company for which Ermotti is working is included in a prepositional phrase (PP). The linguistic pattern "[$_{\text{NP-SUBJ}}$ X] works [$_{\text{PP}}$ for [$_{\text{NP-IOBJ}}$ Y]]"is a very good candidate for a mapping into a relation <X is_employed_by Y>. But clearly the constraints that apply to both "X" and "Y" are, that the first is an instance of a person and the second an instance of a company (*domain* and *range* of the relation).

In this example, the reader could see how the constituent analysis of text, coupled with named entity detection, some lexical semantics and dependency relations, is guiding the ontology population.

In the pseudo formal representation below, we use for ease of presentation, the reader can see how the tools put the information together[1]:

SUBJ(Ermotti) arbeitete TEMP_ADV1(früher) TEMP_ADV2(kurz) für-IOBJ1 (Citigroup)
==> SUBJ-related(job_Position) to IOBJ1
==> INTERVAL1 = TEMP_ADV1 (here OpenIntervalLeft)
==> INTERVAL2 = TEMP_ADV2 (here OpenInterval) & INTERVAL2 ⊆ INTERVAL1

SUBJ(Ermotti) arbeitete TEMP_NP-TEMP1(17 Jahre) PP-TEMP (bis 2004) für IOBJ2 (MerrilLynch)
==> SUBJ-related(job_Postion) to IOBJ2
==> INTERVAL1 = NP-TEMP (here OpenInterval)
==> INTERVAL2 = PP-TEMP (here OpenIntervalLeft, DATE_TYPE = year representation/date representation) & INTERVAL2 = RightParanthesis of INTERVAL1

In this example we can also see that there are at least three syntactic ways to express temporal information; as an Adverb, an NP and a PP.

First the textual analysis gives a linguistic structure to the unstructured text, on the base of which we define a mapping, which associates the name of the person to the person ontology and the name of the company to the company ontology. The relationship "<Errmotti, is_employed_by, Merril Lynch>" can then be associated to the time slice "1987-2004".

From this individual news article under consideration we can not extract information about activities of Ermotti in the time between 2004 and 2005-12-16, but we assume that he had an activity in the banking domain. We can thus automatically query for documents telling us something about "Ermotti" and "Year 2005", in order to "fill the temporal gap" in the information card about Ermotti. The already extracted information and the temporal ontology of MUSING are structuring the semantic content of the query. On this base we found for example an article of the Handelsblatt, published on the 2006-12-06, one year later.

4. Conclusion

In this short paper, we have been showing how Human Language Technology, in close collaboration with Semantic Web resources and tools, can help in creating knowledge bases in the field of Business Intelligence applications, "upgrading" thus the actual strategies implemented in this field, building on quantitative information and statistical models, towards a new generation of semantically driven Business Intelligence methods and tools. We concentrated on the representation and extraction of temporal information, since this is a crucial topic for the applications within the MUSING Project.

5. Acknowledgements

The research described in this paper has been partially financed by the European Integrated Project MUSING, with contract number FP6-027097.

6. Short References

[1] Ivan Terziev, Atanas Kiryakov & Dimitar Manov. D1.8.1 Base upper-level ontology, Guidance1, EU-IST Project IST-2003-506826 SEKT, WP1, D1.8.1, 2005,

[2] Theodore Sider. Four Dimensionalism. Philosophical Review 106, 197–231, 1997.

[3] Gruninger, M., & Fox, M. (1994). *The role of competency questions in enterprise engineering*. Paper presented at the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway.

[1] „OpenIntervalLeft" and similar expressions in the example are classes of our time ontology.

Improving Third Generation Translation Memory Systems Through Identification of Rhetorical Predicates

Ruslan Mitkov [1], *Gloria Corpas* [2]

[1] University of Wolverhampton
[2] University of Malaga
r.mitkov@wlv.ac.uk, gcorpas@uma.es

Abstract

While number of Translation Memory (TM) programs and tools have been developed which are now regarded as indispensable for the work of professional translators, it has been noted that a serious weakness of the current TM technology is the fact that its matching capability is far from perfect. An obvious shortcoming of current TM systems is the fact that they have no access to the meaning of the translated text and operate on its surface form. As a result, they fail to match sentences that have the same meaning, but different syntactic structure. To overcome this shortcoming Pekar and Mitkov (2007) developed the so-called 3rd Generation Translation Memory (3GTM) methodology which analyses the segments not only in terms of syntax but also in terms of semantics. Whereas this technology is a promising way forward, the limitations of current semantic processing may cast a doubt on its use in a practical environment. To enhance the overall low performance of semantic processing tasks, we propose the employment of rhetorical predicates to improve the accuracy of the matching algorithm. The paper will introduce the novel 3GTM developed by us and will show how rhetorical predicates can be used to enhance its performance.

Index Terms: Translation Memory, 3rd Generation Translation Memory, Translation Memory, matching algorithm, semantic processing, rhetorical predicates, Natural Language Processing.

1. Shortcomings of traditional Translation Memory systems and the development of 3rd Generation Translation Methodology

Translation Memory (TM) systems have emerged as highly successful translation aids for more than a decade. TM has proven to be efficient and time-saving technology for voluminous translations especially of technical texts featuring a degree of repetition with previously translated texts. While a number of TM programs and tools have been developed which are now regarded as indispensable for the work of professional translators, it has been noted that a serious weakness of the current TM technology is the fact that its matching capability is far from perfect.

As noted by Mitkov (2005), traditional TM systems would fail to return matches for parts of sentences which would match on their own. By way of example, if *Select 'Shut down'* has been previously translated, that a translation of *Select 'Shut down' from the menu* would not return any partial matches. Similarly, a complex sentence consisting of two previously matched simple sentences, would not offer any matches. As an illustration, if *Select 'Shut down' from the menu* and *Click on 'Shut down'* match, *Select 'Shut down' from the menu and click on 'Shut down'* would not be returned as a match.

A more serious shortcoming of current TM systems is the fact that they have no access to the meaning of the translated text and operate on its surface form. As a result, they fail to match sentences that have the same meaning, but different syntactic structure. The fact that the same semantics can be expressed in a variety of linguistic forms has a number of important consequences for the practical use of TM systems. By way of example, none of the TM systems would be capable of matching *Microsoft developed Windows XP* with *Windows XP was developed by Microsoft* or matching *The company bought shares* with *The company completed acquisition of shares*.

To overcome this shortcoming Pekar and Mitkov (2007) developed the so-called *3rd Generation Translation Memory* (3GTM) methodology[1] which analyses the segments not only in terms of syntax but also in terms of semantics. The authors adopt the so-

[1] SIMILIS, developed by Lingua et Machina (Planas, 2005), was referred to as "Second generation TM software". It does morpho-syntactic analysis (chunks) and holds translation units corresponding to chunks: e.g., as with parallel concordancers, matching at sub-sentence level is possible. By way of example in the translation of 'Les mémoires de seconde génération changent le monde de la traduction' into 'Second generation memories change the translation world', the match of the noun phrases *mémoires de seconde generation* and *second generation memories* is possible. Similar work has been reported in Grönroos and Becks (2005) and in Hodasz and Pohl (2005).

called syntax-driven semantic analysis by performing linguistic processing over trees graphs (Graehl and Knight 2004; Szpektor et al. 2004) followed by lexico-syntactic normalisation. Then similarity between syntactic-semantic tree graphs is computed and matches at sub-sentence level are established, using a similarity filter and node distance filter (Pekar and Mitkov 2007).

Phase 1 of this project covers the implementation and evaluation of the new matching technology through the development of a matching algorithm including the pre-processing modules, the integration with WordNet and the development of a lexical paraphrase module. Phase 2 includes the integration within Wordfast and a translator's evaluation.

Lexical resources such as WordNet are needed to identify automatically synonyms and therefore, to make the match of synonymous expressions possible. By way of example, if 'Unplug the hoover' has already been translated, then of 'Unplug the vacuum cleaner' would show as match too as *vacuum cleaner* would be labelled as a synonym of *hoover*. In addition, generalisations would be possible as WordNet would return *machine* as a superordinate of *vacuum cleaner* so a high match for 'Unplug the machine' would be returned. Similar generalisations would be achieved for patterns with identical or seminal semantic classes such as 'John Smith few to Brussels on February 3rd,' or 'Dr Johnson flew to Rome on 7 January 2006' or even more generally '<person-m> flew to <city> on <date>'. Finally a lexical paraphrase resource is made use of to establish equivalences between expressions such as 'Microsoft developed Windows XP' vs. 'Windows XP was developed by Microsoft' or 'The company bought shares' vs. 'The company completed acquisition of shares'. For more complicated cases such as matching 'As a result of John Smith's resignation, the values of the company shares plummeted' with 'John Smith resigned and this resulted in a sharp decrease of the values of the company shares', textual entailment techniques may be needed.

2. Limitations of the 3rd Generation Translation Memory Methodology

The 3GTM methodology proposed by Pekar and Mitkov (2007) is a promising way forward to ensure that translators have wider range of matches. However, this methodology is still a long way from operating in a practical environment and in particular, from being commercialised. A major issue is whether the new technology would deliver in a robust and scalable way. Even though contingency plans include the implementation of a subset of the techniques, Natural Language Processing (NLP) in general is far from perfect. With particular reference to the semantic tasks involved, the accuracy could be as low as 60% for semantic role labelling or even lower for anaphora resolution (see the last example from the previous paragraph). Therefore, different techniques should be sought to enhance the success rate of semantic processing.

3. A novel rhetorical predicates-driven methodology for 3rd Generation Translation Memory systems

Our presentation will discuss a new project whose objective is to enhance the matching performance when comparing semantically equivalent sentences through the identification of rhetorical predicates[1]. Research in discourse, text generation and text summarisation has already shown that different types of texts feature schemata of rhetorical predicates which account for the stereotypical discourse structure of the specific type of text. By way of example, a research paper normally features among others, rhetorical predicates such as *topic*, *background*, *methodology*, *solution* and *conclusion*. While such a list of rhetorical predicates is genre- or domain-specific, we argue that the identification of rhetorical predicates will assist the establishment of equivalence of sentences. The methodology consists of boosting the confidence/probability that two sentences are semantically equivalent if these sentences are labelled with the same rhetorical predicates or vice versa. The identification of rhetorical predicates will be carried out on the basis of regular expressions containing keywords representative of a specific predicate.

The presentation as well as the final version of the paper will introduce the novel 3GTM methodology developed by us, describe the above experiments and will report on the evaluation of the matching performance of a 3rd Generation Translation Memory system – with and without a module for identification of rhetorical predicates. The experiments will be restricted to specific genres due to the nature of rhetorical predicates. As a convenient background, a comparative assessment of current MT systems (e.g. Déjà vu, Trados, Wordfast, etc.) performance will be also provided. We intend to give the presentation in an easy-to-follow and accessible manner which will be based more on translation/matching examples rather than technicalities, while elaborating on technical issues in the paper itself.

References

[1] Graehl J. and Knight K. 2004. "Training Tree Transducers". *Proceedings of NAACL/HLT-2004*. Boston, MA.

[2] Hodasz G. and Pohl G. 2005. "MetaMorpho TM: A Linguistically Enriched Translation Memory". *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-05).* Borovets, Bulgaria.

[1] Termed also *moves*.

[3] Mitkov, R. 2005. “Panel Discussion: The Future of TM Technology”. *27th International Conference on Translating and the Computer (TC27)*. London, UK.

[4] Pekar V. and Mitkov R. 2007. “New Generation Translation Memory: Content-Sensitive Matching”. *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*, 29-30 September 2006, Bern.

[5] Planas, E. 2005. “SIMILIS - Second generation TM software”. *Proceedings of the 27th International Conference on Translating and the Computer (TC27)*. London, UK.

Language Engineering for Basque in a Visual Communication Technologies Context

Maider Lehr[1], Kutz Arrieta[1], Andoni Arruti[2]

[1]VICOMTech Research Centre,
Mikeletegi Pasealekua 57, 20009, Donostia-San Sebastian
[2]Signal Processing Group, University of the Basque Country,
Manuel de Lardizabal Pasealekua 1, 20018, Donostia-San Sebastian
mlehr@vicomtech.org, karrieta@vicomtech.org, andoni.arruti@ehu.es

Abstract

The integration of language engineering in other applications is gaining support in European research centers and government agencies dedicated to the creation and management of research resources. In this context and given the particular suitability of the Basque Country to understand and promote this type of development and integration, the Basque Government and other institutions are making the necessary efforts and have considered this as one of their most relevant lines of development for the coming years.

In this context, VICOMTech, an applied research center located in Donostia-San Sebastian (Basque Country) has opened a new emerging area in language engineering and intends to integrate in the other areas that the center develops. Therefore, the inclusion of Natural Language Processing devices within applications developed in Digital TV, Multimedia services, Biomedical Sciences, Industrial Applications, and Human-Computer Interaction, will offer added value and contribute to the intelligence of these applications.

This paper is intended to inform the reader about the efforts VICOMTech is making to develop this approach and reports on some of the research already done in the field of speech, in which VICOMTech has already some experience.

Index Terms: language engineering, speech technologies, Basque

1. Introduction

One of the tasks that Language engineering (Computational Linguistics) tries to accomplish is; overcoming the linguistic constraints imposed by the variety of the existing human languages, allowing therefore everyone to use his/her native language to interact with the technology, improving accessibility to the Information Society.

New markets and research interests are opening in the area of Natural Language Processing and Speech. Thanks to the emergence of these new markets this field is making considerable advances. Minority languages cannot afford to stay out of this race, since their survival may depend on this, among other factors. Basque does not count on the resources and devices that have been created for widely spoken languages, but in the last few years Basque has made important advances in the integration to these technologies. Due to the small market for Basque, it is very important, on the one hand, to take actions to raise the demand promoting both the knowledge and the market of existing products. The market must offer in Basque the applications that are available in majority languages such as Spanish, French or English. On the other, we must have a multilingual product to market to the world.

The science, technology and innovation plan for 2007-2010 proposed by the Basque Government makes special mention of Language Engineering. In reference to this, we must mention AnHitz, a strategic project promoted and partially financed by the Basque Government, where VICOMTech [1] is the leader and Elhuyar Fundazioa [2], Robotiker [3], IXA (University of the Basque Country) [4] and Aholab (University of the Basque Country)[5] are partners.

VICOMTech is an applied research center located in the technological park of Miramon in Donostia. It is a non-profit association founded by the INI-GraphicsNet Foundation and Basque Television, Radio and Broadcasting group (EITB) [6].

As mentioned above; one of the recently created strategic research interests in VICOMTech is the promotion of linguistic technologies in multilingual environments, with particular attention to the integration of Basque. AnHitz's goal is the development of linguistic technologies in Basque to enable human-machine interaction, as well as, knowledge management and other applications using this language. The newly created linguistics department is conceived as a transversal technology that should be integrated into and complementing the already existing research fields at VICOMTech.

- Digital TV and Multimedia services: this area specializes in transmission and interactivity standards (DVBT, C,S,H - MHP, etc.), A/V content analysis and management and virtual/augmented reality services for broadcast professionals. The department has some experience in user interfaces for television. Making use of this experience and in order to take advantage of the interactivity offered by the standards used, linguistic technologies will be integrated. First, the know-how in A/V content analysis and management will be complemented with the linguistic tools, developed in order to expand the possibilities for application environments.
- Biomedical applications: this department concentrates on research and development for the healthcare and biotechnology sectors. The main research lines include the most recent advances in image analysis (image processing, segmentation, registration and fusion), visualization (virtual reality and augmented reality), and biomedical information management (transmission, representation, standards and interface). This department is envisioning the integration of linguistic components and

ontologies treatment, along with medical imaging interpretation, to explore Indexation and Retrieval applications for medical and biomedical data.

- Tourism, heritage and creativity: this department designs and implements applications for the creation of interactive digital experiences, providing an added value to the services offered by the tourist, cultural, and creative sectors. Among the key technologies we will mention the following: Virtual and Augmented Reality technologies, mobile applications based on location-based tracking tools for content personalization, semantic-based searching algorithms, content annotation and indexing, and standard based system for multimedia content creation and management. This area has considerable experience in ontology standards treatment and indexing, which has already included linguistic tools for one of its projects with a Basque repository of multimedia patrimonial contents.
- Interaction for education, leisure and e-inclusion: this area develops technologies related to multimodal human - device (PC, PDA, mobile phone or TV) interaction through body and natural language. Here is where speech technologies, face and body animation of the virtual characters, emotional interaction, and natural language processing, are hosted.
- Industrial applications: this department concentrates on industrial applications for VICOMTech's main technologies, such as, interactive 3D computer graphics, Virtual and Augmented Reality, advanced visualization devices, and interactive simulation in environments, for industrial design, manufacturing, and commerce. This area is also, developing an application integrating semantic information and image processing for industrial design, brand, and information monitoring.

2. Speech related research lines

The Speech group in VICOMTech is part of the interfaces research group and Speech is perceived accordingly, as an interface. VICOMTech does not work in the development of synthesizers or recognizers, our goal is to adapt, modify, and/or extend, existing technologies to our context. Given the fact that we work with Basque, we quickly enter the realm of research by having to convert existing technologies to a language with little resources and of a "one of a kind" type of language, quite different from the majority languages.

2.1. Television driven technologies

2.1.1. Subtitling

Subtitling by means of open or closed captions is, for millions of hearing-impaired people, the most useful representation means for speech content in TV and other audio-visual media.

However, most Spanish TV channels provide subtitles as closed captions in teletext format only for some of their prerecorded programs. For live programs, such as sports events and news broadcasts, subtitles are rarely available. Specially trained stenographers and fast typists for live subtitling are expensive.

There are some software solutions for ASR-based live subtitling, e.g. Protitle Live (NINSIGHT) and WinCAPS (Sysmedia) allowing a trained speaker to dictate live subtitles into a trained ASR system ("re-speaking"). Nevertheless, there is no ASR-based system in use for fully automated subtitling.

We studied the feasibility of using dictation for literal transcription of speech for fully automated live subtitling bypassing the re-speaker [7]. To do this, we integrated commercial best-of-its class ASR software with a professional subtitle generator and preprocessing modules we developed . We evaluated the quality of ASR for Spanish, measured the delay between speech and subtitle, and detected particular drawbacks in the components used for subtitling.

Evaluations of the video by means of questionnaires were performed by nine volunteers (aged 25 to 65 years from different social groups) from organizations of deaf and hard of hearing people in Donostia-San Sebastian (Spain).

An objective quality measure of the transcription obtained by the prototype is the word recognition rate, WRR = (H - I) / N, where N is the number of words in the reference, I is the number of the insertions, H is N - (S + D), the number of correctly recognized words, being S the number of substitutions, and D the number of deletions.

Furthermore, we measured the time delay between speech and the appearance of the corresponding subtitle.

We concluded that ASR systems need improvements in:

- Speaker and material independence. It only works well when trained to recognise a single voice (acoustic models) and when trained previously with material related to the contents of the programs (language models and dictionaries).
- Current methods for noise filtering, speech, music and silence detection.
- Speech recognition for multiple speakers.
- Lack of automatic punctuation.

We are, therefore, working on:

- Automatic punctuation. Several algorithms will be developed to segment the audio signal in significant utterances. Three lines will be studied:
  - Detection methods of segmental borders focussing on acoustic features.
  - Detection methods of segmental borders focussing on linguistic features.
  - Detection methods of segmental borders focussing on prosodic features
- Speakers discrimination and identification. Usually, in speaker recognition tasks the speakers to be recognized are known. In other applications, like this one, neither the identity nor the number of speakers are known in advance. The speech signal stream is continuous. Speaker changes must be detected and the stream segmented. This must be done off-line.

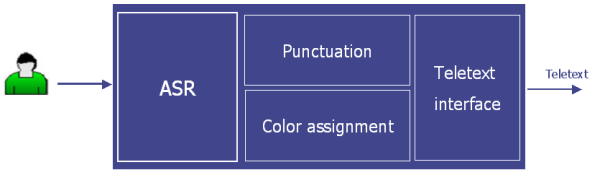


Figure 1: *Subtitling generation system.*

- Automatic generation of captions. We are exploring the possibility of automatically performing all the linguistic

changes that trained subtitlers introduce in the text in order to make it really understandable for hearing-impared viewers.

2.1.2. *Voice Transformation*

Voice Transformation is relevant for a wide range of speech related applications such as:

- Speech processing
  - TTS systems
  - Restoration of old recordings
  - Training of speech recognition systems
  - Speaker verification/ identification systems
- Multimedia/ music
  - Karaoke
  - Voices for virtual environments/ chats
- Dubbing and looping
  - Preservation of the original voice of the actor/ actress
  - Increase of voice registry
  - Looping

Dubbing companies need experienced and qualified people, with some specific features (age, gender, voice depth). Dubbing requires specific skills, with specific registries, a specific voice quality, as well as a good dramatization abilities. So, difficulties to get new voices are considerable. For example, often, adult voices, mostly female adult voices, are used to dub children voices. Dubbing companies must overcome this challenge: to find solutions for this lack of available voice registries. This voice and registry limitations are more noticeable in minority languages.

In this context we proposed to develop a prototype with a friendly interface. The interface provides the possibility of changing a source voice to a target voice in an intuitive and friendly way. We are working with some Basque dubbing companies which work closely with EITB (Basque Television) to create a system capable of generating different voice registries for dubbing in TV and cinema [8].

2.1.3. *Synchronization of audio and animation*

In human-machine interfaces there is a clear trend to merge different possibilities of presenting information, in particular, speech and facial animation. The presence of minority languages in the area of virtual characters is very limited. This is obviously due to the lack of both resources and tailored technology. We have developed a system capable to produce suitable data for the animation of faces from natural voice in Basque using open source technologies [9]. The output of the speech analysis performed matched a set of visemes (visual representation of the phoneme) and phonetic data, corresponding to the lip visualization of the virtual character for each frame of the animation. This output was used to synchronize the animation with on-line audio in real time. The application captures the speech signal from the input through a sound card and identifies the appropriate phonemes. As phonemes are recognized, they are mapped to their corresponding visemes. The virtual character is then animated in real time and synchronized with the speaker's voice. The application consists of three modules Fig. 2:

- The phoneme recognition system.
- The module that sends the input audio to the recognition system.
- The communication interface between the recognition system and the animation platform.

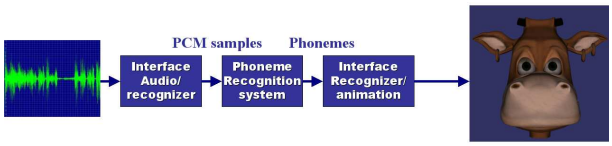


Figure 2: *Facial animation for live speech input in Basque.*

The goal was to obtain a useful and usable application in the television domain, where virtual presenters are more and more common. A quiz type TV program for children using a virtual character was created. This character is presently running in a popular Basque TV (EiTB) program, in which children answer questions and interact with the said character. In this case, a cow. Results are satisfactory for this first version [8].

The animation reacts to the caller's answers, it therefore, needs to run in real time. Lip animation runs as the actress who dubs the virtual character in the program speaks into the microphone. This voice is sent in real time to the software developed in this project. The software analyses the stream and generates the data to synchronize the lips with the audio. This information is interpreted by the animation engine.

2.2. Other applications

2.2.1. *Multimedia Information Retrieval*

The very essence of an information retrieval (IR) system is to satisfy the user's information needs, expressed by a query submitted to the system. There is now widespread use of information retrieval (IR) techniques to access information stored in electronic format. One of the most widely used examples of this is in internet search engines.

There is also a great deal of information in audio format. One such area is the audio associated with radio and television news broadcasts. If these audio sources could be transcribed automatically then the information they contain can be indexed and relevant portions of broadcasts retrieved using conventional IR techniques. In collaboration with EITB, the Basque radio, television and broadcasting group, we are working in a project which consists on developing an information retrieval system to efficiently access multimedia contents.

The approach is to combine speech and more "traditional" IR techniques to get to the desired results, in order to manage and use large catalogs of multilingual multimedia contents. Among other techniques (image analysis, etc.) we are going to use speech recognition systems to convert the spoken word to a text transcription which is then passed on to a regular text-based IR search engine. Speech recognition technologies in Basque are far from the existing technologies for other languages and we are not in the business of developing such technologies. We need to adapt to the situation and, therefore we will proceed in different scenarios, integrating separate applications:

- CLIR: using existing partial Machine Translation applications for Basque we will preserve the ability of the user to perform the query in Basque and receive:

  - Results in Basque and other languages when dealing with parallel text.
  - Results in Basque and other languages (the ones in Basque being of poorer quality) when not dealing with parallel text.
  - Results in languages other than Basque

  In principle, these other languages are Spanish, French and English.
- IR for Basque combining several sources: sound, image recognition and text.
- IR for Basque accessing the Internet or only the Basque Science and Technology repository using speech only as an interface.

We also plan to include some type of mutual feeding of the text, image and audio contents of these catalogs.

2.2.2. *Autopuntuation*

As mentioned before, we are exploring the possibility of generating sentence punctuation automatically combining speech and linguistic technologies. This is quite an ambitious project and our wish would be that, for once, Basque would the first language having an application that other languages lack.

But this is not our only motivation, it is clear that such a development can, in the one hand, be extended to other languages, with promising commercial consequences, and generate knowledge on the linguistic and acoustic aspects of Basque and this type of application in general, in the other.

3. Conclusions

More and more users want and need to interact with devices in more natural and easy ways. Also, the amount of catalogs and knowledge databases has increased. This information is also stored in different formats and languages. This information needs to be stored, managed, retrieved and understood. Natural language Processing is an unavoidable component of the applications of the future. This field has gained great relevance in European research programs. The Basque Government, research centers, universities and public and private companies are showing interest in Speech and Natural Language Processing related applications.

We have tried to present here some of the projects we have worked on, mostly in Speech, and our present and future projects geared towards combining ”forces”: Speech, NLP, Image Analysis, Virtual Characters, Culture, Ambient Intelligence, etc., not only to promote social and industrial progress, but also to reduce the technological gap existing between Basque and other languages.

4. Acknowledgements

The projects mentioned in this paper have been partially supported by the Basque Government through the ETORTEK and INTEK programs of the Department of Industry and by the Spanish Ministry of Industry, Tourism and Commerce.

Furthermore, we would like to thank some our partners and clients such as EITB, Irusoin, Mixer, Baleuko, Talape, Elhuyar, Robotiker, Aholab, IXA, Antena 3 TV, AudioText and Ceapat.

5. References

[1] http://www.vicomtech.org/.

[2] http://www.elhuyar.org/.

[3] http://www.robotiker.com/.

[4] http://ixa.si.ehu.es/Ixa.

[5] http://aholab.ehu.es/aholab/Home/.

[6] http://www.eitb.com/.

[7] Obach, M., Lehr, M., Arruti, A., “Automatic Speech Recognition for Live TV Subtitling for Hearing-Impaired People”, 9th European Conference for the Advancement of Assistive Technology in Europe (AAATE 2007), Assistive Technologies Research Series Volume 20, 2007, pp. 286-291.

[8] http://www.vicomtech.es/castellano/html/videos_demos/index.html.

[9] Lehr, M., Arruti, A., Ortiz A., Oyarzun D. and Obach, M., “Speech Driven Facial Animation using HMMs in Basque”, Text, Speech and Dialogue, Proceedings Lecture Notes in Artificial Intelligence (Springer), 2006, pp. 415-422.

NATIVE LANGUAGE PROCESSING

A language processor for understanding languages compliant with the grammar of Hindi language and extension to a QA system

Anand Bora, Aman Kumar
B.Tech. (2006)
Computer Science & Engineering
SASTRA Deemed University
Thanjavur

September 10, 2006

ABSTRACT

This paper aims at developing a very basic NLP (Natural Language Processing) system for the Hindi language. Also this paper proposes a format which can make the system compatible with the languages supporting the grammar of Hindi Language. This includes languages like Punjabi, Guajarati and the different type of dialects in which Hindi is spoken in different parts of India. The flexibility is possible due to the flexible word structure proposed. Due to the closeness of the system with the Indian Languages, the project has been named as **NATIVE LANGUAGE PROCESSING**. The project is based on XML file access and parsing. In addition to all the conventional parts an NLP system uses, the system has an optional answering part for generating a proper response to the given output. The system has been modelled in such a manner that it can process a given sentence and generate the outputs at different levels of the language processor. Moreover the structure of the system makes it possible to be compatible with most of the modern languages.

1. INTRODUCTION

Proposals for mechanical translators of languages pre-date the invention of the digital computer. The first recognisable NLP application was a dictionary look-up system developed at Birkbeck College, London in 1948. American interest is generally dated to a memorandum written by Warren Weaver in 1949. His idea was simple: given that humans of all nations are much the same (inspite of speaking a variety of languages), a document in one language could be viewed as having been written in code. Once this code was broken, it would be possible to output the document in another language. From this point of view, German was English in code. As a research idea, this caught on quickly. The key developments during the 1980's were in the fields of Augmented Transition network, Case Grammar and Semantic representations. Thus, a way was found to get around the semantic information bottleneck. Though a lot of research has been carried out after that, a generic system which should be capable of processing most of the Indian languages has not yet been developed. We try to find a way of for this by making a generic system

which will be capable of processing most of the Indian languages.

Anusaarka Project

This project is the project meant for Morphological Analysis of Hindi Language. Joint research operations are performed in the field of NLP related to Indian Languages by different professors and scholars under this project. We have taken massive references from this project and project's site. This source can also be used in the morpheme analysis phase.

2. LANGUAGE PROCESSOR STRUCTURE

The project (paper) is based on XML file access and parsing. It consists of all the conventional parts of a language processor *viz.* Lexical Analyzer (for tokenization, word validation and morpheme analysis), Syntactic Analyzer (Syntax Analysis done by Recursive Transition Network Parser), Semantic Analyzer (for POS tagging and disambiguation), Analyzer (for understanding the meaning) and an *optional* answering part for generating a proper response to the given input. The system has been modelled in such a manner that it can process a given sentence and generate the outputs at different levels of the language processor.

The system is a platform independent system. The main programming language used for the development of this system is C++. Different modules generate different outputs which are fed into the final analyzer stage. The important part of the system is the XML file structure. Various Xml files are needed to keep track of the words in the language and also the grammar codes. Since the analyzer system learns new words on runtime, the manipulation of these files is done adequately. In addition, different algorithms have been used to learn new words and reject inadequate words. Learning is done in the conventional manner of asking questions from the user. The system becomes more and more intelligent as it is trained through time. The fundamentals of learning in the project are deriving proper keywords for unknown words from the answers given by the user.

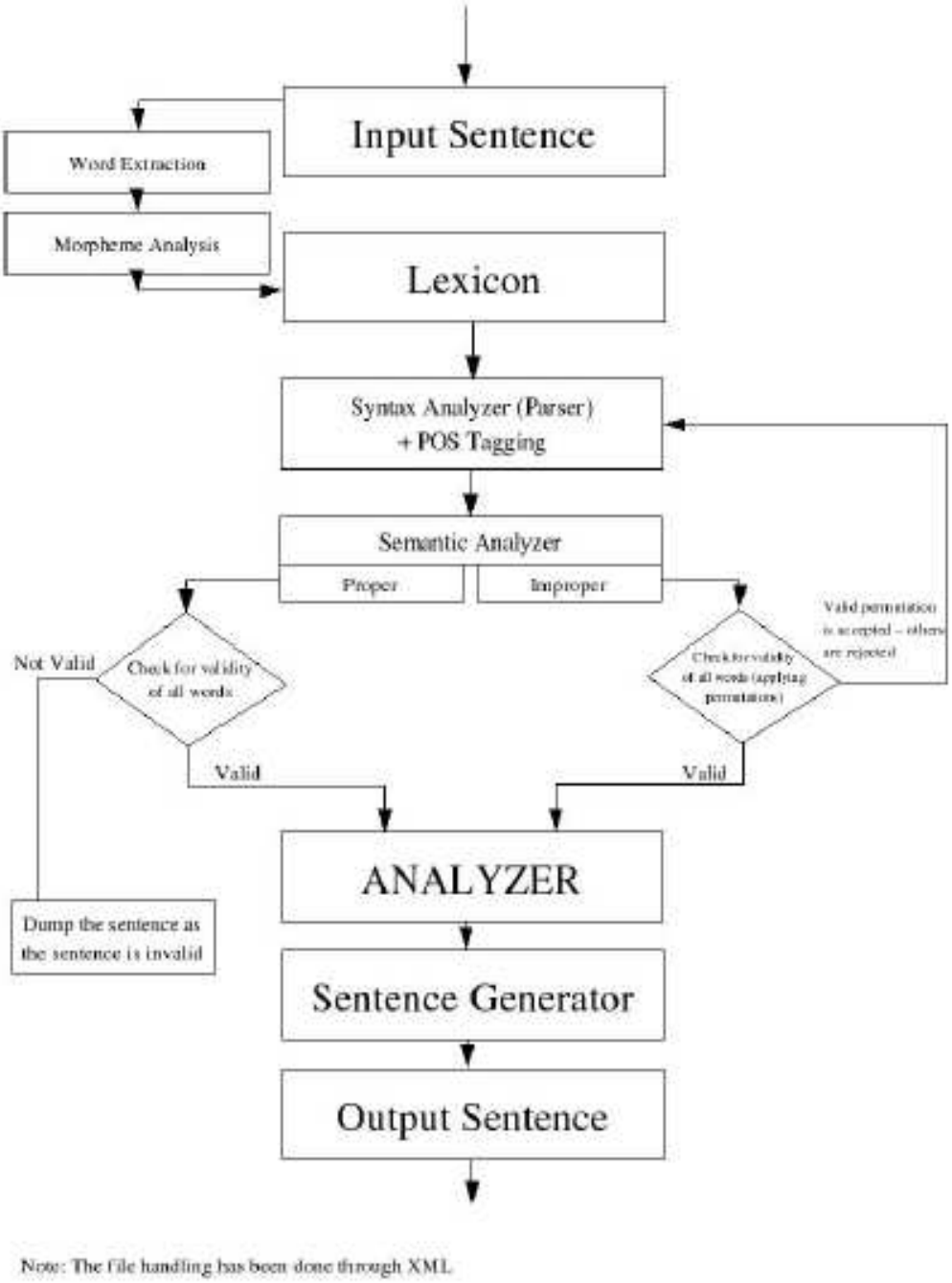


Figure 1.

The modules shown in Fig. 1 are described as follows

.

2.1 PROJECT IMPLEMENTATION DETAILS: Basic implementation

Two methods of morpheme analysis have been implemented, viz., Basic Implementation and the Anusaarka Project[1]. Our basic implementation scheme mirrors that of NLP - Akshar Bharati Book[2] with our modifications. The root words are extracted only from those words that are included inside the given word. Most of the words in our analyzer are mapped in this manner. We have used XML tags for the morphemes as well. This means that for each word there can be a morpheme entry in the XML wordlist file. If the given word is derived from the root word after adding the morpheme, then the word is considered valid and is already available in the wordlist. For any other case, the word is treated as an out of dictionary word. The general structure is shown in the Lexicon discussion section of the article.

The algorithm used for morphological analysis is that used by NLP - A Panian Perspective[2]. It is mentioned here

ALGORITHM

1. L:=empty set
2. If w is in DI with entry b
   then add b to set L.
3. For i:=0 to length w do
   let S=suffix of length i in w
   if reverse(s) in RST
   then for each entry b associated with reverse(s) in RST do
   proposed-root = (w – suffix s) + string added from root
   if(proposed-root is in dictionary of roots)
   then
   construct lexical entry 1 by combining features given for the proposed root in DR
   add 1 to L
   end for each entry
   end for I
4. If L is empty then return (“unknown word w”) else returns (L)

END ALGORITHM

2.2 UNKNOWN INVALID WORD REJECTION

This is a small part in which the system is trained beforehand. The engine is trained for a particular number of words in the language. This wordlist is passed through the engine. Then the transition matrix for the given set of words is formed. This transition matrix gives a pattern for the new words being fed into the system. Apart from this, a recursive filtering module is designed which further enhances the word filtering. This is done by adding a number of test cases.

This part can be modified by using new soft computing techniques. An example for this kind of word rejection is the following: Hindi language cannot have ‘xx’ anytime or anywhere. The transition value is always zero as we have trained the system for ‘n’ number of words. So when this type of string is encountered it is straightaway rejected. It is also passed through this filtering module.

2.3 WORD – LIST ARCHITECTURE

The rich wordlist that has been used in the project is implemented in XML.

Sample tree of the wordlist formed

```
- <wordlist>
    - <m>
        - <mota>
        - <type>
            <adjective />
            <noun />
        </type>
        <gender>M</gender>
        <number>S</number>
        <person>T</person>
        - <morphemes>
        - <i>
            <gender>F</gender>
        </i>
        - <ey>
            <number>P</number>
        </ey>
        </morphemes>
```

First of all indexing has been maintained according to alphabetical order i.e., all the words starting with alphabet 'a' has been tagged inside the alphabet 'a'. Inside the alphabet tagging various other tagging has been maintained such as gender, type, morpheme, noun, adjective, person, and so on. The XML architecture has been implemented in order to make parsing simpler. For the XML architecture of word list please refer to the code section.

2.4 SYNTAX ANALYSIS (PARSER)

Syntax analysis is a process in compilers that recognizes the structure of programming languages. It is also known as parsing. The parser used in the project is a simple *Recursive Transition Network*. This parser though not efficient is enough for parsing basic sentences. We have considered simple sentences and have eliminated the need for complex grammar evaluation. For designing the grammar of basic Hindi Language, we considered different basic sentences which have been enlisted later.

James Allen has described RTN parser beautifully in his book for Natural Language Understanding[3]. This can be described as follows: Simple transition networks are often called finite state machines (FSMs). Finite state machines are equivalent in expressive power to regular grammars, and thus are not powerful enough to describe all languages that can be described by a CFG. To get the descriptive power of CFGs, you need a notion of recursion in the network grammar. A recursive transition network (RTN) is like a simple transition network, except that it allows arc labels to refer to other networks as well as word categories. Thus, given the NP network in figure 1, a network for simple English sentences can be expressed as shown in figure 2. Uppercase labels refer to networks. The arc from S to S1 can be followed only if the NP network can be successfully traversed to a pop arc. Although not shown in this example, RTN's allow true recursion—that is, a network might have an arc labelled with its own name.

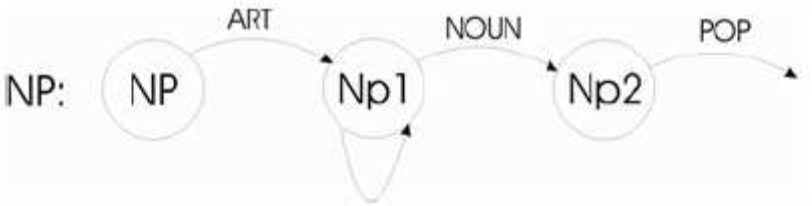


Figure 1

Figure 2

A separate parse code file is maintained which keeps track of the different grammars possible. For every call of the expandable Non Terminal, the RTN calls the related tag recursively. This means that at each call the next word is matched for its properties. If the category of the next word matches the given parse tree formed then the parser proceeds else it pops out returning error. We have used a very basic list of grammar code. This grammar code can be changed dynamically if grammar learning module is also incorporated in the system. The sample parse code structure which we have used is in the following manner:

SAMPLE PARSE CODE LIST

- <codes>
  - <S>
    - <code>
      <NP />
      <VP />
    </code>
    - <code>
      <N />
      <V />
    </code>
    - <code>
      <A />
      <ADJ />
      <N />
    </code>
  </S>
  - <NP>
</codes>

2.5 SEMANTIC ANALYSIS – OUR PROPOSED APPROACH AND PARTIAL IMPLEMENTATION

For any sentence provided, we locate the root verb in the sentence. Then this verb is mapped for the tense form in which it currently exists. This way we can identify the occurrence time of the activity (karaka). Also what has been done can be known by the root word of the verb. In the Anusaarka project[1], this phase is referred as the Vibhakti phase. We have implemented our own algorithm for the same. This root word comes from the XML wordlist provided. Thereafter, the worker (karta of the activity is identified. This can be a pronoun or a noun. We map the person (i.e. first, second or third) for the karta and identify the main worker. Then the karma is identified which can be another noun or pronoun. This is the other person involved in the conversation. This can also be an object. This is identified from the meaning tag of the Xml file. After all the three parts have

been identified, we move on to the karma phase. Though we have already located the root verb, the mapping is yet to be done. Now the system identifies that the karta and karma are related to each other through the karma part.

The system can be explained by the following example:

Raama ne raavana ko tiira se maaraa.
Ram ergative Ravan accus. Arrow instr. Killed.
(Ram killed Ravan with an arrow).

The other important point is the tracking of the postpositions. They are very important at pointing the exact world position and activity scenario. Each postposition (parsarg) has been tracked and the karta and karma are identified adequately. The different postpositions taken into consideration are:
'mein', 'ka', 'ki', 'ke', 'ko', 'se', 'ne', 'par' and 'tak'

For identifying the type of the sentences different signals to the system has been provided. These signals can be enumerated as:

Statement, Negation, Command, Question & Exclamation.

The signals are triggered only when the system identifies proper words related to the given sentence. For instance, a negation statement will always have any of the following three words in the sentence – 'na', 'nahin' or 'mat'. A question will always have words like 'kaun', 'kya' etc. Apart from this a question asked can also be checked by the existence of special symbols in the sentence. An example for this is the '?' in any question. Logical statements are built on the basis of understanding of the parameters talked about earlier. Two important phases of the proposed semantic analyzer are **PROPER** and **IMPROPER**.

The **PROPER** phase parses the sentence in the manner described earlier. The primary condition her is that the sentence must be valid in the parser phase. In other words, the sentence must follow grammar which complies with the available grammar code list.

The **IMPROPER** phase is for understanding those sentences which are structurally not correct but they are semantically correct. We have considered a very basic implementation of this phase too. In this phase, the grammatical construction of sentence is taken. Then different valid permutations are applied on the grammar. Thereafter, the new structured grammar is compared with already available grammar rules. If the rules are available then they are compared and if valid, the sentence is declared valid. Otherwise, the sentence is checked for all possible cases. The sentence is dumped if none of the case matches. The improper phase can also be expanded to incorporate lots of other features.

2.6 PROPOSED PART FOR IMPLEMENTATION ANALYZER PROCESS

This process was earlier intended for both word learning and grammar learning. Word learning has been done in the conventional method of asking questions. The system asks the user about any unknown word. The system maps some keywords from the given response and searches its wordlist for the word. If it finds relevant data available it

stores the word in the Xml file with the meaning tag containing the keywords providing the meaning of the word. This word is stored at the proper tag in the Xml file. This means that the searching is much faster.

This idea can further be classified by the following conversational scenario:

User: main toffee kha raha hoon.
System: mujhe 'toffee' ke bare mein nahin janta.
User: Toffee ek khane ki cheez hai jo ki bahut meethi hoti hai.
[Translation in English]
{
User: I am eating toffee.
System: I don't know about 'toffee'.
User: Toffee is an edible thing which is sweet.
}

In the above case 'khane', 'cheez' and 'meethi' are available in the wordlist. The words have their own respective meanings which provide the required information for the current word being processed. An important point for the analysis is deciding about the noun on runtime. Certain words form different type of relations with real world entities. A typical example can be the 'instance' relation. In the above example toffee will be an instance of object having the different properties. Thus the system will always keep in mind the different aspects about the object. For storing the meaning of the particular word, the word entry is done in the Xml wordlist in the format referred earlier in the lexicon section.

If we talk about making the system generic and extending it to different compatible languages, the constraint which comes initially is the different word forms to be dealt and the different modes of speech. Our system has been basically built for the Hindi Language but this can be modified for different languages by replacing the proper words from different languages in this Expert system.

Grammar learning facilitates learning grammar on the runtime. This is a very difficult part and needs enormous research and development. We also propose a basic grammar learning algorithm. As mentioned earlier we have maintained a parse code xml list for different grammars available in the language. To start with, we first store the structure of the grammar of current sentence in a temporary buffer. Obviously, if we have to learn a new grammar, there will be new words in the sentence. Those words are marked as '*' in the grammar structure stored in the buffer. Words are identified by the conventional question asking method and their category is inserted into the buffer replacing the '*'. This new sentence is considered a valid sentence and the grammar is cross checked in the parser. Grammar Learning can also be done by asking questions. These questions asked from the user will only validate the grammar being provided in the new structure.

Extension of the system to a QA system:
The Analyzer phase returns some words which are send to the Sentence Generator phase. These words are restructured to form a valid sentence in order to give a reply. Though not implemented, the mood and tone of a person in his talk can be tracked and adequate response can be generated.

2.7 SENTENCE GENERATOR

A *Sentence Generator* (SG) constructs sentences, paragraphs, and even papers that fit a prescribed format. The format is specified by a set of rules called a ***grammar***.

We have proposed a very basic sentence generator which generates simple sentences from the grammar rules available. The sentence generator takes some words from the analyzer phase. These words are identified for their categories. Then a basic structure is found out from the grammar rules available which match the current rules. Henceforth, the sentence for the output to the sentence of the input sentence is generated.

This can be further clarified by the following example: Consider the words returned by analyzer phase are – us(that), jagah (place), gayaa (gone).

The system will try to find a valid set of rules and form a valid sentence. Also the system searches for the simplest grammar possible. In the above case, the simplest grammar possible is: Noun+Pronoun+Noun+Verb or Pronoun+Pronoun+Noun+Verb. Here priority is decided on random basis. But in this case the latter case must be followed.

The first word is obviously the system, as it has been modelled. The other three words are then concatenated. The system is very limited and needs to undergo disambiguation. So the output of the sentence generator is “main us jagah gayaa” or “mera us jagah gaya” and so on.

SOURCE CODE EXCERPTS

```
/*Set of standard structures*/
struct types;
struct retxml;
struct morphemes;
/*The Standard XML structure which is retrieved by other modules*/
struct types
{
        char* type;
// The word type (Eg. Noun/Adj./Verb etc.)
        char* meaning;
// The meaning of the word of the given type
        types* next;
// The next type of same word with various
        types()
        {
                next = NULL;
                type = NULL;
                meaning = NULL;
        }
};

struct morphemes
{
        char* morpheme;
// The various morphemes of the same word
        retxml* info;
// The various other infos of the given morpheme
        morphemes* next;
        morphemes()
        {
                morpheme = NULL;
                next = NULL;
                info = NULL;
        }
};
```

3 FUTURE PROSPECTS

NLP has huge prospects in coming time and our system can provide a huge impetus in forming a full fledged NLP system for Hindi Language. Hindi is spoken in five different types in different parts of India. Apart from this there are innumerous dialects which are similar to the language. Our proposed and basic system has been modelled over Xml architecture which means that changing a single word file can change the whole understanding of the system. We have built our own modules and also used some of the resources available on the NET, especially, Anusaarka[1]. Though the system is not a perfect one it gives us a platform for developing fully enhanced NLP systems in the coming time. The applications are enormous and the system expandability is infinite.

REFERENCES

1. Anusaarka Project – http://ltrc.iiit.ac.in/showfile.php?filename=projects/Anusaaraka/index.php
2. Natural language processing – A Paninian perspective
3. Natural language understanding – James Allen
4. www.iiit.net/ltrc/Publications
5. Information Science Institute – www.isi.edu/natural-language
6. Word sense disambiguation – www.senseval.org
7. Open NLP – www.opennlp.sourceforge.net
8. www.citeseer.com
9. www.sciencedirect.com
10. www.au-kbc.org
11. Bauer, Caurie (2004) A glossary of morphology
12. Adaptivity in Question Answering Using Dialogue Interfaces – Silvia Quarteroni and Suresh Manandhar
13. Learning word meaning by instructions – Kevin Knight
14. A plan-based agent architecture for interpreting natural language dialogue - LILIANA ARDISSONO, GUIDO BOELLA AND LEONARDO LESMO
15. Learning word syntactic sub categorizations interactively – Fernando Gomez
16. Knowledge Extraction from Hindi Text - Shachi Dave, Pushpak Bhattacharyya
17. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model – Hua – Ping Zhang, Qun Liu, Xue – Qi Cheng, Hao Zhang, Hong – Kui Yu.

New *"INTERFACE"* Tools for Developing Emotional Talking Heads

Piero Cosi, Graziano Tisato

Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche
piero.cosi@pd.istc.cnr.it, graziano.tisato@pd.istc.cnr.it

Abstract

INTERFACE is a tool for simplifying and automating many of the operations needed for building a talking head. INTERFACE was designed and implemented in Matlab©.and it consists of set of processing tools, focusing mainly on dynamic articulatory data physically extracted by an automatic optotracking 3D movement analyzer. The main reason to implement such a software tool was that of building up the animation engine of LUCIA our emotive/expressive Italian talking head. LUCIA can be directly driven by an emotional XML tagged input text, thus realizing a true audio visual emotive/expressive synthesis. LUCIA's voice is based on an Italian version of FESTIVAL - MBROLA packages, modified for expressive/emotive synthesis by means of an appropriate APML/VSML tagged language. Moreover, by using INTERFACE, it is possible to copy a real human talking by recreating the correct WAV and FAP files needed for the animation by reproducing the movements of some markers positioned on his face and recorded by an optoelectronic device. In this work the latest improvements of INTERFACE will be described and few examples of their application to real cases will be illustrated.

Index Terms: Talking Head, Audio/Visual synthesis, Articulation, Emotions.

1. Introduction

The transmission of emotions in speech communication is a topic that has recently received considerable attention. Automatic speech recognition (ASR) and multimodal or audio-visual (AV) speech synthesis are examples of fields, in which the processing of emotions can have a great impact and can improve the effectiveness and naturalness of human-machine interaction. Viewing the face improves significantly the intelligibility of both natural and synthetic speech, especially under degraded acoustic conditions. Facial expressions signal emotions, add emphasis to the speech and facilitate the interaction in a dialogue situation. From these considerations, it is evident that, in order to create more natural talking heads, it is essential that their capability comprises emotional behaviour.

In our TTS (text-to-speech) framework, AV speech synthesis, that is the automatic generation of voice and facial animation from arbitrary text, is based on parametric descriptions of both the acoustic and visual speech modalities. The visual speech synthesis uses 3D polygon models, that are parametrically articulated and deformed, while the acoustic speech synthesis uses an Italian version of the FESTIVAL diphone TTS synthesizer [1] now modified with emotive/expressive capabilities.

Various applications can be conceived by the use of animated characters, spanning from research on human communication and perception, via tools for the hearing impaired, to spoken and multimodal agent-based user interfaces.

The aim of this work was that of implementing INTERFACE a flexible architecture that allows us to easily develop and test a new animated face speaking in Italian.

2. INTERFACE

INTERFACE [2], whose initial screenshot is given in Figure 1, is an integrated software designed and implemented in Matlab© in order to simplify and automate many of the operations needed for building-up a talking head.

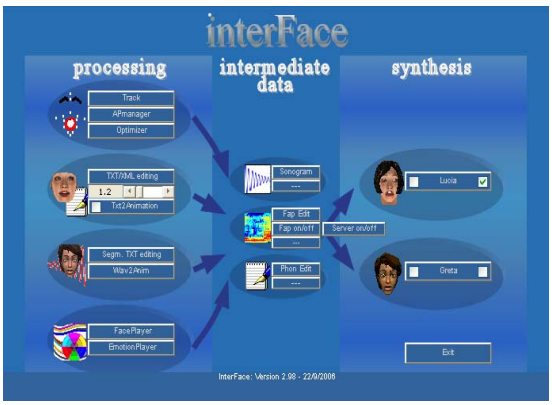


Figure 1: *INTERFACE starting screenshot.*

INTERFACE is mainly focused on articulatory data collected by ELITE, a fully automatic movement analyzer for 3D kinematics data acquisition [3], but it could be also used with similar data captured by analogous hardware instruments. ELITE provides for 3D coordinate reconstruction, starting from 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The 3D data coordinates are then used to create our lips articulatory model and to drive directly, copying human facial movements, our talking face.

INTERFACE was created mainly to develop LUCIA [4], our graphic MPEG-4 [5] compatible Facial Animation Engine (FAE). In MPEG-4 FDPs (Facial Definition Parameters) define the shape of the model, while FAPs (Facial Animation Parameters) define the facial actions [6]. In our case, the model uses a pseudo-muscular approach, in which muscle contractions are obtained through the deformation of the polygonal mesh around feature points that correspond to skin muscle attachments. A particular facial action sequence is generated by deforming the face model, in its neutral state, according to the specified FAP values, indicating the magnitude of the corresponding action, for the corresponding time instant.

As illustrated in Figure1, INTERFACE is divided in three blocks: data "*processing*", editing and control of

"*intermediate data*", and audio/visual "*synthesis*". The "*options*" push-button in the top menu allows the user to configure the different modules and applications and to save initialization parameters in a specific initialization file.

The "*processing*" zone is the fundamental part of INTERFACE, and it contains different modules developed for dealing with data, subdivided according to their type, that is: bimodal data (articulatory data) , XML and textual data, audio and low level (FAP) data.

The functionalities of the "*intermediate data*" zone regard the visualization of the speech waveform, of its relative sonogram, and also the visualization and editing of the phonetic segmentation and of course of the FAP articulatory animation data. An important innovation, introduced in this last version of INTERFACE, is the mechanism of synchronization of real and synthetic animation video thus allowing the user to easily compare them.

As far as "*syntehsis*", the talking faces can be activated on the same computer or they can be started in a client/server modality thus transmitting and receiving the FAP animation data in a local net or within the web via a specific IP address.

The audio synthesis is generated by FESTIVAL [1] coupled with the classical MBROLA engine [7] and also with a new developed SMS (Spectral Modeling Synthesis) engine [8] for Italian.

In summary, INTERFACE handles four types of input data from which the corresponding MPEG-4 compliant FAP-stream could be created:

- **(A) Articulatory data**, represented by the infrared passive marker trajectories captured by ELITE; these data are processed by 4 programs:
  - "***Track***", which defines the pattern utilized for acquisition and implements a new 3D trajectories reconstruction procedure;
  - "***Optimize***", which trains the modified coarticulation model [9] utilized to move the lips of a MPEG-4 compliant talking face;
  - "***APmanager***", which allows the definition of the articulatory parameters in relation with marker positions, and that is also a database manager for all the files used in the optimization stages;
  - "***Sonogram***" a new INTERFACE visualization module which is able to visualize, synchronously with video and articulatory signals, the speech waveform, its corresponding sonogram and pitch (see Figure 2). Moreover with this new feature it is also possible to edit and save articulatory parameters in order manually adjust the final animation when needed. This new feature substitutes the previous "***Mavis***" (Multiple Articulator VISualizer, written by Mark Tiede of ATR Research Laboratories [10]) module.
- **(B) Symbolic high-level TXT/XML text data**, processed by:
  - "***TXT/XMLediting***", a specific XML editor for emotive/expressive tagged text to be used in TTS and Facial Animation output;
  - "***TXT2animation***", the main core animation tool that transforms the tagged input text into corresponding WAV and FAP files. The audio file is synthesized by a FESTIVAL module, which realizes the emotive/expressive vocal modifications. The FAP-stream file, needed to animate MPEG-4 engines such as LUCIA, is obtained by an animation model, designed by the use of *Optimize*;
  - "***TXTediting***", a simple editor for text without any kind of tags, to be used in TTS and Facial Animation output;
- **(C) WAV data**, processed by:
  - "***WAV2animation***", a tool that builds animations on the basis of input WAV files after automatically segmenting them by an automatic ASR alignment system [11];
  - "***WAValignment***", a simple segmentation editor to manipulate segmentation boundaries created by ***WAV2animation;***
- **(D) manual graphic low-level data** , created by:
  - "***FacePlayer***", a direct low-level manual/graphic control of a single (or group of) FAP parameter; in other words, ***FacePlayer*** renders LUCIA's animation, while acting on MPEG-4 FAP points, for useful immediate feedback;
  - "***EmotionPlayer***", a direct low-level manual/graphic control of multi level emotional facial configurations for a useful immediate feedback.

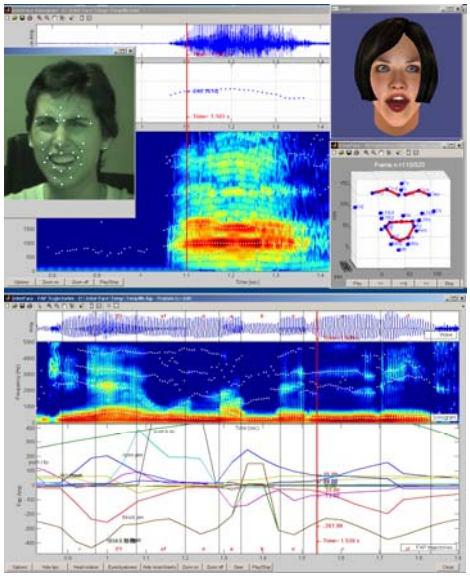


Figure 2: "***Sonogram***" module for INTREFACE.

2.1. "Track"

MatLab© Track was developed with the aim of avoiding marker tracking errors that force a long manual post-processing stage and also a compulsory stage of marker identification in the initial frame for each used camera. Track is quite effective in terms of trajectories reconstruction and processing speed, obtaining a very high score in marker identification and reconstruction by means of a reliable adaptive processing. Moreover only a single manual intervention for creating the reference tracking model (pattern of markers) is needed for all the files acquired in the same working session. Track, in fact, tries to guess the possible target pattern of markers and the user must only accept a proposed association or modify a wrong one if needed, then it runs automatically on all files acquired in the same session. Moreover, we give the user the possibility to independently configure the markers and also the FAP-MPEG correspondence. The actual configuration of the FAPs is described in an initialization file and can be easily changed. The markers assignment to MPEG standard points is realized with a context menu as illustrated in Figure 3. By Track, the

articulatory movements can also be separated from the head roto-translation, thus allowing to realize a correct data driven articulatory synthesis.

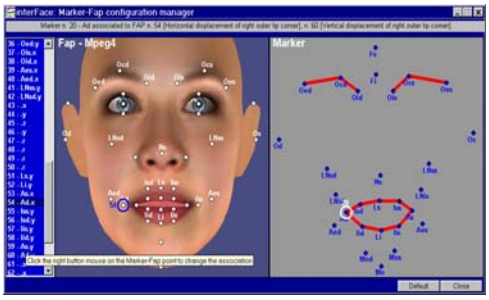


Figure 3: *Marker MPEG-FAP association with the TRACK's reference model. The MPEG reference points (on the left) are associated with the TRACK's marker positions (on the right).*

In other words , as illustrated in the examples shown in Figure 4, for LUCIA, Track allows a true 3D "*data driven animation*" of a talking face, converting the ELITE trajectories into standard MPEG-4 data and eventually it allows, if necessary, an easy editing of bad trajectories.

Figure 4: *Examples of single-frame LUCIA's emotive expressions. These were obtained by acquiring real human movements with ELITE, by automatically tracking and reconstructing them with "Track", and by reproducing them with LUCIA.*

Different MPEG-4 Facial Animation Engines (FAEs) could obviously be animated with the same FAP-stream allowing for an interesting comparison among their different renderings.

2.2. "Optimize"

The Optimize module implements the parameter estimation procedure for LUCIA's lip articulation model. This procedure is based on a least squared phoneme-oriented error minimization scheme with a strong convergence property, between real articulatory data Y(n) and modeled curves F(n) for the whole set of R stimuli belonging to the same phoneme set :

$$e = \sum_{r=1}^{R} \left(\sum_{n=1}^{N} \left(Y_r(n) - F_r(n) \right)^2 \right)$$

where F(n) is generated by a modified version of the Cohen-Massaro coarticulation model [9] as introduced in [12-13].

The mean total error between real and simulated trajectories for the whole set of parameters is lower than 0.3 mm in the case of bilabial and labiodental consonants in the /a/ and /i/ contexts [14, p. 63]. At the end of the optimization stage, the lip movements of our MPEG-4 LUCIA can be obtained simply starting from a WAV file and its corresponding phoneme segmentation information.

2.3. "APManager"

With this tool it is possible to define a certain number of measures or parameters by combining articulatory trajectories given by Track, relative to specific reference points, lines or planes opportunely defined (see Figure 5). APmanager allows also to visualize and modify the patterns of these choosen parameters together with their relative velocity and accelleration and also to extract minimum and maximum points needed to identify and better specify articulatory gestures.

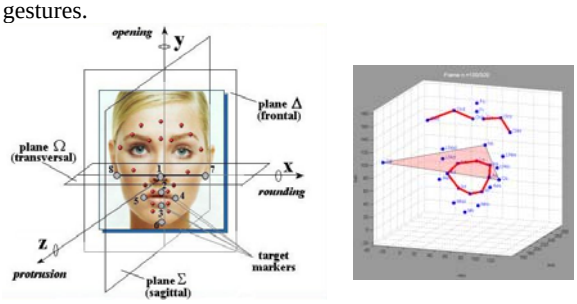


Figure 5: *Reference points and planes used for recording articulatory movements*.

2.4. "TXT/XMLediting"

This is an emotion specific XML editor explicitly designed for emotional tagged text. The APML mark up language [15] for behavior specification permits to specify how to markup the verbal part of a dialog move so as to add to it the "meanings" that the graphical and the speech generation components of an animated agent need to produce the required expressions. So far, the language defines the components that may be useful to drive a face animation through the facial description language (FAP) and facial display functions. The extension of such language is intended to support voice specific controls. An extended version of the APML language has been included in the FESTIVAL speech synthesis environment, allowing the automatic generation of the extended .pho file from an APML tagged text with emotive tags. This module implements a three-level hierarchy in which the affective high level attributes (e.g. <anger>, <joy>, <fear>, etc.) are described in terms of medium-level voice quality attributes defining the phonation type (e.g., <modal>, <soft>, <breathy>, <whispery>, <creaky>, etc.). These medium-level attributes are in turn described by a set of low-level acoustic attributes defining the perceptual correlates of the sound (e.g. <spectral tilt>, <shimmer>, <jitter>, etc.). The low-level acoustic attributes correspond to the acoustic controls that the extended MBROLA synthesizer can render through the sound processing procedure described above. This descriptive scheme has been implemented within FESTIVAL as a set of mappings between high-level and low-level descriptors. The implementation includes the use of envelope generators to produce time curves of each parameter.

2.5. "TXT2animation"

This represents the main animation module. TXT2animation transforms the emotional tagged input text into corresponding WAV and FAP files, where the first are synthesized by the Italian emotive version of FESTIVAL, and the last by the optimized coarticulation model, as for the lip movements, and by specific facial action sequences obtained for each emotion by knowledge-based rules. For example,

anger can be activated using knowledge-based rules acting on action units AU2 + AU4 + AU5 + AU10 + AU20 + AU24, where Action Units correspond to various facial action (i.e. AU1: “inner brow raiser”, AU2: “outer brow raiser”, etc.) [5]. MPEG-4 specifies a set of Face Animation Parameters (FAPs), each corresponding to a particular facial action deforming a face model in its neutral state. A particular facial action sequence is generated by deforming the face model, in its neutral state, according to the specified FAP values, indicating the magnitude of the corresponding action, for the corresponding time instant. In other words, lips are animated by the use of the optimized data driven articulation model, while the full face is animated following knowledge-based rules.

2.6. “WAV2animation” and “WAVsegmentation”

WAV2animation is essentially similar to the previous *TXT2animation* module, but in this case an audio/visual animation is obtained starting from a WAV file instead that from a text file. An automatic segmentation algorithm based on a very effective Italian ASR system [11] extracts the phoneme boundaries. These data could be also verified and edited by the use of the *WAVsegmentation* module, and finally processed by the final visual only animation module of *TXT2animation*. At the present time, the animation is neutral because the data do not correspond to a tagged emotional text, but in future this option will be made available.

2.7. “FacePlayer” and “EmotionPlayer”

The first module FacePlayer lets the user verify immediately through the use of a direct low-level manual/graphic control of a single (or group of) FAP (acting on MPEG4 FAP points) how LUCIA renders the corresponding animation for a useful immediate feedback. EmotionPlayer (inspired by [13]), is instead a direct low-level manual/graphic control of multi level emotional facial configurations for a useful immediate feedback, as exemplified in Figure 6.

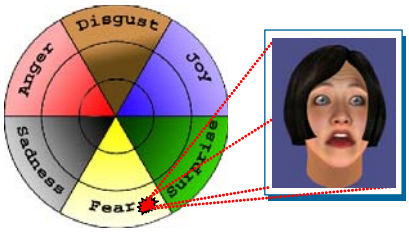


Figure 6: *Emotion Player. Clicking on 3-level intensity (low, mid, high) emotional disc [13], an emotional configuration (i.e. high -fear) is activated.*

3. Conclusions

With the use of INTERFACE, the development of Facial Animation Engines and in general of expressive and emotive Talking Agents could be made, and indeed it was for LUCIA, much more friendly. New visualization tools have been introduced and new evaluation tools will be included in the future such as, for example, perceptual tests for comparing human and talking head animations, thus giving us the possibility to get some insights about where and how the animation engine could be improved.

4. Acknowledgements

Part of this work has been sponsored by PF-STAR European Project IST- 2001-37599 (http://pfstar.itc.it).

5. References

[1] Cosi P., Tesser F., Gretter R., Avesani, C. (2001), “Festival Speaks Italian!”, Proc. Eurospeech 2001, Aalborg, Denmark, September 3-7, 509-512.

[2] Tisato G., Cosi P., Drioli C., Tesser F. (2005), “INTERFACE: a New Tool for Building Emotive/Expressive Talking Heads”, in Proc. INTERSPEECH 2005, Lisbon, Portugal, 781-784.

[3] Ferrigno G., Pedotti A. (1985), “ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing”, IEEE Trans. on Biomedical Engineering, BME-32, 943-950.

[4] Cosi P., Fusaro A., Tisato G. (2003), “LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro’s Labial Coarticulation Model”, Proc. Eurospeech 2003, Geneva, Switzerland, 127-132.

[5] MPEG-4 standard. Home page: http://www.chiariglione.org/mpeg/index.htm

[6] Ekman P. and Friesen W. (1978), Facial Action Coding System, Consulting Psychologist Press Inc., USA.

[7] Dutoit T. and Leich H. (1993), “MBR-PSOLA: Text-To-Speech Synthesis Based on an MBE re-Synthesis of the Segments Database”, Speech Commuication, vol. 13, no. 3-4, 167–184.

[8] Sommavilla G., Cosi P., Drioli C., Paci G. (2007), “SMS-FESTIVAL: a New TTS Framework”, Proc. MAVEBA 2007, Florence, (to be printed).

[9] Cohen M., Massaro D. (1993), “Modeling Coarticulation in Synthetic Visual Speech”, in Magnenat-Thalmann N., Thalmann D. (Eds), Models and Techniques in Computer Animation, Springer Verlag, 139-156.

[10] Tiede, M.K., Vatikiotis-Bateson, E., Hoole, P. and Yehia, H (1999), “Magnetometer data acquisition and analysis software for speech production research”, ATR Technical Report TRH 1999, ATR Human Information Processing Labs, Japan.

[11] Cosi P. and Hosom J.P. (2000), “Performance ‘General Purpose’ Phonetic Recognition for Italian”, Proc. of ICSLP 2000, Beijing, Cina, Vol. II, pp. 527-530, 2000.

[12] Pelachaud C., Magno Caldognetto E., Zmarich C., Cosi P. (2001), “Modelling an Italian Talking Head”, Proc. AVSP 2001, Aalborg, Denmark, September 7-9, 72-77.

[13] Cosi P., Magno Caldognetto E., Perin G., Zmarich C. (2000), “Labial Coarticulation Modeling for Realistic Facial Animation”, Proc. ICMI 2002, Pittsburgh, PA, USA, 505-510.

[14] Perin G. (2000-2001), Facce parlanti: sviluppo di un modello coarticolatorio labiale per un sistema di sintesi bimodale, Thesis, Univ. of Padova, Italy.

[15] De Carolis, B., Pelachaud, C., Poggi I., and Steedman M., “APML (2004), a Mark-up Language for Believable Behavior Generation”, in Prendinger H., Ishizuka M. (eds.), Life-Like Characters, Springer, 65-85.

[16] Ruttkay Z., Noot H., ten Hagen P. (2003), “Emotion Disc and Emotion Squares: tools to explore the facial expression space”, Computer Graphics Forum, 22(1), 49-53.

On Integration of Terminological Data in Translation Systems

Signe Rirdance, Andrejs Vasiljevs

Tilde, Latvia

signe.rirdance@tilde.lv, andrejs@tilde.lv

Abstract

The current translation practice demonstrates lack of integration support between the traditional desktop translation tools and the rich terminological data available on the internet. This article sets the background for development of a new layer of web-based translation tools for automated translation of multilingual terminology, bridging the gap between translation tools and environments and internet term banks. It analyses the experience gained during the EuroTermBank project that proposes solutions to a number of challenges in integration of term banks with translation tools, such as the federation approach to interlinking term banks and the entry compounding approach for visual representation of multiple overlapping terminology entries. The article propones a standards-based approach to ensure data compatibility, and identifies the requirement to support terminology sharing on an interoperable level.

Index Terms: term bases, translation tools, terminology sharing

1. Introduction

In today's translation practice, a significant gap exists between the traditional desktop translation tools and the terminological data available on the internet. Translators spend from 30 up to 60% of total translation time on terminology research, therefore it is vital to ensure that they can use all the required terminology resources in the right format and in the right environment. Currently, translators spend a lot of time inefficiently, searching and processing information from multiple sources and changing its format to the one they require in their work environment. Spending time on technical aspects instead of focusing on true terminology research results in cost inefficiencies and reduced translation quality. Moreover, translation practice often involves redundant work in identifying, creating or compiling the same terminology over and over again, by various translators.

To reach a new level of translation productivity, a layer of new tools and technologies is required that 1) enables consolidation and integration of dispersed terminology resources; 2) provides online access to consolidated multilingual resources through internet term banks; 3) provides tools that connect specific translation environments with terminology resources on the internet; 4) introduces standards that enable terminology interoperability, sharing and reuse.

2. Discussion

This article sets the background for defining a toolset required for integration of terminological data into translation environments. It shortly reviews the state-of-the-art in commercial translation systems and analyzes the experience and best practices from the EuroTermBank project in consolidating diverse terminology resources. It wraps up by introducing a layer of tools being developed to support integration of multiple term banks with a diverse set of typical translation environments.

2.1. Terminology and translation systems

Computer tools and technologies that serve the purpose of assisting human translation are commonly known as CAT or Computer Assisted Translation software, also sometimes referred to as MAHT (Machine Assisted Human Translation) [1].

The basic environment assisting human translation is text processing applications that typically provide very basic CAT features like spell-checking and grammar checking; no terminology support is provided. Microsoft Word provides the additional function of searching in predefined reference resources and sites, provided in English and some other major languages.

Since end of 1980s, translation memory (TM) tools have been developed that utilize alignments or linkages between source and target texts. Often, translation memory tools include a terminology management module that enables the translator to search automatically in a given terminology database for terms appearing in a document, however, this function is limited to searching in a proprietary terminology base format. Examples of TM tools and their terminology management modules are: SDL Trados and SDL MultiTerm, Wordfast, DeJa Vu, Star Transit and others.

There are, however, some major drawbacks of these tools regarding handling of terminology for translation.

The most widely used translation environment tool's terminology module, MultiTerm is a full-fledged terminology management application, and as such, its complexity by far exceeds the complexity required by majority of translators. Translators using SDL Trados usually do not exploit the potential of MultiTerm and refrain from creating and using terminology.

Unlike in translation memory handling, providing efficient terminology recognition requires language-specific support to match inflected term forms with regular forms in the dictionary. Translation tools on the market provide morphology support for only few major languages, which is insufficient in the global multilingual environment.

Most translation tools used by freelance translators provide no support for internet resources or internet-based communication with the language workers' communities on the internet. While server-based work using embedded terminology workflows is supported in systems used by multinational corporations and organizations, their cost is prohibitive for the average industry practitioner.

2.2. Terminology consolidation in EuroTermBank

The goal of EuroTermBank project [1] is to facilitate terminology data accessibility and exchange, by collecting,

consolidating and disseminating dispersed terminology resources through an online terminology data bank. The initial focus of EuroTermBank was to contribute to improvement of the terminology infrastructure in the selected new European Union member countries (Latvia, Lithuania, Estonia, Poland, Hungaria) but project expands its activities to other countries in EU and beyond.

The objective of EuroTermBank is to become the leading site for integration of multilingual terminology resources into the central EuroTermBank database or interlink them via EuroTermBank as a central gateway and single point of service. The data bank works on a two-tier principle – as a central database and as an interlink node or gateway to other national and international terminology banks. Data exchange mechanisms have been developed to establish term import, export and exchange with other terminology databases. EuroTermBank multilingual terminology base is freely accessible online at http://www.eurotermbank.com.

2.2.1. Federation approach in consolidation of term banks

Federation is a new concept in linking portals and also data repositories, which goes far beyond the establishment of pointers or links, but reaches out to the level of semantic interoperability of data and data structures. Especially terminology and other kinds of structure content can be made to enable interoperability in the form of network(s) of federated databases [3].

Semantic interoperability and implementation of the federation principle are essential for the next level of integration of terminological data in translation environments. Consolidation of content, application of unified standards and semantic interoperability significantly eases the task of providing the layer of tools required to seamlessly integrate diverse term banks with diverse translation environments.

Today, however, terminology resources on the internet remain fragmented across diverse term banks and terminology projects. While it is clear that national or institutional terminology can be best identified in the terminology database of the respective institution, a number of user scenarios require consolidation on a multilingual and multinational scale. EuroTermBank not only stores all available terminology content in its database, but also acts as a gateway providing unified access to multiple remote terminology databases.

To ensure the viability of the federated system of terminology databases, inclusion of a termbank in this federated model requires it to be independently supported and maintained both institutionally and technically.

Within EuroTermBank, the mechanism that enables federation of external databases is called interlinking. Interlinking an external database to EuroTermBank enables users to query the external database from EuroTermBank web interface. It is implemented by connecting to the external resource through a web service, ensuring platform-independent interoperable machine-to-machine interaction over a network. Communication is done using XML messages that follow the SOAP-standard, a protocol for exchanging XML-based messages over computer networks, normally using HTTP.

Important steps towards a federated interoperable model of terminology management within an international organization are taking place in ISO. The ongoing project of developing the ISO Concept database envisions a federated approach to development and maintenance of content, as well as public access to ISO terminology, in the form of ISO electronic dictionary [4].

2.2.2. Entry compounding in consolidation of terminology content

Automated entry compounding is an innovative mechanism proposed by EuroTermBank in unification of potentially matching terminology entries from different resources. This novel concept is a cost-efficient solution to consolidated representation of terminology resources. In regards to translation practice, it carries important implications for new web-based approaches to efficient handling of terminology entries from multiple sources, which is a typical translator's scenario that has limited or no support in translation tools.

EuroTermBank data structure is modeled according to concept-oriented approach to terminology. Terminology entry denotes an abstract concept that has designations or terms as well as definitions in one or more languages. If terminology bank contains entries coming from different collections and designating the same concept, there is an obvious interest to merge them into one unified multilingual entry.

For example, if we have term pair *EN tree – LV koks* coming from a Latvian IT terminology resource and another term pair *EN tree – LT medis* from a Lithuanian IT terminology resource we may want to join these two into unified entry *EN tree – LV koks – LT medis*. Such multilingual entry allows to get correspondence between language terms that are not directly available in any terminology resource (in our example, the new term pair *LV koks – LT medis*).

However, merging entries just on the basis of a matching term in one language that is common for these entries will lead to many erroneous term correspondences, due to the frequent ambiguity of terms among subject fields or much rarer cases of ambiguity in the context within one subject field. The only error-free method for merging entries is evaluating whether these entries denote the same concept, however, it is often impossible or very expensive to make comparisons of cross-lingual terminology concepts, especially taken large databases like EuroTermBank thank contain over 1.5 million terms. Therefore, we propose a practical solution by introducing terminology entry compounding, which is an automated approach for matching terminology entries based on available data.

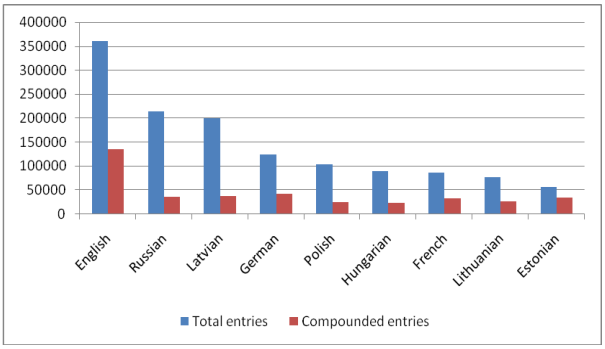


Figure 1. **Total and compounded entries per major languages of EuroTermBank.**

Entry compounding solves the problem of visual representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. At present, the

EuroTermBank database contains over 585,711 term entries with more than 1,500,500 terms. When applying entry compounding, over 135,000 or 23% of entries get compounded (see Figure 1). Hence entry compounding is a considerable aid for the user in finding the required term, for example, in the translation scenario between language pairs for which term equivalence is not established in existing collections.
Further work is planned on evaluating and improving entry compounding results within EuroTermBank, by applying corpus-based context analysis methods.

2.2.3. Application of standards

Integration of a multitude of term banks with a multitude of translation environments is only possible by rigorous implementation of applicable international standards. EuroTermBank project proposes a standards-based approach that is based on the best practice assessment and methodology recommendations developed by the EuroTermBank Consortium [5]. It includes a variety of aspects, from describing terminology collections to defining the data model and ensuring a unified data exchange format.
One of the major tasks in terminology data consolidation is identification and description of terminology resources. Due to a large number of resources to be described and different organizations in several countries involved it is important to use a common format for resource description. For this purpose we propose to use TeDIF format (Betz, Schmitz, 1999).
The Terminology Documentation Interchange Format TeDIF [6] was developed in the framework of the TDCnet project – European Terminology Documentation Centre Network, co-funded by the EU Commission. The TeDIF format was developed with the purpose to establish a common format for bibliographical and factual data related to terminology.
TeDIF provides means to describe bibliographical data like literature (serials, monographs, articles, journals, theses, etc.) and term collections (printed dictionaries, glossaries, thesauri, classifications, terminology databases, etc.).
EuroTermBank experience demonstrates applicability of TeDIF as a standard for terminology resource meta-data description and we recommend this format for other similar activities.
EuroTermBank is an early implementation [7] of the TBX (TermBase eXchange) standard, which enables a standard terminology exchange between term banks. The TBX standard, which is on its way to become an ISO standard in 2009, is based on three ISO standards: ISO 12620, ISO 12200 and ISO 16642. ISO 12620 defines data categories to be used for terminological data storage either in digital or printed format, while ISO 12200 defines MARTIF, an SGML interchange standard for interchange of terminological data and ISO 16642 that defines TMF, the terminological markup framework.

2.2.4. Terminology sharing

A number of emerging areas of activity will require the new tools layer that, first and foremost, connects term banks and translation environments, but also allows publishing and distributing terminology resources that comply with the standards required for this connectivity. Thus, terminology sharing is a phenomenon that involves sharing of non-confidential, non-competing and non-differentiating terminology across various actors – individuals along with companies and language service providers, with the goal to consolidate and promote accessibility to multilingual terminology per vertical industries [8]. Terminology sharing involves returns from streamlined industry terminology, by ensuring reuse of existing terminology assets. For those who share their terminology, it is a way of promoting and disseminating one's well-established terminology, possibly even to the level of de facto industry standard terminology. However, to reap full benefits from the shared terminology, it is essential to ensure integrated access to these terminology resources in translation environments.

3. Conclusions

Currently, the richness of terminology resources on the internet does not translate into the expected increased productivity and quality levels of translation work, and the area of providing tools for integration of terminological data in translation systems is relatively new. Therefore, it is important to establish a few basic principles that enable easier integration of term banks with the variety of translation environments.
The federation principle interlinks independently maintained term banks and provides a consolidated access point for terminology searches by human or machine users.
Terminology entry compounding provides a method that, from the translator's perspective, cuts translation inefficiencies in looking up a multitude of resources.
An underlying principle is a standards-based approach in developing term banks and any internet terminology resources, to ensure that the data can be easily exchanged in various terminology exchange and terminology sharing scenarios.
A new layer of tools and technologies that integrate translation environments with terminological resources on the internet is required to significantly enhance the current productivity of human translation.

4. Acknowledgements

Many thanks to colleagues in all EuroTermBank project partner organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland). EuroTermBank Consortium would also like to acknowledge and thank the European Union eContent program for supporting the EuroTermBank project as well as support from EU Social Fund.

5. References

[1] Hutchins J., "Current commercial machine translation systems and computer-based translation tools: system types and their uses", International Journal of Translation, vol.17, no.1-2, pp. 5-38, Jan-Dec 2005.

[2] Vasiljevs A., Skadins R., "EuroTermBank terminology database and cooperation network", proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, pp. 347-352, 2005.

[3] Galinski C., "New ideas on how to support terminology standardization projects", eDITion, 1/2007.

[4] Weissinger R., ”Integrating Standards in Practice”, *ISO Concept Database* presentation, 10th Open Forum on Metadata Registries, New York, July 2007.

[5] Auksoriute A. et al, “Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project”, Riga, 2006.

[6] Betz A., Schmitz K.-D., “The Terminology Documentation Interchange Format TeDIF”, Terminology and Knowledge Engineering TKE'99, Innsbruck, Wien, pp. 782-792, 1999.

[7] Vasiljevs A., Liedskalnins A., Rirdance S., “From Paper to TBX: Processing Diverse Data Formats for Multilingual Term Bank”, proceedings of the Third Baltic Conference on Human Language Technologies, Tallinn, 2008 (will be published).

[8] Rirdance S., “IP vs. Customer Satisfaction: EuroTermBank and the Business Case for Terminology Sharing”, The Globalization Insider, LISA, 6/2007.

Opentrad: bringing to the market open-source based Machine Translators

Ibon Aizpurua Ugarte [1], *Gema Ramírez Sánchez* [2], *Jose Ramom Pichel* [3], *Josu Waliño* [4]

[1] Eleka Ingeniaritza Linguistikoa, S.L.
[2] Prompsit Language Engineering
[3] Imaxin | Software
[4] Elhuyar Fundazioa
ibon@eleka.net, gema@prompsit.com, jramompichel@imaxin.com, josu@elhuyar.com,

Abstract

Most successful machine translation (MT) systems built until now use proprietary software and data, and are either distributed as commercial products or are accessible on the net with some restrictions. This kind of MT systems are regarded by most professional translators and researchers as closed and static products which cannot be adapted or enhanced for a particular purpose. In contrast to these systems, we present Opentrad, an open-source transfer-based MT system intended for related-language pairs and not so similar pairs. The project is funded by the Spanish government and shared among different universities and small companies. It uses different translation methods according to each language pair. For related-languages it uses shallow-transfer, even though for non-related pairs the system uses deep-transfer. The translation speed obtained is very high because it uses a finite-state transducer technique. The novelty of Opentrad consists of an introduction of open source software-development methodology and interoperability of standards in the field of MT.

Index Terms: machine translation, open source, business

1. Introduction

Most successful MT systems built until now use proprietary software and data, and are either distributed as commercial products or accessible on the net with some usage restrictions. Most professional translators and researchers view them as closed products since they cannot be easily adapted to particular purposes, integrated into other applications or used as resources in research or development projects. Besides, these MT systems mostly use *ad hoc* formats for linguistic data which are unreadable and very hard to maintain or extend.

All these aspects of commercial systems have a negative impact on the development of new techniques or the addition of new language pairs. A better concurrence between developers would had led to a positive motivation to improve existing MT systems, but making new systems from scratch is so costly that usually the primary goals are constricted by what has been already done. For this reason, it seems as if we were constantly reinventing MT and both the techniques and the resulting translations are often very similar to those of ten or even twenty years ago.

Fortunately, in the last decade, propelled by the globalization of the Internet, open-source strategies have established as a sound development practice allowing for reuse of code and data. Under this new situation, developers can now focus on improving and extending available software and data. In order to ease collaborative development, open-source projects are managed on centralized websites which also act as source code repositories. Another fundamental aspect for open-source projects to succeed is the availability of complete documentation describing it.

In the last years, open-source programs and data have also appeared in the field of MT, coexisting with commercial alternatives and bringing new opportunities which are proving very positive on both research and business areas. In this paper we present a real case of these positive effects achieved by Opentrad.

In the Opentrad project two different but coordinated designs have been carried out. The differences are due to the distance between the languages:

- An open-source shallow-transfer MT engine for the Romance languages of Spain (the main ones being Spanish, Catalan and Galician).
- A deeper-transfer engine for the Spanish—Basque pair.

Some of the components (modules, data formats and compilers) from the first architecture will also be useful for the second. Indeed, an important additional goal of this work is testing which modules from the first architecture can be integrated into deeper-transfer architectures for more difficult language pairs.

An overview of two translation method architectures are presented in section 2; section 3 explains available languages for Opentrad; section 4 explains the innovative part in Opentrad; section 5 summarizes real cases and possible

scenarios to apply MT. Finally, section 6 ends the paper with a brief discussion.

2. System architecture

In this section we will describe the two architectures used, Apertium for related-language pairs and Matxin for the Spanish-Basque pair.

2.1. Apertium

In this section we briefly describe Apertium (Armentano-Oller et al. 2006; Corbí-Bellot et al. 2005), an open-source shallow-transfer MT engine, initially intended for related-language pairs (such as Spanish–Catalan, Spanish–Galician, Spanish–Portuguese, Czech–Slovak, Swedish–Danish, Kirwanda–Kiswahili, Bahasa Indonesia–Bahasa Melayu, etc.), but being currently extended to translate between not so related languages (such as Spanish–French and Catalan–English); an early version of this extension is expected to be released by the end of 2006. Apertium's engine, linguistic data, and documentation can be found at the project's website at http://apertium.sourceforge.net.

The open-source MT architecture Apertium is mostly based upon that of systems already developed by the Transducens group at the Universitat d'Alacant, such as the Spanish–Catalan MT system interNOSTRUM (Canals-Marote et al. 2001), and the Spanish–Portuguese translator Traductor Universia (Garrido-Alenda et al. 2004). Both systems are not open-source; however, interNOSTRUM is publicly accessible through the net and used on a daily basis by thousands of users; Traductor Universia was also publicly accessible for some years until it was converted into a full commercial product.

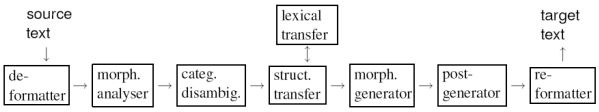


Figure 1: *The eight open-source modules of the Apertium MT system.*

The Apertium MT engine is a classical shallow-transfer or transformer system consisting of the following pipelined modules (see figure 1):

- A de-formatter which separates the text to be translated from the format information (RTF and HTML tags, white spaces, etc.). Format information is encapsulated so that the rest of the modules treat it as blanks between words.
- A morphological analyzer which tokenizes the text in surface forms and delivers, for each surface form, one or more lexical forms consisting of lemma, lexical category and morphological inflection information.
- A part-of-speech tagger which chooses, using a first-order hidden Markov model (Cutting et al. 1992) (HMM), one of the lexical forms corresponding to an ambiguous surface form; this is the only statistical-centered module.
- A lexical transfer module which reads each source-language lexical form and delivers the corresponding target-language lexical form by looking it up in a bilingual dictionary.
- A structural transfer module (parallel to the lexical transfer) which uses a finite-state chunker to detect patterns of lexical forms which need to be processed for word reorderings, agreement, etc. and performs these operations.
- A morphological generator which delivers a target-language surface form for each target language lexical form, by suitably inflecting it.
- A post-generator which performs orthographic operations such as contractions (e.g. Spanish del=de+el) and apostrophations (e.g. Catalan l'institut=el+institut).
- A re-formatter which restores the format information encapsulated by the de-formatter into the translated text.

The modules of the system communicate to each other by using text streams, which allows for easy diagnosis and independent testing. Furthermore, some modules can be used in isolation, independently from the rest of the MT engine, for other natural-language processing tasks. This extra application is also possible thanks to the full separation (or decoupling) of code and data.

2.2. Matxin

The engine is a classical transfer system consisting of 3 main components: analysis of Spanish, transfer from Spanish to Basque and generation of the Basque output.

It is based on the previous work of IXA Taldea (Díaz de Ilarraza et al., 2000) but with new features and a new aim: interoperability with other linguistic resources and convergence with the other engines in the Opentrad project through the use of XML. The previous object-oriented architecture is turning into an open-source one. This way we will be able to use modules which are shared with other engines in the Opentrad project and will comply with its format specifications.

The main modules are five: de-formatter, Spanish analysis based on FreeLing (Carreras et al., 2004), Spanish-Basque transfer, Basque generation and re-formatter.

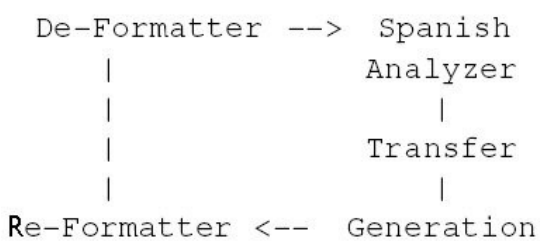


Figure 2: *Matxin MT system's architecture.*

The transfer and generation phases work in three levels: lexical form (tagged as node), chunk and sentence.

No semantic disambiguation is applied, but a large number of multi-word units representing collocations, named-entities and complex terms are being included in the bilingual dictionary in order to minimize this limitation.

2.2.1. Hybridization

In order to improve obtained results with deep-transfer methods we have some projects to develop a hybrid MT system. Currently we have restricted the linguistic field to administrative language and we are applying the following method:

- Firstly the system strips phrases from the text.
- We look for phrases in an example-based MT system and if it finds a match, it will translate the phrase.
- When there is no match, we translate phrases using a statistical MT system. To do so, we only validate sentences that reach a given threshold.
- Finally, if it does not reach that threshold, we will translate the phrase with a deep-transfer method.

We expect to get the results of this hybridization system in early 2008.

2.3. Linguistic data

Apertium's linguistic data (which are fully decoupled from the translation engine) are coded using XML-based formats; this allows for interoperability (that is, the possibility of using the XML data in a set of different scenarios) and for easy data transformation and maintenance. On the one hand, the success of the open-source MT engine heavily depends on the acceptance of these formats by other groups; this is indeed the mechanism by which *de facto* standards appear. Acceptance may be eased by the use of an interoperable XML-based format, and also by the availability of tools to manage linguistic data. But, on the other hand, acceptance of the formats also depends on the success of the translation engine itself. The XML formats for each type of linguistic data are defined through conveniently-designed XML document-type definitions (DTDs) which may be found in the Apertium and Matxin packages.

There are four sets of linguistic data organized at two levels: lexical or morphological level and structural or syntactical level:

- At lexical level morphological and bilingual dictionaries are used following the proposal for the whole Opentrad project.
- At structural level two grammars are being developed: one for structural transfer and other for syntactical generation.

3. Available language pairs

As we see in the section above, we have two translation engines (Matxin and Apertium). However, right now Matxin only works with the Basque-Spanish language pair so we will focus on Apertium.

Apertium's MT engine has been released in two open-source packages: *lttoolbox* (containing all the lexical processing modules and tools) and *apertium* itself (containing the rest of the engine); both are available under GNU GPL license. In addition to these programs, open-source linguistic data are already available for various language pairs:

- The Spanish–Catalan (packaged under the name apertium-es-ca) and Spanish–Galician (package apertium-es-gl) pairs developed under the Opentrad consortium and released under GNU GPL license;
- The Spanish-Portuguese pair (package apertium-es-pt) developed at the Universitat d'Alacant and released under GNU GPL includes the Brazilian variant as well;
- Pilot data for Catalan–French (package apertium-fr-ca) and Catalan–Occitan/Aranese (package apertium-oc-ca) released under GNU GPL;
- Pilot Catalan–English data.
- Spanish–French pair, expected to be released by the end of 2007.
- Pilot for Spanish-English, expected to be released by the end of 2008.

Apertium gives a reasonably good translation quality between related languages (error rates around 5-10 percent in general purpose translations). These results are obtained with the pilot open-source linguistic data already released (having around 10,000 lemmas and less than 80 shallow transfer rules) which might easily improve mainly through lexical contributions from the linguistic communities involved. The Apertium open-source engine itself is being actively developed and contributions to its design may enhance it to perform more advanced lexical and structural processing tasks.

All the available packages and documentation for Apertium are hosted at http://www.sourceforge.net/projects/apertium. Additional information may be found at http://www.apertium.org. Finally, web prototypes for MT systems for all the currently available pairs may be tested on plain texts, RTF and HTML at http://xixona.dlsi.ua.es/prototype.

4. Innovation

Some of the most important innovations of the Opentrad project is its open-source development methodology and introduction of a new business model in MT field.

4.1. Methodology

Open-source software projects are based on collaboration. This way, the source code of these projects is available on the net so that anyone can participate on the development. Many projects are hosted on personal websites but there are other hosting alternatives presenting many advantages such as control of versions, increase of visibility or developers management. These websites allow for centralized collaboration and distribution, and are known as open-source development websites. Two of the most commonly used are:

- **SourceForge.net** is the world's largest open-source software development website, hosting more than 100, 000 projects and over 1,000,000 registered users with a centralized resource for managing projects, issues, communications, and code.
- **Savannah** is a central point for development, distribution and maintenance of open-source software that runs on free operating systems. It hosts more than 2,500 projects and has over 45,000 registered users. It includes issue tracking, project member management by roles and individual account maintenance.

The two Opentrad MT engines are hosted at SourceForge.net and their web sites are: *apertium.sourceforge.net* for Apertium and *matxin.sourceforge.net* for Matxin.

As Opentrad is an open-source project, all software developed can be downloaded. Apart from that, anyone can see its code, change and publish it. This gives new opportunities for research in different MT fields and from the other side, it also gives new opportunities for business as we will explain in the next section.

4.2. Business model

It is not easy to encourage clients to use open-source software: open and free terms are usually perceived as untrustworthy. However, there is a strong reason why clients would prefer open-source to closed-source software: clients who choose open-source software do not see companies distributing them as providers to whom they have a technological dependency, but as technological partners, since clients may feel free to contract services around the open-source system with any other company offering them; therefore, technological dependence, a typical feature associated with closed-source products, is strongly diminished.

Even more interesting for institutions, public entities and large companies is the social action they can contribute to by making open modifications, improving data or adding new functionalities to the open-source software specially developed for them. This gives them a very positive image before their clients and users; they are not only offering a better service, but also benefiting the whole community.

It is also very difficult to convince tech companies to make their software open-source. The point is the change of the business model from a product-selling centered model to a service-offering one. Innovative services around a good open-source software are the main competitive advantages of this business model. Besides, contributions to the open project coming from elsewhere are also contributions that companies can benefit from in order to offer better products and services. This non-controllable aspect of the development makes heavy demands on those companies offering services based on open-source software but, despite this effort, in the current world it is crucial for tech companies to remain constantly updated.

Open-source software pose business challenges for those researchers working on new methods and techniques. Indeed, the number of technological-based spin-offs (here, companies created by researches as a result of a particular research activity) has increased in the last few years.

These companies have not only a product and a catalog of related services to offer, but also the know-how developed during the research work of their members, and, being half-way, they can offer the best services in collaboration with universities and companies.

5. Business real cases and possible scenarios

Open-source software also brings new business models to private companies. Taking the Apertium MT system as an example, companies can offer a wide variety of services around it, such as these:

- installing and supporting translation servers;
- maintaining, adapting and extending linguistic data;
- building data for new language pairs;
- integrating MT systems in multilingual documentation management systems
- Offering full translation services based on MT
- developing new tools for Apertium, etc.

Furthermore, companies and individual translators can adapt linguistic data to restricted language domains or to dialectical varieties in order to ease post-edition or better suit their clients' needs when offering translation services.

An illustrative example of companies benefiting from some of the previous profitable market segments is the case of the three companies (Eleka Ingeniaritza Linguistikoa, Imaxin\| Software and Elhuyar Fundazioa) participating in the

Apertium project as well as the case of a new company named Prompsit Language Engineering, created to exploit the challenges derived from the existence of Apertium.

5.1. Newspaper on-line edition translation

Imaxin software has improved and adapted the initial translation-engine with the Spanish-Galician pair, for the journal “La Voz de Galicia”, the eighth most-read newspaper in Spain. The process to have this machine translation system integrated into the journal editing environment took six months. From then on, the journal's on-line version is available both in Spanish and in Galician. It achieves less than a 5% error rate which results in a good translation quality.

Imaxin has also developed a tool to manage dictionaries for the machine translator so that new words can be easily added into the translator. At this moment, human review is needed in order to eliminate this error rate.

Finally, La Voz de Galicia has decided to leave the dictionaries resulting from this project free. That means that anyone can use the currently largest Galician dictionary.

5.2. Automated full translation service

Elhuyar Fundazioa offers full translation services through Apertium. Translation services are one of the commercial services offered by this foundation. Elhuyar uses Apertium to offer full translation services based on MT: translation using Apertium, linguistic correction and terminology services.

The use of Machine Translation systems integrated on commercial linguistic services gives commercial advantages to a translation service company as Elhuyar:

- Possibility to offer competitive prices for MT language-pairs maintaining linguistic quality
- Specialization in language review instead of human translations
- Integrated translation services: thanks to the best rates in MT language-pairs, we can offer human translation services at a lower price and therefore attract new customers.
- Terminology management services

5.3. Enterprises in internationalization process

In this world of globalization many companies are internationalizing their business, not only spreading their sales areas but also setting new production-plants in countries like the Czech Republic, Romania, Poland or Brazil. In this process enterprises may need to solve communication problems with employees and may also have to integrate in those countries' culture and language.

MT can help totally in this process; human-translation is slow and high-cost to translate piles of documentation generated in the enterprise, while MT systems can translate them fastly. Being open-source anyone who has linguistic and MT knowledge can adapt the system to each organization.

This is one interesting business area because languages like Portuguese (Brazilian) can be translated with very good accuracy.

5.4. Tourism industry

Other interesting scenario is the industry of tourism. Public organizations of tourism do a great effort to offer touristic information in several languages. In the Basque Country for example, they have a web site in Spanish, Basque, English and French. But a lot of tourists come from Catalonia, so to give a better service they can use MT to translate their web page into Catalan. An integration of an open-source MT system to translate web pages is not too difficult and the main task to do would be an adaptation of the dictionaries to the local toponymy, very rich in tourism-related texts.

Local and regional governments are interested in solutions like this to reduce translation costs, but the main obstacle is to find funding opportunities to cover the adaptation of dictionaries.

5.5. Other

Also, some institutions and public entities are juggling with the possibility of installing Apertium as their MT platform to offer on-line translation services. Banks which are present in different heterogeneous linguistic areas have also shown their interest in integrating Apertium-based MT systems in their documentation management systems.

6. Discussion

With the recent trends in open-source software development, new challenges raise for both research institutions and companies. Open-source practices have recently reached the MT arena, therefore introducing new perspectives on MT system development. A new business model, which focuses on the services around translation engines and linguistic data more than on the programs and data themselves, is possible.

In this paper we have presented Apertium, a full open-source MT system with a lot of potentials and introduced the main aspects around the new business model inspired by the Apertium system and the current state of development of an open transfer MT architecture for Spanish-Basque.

We have also presented different areas of business where open-source MT systems and translation-service providers can be very interesting allies.

7. Acknowledgements

Work funded by the Spanish Ministry of Industry, Commerce and Tourism through project Opentrad (FIT-340101-2004-3,

FIT-340001-2005-2, FIT-350401-2006-5, FIT-350401-2007-1) and Basque Government Department of Industry, Commerce and Tourism through project Opentrad (GAITEK IG-2006/00371)

8. References

[1] Corbí-Bellot, A. M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor, A., and Sarasola, K. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. In Proceedings of the 10th European Association for Machine Translation Conference, pages 79–86, Budapest, Hungary.

[2] Alegria I., A. Diaz de Ilarraza, G. Labaka, M. Lersundi. A. Mayor, K. Sarasola, (2005) A FST grammar for verb chain transfer in a Spanish-Basque MT System. Proc. of the Finite State Methods in Natural Language Processing workshop. Helsinki.

[3] Carreras, X., I. Chao, L. Padró and M. Padró (2004). FreeLing: An open source Suite of Language Analyzers, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal.

[4] Garrido-Alenda, A., Gilabert Zarco, P., Pérez-Ortiz, J. A., Pertusa-Ibáñez, A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M. A., and Forcada, M. L. (2004). Shallow parsing for Portuguese-Spanish machine translation. In Branco, A., Mendes, A., and Ribeiro, R., editors, Language technology for Portuguese: shallow processing tools and resources, pages 135–144. Lisbon.

[5] Díaz de Ilarraza, A., A. Mayor, K. Sarasola (2000). Reusability of wide-coverage linguistic resources in the construction of a multilingual machine translation system, in Proceedings of MT 2000 (Univ. of Exeter, UK, 1922 Nov. 2000), .

[6] Forcada, M. L. (2006). Open-source machine translation: an opportunity for minor languages. In Proceedings of Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages).

[7] World Wide Web Consortium (2004). "Extensible Markup Language (XML)", http://www.w3.org/XML/.

Relation Extraction in an Intelligence Context

Bénédicte Goujon

Thales Research & Technology - RD 128
91767 Palaiseau Cedex - France
benedicte.goujon@thalesgroup.com

Abstract

Our aim is to produce structured information from unstructured texts. To do so, we want to automatically extract explicit relations between entities from texts. Our work is constrained by the targeted intelligence domain, where users have no expertise in linguistics and cannot work with linguists for confidentiality reasons. We have developed a first prototype called Sem+ which extracts binary relations between entities from texts. It was mainly used on French corpus, but can be used for English. It was developed in a first time to automatically supply knowledge base. Relations are extracted thanks to patterns that are defined by the end user. For example, *"Henri Konan Bédié a reçu Alassane Dramane Ouattara.*" (Henri Konan Bédié has received Alassane Dramane Ouattara) is a pattern which produces the following relation: CONTACT(Henri Konan Bédié, Alassane Dramane Ouattara). Sem+ uses a learning algorithm, based on the Hearst algorithm, to ease the pattern acquisition. Several evaluations were provided on sale and purchasing relations between companies, and on an Ivory Coast corpus. The good precision and the efficiency of the learning algorithm were motivating to improve the tool. First improvement concerns the verbal pattern management. We add general linguistic knowledge to enhance the number of relations extracted with each pattern. Also we have worked to improve the entity management, in order to identify not only proper names but also nominal expressions related to entities ("*le président ivoiren*" as well as "Laurent Gbagbo"). This work was focused on people category. Now Sem+ is being integrated into platforms. The first one is a decision support platform, where Sem+ extracts relations from texts in order to identify events. The aim of the platform is to send an alarm when several events occur. The objective is to prevent a crisis, and the current study is based on the Ivory Coast crisis of September 2002. Sem+ will also be integrated in a semantic web platform, which contains complementary tools to annotate documents and manage ontologies.
Index Terms: relation extraction, text mining

1. Introduction

The objective of this work is to produce structured information from unstructured texts. To do so, we want to automatically extract structured relations between entities using a method which will be efficient on various domains. The relation extraction task is very hard and has to be restricted to few cases. As we can observe in the last ACE evaluations [1], only one participant proposes the relation extraction on English in 2007. In our approach, we first have worked on sales and purchases relations between companies. We have also studied a corpus dealing with the Ivory Coast situation, containing various named entities (Person, Organization, …) and relations (Location, Contact, …). A first prototype called Sem+ has been developed from these studies on those two different themes.

In this paper, we first describe the Sem+ tool together with its learning algorithm, and a first evaluation. We then present how we have added general linguistic knowledge to improve this tool and the resulting relations. We also present the platforms where Sem+ is used as a relation extraction component for various information treatment needs.

2. Sem+ : our Relation Extraction tool

We present here the most important aspects of our tool Sem+, which was detailed in [2].

2.1. Objectives and context

The principle aim is to analyze unstructured texts to provide to our clients structured information that follows their specific needs. To do so, we want to extract relations between entities from texts.

Thales has developed the Idéliance tool, which is a knowledge management system based on the concept of semantic networks [3]. The conceptors wanted to enable an easy use of the tool, for users without specific notions in knowledge representation. The manipulated knowledge has the format of a triplet "subject / verb / complement", as in "Peter / is from the category / Person", "Peter / is going to / Paris", etc. The main limitation of this tool concerns the knowledge capture. For now, it must be done manually. A great improvement of this tool would be the automation of the capture of all the relations. To do so, we have worked on automatic information extraction.

An important constraint is the specificity of the users: as Idéliance is used in the intelligence domain, the user wants a tool that allows to work alone, without any specific linguistic knowledge. We cannot imagine the intervention of a linguist on sensitive data.

Our objective was to propose a demonstrator allowing an easy acquisition of the relation patterns. Relation patterns are used to extract occurrences of relations from texts. The end user is a domain expert and not a linguist. For the pattern acquisition, we wanted to use an existing learning algorithm, and adapted it if necessary to ease the task. We also wanted to test whether only specific knowledge is sufficient to extract specific information so that the first demonstrator will not manage generic linguistic knowledge.

The work described here was done on French data, but the system is also useful for English. We wanted to obtain a tool providing the less noise, so we have preferred ignoring relations than proposing bad ones. And, as our tool was developed to be plugged with Idéliance, we have only considered as structured information the relations between two entities for the moment. For example, it includes a relation of sales or purchasing between two companies, or a relation of location between a person and a place. Such

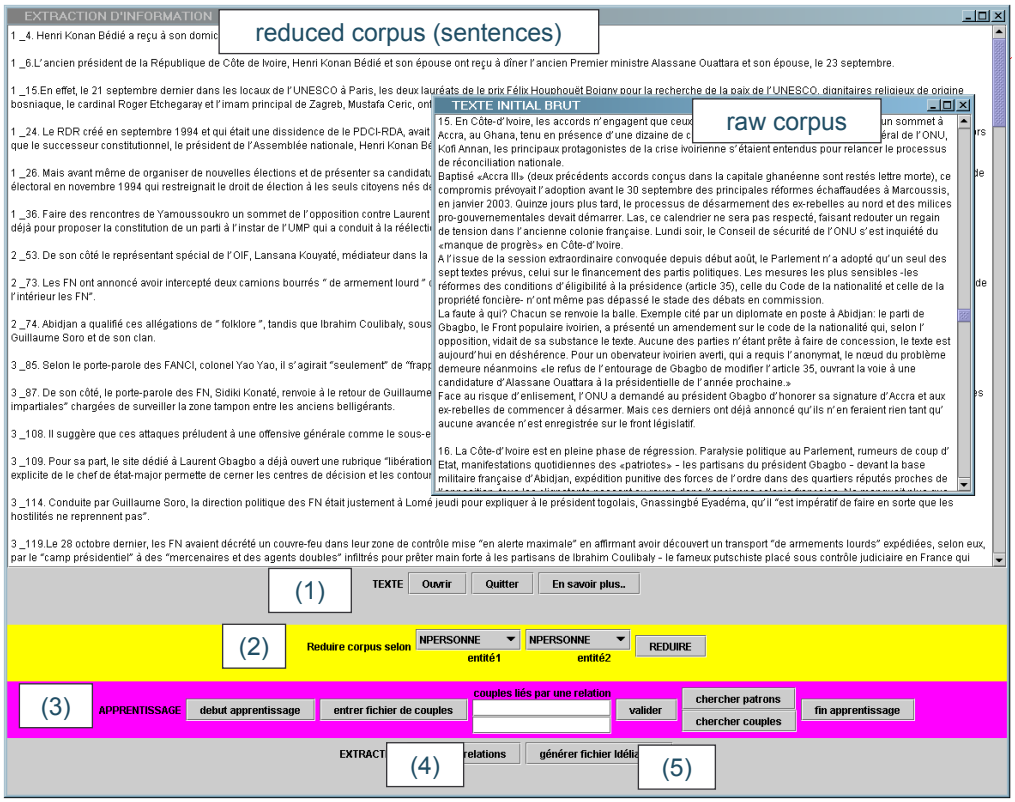


Figure 1: Sem+ Interface.

relations and the related entities are different from those managed in the ACE program [4]. In this campaign, a lot of entity types and subtypes are defined. First, we have worked with Named Entities (Person, Location) which represent a subset of the ACE entity types. Relations are also different from the ACE relations: we have specifically studied explicit relations expressed by verbs or nouns. For us, such relations are very relevant, as they are given by the writer of the text to We have studied two kinds of solutions: frameworks and ACE oriented approaches, in order to compare with our specific need.

T-REX (Trainable Relation Extraction framework) is a general software framework for supervised extraction of entities and relations from text [5]. This framework was designed so as to provide the degree of flexibility required by automatic semantic markup tasks for the Semantic Web. It has been developed as a testbed for experimenting with several extraction algorithms and several extraction scenarios, especially extraction from the web. This framework, like the Gate infrastructure [6], is not an information extraction system, but ease the building of such system. As we were used to manipulate a French linguistic environment (Intex), we have chose it for the first prototype.

SIE (Simple Information Extraction) is an information extraction system based on a supervised machine learning technique for extracting implicit relations from documents [7]. Another extraction method is presented by Zhao and Grishman [8]. It uses a kernel SVM. Those solutions learn off-line a set of data models from a specified labelled corpus, as it is defined in the ACE campaign. Such approaches are not efficient in our context as we can't have pre-tagged corpus dealing with the intelligence domain.

inform on an unusual fact or event. We didn't take into account implicit relations, expressed by word co-occurrences.

2.2. Relation extraction solutions

Several solutions are presented to solve the information extraction, and sometimes the relation extraction from texts.

2.3. Our learning algorithm

We first choose a learning method to build the linguistic patterns, rather than a declarative method and rather than a mathematical approach. The limit of the declarative methods in our context is that linguistic knowledge must be defined by a linguist after a corpus study. With a mathematical approach, the resulting co-occurrences relations are not accurate enough. For example, a relation between two persons can have a lot of meanings (CONTACT, KILL).

After the study of the three following learning systems: Rapier [9], EXDISCO [10] and Prométhée [11], we opt for the development of a specific algorithm. Our algorithm is based on the Hearst algorithm [12] but adapted to the intelligence domain as previously explained. Here are the steps of this algorithm:

1. Selection by the user of a couple of entity categories concerned by the relevant relation;
2. Capture by the user of couples of entities verifying the relation;
3. Automatic recovery of sentences containing these couples, with patterns that potentially describe the relation;
4. Selection by the user of the sentence extracts expressing the relation, capture of the relation information and the automatic transformation of the

extracts into patterns with Intex, a linguistic development environment [13].

5. Use of the patterns: recovery of new couples. Back to the step 3.

In order to simplify the reading of the corpus by the user and the extraction of the couples, the corpus is previously reduced according to the couple of entity categories that is concerned by the relation. At step 4, the user may transform the sentence, if some words are not meaningful, to express the relation by using "**" instead of the optional words.

To illustrate our approach, here is an example. We suppose that the user knows that Henri Konan Bédié and Alassane Dramane Ouattara were in contact recently (2). From this couple, the user obtains the following sentence: *"Henri Konan Bédié a reçu à son domicile parisien Alassane Dramane Ouattara."*[1] (3). The user builds then the pattern "*Henri Konan Bédié a reçu ** Alassane Dramane Ouattara*", and captures additional information: "*Henri Konan Bédié*" is the agent, "*CONTACT*" is the relation expressed by the pattern and "*Alassane Dramane Ouattara*" is the patient (4). Applied on another corpus containing "*Laurent Gbagbo a reçu hier ... Henri Konan Bédié*", this algorithm automatically produces the new relation "*Laurent Gbagbo – CONTACT – Henri Konan Bédié*". From this new couple (5), the system may identify a new sentence (3) that contains potentially another way to express the contact relation (4).

2.4. Details of the Sem+ demonstrator

We have developed the Sem+ demonstrator in order to evaluate the real interest of our algorithm for the intelligence domain. The first Sem+ demonstrator has been developed by Julia Frigière[2]. It is based on Intex [12] and is developed in Java and Perl.

Before the use of Sem+, the user has to construct its own domain-specific dictionaries, which are used to identify the entities. Here is an sample of the user dictionary containing persons:

Gbago,Laurent Gbagbo.NPERSONNE
Henri Konan Bédié,.NPERSONNE
Lansana Kouyaté,.NPERSONNE
Laurent Gbagbo,.NPERSONNE

Here are the steps for using Sem+ (see Figure 1 above).

1. Selection of a corpus, and automatic tagging of the named entities described in the dictionaries (for example, Person.dic contains "Lansana Kouyaté", Location.dic contains "Bouaké").
2. Reduction of the corpus according to a couple of entity categories (Person and Location in our example).
3. Learning step (the two sub-steps below could be applied constantly):
   - Capture of the entity couples to initiate the learning approach / recovery of new couples obtained from the captured patterns.
   - Capture of patterns associated with couples. The capture is easy, and consists of a copy-paste of the sentence extract (pattern) containing the relation, as: "*Lansana Kouyaté s'est rendu à Bouaké*"[3]. The user validates the agent, the patient, and the category of the relation expressed in the extract. For each extract, a generic transducer is automatically created with Intex to obtain for example MOVING(Person, Location) from "Person *s'est rendu à* Location".
4. Use of the pattern set to extract relations on new corpora: "*Thabo Mbeki s'est rendu à Bouaké*" => MOVING(Thabo Mbeki, Bouaké).
5. Production of a file using the Ideliance format, in order to export these relations into the knowledge management system.

2.5. Sem+ evaluations

First evaluations have been done on a financial corpus (Sale and Purchasing relations between Companies), to test the patterns coverage on a new corpus and the efficiency of the learning cyclical algorithm. On a small new corpus, using 45 patterns obtained from the learning corpus, Sem+ has identified 11 relations among 42 relations: the recall was 26% and the precision was 100%. Several identified reasons have caused the low recall. Firstly, only the sentences containing two entities are taken into account, even if sometimes relations may be built on two sentences (with anaphora). Also, as there were no linguistic analysis, the verbal expressions of the pattern were fixed, so "*X a acheté Y*" and "*X achète Y*"[4] are two different patterns. For us it is a correct result, as we want to favour the precision, but we'll have to improve it.

To evaluate the efficiency of the learning algorithm, by using 20 couples that were manually acquired (by reading of the first dispatches), we obtain automatically the capture of 32 patterns. This result is motivating, as it illustrates the efficiency of our method to easily learn new patterns.

Another evaluation has been provided on an Ivory Coast corpus, in order to validate the usability of such an approach by a user without specific linguistic knowledge. This evaluation has been done at the Land Headquarter (S.T.A.T. unit), by the Officer Cadet Cytermann[5]. He has used Sem+ on a corpus composed of journalistic articles related to the Ivory Coast. For this evaluation, no specific criterion have been used to quantify the results. After some technical adjustment and a first test, it has been concluded that Sem+ is easy to use, and that it is efficient in saving time when coupled with the Ideliance system. On this corpus, the learning algorithm was not efficient because of the small size of the acquisition corpus (120 sentences). Also, there were a lot of relations to identify (Meeting between two persons, Appointment between two persons, Moving between a person and a location, Accusation between two organisms, or between persons, etc.). The number of cases for each relation was not large enough for an efficient learning approach. This focuses on the characteristics of the initial corpus, that is essential in a learning approach: if the corpus is not large enough, it may not provide enough relation patterns to automate the information extraction.

---

[1] Henri Konan Bédié has received at his Parisian address Alassane Dramane Ouattara.

[2] *Frigière J. (2004), Information extraction by learning method, Internal report.*

[3] Lansana Kouyaté has gone to Bouaké.

[4] X has bought Y, X buys Y

[5] *Cytermann F. (2005), Évaluation du logiciel Sem+ dans le domaine du renseignement militaire, Training STAT report.*

3. Integration of general linguistic knowledge

As it was shown in the previous evaluations, Sem+ has to be improved in order to extract more relations. As we didn't use general linguistic knowledge in the first version, we have decided to add some to improve our results. We present here an algorithm that was developed to improve the verbal patterns management. We also detailed the focus on the entity management that is done via the anaphora resolving. We then present a future work that deals with the realization degree of each relation occurrences. Those improvements will be evaluated soon.

3.1. Verbal patterns management

During the previous evaluation done by a end-user, we have noticed that the pattern acquisition was a hard task. The use of no other knowledge than the user dictionaries is not sufficient to generate the various verbal expressions in relation with each verb. We propose to improve this by pre-defining the most common verbal expressions in a graph: conjugated verb; "to have", "to be able" or "to go" followed by the infinitive form of the verb; "to have" followed by the past participle form of the verb, eventually with adverbs for each case. We have developed our own code rather than using an existing parser, in order to control everything in Sem+. Here is our algorithm for the enrichment of verbal patterns, after the validation of an extract associated to a relation:

- Identification of the verbal sequence present in the extract;
- Identification of the lemma associated to this verbal sequence;
- Combination of this verb and the generic pre-defined graph containing the most common verbal expressions, and production of the complete verbal pattern;
- Production of the relation pattern, combining most of the variability of the verbal forms and the other words of the extract.

From the English example "X had received Y", we are able to automatically identify all those patterns : "X had received Y": "X is going to receive Y", "X has just received Y", "X will soon receive Y". If a verb is not in the general dictionary, the program just keep the initial extract. Two generic pre-defined graphs are used: one for the active form, and one for the passive form.

This verbal pattern management code has been implemented for French, by using the Delaf dictionary, and for English.

Our system manages patterns containing the verb between entities, as in previous examples, and patterns containing the verb after the entities (as in "*Laurent Gbagbo et Henri Konan Bédié se sont rencontrés*"[1]). Such patterns can contains two optional sequences at the most. We also have added the patterns containing a noun placed before the entities in relation, as is "*La rencontre entre Laurent Gbagbo et Henri Konan Bédié*"[2]. To evaluate this work, we have observed that previously we had to define five patterns for the "*rencontrer*" verb from the acquisition corpus. Now one pattern is enough. On the evaluation corpus, this pattern extracts six relations, which shows its efficiency.

3.2. Entity management

We have focused in a first time our work on the automatic identification of relations, using dictionaries containing proper names to identify the entities. But, this entity identification is not efficient. For example, in our corpus several expressions are used to talk about "*Laurent Gbagbo*": "*Le président Laurent Gbagbo*", "*Le président ivoirien Laurent Gbagbo*", "*Le président*", "*il*"[3], … All those expressions must be associated to Laurent Gbagbo in order to extract most of the relations, as in "*Il a rencontré Henri Konan Bédié*"[4]. In a first time, we have worked on anaphoric nominal expressions related to people. Our aim was to find a simple way to solve the identification of the referent of such expression without needed descriptions by the user. From our corpus, we have observed that the context of proper names was containing a lot of information. For example, if we have in a text "*le président ivoirien Laurent Gbagbo*", we can deduce that the expression "*le président ivoirien*" or "*le président*" can refer to Laurent Gbagbo in specific cases. An algorithm was developed to manage those anaphora[5]. In a first step, a corpus analysis extracts all the expressions preceding proper names and associates their content information (role, country, organization, …) to the corresponding entities. In a second step, a text is analysed and for each expression referring to a person without proper name ("*le président ivoirien*"), the algorithm propose a best candidate, which is referred in the preceding text and which is associated to the same information. This algorithm is currently evaluate.

3.3. Relation realization degree management

During our evaluation with potential users, we have observed that relations extracted from texts were not exact every times, as sometimes the relation has not occurred. We have identified then that an important clue in a relation extraction is the degree of realization of the relation: if we extract: CONTACT(Henri Konan Bédié, Alassane Dramane Ouattara) from the sentence "*Henri Konan Bédié devrait recevoir Alassane Dramane Ouattara*"[6], the user may consider as real a relation that has not occurred yet. Such variations on the meaning are very important in sensitive domain as intelligence or command context. A lot of relations are expressed using conditional tenses or words in order to express the uncertainty of the situation.

We want to propose the use of a specific feature associated to each relation that will express this realization degree. This will allow the distinction between the relations that have occurred (according to the text) using the keyword "certain", that may occur ("not certain"), and that will occur ("not realized"). To do so, we will exploit several linguistic markers: the verb tense (conditional means "not certain", future means "not realized"), the verbal structure ("to go"

---

[1] Laurent Gbagbo and Henri Konan Bédié have met each others

[2] The encounter between Laurent Gbagbo and Henri Konan Bédié

[3] The president Laurent Gbagbo, the president of the Ivory Coast Laurent Gbagbo, the president, he

[4] He has met Henri Konan Bédié

[5] *« Amélioration de la gestion des entités nommées pour l'extraction de relations sémantiques », Elzbieta Gryglicka, training report, 2006.*

[6] Henri Konan Bédié should receive Alassane Dramane Ouattara

followed by an infinitive means “not realized”) and the lexical markers (“may” or “to suppose” means “not certain”).

4. Sem+ integration in plateforms

In parallel with its own improvements, Sem+ is integrated into two platforms. The first integration was done previously for decision support. The second integration is in process and is related to the Semantic Web. We present here the two platforms.

4.1. Integration of Sem+ in an information management platform

The aim of a Command Support System (CSS) is to support decision by providing the Commander an edge over knowledge. In that context, we have built a global platform described in [14] that supports operators in their information analysis task. This platform contains three steps:

1. first, information is extracted from various media (video, text, audio, various signals from various sensors …);
2. second, the information is fused when necessary (if two pieces of information are dealing with a same event);
3. third, an alert is sent when necessary (if the new detected event is similar to important previous events).

Sem+ is integrated in the first phase of this platform. The approach used fo the fusion task is described in [15].

We have began the study of a corpus dealing with the Ivory Coast crisis of September 2002. The main event was on the night of September 18-19, 2002, when as many as 800 disgruntled soldiers took up arms against their country in mutiny. Our corpus contains about 15 000 texts from newspaper or press agency. From this corpus, we would like to identify some precursory events that were at the origin of the crisis, in order to provide the sending of an alarm before such a crisis. To do so, Sem+ will be useful to extract relations between people and/or organizations. By adding date and location information we would like to extract necessary information to describe each event.

4.2. Integration of Sem+ in a Semantic Web platform

The objective of the WebContent French project [16] is to propose a platform for the content management. This platform will be used to automatically discover, understand and structure the information whatever its format, and particularly the documents (XML and HTML), the Web Services and the forms. It contains various components to semantically annotate documents. Several partners have developed components for various tasks dealing with document annotation or knowledge management: document segmentation, lemmatization, parsing, filtering, classification, ontology enrichment, etc.

In this context, Sem+ will be used for the entity extraction and for the relation extraction. To do so, we want to produce relations that will be automatically used to enrich ontology. Also, we would like to include the anaphora resolution algorithm to annotate all the textual occurrences of persons. And the addition of realization degree will be helpful to distinguish events that are not described as “realized”.

This platform will be used by partners for several applications in relation with technological. For Thales, we will use this platform for our study on the Ivory Coast crisis. We will use several services complementary to Sem+ to manage the texts and the knowledge extracted from them.

5. Conclusion

We have presented the first version of our prototype called Sem+ that extract entity relations from texts. It is based on a learning algorithm that helps user without any linguistic knowledge to manage the relations between entities. We have detailed the evaluation of this version. We also have presented some improvements that were provided for the verbal management and for the entity management. Our Sem+ tool is now used into a CSS platform, and will be added to a semantic web platform. We would like to add realization degree information to the relations, and we will have to manage precisely date and location entities, as we are working on person entities to solve anaphoric expressions, in order to extract complete relations.

6. References

[1] *ACE 2007 Automatic Content Extraction Evaluation Offical Results (ACE07) http://www.nist.gov/speech/tests/ace/ace07/doc/ace07_eval_official_results_20070402.htm*

[2] *Goujon B., Frigière J. 2005. Extraction of Relations between Entities from Texts by Learning Methods, in* IST-055 Specialists Meeting on "Information Fusion for Command Support", Netherlands.

[3] *Rohmer J. 2002. Représentation, Fusion et Analyse d'informations mises sous forme de réseaux sémantiques: vers le "calcul littéraire" ?,* Revue REE*, N°7 Juillet 2002.*

[4] *ACE 2005 Evaluation Plan, http://www.itl.nist.gov/iad/894.01/tests/ace/*

[5] *Iria J., Ciravegna F. 2005. Relation Extraction for Mining the Semantic Web,* Dagstuhl Seminar on Machine Learning for the Semantic Web.

[6] *Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.*

[7] *Giuliano C., Lavelli A., Romano L. Simple Information Extraction (SIE), Technical report.*

[8] *Zhao S., Grishman R. 2005. Extracting Relations with Integrated Information Using Kernel Methods, Proceedings of the 43rd Annual Meeting of the ACL, pages 419–426, Ann Arbor, June 2005.*

[9] *Califf M. E., Mooney R. J. 2003. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction, in* Journal of Machine Learning Research 4*, pp177-210.*

[10] *Yangarber R., Grishman R., Tapanainen P., Huttunen S. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction, in* Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Allemagne.*

[11] *Morin E. 1999. Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus,* TKE'99*, Innsbruck, Austria, August 99, pp. 268-278.*

[12] *Hearst M. A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, in* 14TH International

Conference on Computational Linguistics (COLING 1992), *pp. 539-545.*

[13] Silberztein M., INTEX : http://msh.univ-fcomte.fr/intex/

[14] Laudy C, Mattioli J., Museux N., 2005. Cognitive Situation Awareness for Information Superiority, IST-055 Specialists Meeting on "Information Fusion for Command Support", Netherlands.

[15] Laudy C., Ganascia J.-G., Sedogbo C., High-level Fusion based on Conceptual Graphs, Fusion 2007.

[16] WebContent project description (in English) : http://www.webcontent.fr

Speech technology for language tutoring

Helmer Strik [1] *, Ambra Neri* [1] *, and Catia Cucchiarini* [1]

[1] Department of Language and Speech, Radboud University Nijmegen, The Netherlands

h.strik@let.ru.nl, a.neri@let.ru.nl, c.cucchiarini@let.ru.nl

Abstract

Language learners are known to perform best in one-on-one interactive situations in which they receive optimal corrective feedback. However, one-on-one tutoring by trained language instructors is costly and therefore not feasible for the majority of language learners. This particularly applies to oral proficiency, which requires intensive tutoring. Computer Assisted Language Learning (CALL) systems that make use of Automatic Speech Recognition (ASR) seem to offer new perspectives for language tutoring. In this paper we explain how.

Index Terms: Computer Assisted Language Learning (CALL), Automatic Speech Recognition (ASR), language tutoring.

1. Introduction

Language learners are known to perform best in one-on-one interactive situations in which they receive optimal corrective feedback. The two sigma benefit demonstrated by Bloom [2] has provided further support for the advantages of one-on-one tutoring relative to classroom instruction. However, one-on-one tutoring by trained language instructors is costly and therefore not feasible for the majority of language learners. In the classroom, providing individual corrective feedback is not always possible, mainly due to lack of time. This particularly applies to oral proficiency, where corrective feedback has to be provided immediately after the utterance has been spoken, thus making it even more difficult to provide sufficient practice in the classroom.

The emergence of CALL systems that make use of Automatic Speech Recognition (ASR) seems to offer new perspectives for language tutoring. These systems can offer extra learning time and material, specific feedback on individual errors and the possibility to simulate realistic interaction in a private and stress-free environment.

At the same time the increasing mobility in Europe and in the world at large together with the recent emphasis on promoting plurilingualism and linguistic diversity in Europe has led to a situation in many countries in which the demand for language lessons outstrips supply. As a consequence, new methods and technologies that make language learning more efficient and effective are called for.

ASR-based CALL could be employed to develop new methods for teaching literacy, reading, oral proficiency, speaking fluency, and vocabulary. In this paper we first review some studies that have employed ASR for language learning with mixed results. We then go on to consider important aspects of software design and a number of technological challenges. Finally, we draw some conclusions and consider challenges and opportunities for the future.

2. CALL applications

Speech technology is already used in several CALL applications. However, some researchers are skeptical about the usefulness and effectiveness of ASR-based CALL programs: evidence gathered in different lines of research seems to confirm that either speech technology is not mature enough, or ASR-based CALL programs are not effective in improving second language (L2) skills [e.g. 3, 5]. For the sake of our own research, we have studied this literature thoroughly and have gradually acquired the impression that, while it is undeniable that speech technology still presents a number of limitations, especially when applied to non-native speech, part of this pessimism is in fact due to misconceptions about this technology and CALL in general.

2.1. CALL & ASR

ASR dictation packages are being used by a growing number of people working in different branches. These packages offer good performance at a reasonable price and are readily available. It is probably for these reasons that some teachers and CALL practitioners have become interested in these programs as a possible tool to teach L2 skills [3, 5].

Derwing, Munro, and Carbonaro [5] investigated the usefulness of ASR for CALL by evaluating the performance of a standard dictation package, Dragon NaturallySpeaking Preferred, in identifying pronunciation errors in the L2 speech of Cantonese and Spanish learners of English. The authors propose two criteria for establishing the effectiveness of ASR in providing corrective feedback on L2 speech errors. First, the software should be able to recognize the oral language of 'English as a Second Language' (ESL) speakers at an acceptable level. Second, the software's identification of L2 speech errors must resemble that of native, human listeners.

On the basis of their study, Derwing et al. [5] conclude that ASR "cannot be considered to be of benefit to ESL speakers" [5: p. 602], that "the computer's output might be confusing to ESL students" and that "the observed levels would frustrate a user hoping for reliable feedback on intelligibility" [5: p. 600]. However, it is important to stress that the first conclusion does not apply to ASR in general, but to the specific ASR dictation package tested in this study, which was never intended for L2 learning. Analogously, the second and third conclusions are based on the incorrect assumption that the output of a dictation system can be used as a basis for providing feedback to L2 learners. Although the authors clearly state what the domain of their evaluation is in the introduction, they fail to relate their negative results to the characteristics of the specific technology they used, which may lead many to generalize those conclusions to the use of speech technology as a whole for L2 training.

Coniam [3] conducted a study aimed at exploring "the potential of the use of voice recognition technology with second language speakers of English" [3: p. 49] by testing the dictation package Dragon NaturallySpeaking on ten native speakers and ten Cantonese speakers of English. The recognition accuracy of the system was examined for both speakers groups for an excerpt read from a book. Besides, the author compared the output of the recognizer for native and non-native speakers with the original text in an attempt to

identify phonological patterns, e.g. sound substitutions in the speech of the Cantonese speakers, on the basis of the CSR output. The results show that the system's accuracy is higher for native speech than for non-native speech and Coniam [3] concludes that "voice recognition technology is still at an early stage of development in terms of accuracy and single-speaker dependency" [3: p. 49] although it might have potential in the future.

In these studies unsatisfactory results are obtained when standard dictation systems are used for CALL. But dictation systems are not suitable for L2 training, CALL requires dedicated speech technology. Apart from the fact that current dictation packages are usually developed for native speakers, the major problem in using this technology for CALL has to do with the fact that dictation and CALL have different goals which require different approaches in ASR. The aim of a dictation package is to convert an acoustic signal into a string of words and not to identify L2 errors, which requires a more complex procedure. Consequently, the negative conclusions should be related to this specific case and not to speech technology in general.

2.2. Software design

ASR-based CALL systems can recognize what a student actually uttered, to detect errors, and to provide immediate feedback on them. However, the technology needed for such systems is highly complex and still has a number of important limitations that should seriously be reckoned with when designing applications for L2 pronunciation error detection [see e.g. 6, 7].

Another important aspect of system design is pedagogical guidelines. Many commercial CALL systems present fancy looking features that are likely to impress the buyers, but that in fact do not serve any real pedagogical purpose, thus they do not meet the real needs of L2 learners The design of these systems seems to be driven more by a technology push, rather than being based on a comprehensive analysis of the requirements that the system must meet to be effective and efficient. This may in part be due to a difficulty involving different experts in the design phase of a CALL system, or more fundamentally, to the absence of clear pedagogical guidelines that suit CALL. Here we present examples of how inadequacies in system design leading to disappointing performance often end up being unjustly attributed to speech technology.

A fine example of how speech technology can be employed to diagnose segmental errors and provide feedback on them is provided by the ISLE (Interactive Spoken Language Education) project [9, 11]. Within the framework of this project, a system for German and Italian learners of English was developed which provided feedback on pronunciation errors at phoneme and word-stress level. The feedback on phonemes consists in highlighting the grapheme corresponding to the erroneous phoneme in the utterance and by showing on the screen both a frequent word containing the correct target phoneme and another frequent word containing the student's incorrect realization of it. The student can listen to both sounds in a focused way and try to notice the differences. A list of words containing the problematic target phoneme is optionally provided for training.

This system is therefore not only able to determine where a segmental error occurred in an utterance, but also what sound was realized. This approach is based on the belief that a CALL system should not only indicate that there is an error, but also specify where the error is located and how it should be corrected [11]. While this design seems almost ideal, the performance of the system is poor. The authors report that only 25% of the errors are detected by the system and that over 5% of correct phones are incorrectly classified as errors, whereby "students will more frequently be given erroneous discouraging feedback than they will be given helpful diagnoses" [11, p. 54].

However, performance could probably be improved by adopting a slightly different design that takes more account of the limitations of speech technology. Given these limitations, the ISLE system is likely to make mistakes at various stages: in recognizing the utterance, in locating the error, in diagnosing the problem, and thus also in presenting the example words. A slightly less ambitious system that has to make fewer decisions is also likely to make fewer errors. For example, a system that only indicates the part of a word or utterance that was mispronounced, without indicating exactly which erroneous sounds it recognized would be less sophisticated and, probably, less error-prone. Although it is more desirable to provide diagnostic information in Computer Assisted Pronunciation Training (CAPT), we have to conclude that such a system cannot guarantee a satisfactory performance yet with current technology. But, as it turns out, only showing mispronunciations might be sufficient feedback, and a CAPT system that does this may just as well be useful for improving pronunciation as we have shown in our research [12, 13].

Another example concerns some commercial systems which seem driven by technological innovations rather than by pedagogical guidelines. In some of these systems, that advertise themselves by mentioning that ASR is used, the feedback consists of an overall score and a graphical display with waveforms or spectrogram's, usually one window displaying the utterance spoken by the language learner, and another window with a reference utterance. These graphical displays look like an invitation for the student to try to understand how the two are related, but practice with the programs generally does not shed light on this aspect [1, 12]. The student may eventually end up realizing that there is no relation between the two [1, 12], which impoverishes the pedagogical value of this kind of feedback. In addition, the fact that the system shows two comparable displays, one representing the student's utterance and one representing the model utterance, wrongly suggests that the student should produce an utterance that closely corresponds to that of the model. In fact, this is not necessary at all: two utterances with the same content may both be very well pronounced and still have waveforms or spectrograms that are very different from each other. Moreover, waveforms and spectrograms are not easily interpretable for students [1, 12]. Even students with knowledge of acoustic phonetics are likely to find it hard to extract the information needed to improve pronunciation from these displays, since there is no simple correspondence between the articulatory gesture and the acoustic structure in the properties displayed. As many authors have observed, this type of feedback is not easy to comprehend and thus of limited pedagogical value [6, 7, 11, 12]. Despite their little pedagogical value, it might be that these programs are flashy and impressive [1, 12], a factor whose importance should not be underestimated in commercial products.

In these cases, shortcomings in the design of the ASR-based CALL programs contribute to creating the impression that speech technology is to blame. One of the reasons why these systems perform poorly is that they were designed without taking due account of the limitations of speech technology: as a result, programs that are too ambitious given the state of the art of this technology are developed. When

performance turns out to be disappointing, one then gets the impression that ASR as a whole is inadequate for the purpose of automatic training of L2 learners, while a better design with that very same technology might have produced a more effective product. This might in part explain why these systems eventually turn out to be ineffective in improving L2 pronunciation, which then helps create the impression that the disappointing learning results are due to the inadequacy of speech technology for this specific purpose.

3. Technological challenges

In the previous section we have provided some examples of using speech technology in CALL applications for which the results were disappointing. We explained that these disappointing results are to a large extent the result of a combination of factors, not all of which are related to the inadequacy of speech technology. An important factor is speech recognition, but this is only one part of a complex system, for which complex decisions have to be taken. These decisions might not always lead to the desired learning outcomes, as we have seen.

In this section, we intend to show how speech technology, despite its undoubted limitations, can still be employed in a meaningful and useful way in CALL applications. The challenges are the following. We should obviously try to improve the technology. However, since these improvements are likely to be gradual, as they have been during the last decades, we should also try to make optimal use of current technology, taking into account what is possible and what isn't possible with current technology.

3.1. Speech recognition

In the ASR community, it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably [9, 15, 16]. ASR-technology is sensitive to the degree of mismatch between acoustic models and incoming speech, to vocabulary size, and to task complexity. Still, there are some ways in which ASR performance can be improved.

Instead of using an ASR that is trained only on native (L1) speech, an ASR should be used which is trained on non-native (L2) speech (possibly in combination with L1 speech): lexica with non-native pronunciation networks, language models based on words and word orders as spoken by non-natives, and acoustic models that represent the way non-natives pronounce sounds. For the acoustic models there are several possibilities: simply train them on L2 speech [8], use acoustic models of L1 and L2 in parallel [10], or a combination of L1 and L2 models [8, 16], and, optionally, also include intermediate phones.

Besides improving the ASR, one could also make the task of the ASR less difficult. For a constrained task, it is possible to use a constrained lexicon and language model, which will increase the performance. If the number of possible answers is limited, one could even use utterance verification techniques (e.g. confidence measures) to select the utterance that was spoken from the list. In that way, performance could even be improved more. The challenge then is to develop engaging items, for which the possible answers can be predicted. Eliciting speech from the speakers in such a way that the lexicon is known in advance [6, 9] can be achieved in various ways, for instance by asking closed-response questions instead of open-response questions, by having speakers read aloud sentences, or repeat auditorily primed sentences. The choice of the strategy will depend on the type of application.

Such a strategy has been employed in a CAPT system, called Dutch-CAPT, we developed for learning Dutch [12, 13]. This system was based on an already existing multimedia learning system, called 'Nieuwe Buren' ('new neighbors'), which did not make use of speech technology. We selected some of the lessons, added speech technology to it which made it possible to give feedback on the spoken utterances, and developed a new user interface (Fig. 1). Each lesson started with watching a video, followed by some exercises: role-playing (Fig. 1), question answering, and reading. For all items, utterance verification was used for recognition.

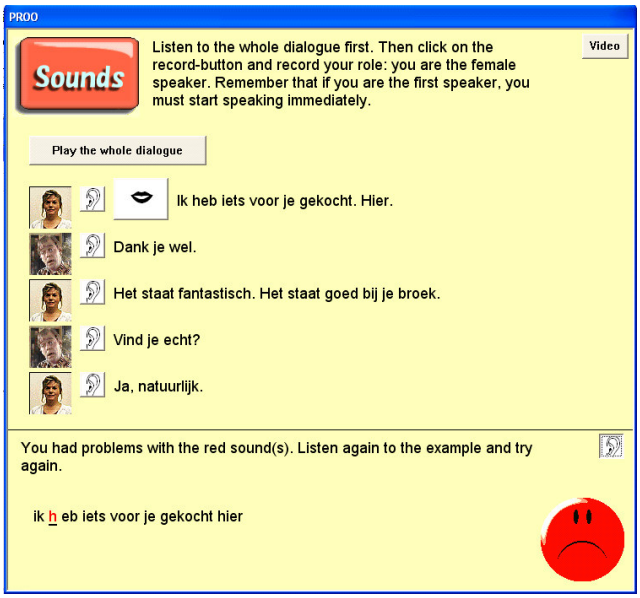


Figure 1. A screen shot of the Dutch-CAPT system.

Our design was based on a thorough study of existing call systems. For instance, we decided to use feedback that is not as detailed as the feedback given in the ISLE system (see section 2.2). In this way we managed to reduce the number of errors in the feedback. An example of the interface of our system is shown in Figure 1: an indication is given of which sounds are not pronounced correctly (red and underlined), and the language learner has the opportunity of listening to his utterance, an example utterance, and try again. When the system was tested, it turned out that language learners who used this system only four times for about 30 to 60 minutes improved more than a control group that did not use the system [12, 13].

3.2. Assessment

Once the speech signal has been recognized, it has to be evaluated. This phase rests on the opposite assumption as that underlying the recognition phase: while good recognition of the students' speech implies that the system be tolerant of discrepancies between the incoming speech and the native speech models, good scoring requires the system to look exactly for those discrepancies. Different terms have been used by the various authors for this stage: pronunciation scoring, pronunciation grading, error detection, error localization, error identification etc. Although these terms are often used interchangeably, they can be used to refer to two different activities, as will be explained below.

In general, error detection indicates the procedure by which a score at a local (phoneme) level is calculated, while pronunciation grading stands for the procedure that is followed

to calculate a global score at the utterance level, which could also be a weighted average of the local scores. Seen in this light, error detection can be considered a specific case of pronunciation grading, but there is more to it than meets the eye. In fact, error detection and pronunciation grading can be viewed as two different tasks, with a different goal and a different output. For grading, more global measures can be used, such as temporal measures [4]. The relation between human and automatic grading becomes better if longer stretches of speech are used, i.e. complete utterances or a couple of utterances.

Error detection can also be carried out for number of utterances in combination; however, for language learning one generally prefers immediate feedback (and not feedback after a number of utterances have been spoken). For pronunciation error detection, some approaches can be used:

1. focus on frequent errors
2. ASR-based metrics
3. acoustic phonetic classifiers

In the first approach, errors frequently made by language learners are explicitly taken into account [10]. For instance, if the sound /r/ is often pronounced as /l/ (e.g. 'angly' instead of 'angry'), then this frequent substitution can be hard-wired in the pronunciation models. If the speech recognizer decides that the best path 'goes through' the /l/ sound, then the system knows that probably this pronunciation error is made.

In the second approach, ASR-based metrics are used, such as posterior probabilities and (log) likelihoods ratios [8 9]. Previous research has shown that these confidence measures can be used for detecting pronunciation errors [8, 9, 16]. A special case concerns the so-called goodness of pronunciation algorithm (GOP) [16], which has been used quite often. We also studied the GOP algorithm in our research, and used it in our Dutch-CAPT system. If trained well, the GOP algorithm works quite well: on average about 80% of the sounds are classified correctly. However, there are large variations between persons and sounds. If specific settings could be used for each person sound combination, better results could be achieved; but currently this is not possible in practice. The challenge here is to find groups for which similar settings perform well.

Acoustic phonetic classifiers are not often used in call applications; still they can be useful [14]. We compared the results of acoustic phonetic classifiers to those obtained with the GOP algorithm, and it turned out that results for acoustic phonetic classifiers were better [14].

As often, a combination of approaches probably will yield the best results. Therefore, the challenge here is to find the proper combination of approaches and settings for which the results are best.

4. Future: challenges and opportunities

In this paper we have shown that ASR, in spite of its limitations, already holds great potential for language tutoring and could be employed for various language learning goals such as literacy development, reading, oral proficiency, and vocabulary learning. It is clear that developing good applications requires mixed expertise: knowledge of speech technology, education / pedagogy, language acquisition, software design and development. Developing good products therefore requires that the right people work together: speech technologists, teaching professionals, software designers and industrial partners (e.g. publishers).

At the same time, a globalized world characterized by increasing internationalization and mobility will continue to require products and material that make it possible to learn new languages efficiently and effectively to guarantee the integration of migrant workers into their new surroundings.

5. References

The references below are listed in alphabetical order.

[1] ALR (1998) Putting Pronunciation Programs Through Their Paces, American Language Review, 2. http://www.languagemagazine.com/internetedition/mj98/eepp56.html (retrieved November 29, 2007)

[2] Bloom, B. S. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 1984, 4-16.

[3] Coniam, D. (1999) Voice recognition software accuracy with second language speakers of English. System, 27, 49-64.

[4] Cucchiarini, C., Strik, H., & Boves, L. (2002) Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America, 111(6), 2862-2873.

[5] Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? TESOL Quarterly, 34, 592-603.

[6] Ehsani, F., & Knodt, E. (1998) Speech technology in computer-aided learning: Strengths and limitations of a new CALL paradigm. Language Learning & Technology, 2, 45-60.

[7] Eskenazi, M. (1999) Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype, Language Learning & Technology, 2, 62-76.

[8] Franco, H., Neumeyer, L., Digalakis, V., & Ronen, O. (2000b) Combination of machine scores for automatic grading of pronunciation quality. Speech Communication, 30, 121-130.

[9] ISLE 1.4 (1999) Pronunciation training: Requirements and solutions, ISLE Deliverable 1.4. Retrieved February 27, 2002, from http://nats-www.informatik.uni-hamburg.de/~isle/public/D14/D14.html.

[10] Kawai, G., & Hirose, K. (1998) A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training, Proceedings of ICSLP, Sydney, Australia, 1823-1826.

[11] Menzel, W., Herron, D., Bonaventura, P., & Morton, R. (2000) Automatic detection and correction of non-native English pronunciations, Proceedings of InSTiL, Dundee, Scotland, 49-56.

[12] Neri, A. (2007) The pedagogical effectiveness of ASR-based computer assisted pronunciation training. PhD thesis, University Nijmegen.

[13] Neri, A. Cucchiarini, C. and Strik, H. (2007) Pronunciation training in Dutch as a second language on the basis of automatic speech recognition, Stem, Spraak- en Taalpathologie, 159-169.

[14] Strik, H., Truong, K., De Wet, F., and Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. Proceedings of Interspeech 2007, Antwerp.

[15] Van Compernolle, D. (2001) Recognizing speech of goats, wolves, sheep and ... non-natives. Speech Communication, 35, 71-79.

[16] Witt, S. (1999) Use of Speech Recognition in Computer Assisted Language Learning. PhD thesis, University of Cambridge.

System ZENON – Semantic Analysis of Intelligence Reports

Matthias Hecking

FGAN/FKIE, Neuenahrer Straße 20, 53343 Wachtberg-Werthhoven, Germany
hecking@fgan.de

Abstract

The new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of Human Intelligence (HUMINT) reports. These reports are good candidates for applying techniques from computational linguistics. In this paper, the ZENON system is described, in which an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. The objective of this research is to realize a navigatable Entity-Action-Network. The information about the actions and named entities are identified from each sentence. These representations can be combined and presented in a network. After a short introduction, the information extraction approach is explained. The ZENON system is described in detail. English HUMINT reports from the KFOR deployment form the basis for the development of the experimental ZENON system. These reports are used to build the KFOR text corpus, which is described as well.

Index Terms: information extraction, text mining, semantic analysis, intelligence, HUMINT reports, text corpus

1. Introduction

The *processing of human language* was identified as a critical capability in many future military applications [1]. Especially the *content analysis* of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept [2, p. 5-15]. The content analysis can be realized through *Information Extraction* (IE) which is a natural language processing technique [3], [4].

We set up the *research project ZENON*[1], in which the information extraction (IE) approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr [5], [6], [7], [8], [9], [10], [11]. The overall objective of this research is to create a *graphically navigatable Entity-Action-Network*. The information about the actions and named entities are identified from each sentence and the content of the sentences are formally represented. These formal representations can be combined and presented in the navigatable network.

2. Information extraction

In the last decades various techniques for processing spoken and written natural languages were developed (e.g. speech recognizer in dictation systems, machine translation, grammar checking). IE is an engineering approach [3] for content analysis of free-form texts based on results of computational linguistics. Each IE system is tailored to a specific domain and task. IE uses a *shallow syntactic approach* [6], i.e. that only parts of the sentences (so-called 'chunks') are processed with finite state automatons or transducers.

During the IE relevant information about the Who, What, When, etc. in natural language texts is identified, collected, and normalized. The relevant information is described through patterns called *templates*. These domain and task specific templates represent the meaning of the relevant information. During the IE task the templates are filled with the extracted information. One possibility to realize the templates is to use *typed feature structures* [7]. Therefore, IE can be seen as the process of normalizing free-form text into a defined semantic structure.

To realize an IE system, language-specific resources (lexicon, grammar) and appropriated parsing software are necessary.

In order to achieve robust and efficient IE systems, domain knowledge must be integrated and shallow algorithms must be used. The domain knowledge is tightly integrated with the language knowledge, e.g., the name 'Leopard' in the lexicon has the categorical information 'tank'. This association between words and semantic information is domain-specific and has to be change for other applications.

The IE is used as the core natural language processing technique in the ZENON project.

3. ZENON system

Starting with English HUMINT reports (and a list of the city names) from the KFOR deployment of the Bundeswehr we have realized in the ZENON system that is able to do a (partial) content analysis of these reports [8]. The content of these KFOR reports are from a wide spectrum. Apart from descriptions of conflicts between ethnic groups, tensions between political parties, information about infrastructure problems, etc. there are also reports, which concern individuals or other entities. Statements of the form *A meets B*, *A marries C*, *A shoots B*, etc. contains information about activities/events and involved entities. This information, completed with location and time data, is combined into a graphically navigatable *Entity-Action-Network* (e.g.; with a person in the center of the network). The intelligence analysts can use this network to navigate through the content of the reports.

3.1. Toolbox GATE

Since most of the reports are in English, GATE (General Architecture for Text Engineering, [12] was selected as the used toolbox. GATE is an architecture, a free open source framework (SDK) and graphical development environment for Natural Language Engineering and offers a lot of processing resources, which are used to realize the natural language processing parts of the ZENON system (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases,

[1] according to: Zenon of Citium, 336 BC - 264 BC, philosopher, founder of the Stoicism

chunk-parsing). The functionality to select, combine and present the extracted information from different sentences and different reports is realized by XSLT (Extensible Stylesheet Language Transformation) filtering and the *Information Extraction Presentation System* (IEPS, [13]).

3.2. ZENON processing chain

In Figure 2 the ZENON processing chain is shown. HUMINT reports are fed into the first sub-component. In this component the natural language text is tokenized (i.e., find words, numbers, etc.), the sentence boundaries are detected, the part-of-speech (i.e., whether it's a noun, a verb, etc.) is determined, simple names of cities, regions, military organizations etc. are annotated (through the Gazetteer), named entities (i.e., complex names of e.g. political organizations, person names, etc.) are recognized and a morphological analysis is done. The result of this sub-component are the annotated sentences of the reports. The second sub-component uses these annotations to extract the action type (e.g., 'kill') starting with the verb of the sentence. If the action type is determined the other parts of the sentence (e.g., subject, object, time expressions) are located and formally represented in *typed feature structures*. These structures are coded in XML (Extensible Markup Language) format and represent the output of the natural language part of the ZENON system. In the third sub-component the extracted content of different reports can be combined and selected according to predefined XSLT sheets. The result of the analysis is presented graphically and can be navigated interactively.

3.3. Named entities

An important processing step during the natural language processing is the recognition of the domain- and application-specific named entities. In the ZENON system transducers for the recognition of the following named entities were developed: *City*, *Company*, *Coordinates*, *Country*, *CountryAdj*, *Currency*, *Date*, *GeneralOrg*, *MilitaryOrg*, *Number*, *Percent*, *Person*, *PoliticalOrg*, *Province*, *Region*, *River*, *Time* and *Title*. An example is shown in Figure 1.

```
Rule: FVGPrePerPasNeg
//Recognizes:       Present Perfect Passive Negative:
//                  e.g. "hasn't been eaten"
//Pattern: (has | have) not been VBN
//Output: VG{adverb, infinitive, neg='yes',
//          tense='PrePer', type='FVG', voice='passive'}
(
        (
                {Token.string == "has"}|
                {Token.string == "have"}
        )
        (NEGATION)
        {Token.string == "been"}
        (ADVS):adverb
        ({Token.category == VBN}):verb
):x ==> {... Java code …}
```

Figure 1: *Verb phrase transducer 'FVGPrePerPasNeg'.*

3.4. Extraction of verb phrases, action types and sentence content

GATE offers various transducers to recognize the English verb groups. We have adapted and extended these transducers to fit our application. In addition to finite and non-finite verbal phrases also modal verb phrases, participles and special composed verb expressions are recognized.

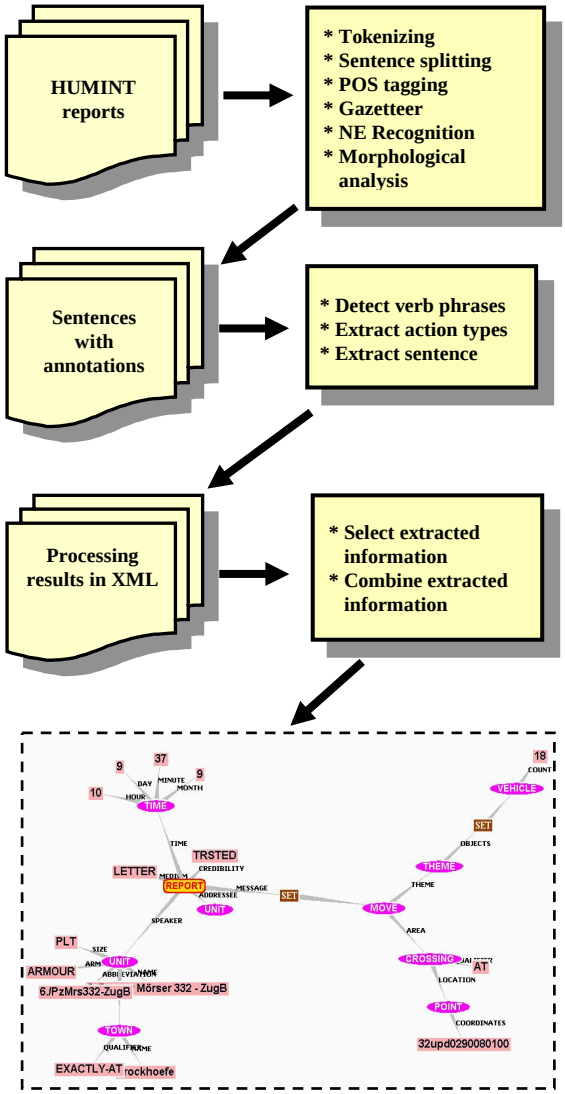


Figure 2: *ZENON processing chain.*

Based on the recognized verb groups different *action types* can be detected (e.g., from the infinitive of 'murder', 'kill', 'decapitate', … the action class 'kill'). After detecting the action type the verb phrase and other parts of the sentence must be combined. In the ZENON project we use the *semantic frames* from the FrameNet project [14] to realize this combination. Semantic frames are schematic representations of situation types (eating, killing, spying, classifying, etc.) together with lists of the kinds of participants, objects, and other conceptual roles that are seen as components of such situations. These semantic arguments are called the *frame elements* of the frame. Figure 3 shows an example. The core (must exist) frame elements for the frame 'killing' are CAUSE or KILLER and VICTIM. In the example

sentence 'John' fills the role KILLER and 'Martha' fills the role VICTIM.

Associated with each semantic frame are examples with typical syntactic realization of the frame elements. These examples and examples from the KFOR reports form the basis to construct the transducers, which produce the sentence content.

Semantic Frame 'killing':	A KILLER or CAUSE causes the death of the VICTIM.
Core frame elements:	CAUSE, KILLER, VICTIM
Non-core frame elements:	DEGREE, DEPICTIVE, INSTRUMENT, MANNER, MEANS, PLACE, PURPOSE, REASON, RESULT, TIME
Example sentence:	[John $_{KILLER}$] DROWNED [Martha $_{VICTIM}$]

Figure 3: *Semantic frame 'killing'.*

During the processing, the associated semantic frame is inferred from the detected action type. With the identified semantic frame the core and non-core frame elements are give. Recognized named entities, POS tagging and expressions from the sentences are used to fill in the frame elements.

3.5. Meaning space navigation

The natural language processing module of the ZENON system creates for the relevant sentences in each KFOR report a formal representation of the content. This contains information chunks about activities, events, entities, times and places. These basic units are selected and combined (e.g., all information about a specific person) through complex filters which can be defined for each scenario in the ZENON system. The filter functionality is realized through *Extensible Stylesheet Language Transformation* (XSLT, [15]). The result of the transformation is in XML format.

The intelligence analyst must be able to access and explore this *meaning space*. Therefore the meaning space is visualized as a *graphically navigatable Entity-Action-Network* by the sub-component IEPS [13]. This is a graphical software tool (see Figure 4) for visualizing information typically extracted from free-form texts by a natural language processing system. Additionally, it offers a framework to organize all the files being employed during the processing in user-defined scenarios and to activate the IE process.

4. KFOR Corpus

4,498 military reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces were used for the realization and optimization of the ZENON system. From these reports 800 were manually annotated and form the *KFOR Corpus*.

This corpus is a specialized micro-text corpus [16]. The corpus covers 886,000 tokens and contains the annotations in different *annotation layers* [11]. The following layers are available:

- *Original markups*: In this layer those parts of the message are annotated that are already formatted (e.g. addressee, topic, source).

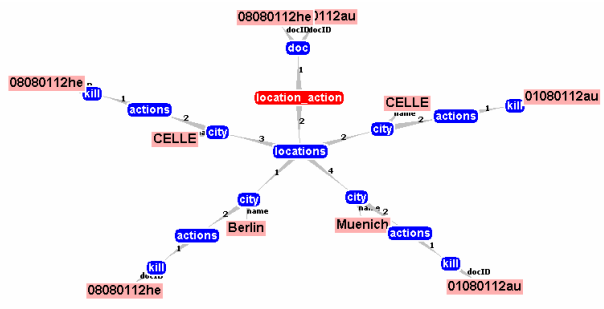


Figure 4: *Location-action meaning space.*

- *Token*: This layer contains the annotations about words, numbers, etc. The part-of-speech information and the lemma are also given.
- *Gazetteer*: In this layer those expressions are annotated that were identified over lists of names (e.g., first names, city names).
- *Sentence*: These annotations refer to sentences and begin and end markers of comments.
- *Named entities*: In this layer the above listed named entities are annotated.
- *Verb group*: The verbal phrases are annotated.
- *Thematic roles*: The syntactic and semantic function of expressions in sentences is annotated [17].

During the creation of the corpus a first version of the annotations was produced automatically. These annotations were then checked manually and corrected. GATE was used for both working-steps.

The corpus contains both syntactic and semantic annotations. The Figure 5 indicates, which annotation layers and annotation types are present, whether they are syntactic or semantic annotation types, and which of the annotation types were manually corrected.

Syntacticl/ semantical	Annotati. layer	Annotation type	Checked manually
syntactical	Original markup	DocID, DTGMeldung, Einsatz, Empfaenger, Hauptthema, Koordinate, Meldung, Meldungstyp, Ort, Quelle, Sachverhalt, Schlagworte, Titel, Unterthema	no
syntactical	Token	Token, SpaceToken	no
semantical	Gazetteer	Lookup	no
syntactical	Sentence	Sentence Comment Split	yes yes no
semantical	NE	City, Company, Coordinates, Colour, CountryAdj, Currency, Date, DocumentID, GeneralOrg, MilDateTime, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time, Title	yes

syntactical	VG	VG	yes
semantical	Thematic Role	ThRo	yes

Figure 5: *Annotation layers and annotation types*.

For each *annotation* the type, the layer, the start- and the end-position and a set of annotation-specific *features* are given. Each feature consists of a name and a value. A feature appears only, if a value is present. In the example

```
City NE xxx yyy {name=BERLIN}
```

the annotation is of type `City`. It belongs to the annotation layer `NE`. The string to which the annotation refers begins in position `xxx` and ends with position `yyy`. The annotation possesses a feature with the name `name` and the value `BERLIN`.

5. Development status

The 1st version of the ZENON system was realized. The system is able to process the action classes KILL, REPORT, KNOW, COMMAND, PROPOSE, EXPLODE and its associated semantic frames. The NE transducers were optimized with the help of the KFOR corpus. The verb group transducers were extended to recognize also modal verb phrases, participles and special composed verb expressions.

For the planned the 2nd version of the ZENON system a HUMINT ontology is under construction. In this new version the information extraction will also be *multilingual*. For this, processing resources to handle the language *Dari* were developed (cf. [18]).

6. Conclusions

In this paper, the ZENON project was presented. In this project an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. First, a short introduction into the information extraction approach was given. Then, the ZENON system was described in detail. The GATE toolbox, the processing chain, and the extraction of named entities, verb phrases, action types and the sentence content was explained. The meaning space navigation was also mentioned. At the end, the KFOR corpus and the development status were presented.

7. References

[1] Steeneken, H. J. M. *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Survey*. NATO, Technical Report, AC/243(Panel 3)TP/21, 1996.

[2] Department of Defense. *Network Centric Warfare – Report to Congress*. 27 July 2001.

[3] Appelt, D. & Israel, D. *Introduction to Information Extraction Technology*. Stockholm: IJCAI-99 Tutorial, 1999, http://www.ai.sri.com/~appelt/ie-tutorial/.

[4] Hecking, M. *Informationsextraktion aus militärischen Freitextmeldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 74, 2004.

[5] Hecking, M. *Information Extraction from Battlefield Reports*. In: Proceedings of the 8th International Command and Control Research and Technology Symposium (ICCRTS), Washington, DC, U.S.A., 2003.

[6] Hecking, M. *Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques*. Paper presented at the RTO IST Symposium on „Military Data and Information Fusion“, held in Prague, Czech Republic, October 20-22, 2003.

[7] Hecking, M. *How to Represent the Content of Free-form Battlefield Reports*. In: Proc. of the 2004 Command and Control Research and Technology Symposium (CCRTS) "The Power of Information Age Concepts and Technologies", June 15-17, 2004, San Diego, California.

[8] Hecking, M. *Domänenspezifische Informations-extraktion am Beispiel militärischer Meldungen*. In: A.B. Cremers, R. Manthey, P. Martini, V. Steinhage (Hrsg.) "INFORMATIK 2005”, Band 2, Lecture Notes in Informatics, Volume P-68, Bonn, 2005.

[9] Hecking, M. *Content Analysis of HUMINT Reports*. In: Proc. of the 2006 Command and Control Research and Technology Symposium (CCRTS) "THE STATE OF THE ART AND THE STATE OF THE PRACTICE", June 20-22, 2006, San Diego, California.

[10] Hecking, M. *Navigation through the Meaning Space of HUMINT Reports*. In: "Proceedings of the 11th International Command and Control Research and Technolgy Symposium”, September 26-28, 2006, Cambridge, UK.

[11] Hecking, M. *Das KFOR-Korpus als Ergebnis semantisch annotierter militärischer Meldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 124, 2006.

[12] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[13] Casals Elvira, X., Hecking, M.. IEPS: *A Framework to Manage and to Visualize Information Extraction Results*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Technischer Bericht FKIE/ITF/2005/2, September 2005.

[14] *FrameNet*. http://framenet.icsi.berkeley.edu/index.php (24.10.2007).

[15] *XSL*. http://www.w3.org/TR/xslt (24.10.2006).

[16] McEnery, T., Wilson, A.. *Corpus Linguistics*. Edinburgh University Press, Edinburgh, 2nd edition, 2001.

[17] Kremer, C. *Eine Untersuchung von Bewegungsverben im KFOR-Korpus im Vergleich zu FrameNet*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 117, 2006.

[18] Schwerdt, C. *Analyse ausgewählter Verbalgruppen der Sprache Dari zur multilingualen Erweiterung des ZENON-Systems*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht, 2007.

New Technologies for Simultaneous Acquisition of Speech Articulatory Data: 3D Articulograph, Ultrasound and Electroglottograph

Mirko Grimaldi[1], *Barbara Gili Fivela*[1], *Francesco Sigona*[1], *Michele Tavella*[2], *Paul Fitzpatrick*[3], *Laila Craighero*[3], *Luciano Fadiga*[3], *Giulio Sandini*[2], *Giorgio Metta*[2]

[1] Centro di Ricerca Interdisciplinare sul Linguaggio, University of Salento, Lecce, Italy
[2] Laboratory for Integrated Advanced Robotics, University of Genoa, Italy
[3] Dep. S.B.T.A., Section of Human Physiology, University of Ferrara, Italy

{mirko.grimaldi, barbara.gili, francesco.sigona}@ateneo.unile.it;
{michele, paulfitz, pasa}@liralab.it; {fdl, crh}@unife.it; sandini@dist.unige.it

Abstract

The study of articulatory features in speech requires highly sophisticated instruments. Most of them are available at the CRIL research centre (*Centro di Ricerca Interdisciplinare sul Linguaggio* – University of Salento – Lecce) where the equipment usually found for articulatory studies in the most advanced international research centres is accessible: 3D articulograph, ultrasound system, electroglottograph, electropalatograph, and aerophone. Some of these instruments have been synchronized within the European CONTACT project, and have already been used for simultaneous recording.

In this paper, we will point out the peculiarity of each instrument, the added information related to the simultaneous recording, and the main steps towards their complete fruition in the specific phonetic-phonology field. In order to do that, we will briefly describe the recording set up and the first data collection, we will focus on a subset of the recordings in order to provide examples of the achieved articulatory-phonetic information, and describe the work-in-progress software analysis environment we use for linguistic purposes.
Index Terms: 3D articulography, ultrasound, electroglottography.

1. Introduction

The study of articulatory features of speech requires the use of an appropriate technology, often specifically developed for this purpose. In CRIL (*Centro di Ricerca Interdisciplinare sul Linguaggio* – University of Salento – Lecce) the main equipment used for articulatory studies in the most advanced international research centres is available: 3D articulograph, ultrasound system, electroglottograph, and aerophone [1], [2]and [3]. Within the European CONTACT project, some of these instruments have been synchronized and are usable to register various material simultaneously [4]. The aim of the CONTACT project is to verify if the development of language can be in parallel linked to the motoric control acquisition, especially the control that is necessary for precision movements [5], [6]. In this perspective, 3D articulograph, ultrasound system and electroglottograph have been synchronized (cf. [6]) in order to acquire articulatory material, which – supplied as input to an automatic speech recognition system – could eventually improve the speech recognition abilities. The material which was registered within the project – that is, a corpus of words and one of pseudowords, read by 9 speakers of the area of Lecce – was chosen according to the main objective of the project, and the methodology was adapted to its purposes. Actually, the registered material is not entirely suitable to verify specific linguistic hypotheses. However, in this paper we will take into account the pseudoword corpus, in order to point out the peculiarity of each instrument and the main aspects of their simultaneous use, and, furthermore, the main steps towards their complete fruition in the specific the phonetic-phonologic field.

Also, we describe the analysis environment and provide examples of the articulatory information which can be obtained by this type of material. The analysis environment – developed in MatLab® [7] – allows to display the wave shape corresponding to the electroglottographic signal, any interval which has been identified thanks to previous segmentation and labelling phases, the pictures obtained simultaneously by the ultrasound equipment and the graphs of the articulograph sensors.

2. Instrumentation

A constellation of hardware and software that allows the simultaneous recording of acoustic and articulatory data has been implemented. It is made up of several interconnected devices; some of those are synchronized in hardware, while the remaining rely on post-processing synchronization (for more details see [6]:49-96).

2.1. 3-D Electromagnetic articulograph

An electromagnetic articulograph (EMA), model AG500 by Carstens Medizinelektronik GmbH, Germany [8], was used to track the movements of a set of sensors glued on the tongue, on the lips, on the front teeth and on a pair of glasses (the latter, used as reference, to compensate head movement), during speech production (Figure 1 shows how the sensors are

Figure 1: *EMA sensors glued on the subjcet's tongue*

positioned onto the tongue). The EMA device is able to determine the Cartesian coordinates x, y and z, as well as azimuth and elevation of up to twelve directionally sensitive magnetic field sensors at a sampling frequency of 200Hz. The measuring sensors are single axis coils. Six reference coils, arranged to form a three-dimensional frame of reference, emit six magnetic fields at different well known frequencies between 7500 and 13750 Hz. During a recording session, the alternating currents induced in the sensors by the magnetic fields of the reference coils are separated by their frequencies, digitized and sent in real-time to the AG500 control PC (using an Ethernet connection). Software provided by Carstens Medizinelektronik GmbH stores the current values on the hard drive of the control PC, making them available for the spatial arrangement determination process.
It is important to stress that the speech signal coming from a microphone is also synchronously recorded with positional data.

2.2. Ultrasound system

The investigation of the tongue motion is also accomplished by means of an ultrasound system, which is both non-invasive and non obtrusive, thus non affecting speech production, and is able to provide the full profile of the tongue dorsum, although the apex and the radix are often occluded during speech production by the jaw and by the hyoid bone respectively.

Medical sonography uses high frequency (2-14 MHz) sound waves emitted from an array of piezoelectric transducers (crystals), multiplexed in time so that only one crystal emits sound waves in a given time interval, while all the remaining crystals are used to convert the received echoes to voltage values. Indeed, when an ultrasound wave reaches an interface between materials with different impedance properties, it is reflected. Voltage values of the received echoes are processed and the image is reconstructed.

As it can be easy verified in the following echography pictures, the brightest area in the achieved images is the subject's tongue. The brighter the pixel is, the higher the density gradient is, thus the acquired echoes. The tongue lamina is clearly visible since a great density gradient exists.

Actually, an Aplio XV machine, by Toshiba Medical System corp. [9] has been used, which also allows exporting ultrasound pictures as a continuous video stream (at 25 Hz) by means of a dedicated S-Video output. The video stream is synchronously acquired together with the audio signal, by means of an external a/v analog-to-digital acquisition card, and then recorder in real-time on a dedicated PC.

2.3. Electroglottograph

Electroglottography is a technique used to register laryngeal movements measuring the change in electrical impedance across the throat during speech production. The EGG device used for the Linguometer is a "Laryngograph Microprocessor" by Laryngograph Ltd, London, United Kingdom [10]. An alternated current generator supplies a high frequency (usually from 300 kHz to 5 MHz) sinusoidal current to a pair of copper-made electrodes, kept by an elastic band on the surface of the throat at the level of the thyroid cartilage. The current amplitude is in the order of few milliampere while the applied voltage is around 0.5 V. When the vocal folds move, a rapid variation of the conductance is observed in the electroglottographic (EGG) signal applied across the larynx.

Driven by a dedicated software, the elettroglottograph acquires both the EGG signal and the speech signal synchronously, at 16 kHz, and sends them to an attached control PC using a standard USB interface.

The system enables the study of the regularity of the vocal folds vibrations, the single opening and closing phases, their correlation and shape.

2.4. Facial expressions recording

During each experiment, facial expressions of the subject have been recorded on a tape, in DV format, by means of a camcorder. The tape content is exported off-line on a PC as a single a/v stream, to be segmented later. At the moment, the actual purpose of facial expression recording is not to be matter of study, but to test the segmentation and the synchronization procedures on a type of signal which does not embed the segmentation pulses in itself, while heavily deploying cross-correlation between the speech component and the speech reference signal ([6] : 58-59).

2.5. Stimuli presentation, signal recording and control

In the data acquisition setup for the CONTACT project, a GNU/Linux-based workstation handles the stimuli (i.e. the corpora) presentation. Furthermore, since all of the signals are acquired as a set of continuous streams (and they need to be segmented off-line in order to make the analysis of single words and pseudowords easier), the stimuli presentation software is also in charge to generate a segmentation signal (a set of well-known acoustic pulses).

The segmentation signal is entered into an audio mixer together with the speech signal coming from a couple of microphones. The resulting audio signal is then entered in a a/v analog-to-digital acquisition card together with the continuous video stream, as depicted before.

Another Microsoft Windows-based PC run the Cartsens Medizinelektronik GmbH software to calibrate and record the data of the articulograph. Furthermore, it run the software required by the laryngograph.

A server shared on the network is used to store the data recorded during the investigations.

Finally, a notebook computer was used to control the recording apparatus and to monitor the execution of the stimuli presentation software.

2.6. Post-processing

As already said, signals are acquired as continuous streams: an off-line software procedure is in charge to segment those streams, i.e., to separate each word the submitted corpus is made of.

Also, data coming from different sources, such as EGG and US data, need to be time-aligned, i.e. for each word/pseudoword, the starting instant for a signal must be related to the starting instant of the others, in order to make it possible to see what happens to all of the signals at a given time interval. This task is accomplished by another software tool, which basically deploys the cross-correlation between the audio signals synchronously recorded together with EGG and US respectively.

Finally, the resulting data are packaged in a Matlab-compliant format, so that they can be analyzed and shared among the scientific community in a *de facto* standard format.

Post processing software includes C/C++, Perl, Matlab programming technologies and run almost full-automatically

(it may requires several hours of computation, depending on a lot of factors such as the CPU speed, the amount of data to be processed, etc.): just a little human control is required at the very beginning of operations.

2.7. Software for data analysis

In order to show the result of the data acquisition procedure, a simple software analysis environment has been set up. This environment is under development, and relies on the main idea to integrate as much as possible already existing software tools. Actually, the environment relies on PRAAT [11], Edgetrak [12], Mplayer [13][13] and a Matlab-based GUI developed at CRIL (let's call it "mygui") which is able to interact with the above-mentioned companions within certain limits, also thanks to the scripting capabilities of the operating system, and the command-line interface of Praat and Mplayer.

Driven by the Matlab GUI commands, Mplayer is used to extract all of the pictures from the ultrasound and facial expression movies, in order to be displayed with ease when requested. Mplayer is also used to build demonstrative a/v movies of the simultaneous playback of all the synchronized signals.

PRAAT is deployed directly by the user at the very beginning of the data analysis stage, to build multiple levels of labels attached to one or more selected time intervals of interest, related to a word/pseudoword: the text file describing labels can be automatically imported in the Matlab environment by means of an ad-hoc script. Even if Matlab is already equipped with a toolbox to generate spectrograms, "mygui" has been designed to drive Praat to generate spectrogram (with superimposed formants) plots and to export them as text files, which are then imported and displayed within "mygui".

Edgetrack is a software that allows to generate a dotted contour of the tongue on the mid-sagittal plane, for a sequence of ultrasound images. The user is asked to select the sequence and help the program tuning some parameters to generate the contour for the first picture: Edgetrak will generate dotted contours for the remaining ones, following the tongue movements picture by picture. The set of contours can be exported in text format: "mygui" can import and use the file to superimpose the contours over the appropriate pictures again.

"Mygui" is the core of the environment, and is used to automatically import and show all of the available and time-aligned information, including EGG signal, facial expression pictures, speech signal, spectrogram and formants, x-y-z coordinates of the EMA sensors, belonging to a selected word/pseudoword. In addition, the user can display single ultrasound pictures, with or without superimposed Edgetrak tongue contours, reference grids and the EMA sensors glued on the tongue dorsum (within the current implementation of the system, the latter feature is a challenge, and actually is just an estimation of the most likely positions: at the moment it should not be considered for quantitative relative measurements between EMA sensors and the tongue dorsum resulting in the ultrasound picture). Of course the user can filter the data in order to show only a subset, or choose a different time interval with respect to the ones associated to the Praat labels at the very beginning of the operations.

3. First data collection and observations

As already mentioned, a corpus of data was recorded within the CONTACT project. The aim was getting articulatory data to train an automatic learning system. Part of these data will be considered here to point out the impact of the simultaneous acquisition of articulatory data on phonetic studies, focusing on specific linguistic hypotheses.

Nine speakers of Lecce Italian were asked to read three times a corpus of 74 words and 68 pseudo-words (both words and pseudo-words were read as declarative sentences; words were also read as questions). Both words and pseudo-words were chosen in order to include all consonantal and vocalic phonemes attested in Italian. In particular, words were mainly stress-initial (e.g., /'matto, 'nome, 'strada/ <mad, name, street>) and were chosen in order to show the various consonants in word initial position, followed by different vowels (e.g., /'matto, 'muffa, 'moro/ <mad, mould, dark >); moreover, instances of words with different stress positions were also inserted (e.g., /mat'tone, pa'pa/ <brick, dad>). Pseudo-words were monosyllables, chosen in order to get data on all the Italian consonants followed by /a, u, i/ vowels (e.g., /'na, 'nu, 'ni, 'λa, 'λu, 'λi, /).

We will focus here on the recordings of two speakers regarding pseudo-words composed by unvoiced alveo-dental/post-alveolar articulation (/t, s, ts, tS/) and a vowel (/a, i, u/). The acoustic signal was first segmented and labelled by means of the software PRAAT [9]. For each monosyllable, both the syllable segment boundaries and the transition(s) between segments were identified (e.g., for [tSu], the transitions between [t] and [S], and [S] and [u]). The inspection of articulatory data was performed by means of the Matlab script described above (see previous section). The data analysed allow us to point out the possible specific contribution of the instruments described above, used simultaneously for linguistic analysis.

3.1. Integrating information on articulatory gestures

The electroglottograph offers an unambiguous indication of the boundaries between voiced and unvoiced segments: the boundary is clearly detectable, independently from the segmentation procedure and the settings for acoustic analysis (e.g., spectrogram settings); in voiced stretches, it offers information on vocal fold vibration during voicing.

The ultrasound system offers information on the morphology of the tongue and integrates the information on specific points obtained by means of the AG500 sensors. In particular, the surface of the whole tongue during the affricate production allowed us to observe the specific transition gesture in the plosive-fricative and fricative-vowel segments. The ultrasound images are definitely useful in relation to the back of the tongue, an area which is particularly difficult to track – see Figure 2. Notice that no AG500 sensor may be glued as back as the point indicated in the figure without causing problems to subjects.

The AG500 offers highly detailed information on specific points located on the articulators, as for both the spatial and the temporal domain. First of all, the system tracks the lip kinematics that would not be available with the other instruments alone – see Figure 3; moreover it integrates the information on tongue morphology obtained by means of the ultrasound system, both in terms of spatial-temporal resolution and in terms of specific points than may be monitored, e.g., the tip of tongue is often missing or difficult to detect in ultrasound images.

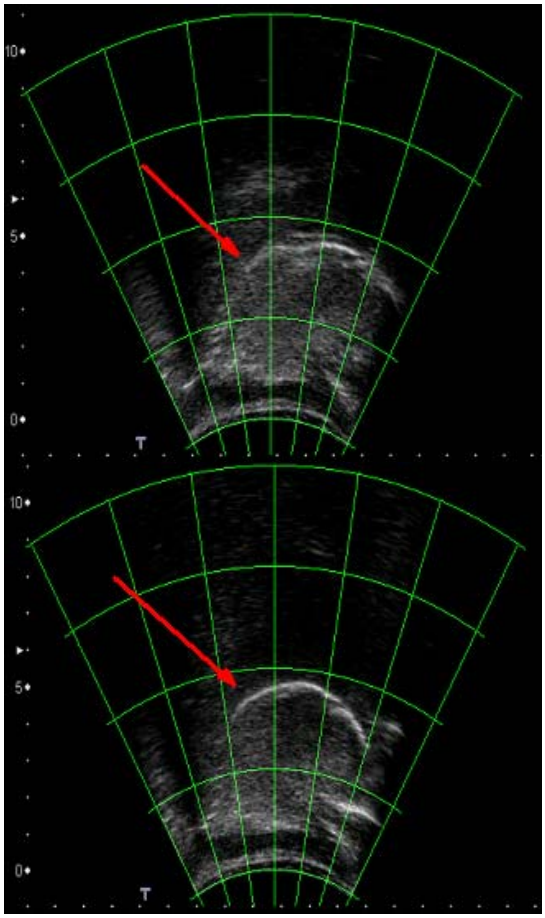


Figure 2*: Tongue position during the production of pseudo-word 'ciu' /tSu/: back of the tongue position (red arrow) at the beginning and at the end (after 120 ms) of the consonant-vowel transition – upper and lower panel, respectively.*

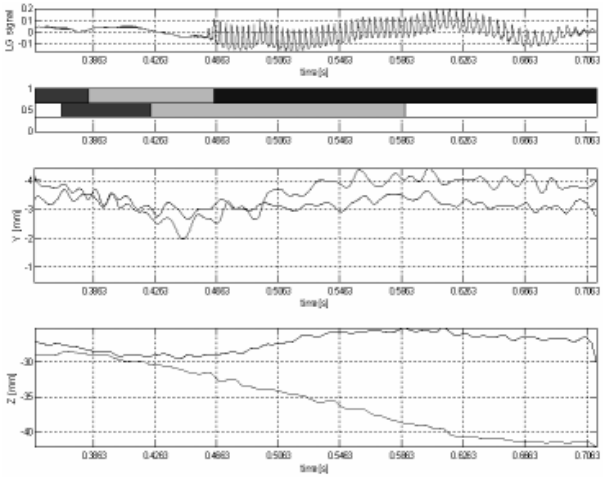


Figure 3 *EGG signal (1st panel from top), segments, transition intervals and slots for ultrasound images (2nd panel), and tracks of AG500 sensors during the production of 'ciu' /tSu/: sensors placed on the lower and upper lips – (3rd panel) – and on the tip and dorsum of the tongue – lower panel.*

3.2. Ultrasound tongue contour and EMA data

One of the possible perspectives of analysis achievable using US and EMA techniques is based on the observation that during speech production the tongue deforms in a complex manner, probably because the gesture is influenced by more than one phone. A problem is deeply related to this: how is this highly deformable tissue of the tongue controlled to obtain fine gestures necessary for speech? Classical hypotheses postulated a tip and body subdivision of the tongue executing quasi-independent motion back and forth and bottom and up directions. However, thanks to recent imaging techniques, it is possible to suppose that tongue deformations are far more complex than the motion of a tongue body and a tongue tip [14], [15]. A more recent hypothesis is that the tongue would be controlled by the synergistic coordination of muscular 'functional segments'. These segments are laid out orthogonally to the longitudinal axis of the vocal tract and composed of multiple muscle systems. From this point of view, the tongue can be divided into quasi-independently controlled functional segments based on regions of the tongue and vocal tract, rather than gross muscle architecture. Instead of entire muscles aligning to execute a gesture, segments would be controlled independently or aggregated into larger units to form coordinative structures determined by language dependent phonetic considerations [16].

Among the muscles that act on the tongue structure, there are three extrinsic muscles that originate on bone structures and insert into the tongue (responsible for the main displacement and shaping of the overall tongue structure): the genioglossus, GG (different fibres produces a forward and upward movement), the styloglossus (raises and retracts the tongue, causing a bunching of the dorsum in the velar region), and the hyoglossus (retracts and lowers the tongue body). The floor of the mouth contains also the geniohyoid (GH) and the mylohyoid (MH) muscles, that elevate the hyoid bone and the base of the tongue. Three additional intrinsic muscles contribute to a minor extent to the sagittal tongue shape. The superior longitudinalis muscle shortens the tongue, and bends its blade upwards. The inferior longitudinalis muscle depresses the tip. The verticalis fibres depress the tongue and flatten its surface ([17]: 171-177; see Figure 4).

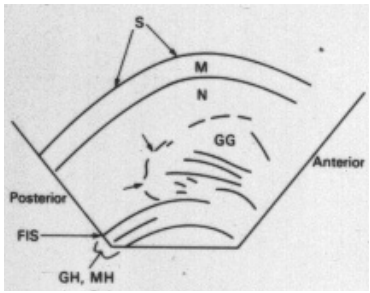


Figure 4: *schematic of US mid-sagittal scan (adapted from [1]: 12). S = Tongue Surface; M = mucosa; N = Superior, Inferiorior logitudinalis, and Vertical, Horizontal network of fibres; FIS = Floor Intermuscular septum.*

Using our data, a preliminary attempt in this direction was made observing the role of tongue muscles and the position of the tongue surface, both in US images and in EMA sensors tacks.

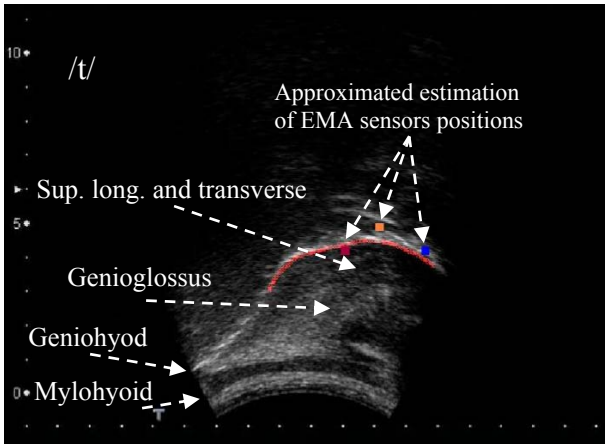


Figure 5: *closure phase and burst in phoneme /t/.*

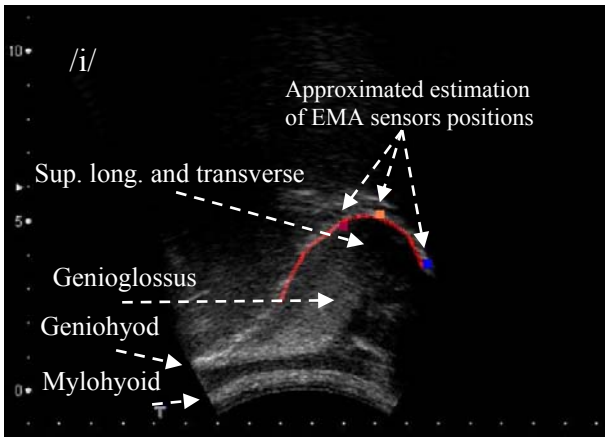


Figure 6 raising and fronting in phoneme /i/.

We can try to analyze a complex segment as /ti/ consisting of an occlusive and an anterior high vowel. We can observe the realization of /t/ in Figure 5 and of /i/ in Figure 6.

It's possible to note that phonemic differentiation in mid-sagittal tongue motions is realized through a functional independence of the tongue muscles. In particular, in the closure phase the expansion and compression of the GG together with superior longitudinal and transverse muscles seem more involved (Figure 5). On the other hand the fronting and raising of the tongue seems to be controlled by the expansion and compression of the posterior fibres of the GG muscle (Figure 6). In the same way GH and MH muscle could be involved to the realization of the Advanced Tongue Root (ATR) feature in vowel /i/. Finally, EMA sensors, whose position on the images is only probabilistic estimated, emphasize the gestural goal reached by the synergistic coupling of the tongue muscles.

4. Conclusions

The most relevant instruments for articulatory studies are available at the CRIL research center, in Lecce, and have been synchronized within the CONTACT project. In the paper we report some notes on the information added by the simultaneous recording, and the main steps towards the complete fruition of the equipment for specific phonetic-phonology purposes. Some examples of the achieved articulatory-phonetic information are provided, and the (currently being developed) analysis environment used for addressing specific linguistic issues is described.

The synchronization of the articulograph, ultrasound and electroglottograph allows us to integrate information on the vocal fold vibration and both internal and external articulators. In particular, the synchronization of the instruments may offer richer information as it enables both to track different points and segments and to track them with different resolution qualities. Moreover, the simultaneous recording by means of the instruments offers different information on the same point/segment. This 'redundancy' may, in any case, be useful to have a clearer idea of articulatory gestures during speech, and to help scholars in the elaboration of more accurate theoretical assumptions about the nature of speech and language.

5. References

[1] Stone, M., (2005), "A Guide to Analyzing Tongue Motion from Ultrasound Images", *Clinical Linguistics and Phonetics*, 19, (6-7), 455-502

[2] Wrench, A.A. (2007), "Advances in EPG palate design", Advances in Speech-Language Pathology, 99, Issue 1 March 2007 , pages 3 – 12

[3] Zierdt, A., Hoole, P., Tillmann, H.G., (1999), "Development of a System for Three-Dimensional Fleshpoint Measurement of Speech Movements", Proc. ICPhS'99, San Francisco. August

[4] Grimaldi, M., Gili Fivela, B., Tavella, M., Sigona, F., Fitzpatrick, P., Metta, G., Craighero, L., Fadiga, L., Sandini, G., (2007). "Synchronized acquisition of Italian speech articulatory data using ultrasound, 3D-articulograph and laryngograph: First results." Presented at *Ultrafest VII,* New York, September 2007 (http://jerome.linguistics.fas.nyu.edu/ultrafest.html), and to be published in Proceedings of AISV "La fonetica sperimentale: metodo ed applicazioni", December 2007.

[5] http://eris.liralab.it/contact/

[6] Tavella, M., (2007), *Simultaneous recording of phono-articulatory parameters during speech production*, unpublished Master Thesis, University of Genova, LiraLab (Laboratory for Integrated Advanced Robotics), 2006-2007.

[7] http://www.mathworks.com

[8] http://www.articulograph.de

[9] http://www.medical.toshiba.com

[10] http://www.laryngograph.com

[11] Boersma and Weenink, University of Amsterdaam. http://www.fon.hum.uva.nl/praat

[12] Li, M., Kambhamettu, C., and Stone, M. (2005) Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 19 (6-7), 545-554.

[13] http://www.mplayerhq.hu

[14] Wilhelms-Tricarico, R. (1995), Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, 97, 3085–3098.

[15] Stone, M. & Lundberg A., (1996), Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America*, 99, 3728–3737.

[16] Stone, M. Epstein, M. A., Iskarous, K., (2004), Functional Segments in Tongue Movement, *Clinical Linguistics and Phonetics*, 18, 6-8, 507-521.

[17] Kent R. D., (1997), *The Speech Sciences*, Singular Publishing, Group, San Diego-London.

XGate and XRG: tools for visually editing, querying and benchmarking XML linguistic annotations.

Francesco Cutugno[1], *Leandro D'Anna* [2]

[1] Department of Physics - NLP Group, University 'Federico II' of Napoli, Italia
[2] Department of Linguistics and Literature, University of Salerno, Italia
cutugno@na.infn.it, ldanna@unisa.it

Abstract

Presently there is a great request in many fields of corpus linguistics of manually annotated texts and transcriptions. XML is rapidly become the principal instrument for linguistic markup even if in many occasions people operating in this field, mainly linguists, are not really experts in managing this technology. Although, at least in principle, many tools are available on the Internet, most of them are not easy-to-use and/or free of charge.
This work presents a set of software tools supporting the activity of producing XML data files with a special attention to linguistic annotation. Two products will be described in details: the first one, XGate, is a program which supports editing and querying of XML files and at the same time implements a semi-automatic method to synchronize such files to a modified DTD or Schema. XML file editing is performed visually, the document is showed in its tree-like form, tags and attributed are filled by the user as in a form. Querying is realized using a further visual interface that implements most of the XPath syntax graphically, furthermore query result can be again queried by means of a tool that automatically analyzes the structure of the former results.
The second program, XRG, is a tool for XML native database benchmark. XRG produces XML files of a given complexity in terms of node numbers, levels of direct and indirect recursion, horizontal width and vertical depth of the tree. Generated file can be queried with a built-in XQuery module and further statistical analyses are possible. If you have an XML DBMS and you want to verify performances, you can use XRG to generate a 'certified' dataset with which to evaluate your system.
Both tool are directed to non-experts, users are not asked to know XML and can visually manipulate their data in most of the software sections. Powerful and user-friendly interfaces have been developed for this aim. XGate and XRG are open software tools and it is possible to download executables (Windows + .NET framework platform only) and source codes for free from the portal of the Italian project 'Parlare Italiano' [1].

Index Terms: XML editing, XML querying, visual interfaces, benchmark.

1. Introduction

In many fields of computational linguistics there is a growing request of corpora including annotations, labeling and markup. At the same time XML is rapidly becoming a standard for the most used procedures for introducing markup to texts and transcriptions. Starting from the efforts made within the TEI framework [2] and following the developments produced in various international projects (as for example TIGER, NITE, AMI [3,4,5]) the necessity to manually add labels and annotations to corpora using XML is growing up even if it is not rare at all that people working in this field still misses specific skills for the use of this technology. There is a large request of tools for XML files managing having user-friendly interfaces; these tools are mostly used within the frame of public research, where economical resources are usually scarce and in which investment on software production cannot be a relevant portion of the overall budget. Another crucial question under discussion in the recent years is how to retrieve information directly from dataset formatted in XML as in many cases these kind of data cannot easily transferred into a relational database management system. Querying and searching into an XML file has been made possible, but not easy, by using specific query languages recently proposed by the W3C consortium. We present here two XML open source tools, XGate and XRG, conceived and projected having in mind some precise constraint: powerful user interfaces both for editing and querying, robustness against errors introduced by the user, support to guided recovery procedures in case of redefinition of the document structure, possibility of hard technical benchmarking both for data and for querying. In the spirit of open source software, these programs are freely downloadable. They run under Windows and require the Framework .NET 2.0.

2. XGate

XGate is a tool born to help linguists to manage with syntactic treebanks as produced during the process of texts manual annotation having as final deliverables XML files. It was conceived to solve all the problems occurring when managing with semistructured data as linguistic treebanks or, in general, with every data represented by a tree and saved in XML. Then XGate can be seen as a simple, efficient and intuitive tool to edit and query XML files. It is divided into two main modules plus some additional feature supporting and facilitating the annotation process. The two main sections are: Editor and Query Manager, while the additional features are a wizard for resyncronizing XML documents when DTD changes according to variations required by the user, an XML-Schema Manager, a tool for deriving a valid DTD from an XML document and a Notepad.

2.1. XGate: Editor

XML files are not easy to read for a non expert user. XGate proposed a visual interface into which the XML document is showed as a hierarchical tree. Our program analyses the DTD related to the document to be created and uses this knowledge to directly support the annotation process suggesting relations between tags and attribute values when these are to be choosen in a closed set.

The interface presents many frames in which all knowledge required for texts annotation is included, it is projected in such a way that the user is accmpanied step-by-step during the annotation process, XML syntax is verified on the fly and validation process takes place every time the file is saved.
In fig. 1 the Editor interface is showed.

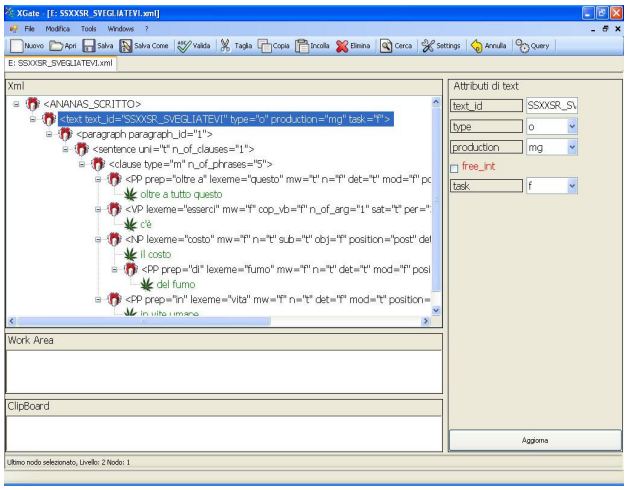


Figure 1*: XGate: Editor Interface.*

2.2. Query manager

Querying Treebanks in particular and XML native database in general, is a complex task. In many case query results are conditioned by the level of articulation that the hierarchy of tags in the document imposes, while at the same time, closed loops and recursion - typical in an XML document related to liguistic annotations - make the information retrieval process ambiguous and depending on the path choosen to visit the database.
To partly recover this problem W3C proposes two standards for querying XML documents: XPath and Xquery [6,7]. The first defines a query language based on paths on the tree structure. User defines a template for a possible sequence path in the documents with well defined constraints on the tags and the attributes, specifies the class of possible entry points in the document from where the search process can start and the query returns all the nodes or the subtrees satisfying the request.
XQuery adds to XPath the usual programming statements as if-then-else, loops, variable definition and assignment in order to improve possibilities offered by the language for querying. XGate implements an user friendly interface to produce single XPath queries. According to the analysis of the user requirements it has been decided not to implement XQuery in this tool as it would have required the user possess at least basic programming skills. Luckyly, most of the user requests can in any case reconducted to a sequence of XPath expressions. XGate implements a visually guided query construction, user does not need to know XPath syntax and is guided through the process of retriving information on the base of a preanalysis of the document under examination of the related DTD.
Fig. 2 shows an example of the Query interface.

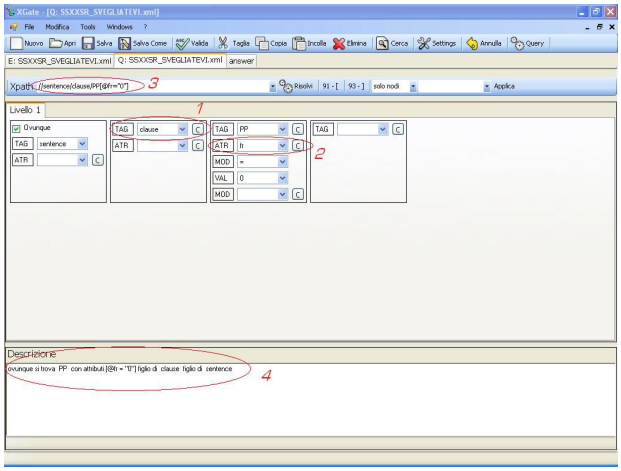


Figure 2: *XGate: Query interface.*

The output from the query is automatically formatted in a new XML document. The user can, at this point, formulate a new query on this output making once again use of the interface, XGate automatically determines a possible DTD for this new document (this process is obviously mandatory if you want to reuse the interface), and allows to restart the querying process. In this way it is easy to increase the level of data selection capacity even not using XQuery that would have been the preferred choice in case of an user with programming skills.

2.3. XGate - Additional features

2.3.1. XML to DTD

As formerly described, within XGate it is often necessary to indirectly obtain the structure of the XML document under examination. This can be made making use of a tool that finds a possible DTD for a given XML document. As it is well known, this process is not unique as a single XML document can be related to an infinite class of DTDs depending on some factors like undeclared tag dependencies, or implied attributes role, closed vs open set values for attributes and so on. Our tool produces instances of possible DTDs according to constraints defined by the user and mainly balanced to permit most of the automatic analyses performed by the system on the XML documents.

2.3.2. Notepad

It is in any moment possible to visualize the textual aspect of the XML document using an internal notepad. This tool is also useful to pre-load a text or a transcription to be annotated.

2.3.3. Statistics on docs

A very useful tool for linguists permits listing and counting of all tags, attribute and values encountered within the XML document loaded into the interface.

2.3.4. XSL editor

We included a third party open source product into XGate, namely XSDEditor [8]. This tools permits the visual organization of the tagset and of attributes and is really useful in the phase of database project.

2.3.5. XML - DTD re-synchronizing Wizard

Linguists often change the metadata structure during the course of annotation if they encounter phenomena that were not foreseen. In this case the already made work is to be re-controlled and eventually corrected. At the same time XGate automatically verifies files wellformedness and checks validity against the assigned DTD. When a file is not valid both because of an error in the annotation or because of a structural (and pre-ordered) change in the DTD, it is possible to modify it in order to be a valid file. When possible the modifications are realized automatically without requiring a user feedback. If this process finishes and the document is not valid, the module begins an assisted correction in order to satisfy the DTD file asking the user to choose changes to be made in the file in every miscorresponding tree node.

3. XRG

XML native databases are characterized by the presence of closed loops and recursion. Under these conditions a given query on the data, differently from what happens with relational database, can produce different outputs depending on the entry point in the document and on the level of recursion that is present in the semistructured dataset.
In this view it is really important, in order to evaluate the real power of expressivity of your queries to test them on a certified dataset before testing them on real data.
Xml recursive generator (XRG) is an automatic tool for creating Xml files using a pseudo random generation. The Xml files is generated by XRG in a controlled mode starting from a given DTD file and setting the level of document 'complexity' assigning values to a set of parameters by means of a very easy visual interface. This process can be divided in two stage: a first stage in which a deep analysis of DTD is performed and a second stage where the generation of XML files takes place together with the possibility to evaluate and compare difference among different runs on the same DTD and with the same number of tag. Finally it is possible to query the generated XML file using the standard W3C query language Xquery.

3.1. Pre-analysis stage

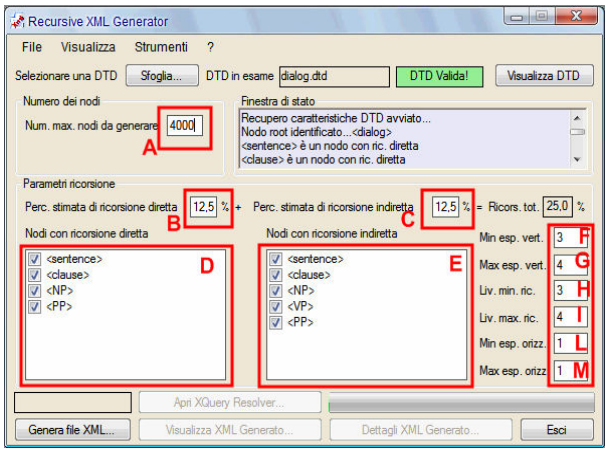


Figure 3: XRG *GUI. In rectangles forms to set parameters.*

First of all the DTD file was checked to verify its validity. Then all possible tag elements in DTD are analyzed to determine if they have a simple or recursive structure.

For those that have a recursive structure, it is determined if they have a direct recursion (eg. TAG1 can include TAG1 between possible sons) or indirect recursion (eg. TAG1 can have a TAG1 as possible nephew or successive descendants). It is important to notice that a single tag can have both direct and indirect recursion.
By mean of a user friendly GUI (see rectangles in fig. 3) it is possible to choose these parameters:

A. The overall number of elements that would be insert in XML file;
B. Maximum percentage of tags with direct recursion;
C. Maximum percentage of tags with indirect recursion;
D. Which tags from the list of element with direct recursion to use in generation;
E. Which tags from the list of element with indirect recursion to use in generation;
F. The minimum height (vertically) of XML file seen as a tree;
G. The maximum height (vertically) of tree;
H. The minimum depth where it is possible insert recursive elements;
I. The maximum depth where it is possible insert recursive elements;
J. The minimum width (horizontally) of tree;
K. The maximum width (horizontally) of tree;

During the parameters setting operation, an automatic check discards value erroneously inserted.

3.2. Generation stage

The generation engine is based on a sequential algorithm that controls step by step that the partially generated tree satisfies all structural constraints imposed by DTD.
This approach permits to obtain good performance both in time and in number of elements even on computers with poor performance. Obviously the generation time increase whit the number of elements in the tree with an experimental exponential trend for our module. It is interesting to note that noticeable changes in the structure of tree without modify the overall number of elements, causes little change in the generation time.
XRG has been investigated even in terms of temporal complexity in relation to the number of generated XML nodes. In fig.5 such performances are showed.

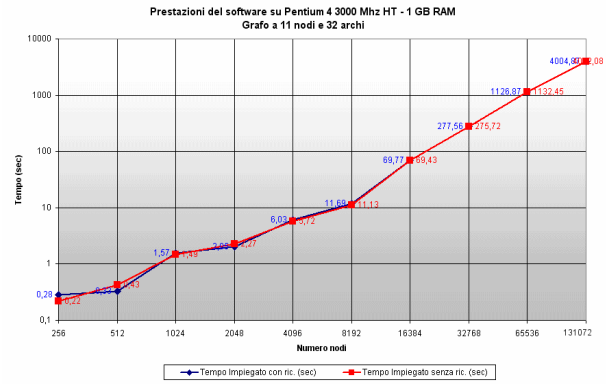


Fig.5: *temporal complexity for XRG (double log scale, xaxis: number of generated node, y-axis: sec)*

After a quick generation (e.g. 16384 elements in about 70s with a CPU PENTIUM4 3000Mhz and 1 Gbyte), it is possible save the XML file generated and make accurate evaluation of the output.

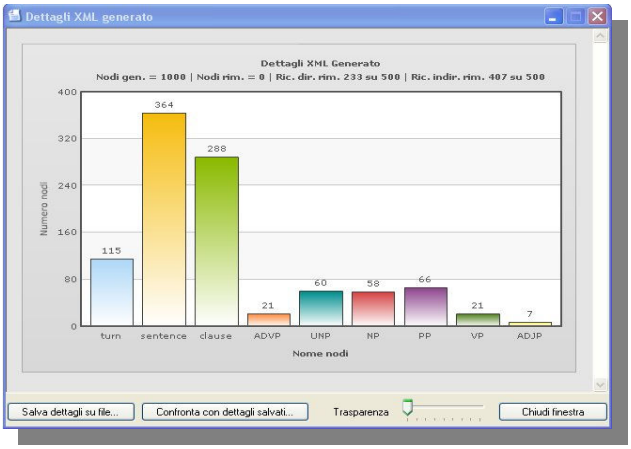


Figure 4 XRG*: Graphical evaluation of output.*

To do this a graphical tool for statistics (see fig. 4) permits to analyze the distribution of the tags in the XML file.
All the statistics can be saved both as text and as image. It can be useful to make further comparisons between two file generated with the same parameter set.
Finally, in order to realize more complex analysis on the XML file, it is possible to use a query module that permits to realize XPath 2.0 and XQuery 1.0 queries and it is based on the Saxon processor [9].

4. Cross controlling

To evaluate the correctness of the generated XML file and the related statistics computed by XRG, we have used the query module of XGate.
These cross controls can be conceived in two ways:

1) It is possible to check if percentages of direct and indirect recursion were satisfied into the generated XML document.
In order to do this we have imposed for parameters B) and C) (see figure 3) the same value chosen as higher as possible in relation to the constraints derived by the assigned DTD. Furthermore we have assigned all the recursion to only one tag for every type of recursion.
Then using a simple XPath instruction like: **//TAG1/TAG1** we have verified the overall number of direct recursions.
In order to verify the number of indirect recursions, we used the XPath instruction: **//TAG2/*/TAG2**

2) with the second test we checked if the XML files generated with XRG were correctly created according to the tree-like structure depicted in the document design made by means of the DTD. For this aim we assigned the same percentage for direct and indirect recursion using only one tag for every type of recursion. Then we have imposed that the positive value of parameters H) is 'value1'.
To check that Xml files respect this ties we use these XPath instructions:

//TAG1[@deep<value1] for direct recursion
and //TAG2[@deep<value1] for indirect recursion.

Many other check controls can be made changing opportunely parameters in XRG and using the power of expressivity of the XPath commands.
In both cases requirements expressed in XRG perfectly match with the verification made by XGate.

5. Conclusions

The tools have been produced within a national project (see next section) within which an observatory on researches regarding spoken Italian has been realized. A web portal has been designed and data, tools, corpora and other resources regarding Italian language were collected. Within this project, the development of XGate and XRG has been realized in conjunction with a group of users ranging from more to less XML experts, their feedback has been constantly taken into consideration in all phases of the project. XGate has been tested on two different corpora: a syntactic treebank, ANANAS and a corpus describing dialogue pragmatics, PraTID [both available in 1]. The first corpus presented an high level of complexity due to the high number of attributes for each tag and to the presence of direct and indirect recursion in many levels of the document structure; furthermore , during the construction of the corpus, authors decided in various occasions to modify the DTD to introduce labels and to describe phenomena not initially foreseen. PRaTID was made with a simpler XML structure, for this corpus times for labelling were evaluated. Results reports an high grade of satisfaction by all the users. As already seen in the previous paragraph, XRG has been widely tested for benchmark in cross validation with the XGate query generator section.

6. Acknowledgements

This work has been funded by the Project 'Parlare Italiano' (PRIN 04 and PRIN 06 -MIUR Italy, National coordinator Miriam Voghera). LDA and FC have projected the tools and have defined the overall software architecture and fixed requirements and constraints. XGate has been written by Gennaro Cavezza, Emilio Diana and Antonio Vuolo. XRG has been written by Raffaele Liguoro. Miriam Voghera, Renata Savy, Giusy Turco, Simona De Leo hardly tested the programs. Annamaria Landolfi and Carmela Sammarco wrote manuals and docs.

7. References

[1] Parlare Italiano: http://www.parlaritaliano.it
[2] TEI: http://www.tei-c.org.
[3] The TIGER corpus: http://www.ims.uni-stuttgart.de/ /projekte/TIGER/TIGERCorpus
[4] J. Carletta, S. Evert, U. Heid and J. Kilgour, "The NITE XML Toolkit: Data Model and Query Language", Language Resources and Evaluation, 39(4):313-334, 2005. http://nite.nis.sdu.dk/
[5] AMI: http://www.amiproject.org/
[6] Xml path language (Xpath) 2.0 W3C recommendation: http://www.w3.org/TR/xpath20/
[7] Xquery 1.0 XML query language W3C recommendation: http://www.w3.org/TR/2007/REC-xquery-20070123
[8] XDSEditor: http://www.codeproject.com/dotnet/xsdeditor.asp
[9] Saxon XQuery processor: http://saxon.sourceforge.net

A Calendar Interface in French: XIPAgenda

Claude Roux
Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan, France
Claude.roux@xrce.xerox.com
+33 4 76 61 51 38

ABSTRACT
In this paper we describe a French language interface to a calendar system. The system has been successfully implemented using state-of-the art technologies in parsing. We show how temporal expressions are analyzed and how this interface simplifies the task of setting your agenda.

Author Keywords
NLP, French, calendar, parser, XIP, temporal expression.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Most companies rely today on calendar software to schedule meetings and events. The use of these agendas has become so ubiquitous that most people take them for granted. However, setting a new event, even though most GUI are pretty simple to use, has become a bit of a tedious task as the number of options has steadily but incredibly increased over the last years. Of course, adding a new meeting to your agenda does not require a full week training. It takes less than a minute for most users to fill in the necessary template to add a meeting or a rendezvous. However, a current calendar system involves some clicking and some typing that may prove cumbersome. In a typical software from a well-known company from Seattle, between seven to ten different manipulation are needed in order to correctly insert one single appointment. The user must first select the day, then the hour, then split his/her message into a title, a location, a list of attendees, set the correct duration. Again, this list of tasks is neither complex nor difficult, but since the calendar template over the years has gained in options what it has lost in clarity, it definitely requires some careful writing. It is a well known problem that mouse manipulations are usually slower than keyboard strokes, as using a mouse requires moving a hand off the keyboard, which stops the writing flow. The most efficient way would be to speak to your computer, using a voice recognition system, to automatically update your agenda. Unfortunately, the state of voice recognition technology is far from offering such a feat. Furthermore, the use of voice recognition software in an open space does pose some etiquette problems, which have not been solved up to now. If we cannot talk to our computer, we can still write to it.

The goal of this article is to describe how a natural language interface with a broad coverage French grammar can improve the communication of a human being with his/her machine. The principal difficulty in this task is the parsing and the understanding of the temporal expression which has been typed in by the user. As most experiences in the domain of NLP interfaces have shown in the past, the patience of users with a system that would not analyze most of her/his propositions usually thins very quickly. We will describe our system: *XIPAgenda*, which implements a natural language interface to a calendar. *XIPAgenda* is used both for instantiating and querying the agenda.

AN AGENDA ITEM?

XIPAgenda purpose is to allow a user, with no particular training, to be able to add a meeting in a calendar, using only her/his knowledge of her/his own language. However, before describing in more details our system, it would be interesting to study what a calendar item looks like.

An agenda item can be modeled as a function: $\Psi(\theta_0,\theta_1,\tau_0,\tau_1,\delta_0,\delta_1,\lambda,\rho,\mu)$, where:

1. θ_0,θ_1 are initial and final dates.
2. τ_0,τ_1 are initial and final times
3. δ_0, δ_1 is first the time lag from a referent time and the duration of the event.
4. λ is a location
5. ρ is a list of persons (*attendees*)
6. μ is the topic of the event

The goal of our system is to map a sentence over this representation.

Unfortunately, while you may find in a template a specific slot for each of these data, a sentence rarely split into any of these categories easily. Most of these data might be either absent or worse implicit.

Example
The meeting on technology transfer will take place in 20mn, starting at 10:00AM and will last half an hour, in the Everest room.

If we analyze this sentence, we can compute the following dates and times. First, the referent date will be today, as

none is provided in the sentence. We will suppose that θ_0,θ_1 are then seeded with today's date. Second, since, no exact time has been provided τ_0,τ_1 will be initialized with the time the message was written. Third, we have two durations, which have been provided. The first one *20mn* maps over δ_0 and will be used to compute the exact starting date of the meeting. The second duration *"half an hour"* is mapped over δ_1 and will be used to compute the exact date of the end of the meeting. The location λ is "*Everest room*". The list of persons ρ is empty and the topic μ will be restricted to "*meeting on technology transfer"*, since no other information is available. As we can see on this simple example, the definition of a simple meeting can prove quite complex to process.

Temporal Expression

Frank Schilder [1] gives a list of different sorts of temporal expressions that one may find in a text:

- Explicit, *the exact date is provided*
- Indexical: *tomorrow, yesterday*
- Duration: *for two hours*
- Vague: *in a few days*

An appointment statement might fall in each of these categories. However, we will discard the *vague* one as we hope users to set precise appointments.

At first glance, the limitation of our system to only simple agenda statements, one sentence at time, would sound as too simple a task. It would be reasonable to think that extracting temporal expressions from texts would not require very complex grammars. Yet, in evaluation campaigns such as TREC or MUC, the existence of *when* questions such as:

- *When did Napoleon die?*
- *When was President Reagan first elected?*

have shown that this job was far from being trivial. It has triggered a renewed interest in this task and a variety of new systems to detect and solve these expressions has emerged. Different approaches exist. For instance, since machine learning techniques have become prevalent in the last years in the NLP domains, some serious attempts have been made to build temporal corpora such as TIMEBANK (see [2]). This corpus has been annotated with TimeML (see [4]), which is an XML annotation schema, based on TIMEX2. TIMEX2 (see [5]) provides a descriptive language to make some specific temporal data more explicit. In other words, the goal is to transform a natural language time description into a tractable item. The goal is to translate the time descriptions into actual dates.

Example

The following sentence has been flagged according to the TIMEX2 annotation schema:

In January, the weather was very mild.

In <**TIMEX VAL= "2007-01-XX-X"**>*January*</TIMEX>, *the weather was very mild.*

The TimeML annotation schema is much more ambitious than TIMEX2 as it also provides a structure in which events are annotated and linked together through temporal expressions. The TIMEBANK has been used to train machine learning algorithms as in [7] approach, in which their system is trained to *temporally order* events in documents. However, the use of machine learning techniques to build temporal expression extractors would be rather difficult to use in an agenda application. First, the training requires hundred of sentences, already annotated, second the correction of wrong analyses usually imposes the annotation of more sentences with a re-training of corpus, with limited confidence that the new model is effectively corrected. Unfortunately, there is no available corpus in French to train a system on temporal expression, which seriously hampers the use of these methods in our case. Another interesting system, which has been developed for the German language, is COSMA (see [8]). COSMA is a language server, which analyzes the content of e-mails and detects any agenda items. It then tries to set an appointment that would be suitable to all participants. The system is based on a full-fledged grammar, which is composed of rules. The use of a symbolic grammar is both lighter in terms of computing power (statistical models tend to be quite greedy when running) and simpler to correct. This assumption is especially true when the goal is to detect a subset of natural language expressions, which can be described *in extenso*. Yet, we should keep in mind that a limited set of potential expressions does not translate into a simple list of regular expressions. For instance, Google Calendar provides a tool which accepts agenda description in plain English. The analysis of the sentences is made with a set of regular expressions that are extremely sensitive to variations, and often fail to provide the correct output. We have used the Xerox Incremental Parser (e.g. XIP) with an updated grammar for the French language, in order to meet both our requirements in terms of coverage and efficiency.

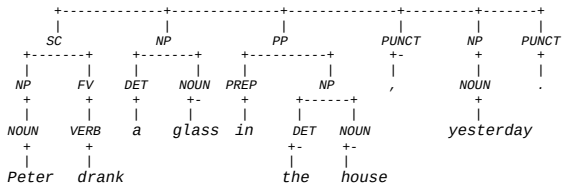
THE PARSING ENGINE: XIP

The Xerox Incremental Parser (e.g. XIP), (see [9,10]) is a symbolic parser which has been developed at the Xerox Research Centre Europe (e.g. XRCE) in the last years. This parser provides a broad coverage grammar for at least seven languages. The output of XIP is composed of a syntactic node tree (also called chunk tree) and syntactic functions (or dependencies). A syntactic function may link together two or more chunk nodes, or simply defines the nature of one single node.

Example

Peter drank a glass in the house, yesterday.

The output of the English grammar is the following chunk tree:



And a list of syntactic functions:

- SUBJECT(*drink,Peter*)
- OBJECT(*drink,glass*)
- PERSON(*Peter*)
- LOCATION(*house*)
- DATE(*yesterday*)

XIP extracts the different dates, locations and person names which are present in a text. It also extracts some important syntactic relations such a SUBJECT and OBJECT. The example here is in English, but a sentence in French would yield similar syntactic functions. The XIP engine is not new to temporal extraction or question/answering techniques and it has been successfully used in different campaigns such as EQUER (see [11]) or TempEval (see [12]). Thanks to these campaigns, we have at our disposal a large grammar for French, with some specific set of rules dedicated to temporal extractions. Furthermore, a large amount of work has been dedicated to the definition of specific sub-grammars embedded within regular grammars to add name entity detection for languages such English and French (see [14]). Thus, the analysis of a sentence as below:

- *Pierre Dupont s'est rendu à Paris en 2001 (Pierre Dupont traveled to Paris in 2001)*

already yields the following list of dependencies:

- PERSON(*Pierre Dupont*)
- LOCATION(*Paris*)
- DATE(*2001*)

TEMPORAL GRAMMAR

Yet, this list of functions is insufficient to map any sentences to an agenda. The DATE function lacks many features which would help our system to translate them into exact dates. Furthermore, we also need to detect times and durations, which would be important to any appointment definitions. Let's examine the following sentences:

- Meeting in two hours *time lag*
- Meeting at 2:00PM *exact time*
- 30mn meeting in one hour *duration and time lag*

A study of these sentences shows how versatile the description of meetings can be.

The French grammar has been modified in order to enrich it with new functions that are mapped over our agenda item function: $\Psi(\theta_0,\theta_1,\tau_0,\tau_1,\delta_0,\delta_1,\lambda,\rho,\mu)$.

The French grammar, which has been re-designed for our task, yields now the following dependencies: *DATE, TIME, DURATION, LOCATION, PERSON, TOPIC.* Furthermore, each of these functions can be enriched with a list of specific features that gives some more details.

Features

A date, for instance might have many different formats:

- *the 15th*
- *April, the 3rd*
- *2007-09-05*
- *On Monday*
- *Tomorrow*

Each of these dates requires a different set of features in order to refine their description. These features are the following: *YEAR, MONTH, MDIGIT, DDIGIT, DAY, SHIFT.* The difference between *MONTH* and *MDIGIT* is simply that in one case, the month is a word, while in the second case it is a number. The analysis of the above strings gives the following results:

- DATE_DDIGIT(*15th*)
- DATE_MONTH_DDIGIT(*April, the 3rd*)
- DATE_YEAR_MDIGIT_DDIGIT(*2007-09-05*)
- DATE_DAY(*Monday*)
- DATE_SHIFT(*tomorrow*)

The grammar can also set two more features: *START* and *END* which corresponds to the initial and the final dates in the sentence. These two features are also associated to the *TIME* function in the same fashion.

Example

- *Meeting tomorrow from 10:00AM to 11:00AM, in room 20.*

The analysis of the above sentence will give the following functions:

- DATE_SHIFT(*tomorrow*)
- TIME_START(*10:00AM*)
- TIME_END(*11:00AM*)
- LOCATION(*room 20*)

A similar French agenda description would give the following analysis:

- *Réunion de deux heures, lundi à 13h00, en salle 12. (Two hours meeting, Monday at 1 :00PM, in room 12)*

The resulting functions are:

- DATE_START_DAY(*lundi*)
- DURATION (*two hours*)
- TIME_START(*13 h00*)
- LOCATION(*salle 12*)

As we can see on this example, the grammar has taken some decisions about the *START* feature added to the functions: *DATE* and *TIME*. The strategy goes as follow: *when only one date or one time is found in a sentence, then we add the feature* START *to their corresponding functions.*

Time Calculus

The grammar does not compute any actual dates; this job is passed to a specific piece of program, which in our architecture has been written in Python. The role of the grammar is to detect linguistic occurrences of date and time in a sentence, not to compute any actual values. However, thanks to the different features that the grammar adds to the different functions, the definition of a date is quite simple. Each feature leads to a specific sub-case of date computing, which makes the process quite straightforward. For instance, on a *DATE_DAY(Monday)* instance, the program will jump to the *day* resolution part of our program to determine the exact date of that *Monday*. Missing elements from our agenda function Ψ, will be replaced with today's values, or will be computed on the fly. The different sorts of duration, time start or time end will then be mixed together to compute the correct values. The purpose of this piece of program is to calculate two date representations, whose format will be compliant with our agenda software. Usually, a complete date format on a computer will involve year, month, day, hour, minute and second values, which will be merged into one single string. The program is left with the task of filling in the blank: the main implicit datum is that the referent date is today.

Location and Persons

The location extraction is part of the generic French grammar. It is automatically detected in a statement and a LOCATION function is yielded, which the calling program will use to set the location field in the agenda template. We have a similar behavior for the PERSON function, which is also part of the regular French grammar. The values of these functions will be passed to our program, which will set the corresponding fields in our agenda application.

TOPIC

The goal of an agenda is to inform a user of an imminent appointment on time. Usually, a small window pops up, which displays a few lines to describe the next appointments. However, the only data available to define μ in our Ψ function is the initial statement. In a preliminary prototype, the subject line was the statement itself, a solution that was far from being satisfactory. The title was both too long and too redundant with the agenda itself. The other solution is to remove all redundant information that is already known to the system, such as dates, times and the locations.

Example

- Réunion, le 15 novembre à 18h00, en salle Mont-Blanc, pour proposer une offre à Metal Corp. *(Meeting on November the 15th at 6 :00PM, in Mont-blanc room to propose an offer to Metal Corp.)*

The objective is to remove from the above sentence all the words that would be related to locations or temporal expressions, in order to store the following line:

- Réunion pour proposer une offre à Metal Corp (*Meeting to propose an offer to Metal Corp)*

This operation is very similar to summarization techniques such as the sentence compression method used in [13]. The sentence compression consists in removing sub-clauses and other less essential information from a sentence in order to produce a shorter sentence, which still retains most of its original meaning. The operation consists in deleting specific phrases (or chunks) instead of removing independent words. A chunk such as *à 18h00* has a certain autonomy within a sentence, which a word such as *18h00* does not have. This autonomy means that removing one sole word will unbalance the sentence, while removing the chunk will have a certain impact on the sentence meaning without destroying its readability.

Example

If we compare the two sentences below:

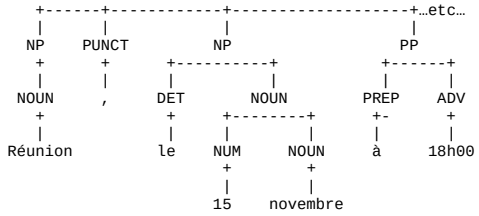
- **Réunion, le 15 novembre* ***à****, en salle Mont-Blanc, pour proposer une offre à Metal Corp.*
- *Réunion, le 15 novembre, en salle Mont-Blanc, pour proposer une offre à Metal Corp.*

In the first sentence, the time *18h00* was removed, while leaving the preposition *à*. This sentence is ungrammatical. The second sentence, where the full chunk was removed is still a valid French sentence.

Chunk Tree

The use of a parser is absolutely central to this task. In XIP, since the building of the chunk tree is preliminary to all others operation, we have enough material to proceed to the compression. Once, the syntactic functions have been computed, the nodes involved in temporal expressions are marked with a specific feature.

Example
The chunk tree of our previous sentence is the following:



We have only represented on this tree the temporal nodes that we want to remove.

The grammar has also yielded the following functions:

- DATE_MONTH_DDIGIT(*15 novembre*)
- TIME(*18h00)*

The compression in this case consists in marking the nodes that are involved in these temporal expressions, together with their top nodes: *NP* and *PP*. Once all these nodes have been discovered and isolated, they can be safely removed from the sentence. This method allows us to generate a short but informative subject line.

QUERIES

Most calendars provide a search engine, which is usually limited to only keyword search. This search possibility is sometimes enhanced with some categories, which can be used to sort out appointments according to their content, such as *business, birthday, meeting etc.* This list is usually restricted to a few domains and cannot be easily expanded. Furthermore, a message can only be categorized with one category at a time.

Ontologies

The only way to give a message some background is to enrich it with more information so that the search engine will be able to find similar items. This enrichment is done through an ontological description of what a calendar action is. For instance, a *meeting* in French can be described in many different ways:

- It could be another noun such as: *réunion* or *rendez-vous.*
- It could be a verb such as *voir (see), discuter (discuss), parler (speak).*
- It could also be an expression such as: *Donner une presentation (give a presentation)* or *Organiser un événement (to organize an event).*

XIPAGENDA has been enriched with an ontological description of most of these words. Whenever, an appointment statement is processed, the system automatically analyzed each word and returns a list of related words. The calling program uses this list to build a *body* in which hyperonyms and hyponyms of these words are automatically appended to the initial statement. This enlarged *body* is then stored together with the other information into the calendar. Thus a simple sentence such as: *Meeting with Peter tomorrow,* is automatically enriched with the following elements: *reunion, appointment.* When a user looks for a given appointment, it can search for *meeting* or *appointment* even though these words were not present in the message in the first place. This enrichment allows a limited calendar application to become *semantically aware*, despite the fact that it only provides a simple search interface.

Querying

Querying a calendar is very similar to defining a statement, to one exception: The dates in this case could be in the past. Furthermore, while the duration in most agenda statements is within a day, a query might cover a full month or even a year.

Example
Below is a list of questions about past statements:

- Réunion en janvier? *(meeting in January)*
- Présentation en 2007? *(meeting in June)*
- Conférence la semaine dernière ? *(conference last week)*

In a first version of our prototype, once the initial and final dates had been computed, we would build a query in which the keywords and these dates were grouped into one single instruction that was sent to the calendar application. However, this method proved very slow and was replaced by a complete download of all agenda items for a given period of time into a python dictionary, which was then indexed according to different strings computed on:

1. Year
2. Year/Month
3. Year/Month/Day
4. Year/Month/Day/Hour
5. Week number
6. Words from the statements
7. Statement strings

The reason behind this architecture is that querying for a period of time or for an exact date corresponds to one single access to that list. For instance, a query on *last week*, would automatically return the list of appointments recorded for "*current week -1*". The list was quite tractable as the number of agenda items for one single person rarely exceeds a few hundred lines a year.

Querying this structure with a sentence is done in a four steps:

1. The sentence is parsed
2. The temporal information is computed
3. The system searches our python dictionary for each word from the statement, for each location and for each temporal data. Each element in this dictionary is a list of appointments.
4. Only elements common to all lists are returned as a result.

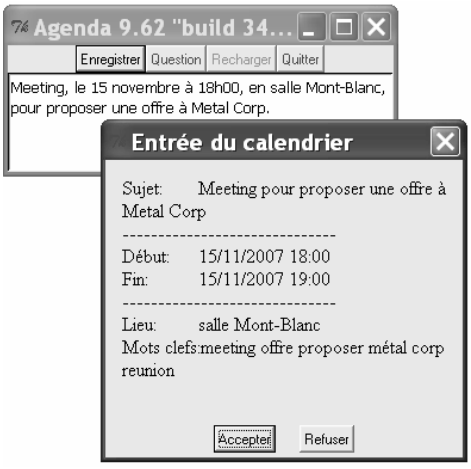
Example
A query such as:

- Conférence la semaine dernière ? *(conference last week)*

Corresponds to a search on : *"week number"* and *conference.*

ARCHITECTURE

The *XIPAGENDA* is written in python, using the basic *Tkinter* graphical library to handle windows. XIP provides both an internal and external python API. The internal API allows a rule, anywhere in the grammar, to call a python procedure, while the external API transforms XIP into a python library that can be freely imported into a python program. Below is an example of the window which is used to enter an appointment. The user might press *enregistrer (record)* to add a new item in his agenda or he might press *Question* to query the calendar application. If the statement ends with a question mark: "?", the system will automatically consider the statement as a query, even though *enregistrer* is clicked. A small window pops up with a summary of the information that was extracted from the statement by the grammar. The user might then decide to store this item to his/her calendar application.



The summary proposes a *subject* line, which is a compression of our initial statement. The two other lines are the beginning and the end of the appointment. The *lieu* is the location of the meeting. The *Mots Clefs* (keywords) section is a list of all potential important words from the statement together with some synonyms.

We have linked our application to *Outlook*© through a COM interface, but our solution could easily be adapted to other applications as the python program first builds a specific calendar object that is then mapped over the COM *Outlook*© interface.

EVALUATION

The evaluation was made on 104 different sentences which were mainly extracted from e-mails. Below is a sample of some of these sentences:

- Voyage du 15 au 18 mars (*Travel from the 15th to the 18th)*
- Dîner, le 2 à 19h00, dans l'auberge Napoléon, 7 rue Montorge (*Dinner, the 2nd at 7 :00PM, in the Auberge Napoléon, 7 rue Montorge)*
- Retour à Grenoble, le 23/02/2007 à 19h38 (*return to Grenoble, the 23rd at 19 :38)*
- Conférence téléphonique, demain à 9h00, dans le bureau de Laurent, pour le projet ALPHA (*Phone conference, tomorrow at 9 :00AM, in Laurent office, for the ALPHA project*)
- Réunion en Mont Blanc dans deux heures (*Meeting in Mont-Blanc in two hours)*
- Départ de Bruxelles le 15 février 2007 à 18h00, arrivée à Grenoble le 16 à 20h00. (*Departure from Bruxelles, February the 15th 2007 at 6 :00PM, arrival at Grenoble the 16th at 8 :00PM)*
- Conférence dimanche en huit en salle 12 *(Conference, Sunday after next Sunday, in room 12)*

The grammar had been designed over a different set of sentences. The issue in testing the system was that errors could stem from different parts of the system. The grammar could miss some elements from the sentence, or the python calling program could wrongly compute some specific dates. The definition of recall and precision was a real issue. Our preliminary idea was to compute precision and recall, sentence by sentence. If precision could be defined as the number of sentences that had a correct summary, we had more trouble to define a recall, as all sentences did receive a parse, even though it proved wrong in some cases.

We decided to compute our precision and recall using a different approach. We manually counted the number of potential dates, durations and locations that the whole corpus contained and we checked how many dates, times, durations and locations were actually extracted from our corpus by our system.

- The recall was defined as the number of dates, times, durations and locations extracted by

XIPAGENDA over the corpus versus the number that we had manually extracted.

- The precision was the number of correct dates, times, locations and durations that was computed by *XIPAGENDA*.

Our corpus contained 104 sentences with 259 actual definitions:

- 98 dates
- 81 times
- 25 durations
- 55 locations

The system then computed the exact initial and final dates for each of these definitions using both dates, times and durations. The recall and precision was computed on these pairs, together with the location data.

The recall was very high, which was a proof of the broad coverage of the grammar: *95%*

The precision was also very high, as the system did not try to cope with anything but specific temporal and location information in the different sentences: *94%*.

However, the number of sentences that were correct was only 85%. We rejected any sentences that would contain one or more errors.

We did not try to extract proper names in our experimentation, as our calendar system was not designed to use them. We decided to simply discard them for latter trials.

The system failed on certain date format. In French, the tradition is to use a pattern such as: DD/MM/YY, however, certain people use other formats such as: YYYY/MM/DD, which XIPAGENDA could not properly analyze.

The system also failed on sentences such as:

- En déplacement, les 14, 15, 16 et 17 septembre (*Away on September the 14th,15th, 16th and 17th)*

In the above sentence, the grammar could not cope with a list of days.

Finally, we also rejected one sentence whose *subject* line had been wrongly computed, due to an unexpected lexicon entry. *Pierre*, which is a very common first name in French, had been recorded in the lexicon with an unanticipated temporal feature, which eventually modified the full analysis of the statement: the word *Pierre* is used in the French expression: *l'Age de Pierre*, which means *Stone Age*.

- Voir Pierre, demain à 15h00 (*see Pierre, tomorrow at 3 :00PM)*

The final stage of the analysis failed to compute the exact date, which *Pierre* was supposed to be, while correctly computing the initial and final dates.

It should be noted that the grammar has been amended since and covers now almost 98% of all these sentences.

CONCLUSION

We have tried with *XIPAGENDA* to design a system which would both robust and utilizable. The robustness is a real challenge as each new user has her/his own way to describe a calendar event. However, our grammar is now quite reliable and thanks to the relative closure of this domain, covers most of the temporal expressions currently used in French. Our system proposes some specific features such as the possibility of searching the calendar with natural language queries, the compression of the sentence to build a comprehensible subject line and finally the integration of ontologies to enlarge the query coverage. Even though *XIPAGENDA* still presents some glitches, it has proven that the use natural language in such an application was feasible. The possibility to tell your computer to do things is an old dream, which when it works, even on such a small subset, is still enthralling. Most products today consist of an all-in-one application where e-mails, calendar, to-do list, and contacts are merged into one single program. However, even though this integration is usually well made, the use of these tools requires jumping from one window to another tab. Adding a contact or sending an e-mail involve filling in templates that have become more and more complex. Again, these applications are far from being excruciatingly complex and most people use them seamlessly. *XIPAGENDA* in our vision is only one step further to the complete control of these tools in natural language. We think that a NLP central could be designed which would allow people to add new contacts with a simple sentence, leaving the task of finding who is who and where (s)he lives to the computer. E-mails and to-do lists could also be managed with simple statements that would be recognized by the machine. Furthermore, NLP can also apply to the content of messages or e-mails. With such a NLP central, it becomes possible to spot dates and addresses in documents, and automatically proposes meeting or new contacts to a user. When such an application is put together, all textual information originating either from the user's commands or from external mails is linked into one single net of information, whose querying might provide in-sight on one's work which would be difficult otherwise.

REFERENCE

[1] Frank Schilder, Extracting meaning from temporal nouns and temporal prepositions, *ACM Transactions on Asian Language Information Processing (TALIP)*, (2004), 33-50.

[2] Pustejovsky, J., P. Hanks, R. Saurı, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D.Day, L. Ferro and M. Lazo. 2003b. The TIMEBANK Corpus. *Proceedings of Corpus Linguistics (2003)*: 647-656.

[3] Pustejovsky, J., R. Saurı, J. Castano, D. R. Radev, R. Gaizauskas, A. Setzer, B. Sundheim and G. Katz., Representing Temporal and Event Knowledge for QA Systems. Mark T. Maybury (ed.) *New Directions in Question Answering*. MIT Press, Cambridge, (2004).

[4] Pustejovsky, J., Saurı, R., Setzer, A., Gaizauskas, R., and Ingria, B., TimeML Annotation Guidelines. http://www.cs.brandeis.edu/~jamesp/arda/time/documentation/AnnotationGuideline-v0.4.0.pdf, (2002).

[5] Mani, I., Ferro, L., Sundheim, B., and Wilson, G., Guidelines for Annotating Temporal Information. *In Proceedings of the Human Language Technology Conference*, 2001.

[6] Dawson, F. Stenerson, D.. Internet Calendaring and Scheduling Core Object Specification (iCalendar), RFC2445, Internet Society, (1998)

[7] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. *Machine Learning of Temporal Relations. Proceedings of ACL'2006*, (2006)

[8] S. Busemann, T. Decleck, A. K. Diagne, L. Dini, J. Klein, and S. Schmeier. Natural Language Dialogue Service for Appointment Scheduling Agents. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997, 25-32.

[9] Claude Roux. Phrase-driven parser. In *Proceedings of VEXTAL'99*, Venezia, Italia. VEXTAL'99, (1999)

[10] Salah Aıt-Mokhtar, Jean-Pierre Chanod and Claude Roux. Robustness beyond shallowness: incremental dependency parsing. *Special issue of the NLE Journal*, (2002).

[11] B. Grau, A.-L. Ligozat, I. Robba, A. Vilnat, and L. Monceaux. Frasques: A question-answering system in the EQUER evaluation campaign. *In LREC 2006, Genoa, Italia*, May 2006.

[12] Caroline Hagège and Xavier Tannier, XIP temporal module for TempEval campaign, Proceedings of SemEval workshop at ACL 2007 Prague, Czech Republic, (2007), p. 492-495

[13] Vandeghinste, Vincent and Yi Pan. Sentence compression for automated subtitling: A hybrid approach. In *Proceedings of the ACL Workshop on Text Summarization*. Barcelona, Spain, (2004), 89-95.

[14] Caroline Brun, Caroline Hagège, Intertwining Deep Syntactic Processing and Named Entity Detection, *LECTURE NOTES IN COMPUTER SCIENCE, Springer, (2004).*

Acquiring Legal Ontologies from Domain-specific Texts

Felice Dell'Orletta[1], *Alessandro Lenci*[2], *Simonetta Montemagni*[1], *Simone Marchi*[1], *Vito Pirrelli*[1], *Giulia Venturi*[1]

[1]Istituto di Linguistica Computazionale, CNR, Pisa, Italy
[2]Department of Linguistics, University of Pisa, Italy

Abstract

The paper reports on methodology and preliminary results of a case study in automatically extracting ontological knowledge from Italian legislative texts in the environmental domain. We use a fully-implemented ontology learning system (T2K) that includes a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning. Tools are dynamically integrated to provide an incremental representation of the content of vast repositories of unstructured documents. Evaluated results, however preliminary, are very encouraging, showing the great potential of NLP-powered incremental systems like T2K for accurate large-scale semi–automatic extraction of legal ontologies.

Index Terms: ontology learning, document management, knowledge extraction from texts, Natural Language Processing

1. Introduction

The last few years have witnessed a growing body of research and practice aimed at developing legal ontologies for application in the law domain. A number of legal ontologies have been proposed in a variety of research projects, mostly focusing on upper level concepts hand-crafted by domain experts (see [12], for a recent survey). It goes without saying that realistically large knowledge–based applications in the legal domain will need more and more comprehensive ontologies, incrementally integrating continuously updated knowledge. In this perspective, techniques for automated ontology–learning from texts are expected to play an increasingly more prominent role in the near future.

To our knowledge, however, relatively few attempts have been made so far to automatically induce legal domain ontologies from texts. This is the case, for instance, of [10], [11] and [14].The work illustrated in this paper represents another attempt in this direction. It reports the results of a case study carried out in the legal domain to automatically induce ontological knowledge from texts with an ontology learning system, hereafter referred to as T2K (TexttoKnowledge), jointly designed and developed by the Institute of Computational Linguistics (CNR) and the Department of Linguistics of the University of Pisa. The system offers a battery of tools for Natural Language Processing (NLP), statistical text analysis and machine language learning, which are dynamically integrated to provide an accurate representation of the content of vast repositories of unstructured documents in technical domains (DellOrletta et al., 2006). Text interpretation ranges from acquisition of lexical and terminological resources, to advanced syntax and ontological/conceptual mapping. Interpretation results are annotated as XML metadata, thus offering the further bonus of a growing interoperability with automated content management systems for personalized knowledge profiling. Prototype versions of T2K are currently running on public administration portals and have been used for indexing Elearning and Ecommerce materials. In what follows, we report some ontology learning experiments carried out with T2K on Italian legislative texts.

2. From Text to Knowledge

Technologies in the area of knowledge management and information access are confronted with a typical acquisition paradox. As knowledge is mostly conveyed through text, content access requires understanding the linguistic structures representing content in text at a level of considerable detail. In turn, processing linguistic structures at the depth needed for content understanding presupposes that a considerable amount of domain knowledge is already in place. Structural ambiguities, long-range dependency chains, complex domain-specific terms and the ubiquitous surface variability of phraseological expressions require the operation of a battery of disambiguating constraints, i.e. a set of interface rules mapping the underlying conceptual organization of a domain onto surface language. With no such constraints in place, text becomes a slippery ground of unstructured, strongly perspectivized and combinatorially ambiguous information bits.

There is no simple way around this paradox. Pattern matching techniques allow for fragments of knowledge to be tracked down only in limited text windows, while foundational ontologies turn out to be too general to make successful contact with language variability at large. The only effective solution, we believe, is to face the paradox in its full complexity. An incremental interleaving of robust parsing technology and machine learning techniques can go a long way towards meeting this objective. Language technology offers the jumping–off point for segmenting texts into grammatically meaningful complex units and organizing them into non recursive phrasal "chunks" that require no domain–specific knowledge. In turn, chunked texts can sensibly be accessed and compared for statisticallysignificant patterns of domain-specific terms to be tracked down. Surely, this level of paradigmatic categorization is still very rudimentary: at this stage we do not yet know how chunked units are mutually related in context (i.e. what grammatical relations link the min texts) or how similar they are semantically. To go beyond this stage, we suggest getting back to the syntagmatic organization of texts. Current parsing technologies allow for local dependency relations among chunks to be identified reliably. If a sufficiently large amount of parsed text is provided, local dependencies can be used to acquire a first level of domain-specific conceptual organization. We can then use this preliminary conceptual map for harder and longer dependency chains to be parsed and for larger and deeper conceptual networks to be acquired. To sum up, facing the bootstrapping paradox requires an incremental process of annotation-

acquisition-annotation, whereby domain-specific knowledge is acquired from linguistically-annotated texts and then projected back onto texts for extra linguistic information to be annotated and further knowledge layers to be extracted.

To implement this scenario, a few NLP ingredients are required. Preliminary term extraction presupposes postagged texts, where each word form is assigned the contextually appropriate part-of-speech and a set of morpho-syntactic features plus an indication of lemma. Whenever more information about the local syntactic context is to be exploited, it is advisable that basic syntactic structures are identified. As we shall see in more detail below, we use chunking technology to attain this level of basic syntactic structuring. NLP requirements become more demanding when identified terms need be organised into larger conceptual structures and connected through long-distance relational information. For this purpose syntactic information must include identification of dependencies among lexical heads. The approach to ontology learning adopted by T2K exploits all these levels of linguistic annotation of texts in an incremental fashion. Term extraction operates on texts annotated with basic syntactic structures (so–called "chunks"). Identification of conceptual structures, on the other hand, is carried out against a dependency-annotated text.

3. T2K architecture

T2K is a hybrid ontology learning system combining linguistic technologies and statistical techniques. T2K does its job into two basic steps:

1. extraction of domain terminology, both single and multi–word terms, from a document base;
2. organization and structuring of the set of acquired terms into proto–conceptual structures, namely
   - fragments of taxonomical chains, and
   - clusters of semantically related terms.

Figure 1 illustrates the functional architecture of T2K:

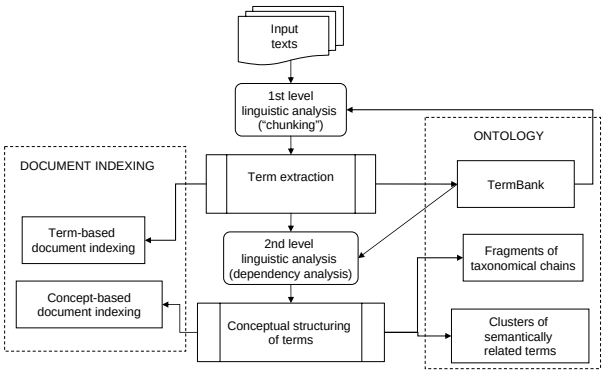


Figure 1: *T2K architecture.*

The two basic steps take the central pillar of the portrayed architecture, showing the interleaving of Natural Language Processing (NLP) and statistical tools. Acquired results are structured in the ontology box on the right–hand–side of the diagram, whose stratified organization is reminiscent of the hierarchical cascade of knowledge layers in the "Ontology Learning Layer Cake" by [6], going from terminological information to proto–conceptual structures corresponding to taxonomical and non-hierarchical relationships among terms. Acquired knowledge is also used for document indexing, on the basis of extracted terms and acquired conceptual structures. In what follows we focus on the ontology learning process.

3.1. Term extraction

Term extraction is the first and most–established step in ontology learning from texts. For our present purposes, a term can be a common noun as well as a complex nominal structure with modifiers (typically, adjectival and prepositional modifiers).

T2K looks for terms in shallow parsed texts, i.e. texts segmented into an unstructured (non-recursive) sequence of syntactically organized text units called "chunks" (e.g. nominal, verbal, prepositional chunks). Candidate terms may be one word terms ("single terms") or multi–word terms ("complex terms"). The acquisition strategy differs in the two cases.

Single terms are identified on the basis of frequency counts in the shallow parsed texts, after discounting stop–words. The acquisition of multi–word terms, on the other hand, follows a two–stage strategy. First, the chunked text is searched for on the basis of a set of chunk patterns. Chunk patterns encode syntactic templates of candidate complex terms: for instance, adjectival modification (e.g. *organizzazione internazionale* 'international organisation'), prepositional modification (e.g. *commercializzazione di autovetture* 'marketing of cars'), including more complex cases where different modification types are compounded (e.g. *commercio di prodotti fitosanitari* 'trade of fitosanitary products'). Secondly, the list of acquired potential complex terms is ranked according to their log–likelihood ratio [8], an association measure that quantifies how likely the constituents of a complex term are to occur together in a corpus if they were (in)dipendently distributed, where the (in)dependence hypothesis is estimated with the binomial distribution of their joint and disjoint frequencies.

Recognition of longer terms is carried out by iteratively applying the extraction process to the results of the previous acquisition step. This means that acquired complex terms are projected back onto the original text and the acquisition procedure is iterated on the newly annotated text. The method proves helpful in reducing the number of false positives consisting of more than two chunks [4]. Interestingly, the chunk patterns used for recognition of multi–word terms need not necessarily be the same across different iteration stages. In fact, it is advisable to introduce potentially noisy patterns (such as, for example, coordination patterns) only at later stages.

The iterative process of term acquisition yields a list of candidate single terms ranked by decreasing frequencies, and a list of candidate complex terms ranked by decreasing scores of association strength. The selection of a final set of terms to eventually be acquired requires some threshold tuning, depending on the size of the document collection and the typology and reliability of expected results. Thresholds define *a)* the minimum frequency for a candidate term to enter the lexicon, and *b)* the overall percentage of terms that are promoted from the ranked lists.

3.2. Term organization and structuring

In the second extraction step, proto–conceptual structures involving acquired terms are identified. The basic source of in-

formation is no longer a chunked text, but rather a dependency–annotated text, including information about multi–word terms acquired at the previous extraction stage.

We envisage two levels of conceptual organization. Terms in the TermBank are first organized into fragments of head–sharing taxonomical chains, whereby *commercio dei medicinali* 'trade of medicines' and *commercio elettronico* 'elettronic trade' are classified as co–hyponyms of the general single term *commercio* 'trade'. In this way, single and multi–word terms are structured in vertical relationships providing fragments of taxonomical chains.

The second structuring step consists in the identification of clusters of semantically related terms, carried out on the basis of distributionally–based similarity measures. This involves use of CLASS, a distributionally–based algorithm for building classes of semantically–related terms [1]. According to CLASS, two terms are semantically related if they can be used interchangeably in a statistically significant number of syntactic contexts. For all terms (both single and complex) in the TermBank, we extracted from the dependency–annotated text all relations involving these terms in the text. For each term, we selected all the grammatical dependencies it is involved in, and identified (after discarding auxiliary and commonest verbs) the most meaningful (i.e. selective) verbs as resulting from the log–likelihood ratio association measure. The cluster of terms semantically related to a given term is finally determined by computing all the similar terms with respect to each meaningful verb and by grouping the highest ranked terms obtained from the computation on different verbs.

4. Ontology learning from legislative texts: a case study

In this section we summarise the results of a case study carried out on a corpus of legal texts in the environmental domain (Venturi, 2006).

4.1. Corpus description and preprocessing

The corpus consists of 824 legislative, institutional and administrative acts in the environmental domain, for a total of 1.399.617 word tokens, coming from the BGA (*Bollettino Giuridico Ambientale*) database edited by the Piedmont local authority for environment.[1] The corpus includes acts released over a nine years period (from 1997 to 2005) by three different agencies: the European Union, the Italian state and the Piedmont region. It is a heterogeneous document collection including legal acts such as national and regional laws, european directives, legislative decrees as well as administrative acts such as ministerial circulars, decisions, etc.

4.2. The legal–environmental TermBank

Table 1 contains a fragment of the automatically acquired TermBank. For each selected term, the table reports its prototypical form (in the column headed "Term") and its frequency of occurrence in the whole document collection. The choice of representing a domain term through its prototypical form rather than the lemma exponent follows from the assumption that a bootstrapped glossary should reflect the actual usage of terms in texts. In fact, domain-specific meanings are often associated with a particular morphological form of a given term (e.g. the plural form). This is well exemplified in Table 1 where the

ID	Term	Freq
2192	acqua calda	11
974	acqua potabile	36
501	acqua pubblica	121
47	acque	1655
2280	acque costiere	10
2891	acque di lavaggio	6
2648	acque di prima pioggia	8
3479	acque di transizione	5
1984	acque meteoriche	12
1690	acque minerali	16
400	acque reflue	231
505	acque sotterranee	120
486	acque superficiali	131
2692	acque utilizzate	8

Table 1: A fragment of the automatically acquired TermBank

acquired terms headed by *acqua* 'water' can be parted into two groups according to their prototypical form: either singular (e.g. *acqua potabile* 'drinkable water') or plural (e.g. *acque superficiali* 'surface runoff'). Note, however, that reported frequencies are not limited to the prototypical form, but refer to all occurrences of the abstract term.

Most notaby, the acquired TermBank includes both legal and environmental terms. The two classes of terms show quite different frequency distributions and turn out to be differentially sensitive to varying frequency thresholds (see Section 3.1). Evaluation of acquired results was carried out with respect to the most conservative TermBank of 4.685 terms, obtained by setting a high minimum frequency threshold (7). Due to the heterogeneous nature of acquired terms, belonging to both the legal–administrative and environmental domains, different resources were taken as evaluation standards: the *Dizionario giuridico* (Edizioni Simone) available online[2] was used as a reference resource for what concerns the legal domain (henceforth referred to as Legal_RR), and the *Glossary of the Osservatorio Nazionale sui Rifiuti* (Ministero dell'Ambiente) available online[3] for the environmental domain (henceforth referred to as Env_RR), which contain respectively 6.041 and 1.090 terminological entries recorded in their prototypical form. For evaluation purposes, more charitable matching metrics between acquired and target terms were considered than full matching, namely:

1. the acquired and target term can appear in different prototypical forms (e.g. *accordi di programma* 'programmatic agreement/PLUR' vs. *accordo di programma* 'programmatic agreement/SING', or *acquisizione dati* 'data acquisition' vs. *acquisizione* ***di*** *dati* 'acquisition of data');
2. the target term can be more general than the acquired one: for example the T2K term *abrogazione di norme* 'repeal of rules' is a good match of Legal_RR *abrogazione* 'repeal';
3. the target term can be more specific than the acquired one: e.g. T2K *agente di polizia* 'policeman' (T2K) is matched against *agente di polizia giudiziaria* 'prison guard' attested in Legal_RR.

Finally, criteria 2 and 3 above can combine with 1. Results are summarised as follows: we found 51% of either full

[1] http://extranet.regione.piemonte.it/ambiente/bga/

[2] http://www.simone.it/cgi-local/Dizionari/newdiz.cgi?index,5,A

[3] http://www.osservatorionazionalerifiuti.it/ShowGlossario.asp?L=Z

or partial matches between the T2K glossary and the reference resources. 89% of the matches covered legal terms and 34,5% environmental ones. 23,5% were found to match entries in both legal and environmental resources. What about the remaining 49% mismatches? How many of them can be considered out–of–dictionary hits? To answer these questions, we selected two additional terminological resources available on the Web: the list of keywords used for the online query of the *Archivio DoGi (Dottrina Giuridica)*[4] for the legal domain, and the thesaurus *EARTh (Environmental Applications Reference Thesaurus)*[5] for the environmental domain. Results are quite encouraging: inclusion of these richer reference resources increases the percentage of matches up to 75,4%. The same percentage goes up even further (83,7%) if we include terms which, in spite of their absence in the selected reference resources, were manually evaluated as domain–relevant terms (see e.g. *anidride carbonica* 'carbon dioxide' in the environmental domain or *beneficiari* 'beneficiary' in the legal one).

5. Conclusions and further directions of research

We reported encouraging results of the application of an automatic ontology learning system, T2K, on a corpus of Italian legislative texts in the environmental domain. Our work shows that the incremental interleaving of robust NLP and machine–learning technologies is key to any attempt to successfully face what we termed the acquisition paradox. By bootstrapping base domain–specific knowledge from texts through knowledge–poor language tools we can incrementally develop more and more sophisticated levels of content representation. In the end the purported dividing line between language–knowledge and domain–specific knowledge proves to be untenable in language use, where language structures and bits of world–knowledge are inextricably intertwined.

There is an enormous potential for this bootstrapping technology. Acquired TermBanks can be transformed into semantic networks linking identified legal and environmental entities. Current lines of research in this direction include a) semi–automatic induction and labelling of ontological classes from the proto–conceptual structures identified by T2K, and b) the extension of the acquired ontology with concept–linking relations (Venturi, 2006). Furthermore, estabilishing the domain relevance of each acquired term represents a central issue in dealing with domain–specificity. By comparing TermBanks automatically extracted from different legislative corpora, we can be successful in classifying the terms belonging to their intersection as specific of the shared domain (in line with the contrastive approach to term extraction proposed by [5]).

6. References

[1] Allegrini, P., Montemagni, S. and V. Pirrelli. Example-Based Automatic Induction Of Semantic Classes Through Entropic Scores. *Linguistica Computazionale*, 1-43: 2003.

[2] Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Nave Interplay. In *Proceedings of the International COLING-2002 Workshop "Grammar Engineering and Evaluation"*, Taiwan 2004.

[3] Bartolini, R., Lenci, A., Montemagni, S. and V. Pirrelli. Hybrid Constrains for Robust Parsing: First Experiments and Evaluation. In *Proceedings of LREC 2004*, Lisbon 2004.

[4] Bartolini, R., Giorgetti, D., Lenci, A., Montemagni, S. and V. Pirrelli. Automatic Incremental Term Acquisition from Domain Corpora. In *Proceedings of the 7th International conference on "Terminology and Knowledge Engineering" (TKE2005)*, Copenhagen 2005.

[5] Basili, R., Moschitti, A., Pazienza, M.T. and Zanzotto, F.M. A contrastive approach to term extraction. In *Proceedings of the 4th Conference on Terminology and Artificial Intelligence (TIA2001)*, Nancy, France, 2001.

[6] Buitelaar, P., Cimiano, P., and B. Magnini. Ontology Learning from Text: an Overview. In Buitelaar et al. (eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* (Volume 123 Frontiers in Artificial Intelligence and Applications): 3–12, 2005.

[7] Dell'Orletta, F., Lenci, A., Marchi, S., Montemagni, S. and V. Pirrelli. Text-2-Knowledge: una piattaforma linguistico-computazionale per l'estrazione di conoscenza da testi. In *Proceedings of the SLI-2006 Conference*: 20–28, Vercelli 2006.

[8] Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*: 19(1), 1993.

[9] Federici, S., Montemagni, S. and V. Pirrelli. Shallow Parsing and Text Chunking: a View on Underspecification in Syntax. In *Proceedings of the Workshop On Robust Parsing*, in the framework of the European Summer School on Language, Logic and Information (ESSLLI-96), Prague 1996.

[10] Lame, G. Using NLP techniques to identify legal ontology components: concepts and relations. *Lecture Notes in Computer Science*, Volume 3369: 169–184, 2005.

[11] Sais, J. and P. Quaresma. A Methodology to Create Legal Ontologies in a Logic Programming Based Web Information Retrieval System. *Lecture Notes in Computer Science*, Volume 3369: 185–200, 2005.

[12] Valente, A. Types and Roles of Legal Ontologies. *Lecture Notes in Computer Science*, Volume 3369: 65–76, 2005.

[13] Venturi, G. L'ambiente, le norme, il computer. Studio linguistico–computazionale per la creazione di ontologie giuridiche in materia ambientale. Degree Thesis, Manuscript, December 2006.

[14] Walter, S. and M. Pinkal. Automatic extraction of definitions from german court decisions. In *Proceedings of the COLING-2006 Workshop on Information Extraction Beyond The Document*: 20–28, Sidney 2006.

[4] http://nir.ittig.cnr.it/dogiswish/dogiConsultazioneClassificazioneKWOC.php

[5] http://uta.iia.cnr.it/earth.htm#EARTh%202002

Advances in NLP applied to Word Prediction

Carlo Aliprandi[1], *Nicola Carmignani*[2], *Nedjma Deha*[2], *Paolo Mancarella*[2], *Michele Rubino*[2]

[1]Synthema Srl – Pisa, Italy
[2]Department of Computer Science – University of Pisa, Italy
carlo.aliprandi@synthema.it, {nicola, deha, paolo, rubino}@di.unipi.it

Abstract

Presenting some recent advances in word prediction, a fluorishing research area in Natural Language Processing, we describe FastType, an innovative word prediction system that outclasses typical limitations of standard techniques when applied to inflected languages. FastType is based on combined statistical and rule-base methods relying on robust open-domain language resources, that have been refined to improve Keystroke Saving. Word prediction is particularly useful to minimise keystrokes for users with special needs, and to reduce misspellings for users having limited language proficiency. Word prediction can be effectively used in language learning, by suggesting correct words to non-native users. FastType has been tried out and evaluated in some test benchmarks, showing a relevant improvement in Keystroke Saving, which now reaches 51%, comparable to what achieved by word prediction methods for non-inflected languages.

Index Terms: Word Prediction, Natural Language Processing (NLP), Augmentative and Alternative Communication, Computer Aided Language Learning, Speech and Natural Language Interfaces, Assistive Technology

1. Introduction

This paper describes an innovative approach to Word Prediction, presenting recent results achieved for inflected languages.

Word Prediction is the task of guessing words that are likely to follow a given fragment of text. A Word Prediction software is a writing support: at each keystroke it suggests a list of meaningful predictions, amongst which the user can possibly identify the word he is willing to type. By selecting a word from the list, the software will automatically complete the word being written, thus saving keystrokes.

Word prediction is facing a very ambitious challenge, as several typical complex problems arising when dealing with Natural Language are to be faced. The inherent amount of arising ambiguities (lexical, structural and semantic ambiguities but also pragmatic, cultural and phonetic ambiguities for speech) are complex problems to be solved by a computer. Many research efforts have been experimented and several core NLP tasks have been employed as, for example, Language Modeling, Part-of-Speech (POS) Tagging, Parsing and Lemmatisation.

Word prediction has been widely adopted in Augmentative and Alternative Communication (AAC) systems [1], becoming an essential aid for people with motor or cognitive disabilities, in order to reduce the typing effort and to assist learning or language impairments. Indeed, writing text for work, study or communicating is, according to a survey we conducted (as described in [2]), the most frequent and time-consuming activity for most computer users. Therefore a word predictor would be useful to a very large number of computer users, both disabled and not.

FastType is designed to predict words for inflected languages, that is languages that have a large dictionary of word forms with several morphological features, produced from a root or lemma and a set of inflection rules. The degree of inflection of a language may vary from very high (e.g. Basque), to moderate (e.g. Spanish, Italian, French), to low (e.g. English). The large number of word forms makes word prediction for inflected languages a hard task. As word prediction operates at typing time, any NLP task that can be applied, unlike common NLP analytics which processes complete sentences, has to cope with the further problem of sentence incompleteness.

To make word prediction as simple and immediate as possible, we have implemented DonKey, a new human-computer interface. DonKey improves the original, naive, interface of FastType, allowing the user to benefit from automatic word prediction in any desktop application. In addition to re-designing the user interface, the underlying prediction engine has been enhanced: we added new resources, like the word and Part-of-Speech *n*-gram Language Models, and implemented more efficient prediction algorithms.

Thanks to the upgrades, performances are greatly improved. Keystroke Saving reached 51% and is now comparable to the one achieved with state-of-the-art methods for non-inflected languages.

2. State of the Art on Word Prediction

Word prediction is a research area where a very challenging and ambitious task is faced, basically with methods coming from Artificial Intelligence, Natural Language Processing and Machine Learning.

The main goal of word prediction is guessing and completing the word a user is willing to type. Word predictors are intended to support writing and are commonly used in combination with assistive devices such as keyboards, virtual keyboards, touchpads and pointing devices. Another potential application is in text-entry interfaces [3] for messaging on mobile phones and typing on handheld and ubiquitous devices (e.g. PDAs or smartphones).

Prediction methods have become quite known as largely adopted in mobile phones and PDAs, where *multitap* is the input method. Nuance T9 (formerly Tegic Communications T9)[1] and Zi Corporation eZiText[2] are commercial systems that adopt a very simple method of prediction based on *dictionary disambiguation*. At each user keystroke the system selects the letter between the ones associated with the key guessing it from a dic-

[1]http://www.nuance.com/t9/
[2]http://www.zicorp.com/eProducts/ZiPredictiveTextSuite/

tionary of words: hence they are commonly referred to as letter predictors. Letter predictors bring a Keystroke Saving (KS) but it has been proven to be not completely free from ambiguities that are more frequent for inflected languages. So it is not surprising that these methods had a great success for non inflected languages such as English: the limited number of inflectional forms lead to very high KS that, at the moment, are above 40%.

Word prediction is a more sophisticated technique within recent research. Differently from letter predictors, word predictors typically make use of language modelling techniques, namely stochastic models that are able to give context information in order to improve the prediction quality.

FASTY [4] is a statistically based adaptive word prediction program. The FASTY Language Model utilizes word unigrams, word bigrams, POS trigrams and the probability distribution $P(t|w)$, i.e. the probability that POS tag t occurs with a given word w.

Most of the literature related to word prediction concerns non-inflected languages [5]. In [6] and [7] Language Models and prediction techniques are presented that allow the user to save more than 50% of keystrokes. The contribution of the system presented in this paper is the adaptation and improvement of these techniques for inflected languages.

The language that the system has to model influences the prediction techniques; inflected languages pose a harder challenge to prediction algorithms, since they have to deal with a usually high number of inflected forms that dramatically decrease Keystroke Saving [8]. To simplify the task of predicting the correct form, some techniques [9] provide a two-step procedure, choosing first only among word "roots", and proposing all the possible word forms only when the user selects a root. FastType relies instead on Part-of-Speech (POS) and related morpho-syntactic information to provide a one-step procedure, presenting to the user a list of word forms. This procedure, combined with on-the-fly POS tagging, enables FastType to boost performances, cutting off of the prediction list all words whose gender, number, tense or mood are not consistent with the sentence context. The prediction list becomes also a "guide tool" to write syntactically correct sentences.

3. Description of the Word Predictor

Figure 1 shows the three main components of the FastType system: the *User Interface*, the *Prediction Engine* and the *Linguistic Resources*.

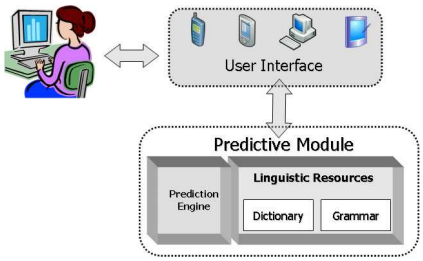


Figure 1: *FastType Architecture.*

The *Prediction Engine* is the kernel of the *Predictive Module* since it manages the communication with the User Interface, keeping trace of the prediction status and of the words already typed. At each keystroke it predicts suggestions, in the form of a list of word completions, by assuring accordance (gender, number, person, tense and mood) with the syntactic sentence context.

All the prediction functions are now encapsulated into a separate library, the *Predictive Module*, available also for integration with others applications. The *Predictive Module* provides core functionalities, such as the morpho-syntactic agreement and the lexicon coverage, efficiently accessing the *Linguistic Resources*, as the Language Model and very large lexical resources. We added new resources, namely POS *n*-grams and Tagged Word (TW) *n*-grams to our Language Model, improving the quality of morphological information available for the Prediction Engine. The tagged word *n*-gram model extends the typical word *n*-gram model [10] by adding POS information. For example a word bigram (w_{i-1}, w_i) is extended to a Tagged Word bigram (w_{i-1}, w_i, t_i), where t_i is the POS of w_i.

We introduce a new prediction algorithm for the Italian language based on Linear Combination [11]. The approach closest to ours is the one presented in [6], that is a linear combination algorithm combining POS trigrams and simple word bigrams. Our algorithm extends this model to cope with inflected languages, by combining POS *n*-gram models with tagged word *n*-gram models.

The Italian POS *n*-grams, approximated to $n = 2$ (bigrams) and $n = 3$ (trigrams) and tagged word *n*-grams, approximated to $n = 1$ (unigrams) and $n = 2$ (bigrams) have been trained from a large corpus created from newspapers, magazines, documents, commercial letters and emails.

The POS trigram model determines the most likely POS tags for the current word, given the two previous POS tags, if necessary backed up by POS bigrams. The TW bigram model establishes the most likely words given the immediately previous word. The probability S for the current word is the result of a weighed linear combination of the models:

$$S = \alpha \cdot \mathbb{P}(w_i \mid (w_{i-1}, t_i)) + \beta \cdot f(t_i, t_{i-1}, t_{i-2}) \quad (1)$$

where $\mathbb{P}(w_i \mid (w_{i-1},\ t_i))$ is the probability of the TW bigram (w_{i-1}, w_i, t_i), i.e. the probability of the next word being w_i, given that the previous word is w_{i-1} and the next word should have the t_i POS,

$$f(t, t', t'') = \begin{cases} \mathbb{P}(t \mid t', t'') & if\ \mathbb{P}(t \mid t', t'') > \vartheta \\ \mathbb{P}(t \mid t') & \text{otherwise} \end{cases} \quad (2)$$

and ϑ is the threshold empirically set. α and β are the coefficients of the linear combination and their sum must be 1 ($\alpha + \beta = 1$).

Donkey, the new FastType user interface (shown in Figure 2), is very simple and particularly easy to use. The system provides the user with a list of ranked suggestions. The user accept a word either by selecting the related function key (F1, F2, F3, and so on) or by using the pointing device (e.g. a traditional mouse or an eye tracker) to click the corresponding button. In this way the user can continue to write, looking for suggestions in the list and choosing the desired word that will be automatically inserted into the text.

Since there is a typical cognitive load associated to the interaction with word prediction systems due to the disability level or the limited language proficiency, Donkey can be adapted to the user needs. Donkey configuration utility provides a set of options that allow the user to personalize the word predictor functionalities, such as dimension, font, capitalization of the text in the suggestion list or its length.

Figure 2: *The User Interface.*

The length of the suggestion list influences the time and the effort required to search and select the right word. In consequence, the user can customize the number of suggestions in a range of 1-10. We have limited the number of suggestions presented by Donkey to 10, since a user can notice at a glance a word appearing in a smaller list, rather than in a larger list. Indeed, the larger the list, the higher the level of concentration required to read all the suggestions.

Donkey can be adapted even further in order to achieve a better interaction for blind or visually impaired users: for example Text-to-Speech options are available for reading words in the suggestion list or the selected word.

4. Evaluation

As described in [5], it is difficult to find appropriate metrics to measure prediction activities. In particular, a metric may be of more pertinence than another if there is an impairment in the user abilities. Thus we performed a general evaluation of the system, using different evaluation metrics.

Keystroke Saving (KS): being $c_1 \ldots c_n$ is an evaluation metric largely adopted in literature and provides a significative-for-all measure of efficacy. Keystroke Saving (KS) estimates the saved effort percentage and is calculated by comparing two kinds of measures: the total number of keystrokes needed to type the text (K_T) and the effective number of keystrokes using word prediction (K_E). Hence,

$$KS = \frac{K_T - K_E}{K_T} \cdot 100$$

There are two additional metrics we use to evaluate FastType prediction accuracy: Keystrokes Until Completion (KUC) and Word Type Saving (WTS).

Keystrokes until Completion (KUC): being $c_1 \ldots c_n$ the number of keystrokes for each of the n words before the desired suggestion appears in the prediction list,

$$KUC = \frac{(c_1 + c_2 + \ldots + c_n)}{n}$$

Word Type Saving (WTS): the percentage of time the user saves with FastType. Being T_n the time needed to write a text without FastType and T_a the time needed to write the same text with FastType,

$$WTS = \frac{(T_n - T_a)}{T_n} \times 100$$

To measure FastType performance improvements with the new linear combination algorithm we ran trials on the same test set presented in [12]. The test set was a subset of 40 texts disjoint from the training set. We developed a new test bench, performing different trials to experimentally determine the optimal value for α and β. The nutshell of the test bench is a 'simulated user' typing the test set and acting as a user that always selects the correct suggestion when predicted. We then measured the KS varying values for α and β. We ran trials increasing α by 0.1 from 0.1 to 0.9, empirically isolating the value of α producing the best KS.

Table 1: *Performance Measurement Results*

L	***KS***	***KUC***	***WTS***
5	46.79%	2.55	25.36%
10	51.16%	2.34	28.66%
20	55.13%	2.06	29.19%

A parameter that can greatly influence performance measurements is the length L of the prediction list, so we ran three trials on the test set with $L = 5$, $L = 10$ and $L = 20$. As we can see in Table 1, the increase in KS, WTS and KUC between $L = 5$ and $L = 10$ is way more relevant than the increase between $L = 10$ and $L = 20$.

The average KS is between 46.79% and 55.13%, marking a sensible improvement if compared with our previous results. Performances are significantly good for WTS, meaning that -at a standard speed and without any added cognitive load- saving in time is average around 29%. Particularly significant is also the KUC, meaning that the correct word is suggested after an average of 2.5 keystrokes for $L = 5$, 2.3 keystrokes for $L = 10$ and 2 keystrokes for $L = 20$.

Figure 3 presents a sample text: predicted keystrokes, blue marked, are 175 out of a total of 349 keystrokes, thus producing, in this case, a KS of about 50%.

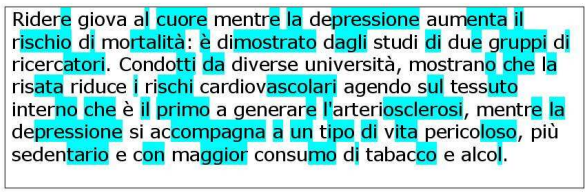


Figure 3: *Word Prediction results.*

Performances are comparable with existing works on non-inflected languages, as in [6] and [7], since with $L = 10$ FastType KS rises to 51%.

5. Conclusions

In this paper, we have presented the FastType system and its new human-computer interface, DonKey. DonKey allows the user to benefit from automatic word prediction in any desktop application.

We also described recent enhancements we introduced to the FastType system. By making use of POS tags we built a new Language Model and we refined the prediction algorithm.

We have evaluated FastType performance enhancements for an inflected language, i.e. Italian. According to our tests word prediction reaches a Keystroke Saving up to 51% for a standard prediction list of length 10. Keystroke Saving is now comparable to the one achieved by other systems for non-inflected languages, thus outclassing some typical word prediction limitations.

Our conclusions are consistent with state of the art literature, for example with [8], who claimed that a word prediction method without syntactic information are not applicable to inflected languages. We additionally enriched the Language Model with morpho-syntactic information and provided the prediction method with an on-the-fly Part-of-Speech word tagger and large lexicon dictionaries.

For future work we have plans for running field tests with disabled people, in order to improve DonKey usability in daily tasks as writing texts or emails and communicating. We have also plans for designing and developing a prototype version for PDAs and smartphones.

In conclusion, FastType has peculiarities and potential advantages since, using very large lexical resources and statistically based techniques, an effective word prediction can be performed in real domains.

We believe that the application of this technology is wide and we are working to bring the benefits of fast text typing to virtual keyboards and portable devices like smartphones and PDAs.

6. Acknowledgements

The FastType project is partially funded by the Fondazione Cassa di Risparmio di Pisa.

7. References

[1] A. Copestake, Augmentative and Alternative NLP Techniques for Augmentative and Alternative Communication, *Proceedings of the ACL Workshop on NLP for Communication Aids*, 37–42, 1997.

[2] C. Aliprandi, N. Carmignani, P. Mancarella and M. Rubino, A Word Predictor for Inflected Languages: System Design and User-Centric Inferface, *Proceedings of the 2^{nd} IASTED International Conference on Human-Computer Interaction*, 2007.

[3] I. MacKenzie, J. Chen and A. Oniszczak, Unipad: Single-Stroke Text Entry with Language-based Acceleration, *Proceedings of the Fourth Nordic Conference on Human-Computer Interaction*, 78–85, 2006.

[4] H. Trost, J. Matiasek and M. Baroni. The language component of the FASTY text prediction system. *Applied Artificial Intelligence*, 743-781, 2005.

[5] N. Garay-Vitoria and J. Abascal, Text Prediction Systems: A Survey, *Universal Access in the Information Society*, 4(3), 188–203, 2006.

[6] A. Fazly and G. Hirst, Testing the Efficacy of Part-of-Speech Information in Word Prediction, *Proceedings of the 10^{th} Conference of the EACL*, 2003.

[7] S. Palazuelos-Cagigas, J. Martín-Sánchez, L. Hierrezuelo Sabatela and J. Macías Guarasa, Design and Evaluation of a Versatile Architecture for a Multilingual Word Prediction System, Proceedings 10^{th} International Conference on Computers Helping People with Special Needs, 894–901, 2006.

[8] K. Tanaka-Ishii, Word-based Predictive Text Entry Using Adaptive Language Models, *Natural Language Engineering*, 13(1), 51–64, 2007.

[9] N. Garay-Vitoria and J. Abascal, Word Prediction for Inflected Languages. Application to Basque Language, *Proceedings of the ACL Workshop on NLP for Communication Aids*, 29–36, 1997.

[10] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, 1999.

[11] N. Deha, FastType: Predizione di Parola basata su Modelli Statistici in un Ambiente di Scrittura Assistita, Master's Thesis in Computer Science, Department of Computer Science, University of Pisa, 2007.

[12] C. Aliprandi, N. Carmignani and P. Mancarella, An Inflected-Sensitive Letter and Word Prediction System, *Proceedings of the International Conference on Interactive Computer Aided Learning*, 2006.

[13] J. Arnott, A. Newell and N. Alm, Prediction and Conversational Momentum in an Augmentative Communication System, *Communications of the ACM*, 35(5), 46–57, 1992.

[14] C. Aliprandi, N. Carmignani, N. Deha, P. Mancarella and M. Rubino, FastType, a Word Predictor for Inflected Languages: Syntactic Prediction Features and User-Centric Interface, *Proceedings of the 9th European Conference for the Advancement of Assistive Technology*, 378–382, 2007.

An Online Linguistic Journalism Agency – Starting Up Project

Annibale Elia [1], *Ernesto D'Avanzo* [1], *Tsvi Kufik*[3], *Giovanni Catapano* [1], *Mara Gruber* [2]

[1] CLiCLab, Department of Communication Sciences, University of Salerno, Fisciano (SA), Italy
[2] Istituto Italiano di Scienze Umane, Napoli, Italy
[3]Management of Information Systems Department, University of Haifa, Haifa, Israel
aelia@unisa.it, edavanzo@unisa.it, tsvikak@mis.hevra.haifa.ac.il, gcatapano@unisa.it, gruber@fbk.eu

Abstract

The Web provides easy access from everywhere to every kind of information. It is becoming a substitute source for news, instead of traditional media such as newspapers, radio and television. However, with the ease of access and the tremendous amounts of information available online, finding relevant information is not an easy task. This paper reports on an under development joint research project which aims at the development of an online Journalism Agency that makes use of Natural Language techniques in order to provide trainee and professional journalists with topical summaries of information relevant to their interest on different channels (Web, PDAs, etc.). Our system obtained good results of the linguistic quality of the summaries in international summarization campaigns. With such a background we are quite optimistic for the future development of the project.

Index Terms: Automatic Text Summarization, Machine Learning, Linguistic Analysis, Web Mining

1. Motivation and Background

A report by Forrester [1] provoked many debates concerning the Web as substitute source for news, taking the place of traditional media as newspaper or television. This seems to be a growing trend as the report pointed out. New York Times' Web site for example counts about 20 million logs monthly meanwhile the printed edition is over one million daily. The huge number of news sources available online, and the easy, fast and free access to many news Web sites are only some reasons that motivate this trend.

However, the more information is available the harder it is to access the really relevant information. A reader interested in a given topic must read a large number of news items before he/she can satisfy his/her information need, facing the well known problem of the *information overload*.

To assist the *information consumer* in coping with this problem an out-and-out call to arms was raised, bringing together people from different disciplines in order to design and develop methods and techniques able to automatically, succinctly, and efficiently summarize these huge volumes of information available.

The "consumption" of information and accessing it is also a hard problem for information professionals such as journalist that are required on a daily basis, or even in real-time, to digest volumes of news and summarize them for a briefing with their editorial unit or to be published on the online or printed version of their newspaper.

The Department of Communication Sciences, at the University of Salerno, founded last year a Journalism School. Immediately, a lot of interests, both professional and scientific arose around the School.

One of the main goals of the Scientific Board of the School is the integration and the technological transfer from the Department of Communication Sciences. To this end this ongoing research project has been launched with the aim of building an online platform, able to support students of the school (that are journalist trainees), and the academic staff in their daily task of consumption and, at the same time, production of information. The Department developed linguistic resources as DELAS-DELAF and DELAC, two Electronic Dictionaries for single and compound terms containing about 500,000 lexical entries and local grammars. Based on these resources several Text Processing tools were built, that allow these resources to be easily exploited for Web Mining tasks.

During the past summer the Department of Communication Sciences launched CLiCLab (Computational Linguistics and Complexity Laboratory) an interdisciplinary lab with interests ranging from Computational Linguistics to Web Mining and Computational Models up to Cognitive Systems. The lab brings together people with different backgrounds, coming from different Departments and Universities worldwide.

CLiCLab works as a catalyst in order to start up this joined project between the Department of Communication Sciences and the Journalism School. Among its main activities CLiCLab developed Text and Web Mining tools acknowledged in international competitions where they were tested.

Summing up, we believe that the development of the Automatic Journalism Agency will bring the latest text mining and summarization research results and tools to the doorstep of future journalists, allowing them to better exploit the new opportunities offered by the Web.

2. Related Work

Among worldwide news Web services, *Google News* and *AltaVista News* are the most popular examples of news services that present clusters of related documents. However, none of these services is equipped with an overall summary of the whole cluster. And even if there are summaries for each document, the majority of these are created by extracting the top few sentences, with the risk of loosing important topics appearing later in the document.

Radev and co-workers developed NewsInEssence [2] a summarization system that, given a user's topic, acts as a users's agent in order to gather and summarize online news articles. The system can gather clusters of related stories from different sources on the Web and generates summaries of the whole cluster, emphasizing its most important content. NewsInEssence allows users to create personalized clusters of summaries. Cluster summaries created by NewsInEssence,

however, are not so readable and this is the major drawback of the system.

McKeown and co-workers designed Newsblaster, another system for online news summarization [4]. The systems provides updates of the news daily, crawling news sites, filtering out news from non-news, grouping news into stories on the same event, and, finally, generating a summary of each event. The crawling made by Newsblaster considers a tunable number of news sites among which CNN, Reuters, Fox News, NY Post, etc. For each page considered if the amount of text is greater than a constant (usually about 500 characters) then it is assumed to be a news article.

Since its launch Newsblaster got some changes. The interface shows a "close up" frame where there are the most important news consisting of a cluster summary about a hot topic for which have been gathered news in the past few days. At the time of this writing the close up news is titled "Ahmadinejad: Annapolis failed, Israel doomed to collapse". The news belongs to *World* category, one of the six appearing on the home of the system (other five categories are *U.S., Finance, Entertainment, Science/Technology, Sports*) and is created summarizing 32 sources articles (Washington Post, L.A. Times, nytimes.com, etc.) that are listed at the end of the page containing the summary. The summary is completed with a list of keywords and an *event tracking* to record the story's development in time. At the core of Newsblaster there are two main components: the organizer of stories, the clusterer and the multidocument summarizer. The former component uses agglomerative clustering with a groupwise average similarity function and linguistic features, such as terms, noun phrase heads and proper nouns. The summarizer component is the Columbia Summarizer made of different summarization strategies chosen depending on the typology of document sets (i.e. *single-event* documents, *person-centered* documents and *multi-event* documents). For *single-event* document summarization Newsblaster uses MultiGen a system that makes use of machine learning and statistical techniques to extract similar sentences (a set of similar sentence is called *theme*). Afterward, an alignment of parse trees finds the intersection of similar phrases within sentences. Language generation techniques are then used to cut and paste together similar phrases from the theme sentences. A theme corresponds roughly to one sentence of the summary.

Biographical document are summarized using DEMS that uses frequencies of concepts (set of synonyms) combined with global information about what words are likely to appear in a lead sentence, to decide if an article sentence should be included in the summary

3. The Start Up Project: a Linguistic Journalist Agency

The project of our news agency is composed of several modules that can be grouped in three main components:

- Topical Web crawler able to gather related news on the Web
- Summarizer able to create multi-document summaries that gets the news clusters coming from the previous module and creates summaries using Natural Language Processing techniques
- A Web platform able to deliver summaries created to different channels (e.g., email, Web, PDA's)

The whole platform is equipped with an agent like behavior. The system infers users (journalists) preferences. Then, based on these it runs the topical crawler to gather new clusters of news. In the following we describe each of these components. Some of them are mature methodologies that we developed and evaluated as standalone applications. Others are under development and need to be evaluated. An overall evaluation of the agency, involving professional and trainee journalists, is planned after a first prototype will be released (September 2008).

3.1. Crawling

A web crawler/Spider is an automated script which automatically browses the World Wide Web. The crawler is mainly used by sites for providing updated data from web pages; it creates a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

For our purposes in this early stage we immediately played with topical crawlers [5], a method and a system for directing Web crawling to a topic, using a focused search engine that produces a specialized collection of documents and document ranking. The method includes the following steps:

- receiving a user's request query that includes one or more words, phrases or documents, for defining the topic
- generating an affinity list which is a ranked list of terms, phrases, documents or set of documents related to the query that are derived from statistics about the document collection locating and retrieving seed documents, that includes relevant and irrelevant information
- training a binary classifier using seed documents to define documents
- causing a web spider to locate and retrieve documents related to the user's query
- ranking URLs associated with the documents

An important aspect to consider when applying crawlers, especially topical crawlers, is the nature of the crawl task. Crawl characteristics such as queries and/or keywords provided as input criteria to the crawler, user-profiles, and desired properties of the pages to be fetched (similar pages, popular pages, authoritative pages etc.) can lead to significant differences in crawler design and implementation.

In our experiments we have available profiles (as a set of keywords) acquired from trainee journalists. After a preliminary set of experiments where keywords were manually provided we installed cookies on the client (a client for each trainee) in order to automatically acquire user preferences during her/his navigation.

3.2. Document Summarization

This is the main components of the platform. In this preliminary stage to let us end the overall development cycle we are using LAKE (Linguistic Analysis based Keyphrase Extractor) a tool that worked as a single-document and multi-document summarization obtaining encouraging experimental results.

LAKE is a keyphrase extraction system based on a supervised learning approach that applies linguistic processing on documents. In the past the system used Naïve Bayes algorithm [6] as the learning method and *TF* × *IDF* term weighting with the *position* of a phrase as features. For this year competition we have used a Support Vector Machine (SVM) as a learner [7]. Unlike other keyphrase exctraction systems LAKE chooses the candidate phrases using linguistic knowledge. The candidate phrases generated by LAKE are sequences of Part of Speech (PoS) containing Multiword Expressions (ME) and Named Entities (NE). Extraction is driven by a set of "patterns" which are stored in a pattern database; once there, the main work is done by the learner

device (i.e., the SVM). The linguistic database makes LAKE unique in its category.
LAKE is based on three main components: the Linguistic Pre-Processor, the candidate Phrase Extractor and the Candidate Phrase Scorer. In the following sections there is a brief description of the system. For a more detailed description the reader is referred to previous publications.

Linguistic Pre-Processor

Every document is analyzed by the Linguistic Pre-Processor following three consecutive steps: Part of Speech (PoS) analysis, Multiword Expressions (ME) recognition and Named Entities (NE) recognition.

Candidate Phrase Extractor

Syntactic patterns have a twofold objective:

- focusing on uni-grams and bi-grams (for instance Named Entity, noun, and sequences of adjective+noun, etc.) to describe a precise and well defined entity;
- considering longer sequences of PoS, often containing verbal forms (for instance noun+verb+adjective+noun) to describe concise events/situations.

Once all the uni-grams, bi-grams, tri-grams, and four-grams are extracted from the linguistic pre-processor, they are filtered with the patterns defined above. The result of this process is a set of keyphrases that may represent the current document.

Candidate Phrases Scorer

Candidates keyphrases identified in the previous step are scored in order to select the most appropriate phrases as representative of the original text. The score is based on a combination of *TF ×IDF* and *first occurrence*, i.e. the distance of the candidate phrase from the beginning of the document in which it appears.
However, since candidate phrases do not appear frequently enough in the collection, it has been decided to estimate the values of the TF×IDF using the head of the candidate phrase, instead of the whole phrase. According to the principle of headedness (Arampatzis et al., 2000), every phrase has a single word as head. The head is the main verb in the case of verb phrases, and a noun (last noun before any post-modifiers) in noun phrases. As learning algorithm, it has been used an SVM provided by the WEKA package (Witten and Frank, 1999).
The classifier was trained on a corpus with the available keyphrases. From the document collection we extracted all nouns and verbs. Each of them was marked as a positive example of a relevant keyphrase for a certain document if it was present in the assessor's judgment of that document; otherwise it was marked as a negative example. Then the two features (i.e. TF×IDF and first occurrence) were calculated for each word. The classifier was trained using this material and a ranked word list was returned. The system automatically looks in the candidate phrases for those phrases containing these words. The top candidate phrases matching the word output of the classifier are kept. The model obtained is reused in the subsequent steps. When a new document or corpus is ready we use the pre-processor module to prepare the candidate phrases. The model we got in the training is then used to score the phrases obtained. In this case the pre-processing part is the same. Using the model thus obtained, we extracted nouns and verbs from documents, and then we kept the candidate phrases containing them.

The Lake system uses two parameters for controlling its work: one is the maximum number of words allowed in a keyphrase and the second is the maximum number of keyphrases to be extracted from a document.
These parameters are used for creating from a set of documents a brief, well-organized, fluent summary addressing a need for information expressed in a specific topic, at a level of granularity specified in the user profile (DUC-2005 definition).
Lake is required to select the most representative keyphrases that have the highest *relevance* and *coverage* scores of a set of document, given the topic and profile.
The *relevance* of a keyphrase list kl_j with respect to a cluster C_j is computed considering the frequency of the keyphrases composing the list. The intuition is that keyphrases with higher frequency bring the more relevant information in the cluster:

$$relevance(kl_j) = \frac{\sum_{w=1}^{n} freq(w, kl_j)}{freq(w, C_j)}$$

where $freq(w,kl_j)$ is the count of a word w in a certain document and $freq(w,C_j)$ is the count of w in all the documents in the cluster C_j.
The *Coverage* of a keyphrase list kl_j is an indication of the amount of information that the keyphrase list contains with respect to the total amount of information included in a cluster of documents:

$$\mathrm{cov}erage(kl_j, C) = \frac{length(kl_j)}{\max length(kl_j, C)}$$

where $length(kl_j)$ is the number of keyphrases extracted from document j and $maxlength(kl_j,C)$ is the length of the longest keyphrase list extracted from a document belonging to cluster C_j. The intuition underlynes that the longer the keyphrase list, the more is its coverage for a certain cluster.
Relevance and *Coverage* are combined according to the following formula:

$$rep(kl_j) = relevance(kl_j, C) \times \mathrm{cov}erage(kl_j C)$$

which gives an overall measure of the representativeness of a keyphrase list for a certain document with respect to a cluster.
Finally, the keyphrase list which maximize the two parameters is selected as the most representative of the cluster and each keyphrase is substituted with the whole sentence in which it appears.

3.3. Experiments

Document Understanding Conferences (DUC) is a series of text summarization campaigns presenting text summarization competitions results. LAKE participated at DUC since 2004 while obtaining encouraging results every time. For brevity, we only report the linguistic quality of LAKE's summaries

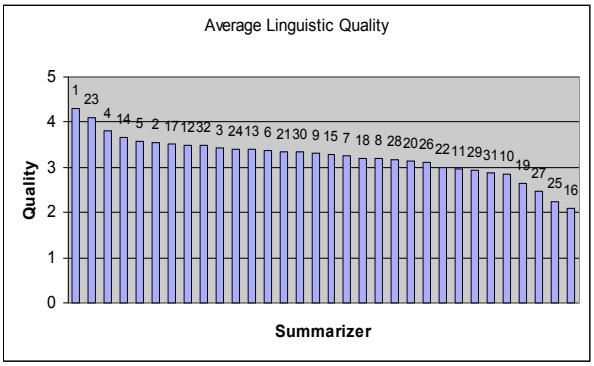


Figure. 1: Average linguistic Quality.

Linguistic quality assesses how readable and fluent the summaries are, without comparing them with a model summary. Five Quality Questions were used, and all questions were assessed on a five-point scale from"1" (very poor) to "5" (very good). Being a linguistically motivated summarizer, LAKE is expected to perform well at the manual evaluation with respect to language quality and responsiveness. Regarding language quality, as can be expected, LAKE scored relatively high – it was ranked 6th out of the 30 systems for average language quality (see figure 1), with an average value of 3.502 compared to 3.41 – the overall average - and 4.23 which was the highest score of the baseline system (no 1) and very close to the second baseline system (no 2) that scored 3.56. However, we should note that most of the systems scored between 3.0 and 4.0 for linguistic quality, so the differences were relatively small. Compared to 2006, Lake got a little lower score (3.5 compared to 3.7), and was ranked relatively lower (3rd in 2006).

Figure 2

3.4. Web Interface

Figure 2 shows the Web interface at very beginning stage. The output of the system will be more elaborated like Newsblaster's one, reporting the sources of the summaries and some more keywords of the final summary displayed on the PDA or mobile of the user.

4. Conclusions

We presented an initial idea for a prototype of an online news agency platform. The prototype will be the product of a joint research project between the Department of Communication Sciences and the Journalism School at the University of Salerno.

The prototype aims at assisting and supporting trainee journalists in their daily search of information avoiding/reducing the Information Overload problem. In fact, even if there are many popular news services available online (e.g. Google News) they all suffer from one major drawback - that the user has to open/read a lot of documents before she/he can find the information satisfying her/his information need. Moreover, current automatically generated snippets/summaries of document only represent the initial parts of them, with the risk of losing a lot of relevant information.

The methodology proposed allows the acquisition of user (journalist) preferences as a background process and gathering the information on an ongoing or daily baseis, based on these preferences.

Information gathered and summarized using machine learning and Natural Language Processing techniques will delivered using different channels: Web, email, PDAs, so users will be able to access it literally every time and everywhere

5. Acknowledgements

We thank Biagio Agnes, Lillo D'Agostino and Pino Blasi respectively Director, President of Scientific Board and Coordinator of the Journalism School for their collaboration and helpful advices in this start up process. We also thank Mario Monteleone that let us to play with DELAS-DELAF and DELAC dictionaries and other resources.

6. References

Kelley, C.M., DeMoulin, G., The Web cannibalizes media. Technical Report, The Forrester Group. May 2002

Radev, D., Otterbacher, J., Winkel, A., and Blair-Goldensohn, S., NewsInEssence: summarizing online news topics, Commun. ACM, Vol. 48, 2005, ACM, New York

Allen, L., Charron, C., and Roshan, S. Re-engineering the news business. Technical Report, The Forrester Group, June 2002.

McKeown, K. R., Barzilay, R., Evans, E., Vasileios, H., Klavans, J. L., Nenkova, A., Sable, C., Schiffman, B. Sigelman, S., Tracking and summarizing news on a daily basis with Columbia's Newsblaster, Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., 2002, San Francisco, CA.

Pant, G., Srinivasan, P., Menczer, F., Crawling the Web. In M. Levene and A. Poulovassilis, eds.: Web Dynamics, Springer, 2004

Mitchell, T. 1997. Machine Learning. McGraw-Hill.

Cristianini, N., Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

Boosting the Recall of Descriptive Phrases in Web Snippets

Alejandro Figueroa[1]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz - DFKI, Saarbrücken, Germany
figueroa@dfki.de

Abstract

WebQA is a Web Question Answering System[1] which is aimed at discovering answers to natural language questions on the web. One of its major components is the module that answers definition questions. This module searches for answers by means of a query rewriting strategy, which considerably boosts the recall of descriptive utterances. This study compares three different search strategies and deals at greater length with the challenges posed by the assessment of web-based definition Question Answering Systems.

Index Terms: Question Answering, Definition Questions, Web Mining, Search

1. Introduction

WebQA is built on top of commercial search engines like MSN Search and Yahoo. WebQA is part of sustained efforts to implement a system which extracts answers to factoid [3] and definition [4], as well as list questions [5] **exclusively** from the brief descriptions returned by these search engines, called *web snippets*.

The reason to use web snippets as an answer source is fourfold: (a) they are computed at high speed by current commercials search engines, and therefore provide a quick and contextualised response, (b) to take advantage of the current power of indexing of vanguard search engines, (c) to the user, web snippets are the first view of the response, thus highlighting answers would make them more informative, and (d) to avoid, or at least lessen, the retrieval and costly processing of a wealth of documents. In particular, web snippets have proven to be promising for answering difficult queries like definitions questions (such as "*Who is Allen Iverson?*" or "*What are fractals?*"). This sort of query is particularly important, because 27% of the questions of real user logs are a request for a definition. In order to satisfactorily answer definition questions, Question Answering Systems (QAS) must take answers from several documents and afterwards, discriminate senses, merge answers, remove redundancy, and eventually generate a final output for the user. This study focus its attention on the first step: the search or retrieval of definition answers.

2. Related Work

QAS are usually assessed in the context of the Question Answering track of the Text REtrieval Conference (TREC). During the last years, the thoroughness and difficulty of this assessment has been systematically increased by making allowances for more challenging queries, such as definition questions.

In TREC, the target collection is the AQUAINT corpus. QAS make use of several external resources of definition information in order to successfully discover the right answers in this corpus. QAS then identify descriptive phrases by projecting the obtained nuggets into the corpus. In this way, they also filter out some misleading and spurious nuggets taken from these external sources. In the jargon of definition questions, a nugget is a piece of relevant or factual information about the particular topic of the question (a. k. a. the *definiendum*).

For instance, [6] introduced a method for answering definition questions that was assisted by a wrapper for the online Merriam Webster dictionary, which retrieved about 1.5 nuggets per question. These nuggets were used as query expansion terms for retrieving promising documents from the collection afterwards. Additionally, they automatically constructed offline an extremely large relational database containing nuggets about every entity mentioned in the AQUAINT corpus. These nuggets were accordingly taken from every article within it, and therefore, answering definition questions consisted of a simple lookup for the *definiendum*. Since nuggets often seem odd and out of place without their context, [6] expanded them to surround one hundred (non-white-space) characters in order to enhance readability.

Another example, is the strategy proposed by [2], which took advantage of external resources like WordNet glossaries, online specific resources (e.g., Wikipedia) and web snippets for learning frequencies and correlation of words, especially with the *definiendum*. One of their findings was that definitional web-sites greatly improve the performance, leading to few unanswered questions: Wikipedia covered 34 out of the 50 TREC–2003 definition queries and biography.com 23 out of 30 questions regarding people, all together provided answers to 42 queries. They additionally found that web snippets, though they yielded relevant information about the *definiendum*, were not likely to supply descriptive utterances, bringing about only a marginal improvement.

Another method that takes advantage of web snippets was presented in [1]. This method uses a centroid vector that considers word dependencies learnt from the 350 most frequent stemmed co-occurring terms taken from the best **500 snippets** retrieved by Google. These snippets were fetched by expanding the original query by means of a set of five highly co-occurring terms. These terms co-occur with the *definiendum* in sentences obtained by submitting the original query plus some task specific clues, e.g.,"*biography*". As a result, this query expansion technique improved the $\mathcal{F}(5)$ score of their system from 0.511 to 0.531. They concluded that the use of multiple search engines would help to fetch more sentences containing the *definiendum*.

The module of WebQA that answers definition questions was described firstly in [4]. Contrary to QAS in TREC, WebQA searches for definition sentences only on the web, in particular in web snippets. The advantage of descriptive phrases extracted from web snippets is that they provide an adequate unit of contextual information [4], being comparable in size with the enhanced nuggets obtained by [6]. For the purpose of markedly

[1]http://experimental-quetal.dfki.de/

increasing the recall of definition sentences within web snippets, WebQA biases the search engine in favour of some lexico-syntactic structures that often convey definitions by means of a purpose-built query rewriting strategy. As a result, WebQA finished with $\mathcal{F}(5)$ score of 0.53 for the TREC 2003 data-set, which is "competitive" with the best systems, which achieve a value between 0.5 and 0.56 [1, 6, 7, 8].

However, a key point for correctly interpreting these results is the completeness of the assessor's list. It is known that systems in TREC were able to find relevant nuggets, which were not included in this list (cf. [6] for details). In the case of web-based systems like WebQA, this vital fact is more likely to happen, because they discover many additional nuggets seen as relevant by the user, but excluded from the assessor's list. This exclusion actually brings about a decrease in the $\mathcal{F}(5)$ score, because these extra nuggets enlarge the response without increasing precision. Moreover, WebQA must determine exclusively from the context whether or not a certain nugget conveys definition information; this means it lacks a target corpus that could act like a filter for some spurious and misleading answers. This kind of evaluation is, nonetheless, the unique current way to have an objective reference to the performance of several systems.

This study shows two search strategies that boost the recall of sentences that convey definitions, and consequently, they better the performance of the definition module of WebQA. These strategies: (a) take into consideration the prior knowledge provided by Google n-grams while rewriting the query, and (b) take up the suggestion of [1] by adding an extra search engine[2]. Another thing minutely examined in this work, is the impact of the assessor's list on the evaluation of web-based definition QAS.

3. Mining the Web for Definitions

Currently, the definition component of WebQA makes use of ten purpose-built search queries, which are based on some local lexico-syntactic constructions that often convey definitions (cf. [4] for details). These ten search queries help WebQA to substantially increase the recall of descriptive utterances within web snippets (δ stands for the *definiendum*):

q_1="δ"
q_2="δ is a " $\vee$ "δ was a " $\vee$ "δ were a " $\vee$ "δ are a "
q_3="δ is an " $\vee$ "δ was an " $\vee$ "δ were an " $\vee$ "δ are an "
q_4="δ is the " $\vee$ "δ was the " $\vee$ "δ were the " $\vee$ "δ are the "
q_5="δ has been a " $\vee$ "δ has been an " $\vee$ "δ has been the " $\vee$ "δ have been a " $\vee$ "δ have been an " $\vee$ "δ have been the "
q_6="δ, a " $\vee$ "δ, an " $\vee$ "δ, the " $\vee$ "δ, or "
q_7=("δ" $\vee$ "δ also " $\vee$ "δ is " $\vee$ "δ are ") $\wedge$ (called $\vee$ nicknamed $\vee$ "known as")
q_8="δ became " $\vee$ "δ become " $\vee$ "δ becomes "
q_9="δ which " $\vee$ "δ that " $\vee$ "δ who "
q_{10}="δ was born " $\vee$ "(δ)"

The drawback to this query rewriting strategy is that these search queries are statically built, causing that two promising lexico-syntactic clauses could be submitted in the same query, lessening the retrieval of descriptive phrases. A good illustrative example is δ="*Allen Iverson*" and q_2. In this case, "*Allen Iverson is a*" and "*Allen Iverson was a*" are two clauses likely to yield definitions. Consequently, they should be separately submitted in order to avoid weakening the recall. Further, clauses such as "*Allen Iverson were a*" and "*Allen Iverson are a*" only bring about misleading sentences:

[2] http://www.yahoo.com/

*- Cheers for visiting **Allen Iverson were a** slap in the face to the Clippers.*
*- Carmelo Anthony and **Allen Iverson were a** combined 1 for 10 in the third, when Denver committed 9 turnovers.*

Analogously, a set of unpromising lexico-syntactic patterns can be set in the same query and hence, bring about an unproductive retrieval, diminishing the number of descriptive utterances. Nevertheless, these patterns observe a local lexico-syntactic dependency with the *definiendum*, specifically, they are unlikely to contain additional words in between. This is an important fact, because off-line n-grams counts supplied by Google can be used to transform this static query construction into a more dynamic one. In our working example, an excerpt of Google 4-grams counts is as follows:

Allen Iverson is a 209
Allen Iverson is an 68
Allen Iverson is the 425
Allen Iverson was a 57
Allen Iverson was the 101

The first beneficial aspect of Google n-grams is that, in some cases, the grammatical number can be inferred. In particular, in the case of "*Allen Iverson*", singular lexico-syntactic clues are most promising. However, it is not always possible to draw a clear distinction. A good example is "*fractals*":

fractals are a 176 (e.g. "Fractals are a powerful tool...")
fractals are an 86 (e.g. "Fractals are an exquisite...")
fractals are the 215 (e.g. "Fractals are the place...")
fractals is a 124 (e.g. "Fractals is a new branch of...)
fractals is the 148 (e.g. "Fractals is an innovative...)

Then, a strategy was designed (S-I), which selects a grammatical number whenever more than three keywords corresponding to one grammatical number exist, and zero to the another. The second favourable aspect is that the frequencies give hints about the hierarchy within the lexico-syntactic patterns. S-I takes advantage of this hierarchy for configuring the ten queries. First, the search queries q_7 and q_{10} are merged into one query q_7'. This query is composed of the following clauses:

"δ also called ", "δ also nicknamed", "δ also known", "δ is called", "δ stands for", "δ is known", "δ are called", "δ are nicknamed", "δ are known", "δ was born", "δ was founded", "δ was founded", "δ is nicknamed"

Accordingly, q_7' consists merely of the clauses that can be found in Google n-grams. If any clause cannot be found, q_7' is set to $\emptyset$. In any case, q_{10}' remains as $\emptyset$. It is worth pointing out that, the term "*stands for*" replaces the parentheses in q_{10}. Second, $q_5' = q_5$, $q_6' = q_6$ and $q_8' = q_8$ as well as $q_9' = q_9$. Additionally, the q_1' is set to $\emptyset$. Third, the clauses included in the queries q_2 and q_3, as well as q_4, are dynamically sorted across the available queries, as highlighted in table 1.

Table 1: Dynamic queries (grammatical number known).

$q_7' = \emptyset$	$q_7' \neq \emptyset$
q_1':"δ R_1" q_2':"δ R_2" q_3':"δ R_3" q_4':"δ R_4" q_5':"δ R_5" q_7':"δR_6"	q_1':"δ R_1" q_2':"δ R_2" q_3':"δ R_3" q_4':"δ R_4" q_5':"δ R_5" $\vee$ "δ R_6"

where R_1 and R_6 correspond to the highest and lowest frequent lexico-syntactic patterns according to Google frequency counts. In the case that the grammatical number cannot be distinguished, the queries are as follows:

$q_1^{'}$:"δ is a" $\vee$ "δ were an" $\vee$ "δ was the"
$q_2^{'}$:"δ was a" $\vee$ "δ are an"
$q_3^{'}$:"δ are a" $\vee$ "δ was an" $\vee$ "δ were the"
$q_4^{'}$:"δ were a" $\vee$ "δ is an"
$q_{10}^{'}$:"δ is the" $\vee$ "δ are the"

In the case $q_{10}^{'} = \emptyset$, the following queries are reformulated:

$q_1^{'}$:"δ is a" $\vee$ "δ were an"
$q_3^{'}$:"δ are a" $\vee$ "δ was an"
$q_7^{'}$:"δ was the" $\vee$ "δ were the"

Every query is eventually surrounded with the feature "*inbody:*" in order to avoid matching a clause with the title of a web page.

4. Experiments

S-I and the static query rewriting strategy (S-O) were assessed by means of the defintion question set supplied by TREC 2003. Following the suggestion of [1], S-I was additionally tested together with the use of an extra search engine (S-II). Figure 1 compares the $\mathcal{F}(5)$ score per question for the three strategies.

WebQA with the static query rewriting finished with an average $\mathcal{F}(5)$ score of 0.5472, while the dynamic query rewriting improved the average value to 0.5792, and this rewriting along with an additional search engine, improved to 0.5842. Here, the first aspect to point out is the increase to 0.5472 with respect to the $\mathcal{F}(5)$ value (0.53) reported in [4]. We interpret this increase as a change in the fetched content from the web. As well as that, it is worth remarking that S-I obtained an improvement without increasing the number of submitted queries, whereas the marginal increase achieved by S-II with respect to S-I, is at the expense of sending ten extra queries to the additional search engine. It is also worth noting that each submission is done as described in [4], and hence, S-O and S-I fetch a maximun of 300 snippets, while S-II 600. These sets of snippets are comparable in size with the 500 snippets retrieved by [1]. Overall, the $\mathcal{F}(5)$ values, achieved by WebQA with our rewriting strategies incorporated, are "competitive" with the best definition QAS. These systems obtain a value between 0.5 and 0.56 [1, 6, 7, 8].

S-O and S-I scored zero for four different *definiendums*, despite the "okay" nuggets found by both systems. In fact, if a system does not discover any nugget assessed as "vital", it finishes with a $\mathcal{F}(5)$ value equal to zero. For instance, S-II scored zero for three questions; in particular, for the following output concerning "*Albert Ghiorso*":

- said ***Albert Ghiorso, a*** *veteran Berkeley researcher, who holds the Guinness world record.*
- ***Albert Ghiorso is a*** *nuclear scientist at Lawrence Berkeley National Laboratory in Berkeley, Calif.*
- That's what Berkeley Lab's ***Albert Ghiorso, a*** *man who has participated in the discovery of more atomic elements than any living person, told the students and teachers who packed.*
- ***Albert Ghiorso is an*** *American nuclear scientist who helped discover several elements on the periodic table.*

The "okay" nugget is underlined that matches the assessors' list provided by TREC 2003:

vital designed and built cyclotron accelator
okay nuclear physicists/experimentalist
vital co-creator of 12 artificial elements
vital co-discovered element 106

Like [6] also noticed, "okay" nuggets, like *nuclear physicists/experimentalist* can be easily interpreted as "vital". For example, if one considers abstracts supplied by Wikipedia as a third-party judgement, at the time of writing, one finds:

- ***Albert Ghiorso*** *(b. 15 July 1915) is an American nuclear scientist who helped discover numerous chemical elements on the periodic table.*

Further, some relevant nuggets, including *veteran Berkeley researcher*, are unconsidered, enlarging the response, and thus decreasing the $\mathcal{F}(5)$ score. We hypothesise that a nugget can be seen as "vital" or "okay" according to how often its **type** (birthplace, birthdate, occupation, outstanding achievement) occurs across abstracts and/or bodys of online encyclopedias, such as Encarta or Wikpedia. We deem that this sort of **type-oriented** evaluation would be more appropriate to web-based definition QAS. Only in one *definiendum* were the three strategies unable to discover any nugget in the assessor' list: "*Abu Sayaf*". The reason is uncovered when the following frequencies on Google n-grams are checked:

Abu Sayyaf 96204
Abu Sayyafs 89
Abu Sayaf 1156
Abu Sayaff 3205

In this case, the spelling of the *definiendum* in the query is unlikely to occur in the web, causing an $\mathcal{F}(5)$ equals to zero. Conversely, when WebQA processes "*Abu Sayyaf*", the scores obtained by each method are: 0.844 (S-O), 0.8794 (S-I) and 0.8959 (S-II). Accordingly, the new average $\mathcal{F}(5)$ values are: 0.564 (S-O), 0.59679 (S-I) and 0.602 (S-II).

Another complicated problem is that the list of the assessor is aimed predominantly at one possible sense of the *definiendum*. Therefore, discovered descriptive utterances concerning additional senses, similar to the uncosidered nuggets, bring about a decrease in the $\mathcal{F}(5)$ value. To illustrate this, a descriptive sentence found by S-II regarding "*Nostradamus*":

- ***Nostradamus is a*** *neural network-based, short-term demand and price forecasting system, utilized by electric and gas utilities, system operators and power pools, electric...*

Indeed, it is highly frequent to find ambiguous terms. For example, Wikipedia contains more than 19000 different disambiguation pages. In this case, the list of the assessor only accounts for the reference to the French astrologer/prophet. When sentences concerning other senses are manually removed, the $\mathcal{F}(5)$ values for this concept increase as follows: from 0.5871 to 0.5936 (S-O), from 0.9028 to 0.9182 (S-I) and from 0.8977 to 0.9167 (S-II). Obviously, a more noticeable difference is due to *definiendums* with more senses such as "*Absalom*".

Another difficulty that QAS encounter when they extract definition phrases from the web, is that opinions are also given like definitions. A good example is given by the *definiendum* "*Charles Lindberg*":

- ***Charles Lindberg was a*** *true American hero.*

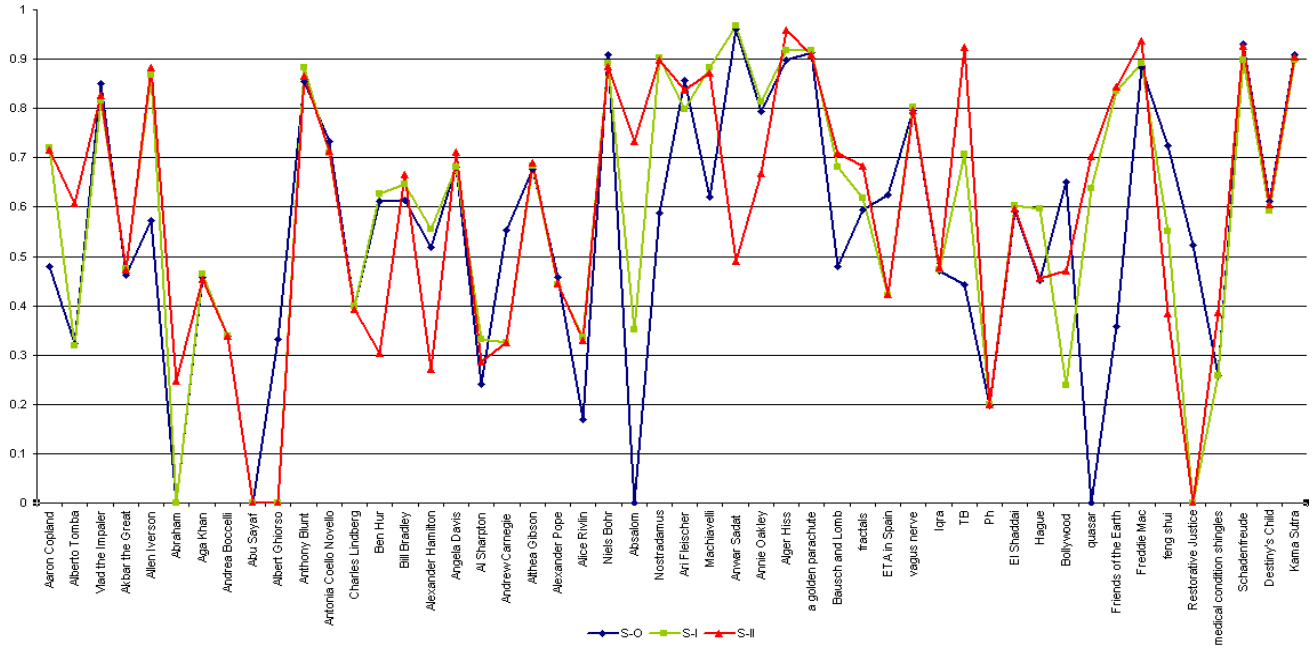


Figure 1: Comparison between F(5) scores obtained by each strategy for each *definiendum* in the TREC 2003 question-set.

This sentence does not syntactically differ from the definition "*Charles Lindberg was a famous American pilot.*" We envisage that a large-scale redundancy and the use of opinion mining techniques would help to discriminate opinions from facts.

Our ongoing research is aimed at incorporating more linguistic information into the query rewriting strategy. Specifically, promising verb phrases can be interpreted as definition lexico-syntactic patterns, and therefore, appended to the "*definiendum*". These verb phrases can be determined by means of retrieved descriptive sentences, a chunker, and the corresponding recalls can be estimated by inspecting the frequency of these new clauses on Google n-grams. This sort of strategy would help to fetch more and diverse descriptive information about the *definiendum*.

5. Conclusions

This study compares three query rewriting strategies that are aimed at boosting the recall of descriptive sentences in web snippets and consequently, at improving the performance of definition QAS. One interesting finding is that Google n-grams can be used particularly for optimising the retrieval of defintions in web snippets, and accordingly, they can also assist QAS in fetching more promising full-documents.

This paper additionally discusses the major challenges posed by web-based definition QAS, and it sketches accordingly some directions that could help to face these challenges.

6. Acknowledgements

The work presented here was partially supported by a research grant from the German Federal Ministry of Education, Science, Research and Technology (BMBF) to the DFKI project HyLaP (FKZ: 01 IW F02) and the EC-funded project QALL-ME.

7. References

[1] Chen, Y., Zhon M. and Wang, S. "Reranking Answers for Definitional QA Using Language Modeling", Proceedings of the Coling/ACL-2006, pp. 1081–1088.

[2] Cui, T.S.C.H., Kan M.Y. and Xiao J. "A comparative study on sentence retrieval for definitional question answering", SIGIR Workshop on Information Retrieval for Question Answering (IR4QA), 2004.

[3] Figueroa, A. and Neumann, G. "Language Independent Answer Prediction from the Web", Proceedings of the FinTAL 5th International Conference on Natural Language Processing, August 23-25 in Turku, Finland, 2006.

[4] Figueroa, A. and Neumann, G. "A Multilingual Framework for Searching Definitions on Web Snippets", KI 2007: Advances in Artificial Intelligence', LNCS, Volume 4667/2007, p. 144-159.

[5] Figueroa, A. and Neumann, G. "Mining Web Snippets to Answer List Questions", In AI07: the 2nd International Workshop on Integrating AI and Data Mining, 2nd-6th December 2007, Gold Coast, Queensland, Australia.

[6] Hildebrandt W., Katz B. and Lin J. "Answering Definition Questions Using Multiple Knowledge Sources", Proceedings of HLT-NAACL 2004, pp. 49–56.

[7] Voorhees, E., M. "Evaluating Answers to Definition Questions", Proceedings of HLT-NAACL 2003, pp. 109–111, 2003.

[8] Xu, J., Licuanan, A. and Weischedel, R. "TREC2003 QA at BBN: Answering definitional questions", Proceedings of the Twelfth Text REtrieval Conference, 2003.

COLDIC a generic tool for the creation, maintenance and management of Lexical Resources

Núria Bel, Sergio Espeja, Montserrat Marimon, Marta Villegas

Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona, Spain
nuria.bel, sergio.espeja, montserrat.marimon, marta.villegas@upf.edu

Abstract

Although most of the Language Technologies applications need to develop and maintain large lexica, there has been a lack of generic tools for its creation, maintenance, and management which are independent of particular applications, and are well equipped for supporting lexicographic work. The most important obstacle to such generic tools was the proliferation of lexical models and formats: each application defined what information was required and how it should be declared.

The definition of standards for lexical encoding, as the one being developed in the Lexical Markup Framework (LMF, supported by the ISO and the e-content project LIRICS) will open the room for generic tools which are feasible and useful. Lexical management platforms can be tuned to the standard model and format, in order to create, merge or to maintain resources which can be used to feed different tools.

Besides, the existence of such standards can also enable the integration of high level supporting lexicographical tools, such as automatic acquisition, creation of analytical tools for corpus data assessment, etc.

We present in this paper a first approach for such a generic tool crucially based in the LMF model. COLDIC is a lexicographical management platform intended to be a generic tool independent of a particular technology and/or application.

1. Introduction

Computational lexica are repositories of information about words in particular languages that are traditionally developed for specific applications and tools. The quality of these resources is critical for the performance of the tool. For instance, Briscoe and Carroll (1993) observed that half of parse failures on unseen test data were caused by inaccurate lexical information, and Baldwin et al. (2004) identified that in parsing 20,000 strings from British National Corpus (BCN) a 40% of grammar failures were due to missing lexical entries, with a grammar dictionary of about 10,500 lexical entries.

Lexica are normally created and maintained by lexicographers that get little support. The information found in traditional dictionaries and terminological glossaries is not directly reusable in the encoding of word formal properties, like part of speech, gender, inflection paradigm, syntactic valency, semantic information etc. Besides, lexicon has to be tuned to specific domains, and the lexicographer must check whether a particular word is in the lexicon, and also that the encoding covers the use of the word in this new domain. The tuning of a lexicon to a new domain can involve the encoding of 4,000 to 20,000 entries.

Despite of the amount of work involved in manually crafting a lexicon, and the importance of the task, there has been a complete lack of generic tools that focus on supporting the lexicographer. For any particular project or application to develop a sophisticated lexicographical tool is usually out of its scope (and budget).

The main objective of COLDIC is to offer lexicographers working on lexica for Language Technology a tool that is particularly suited for lexical development tasks and that can be tuned and adapted to any application and model. The LMF model (Francopoulo et al. 2006) has been the basis for defining a database internal structure that guarantees the coverage of a large number of applications and languages. COLDIC also follows LMF to guarantee the interoperability for data exchange and access to other resources via web services[1].

Our tool has taken into account some basic factors to adequate the design and facilities to lexicographers, who can find an expressive enough framework to handle complex operations.

1) No technical background on databases should be required. The information to be encoded is declared in a DTD like file, which the system reads to build the database.
2) No previous knowledge on interfaces should be required. The system offers an LMF model based interface that guides the lexicographer to easily build queries and forms for introducing new data.
3) Information to be displayed in forms should be adjustable to different needs, and the forms can be fixed so to show or hidden particular features, as well as to use defaults for encoding of predefined groups.
4) Complexity of the model should be compensated by graphical views of the contents of the database.

[1] A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

2. Lexical Markup Framework: LMF

The Lexical Markup Framework was a joint ISO TC37/SC34 and LIRICS (EU e-content project) initiative to build a standardized abstract framework for the construction of computational lexica. The aim of LMF was to create a metamodel inclusive of the specificities of main lexicographical practices. This metamodel provides the user with a representation of lexical objects, the structure of the information underlying its description and its use. Used as a standard for the creation and use of electronic lexical resources, LMF is the basis for a real exchange of data between and among these resources, and for the use of these resources in remote, distributed applications based on web services. The ultimate goal of LMF is to facilitate true content interoperability across all aspects of electronic lexical resources.

LMF is defined in the LMF core package, together with two extensions: a Machine Readable Dictionary one and a NLP lexical resources one. While the core package describes the basic hierarchy of information of a lexical entry, the extensions (that use of LMF core components) address specific requirements for particular functionalities. LMF has been devised to handle only the structure of the lexical entry. Linguistic features that describe lexical items are defined in the ISO 12620 Data Category Registry[1], further guaranteeing standardization and interoperability.

3. COLDIC

COLDIC is a lexicographic platform for the creation and management of NLP lexica. One of its main features is that it provides an interface not only for human users, but also for consultation and information delivery asked by remote systems via web services. This interface is automatically generated at the same time than the database and needs no previous knowledge of web services by the user.

The other main features of the platform are:

- Reading and parsing of a LMF compliant lexical model DTD and generation of a relational database that can be managed with a core web based interface.
- Offering of a query builder tool that supports the user in the creation of content based queries, with advanced features as macro like queries with parametrized arguments.
- Automatic generation of a graphical view of the lexical model that is used as a support in the query and form builder tools.
- Support in the creation of encoding forms to assist lexicographers in the introduction of new data, search and validation of encoding features via Data Category Registration remote look up.
- Automatic creation of a number of standard and lexically oriented web services by exploiting the interoperability capabilities of the LMF standard and implementing the LMF for lexica API.

[1] The Data Category Registry can be accessed at http://syntax.inist.fr/

The creation of the database and its maintenance demand no specific training in databases as the LMF schema implemented in a XML Document Type Definition (DTD) is taken by the system to configure the platform. Special building functions support the user in the most common tasks of the lexicographic work, and the tool can be supported by analytical information about entries got from a pre-defined corpus.

Because user requirements might vary according to different needs, COLDIC comes equipped with a set of basic functionalities:

- Query and browse facilities by means of user build forms,
- import, export and migration of data,
- easy encoding of new data by means of user build forms, and direct access to different sources of information, such as the Data Categories repository,
- test and validation of both the data and the model, and
- lexicographic tools such as type definition, class extraction, evaluation, validation and statistical facilities.

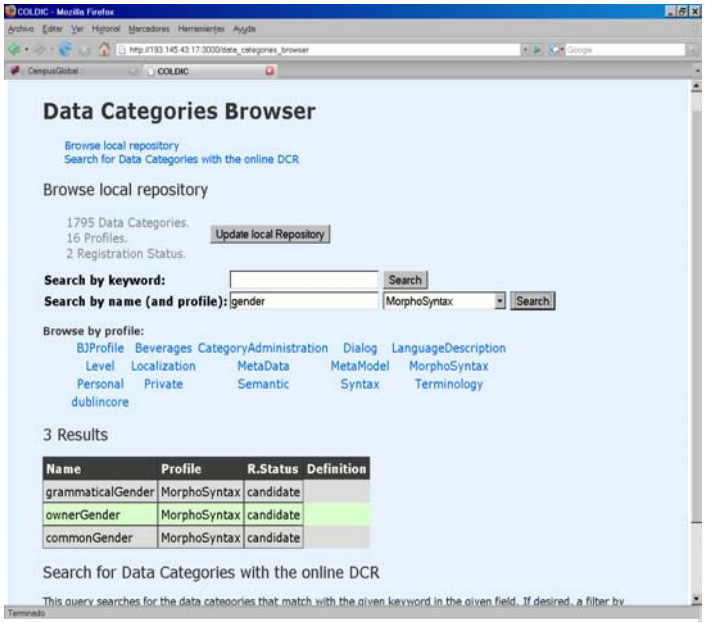


Figure 1. *Window for searching at the DataCat repository. Lexicographer is supported with direct access and search/browse facilities for finding the standard DataCat for describing entries*

COLDIC consists of the following modules:

- The application generation module, which handles the creation of a relational database in terms of a LMF compliant DTD.
- The administration module, which handles profiles, imports and exports data.
- The core interface module, a graphical interface with facilities for building and executing queries and forms.
- The web services module, which offers a LMF compliant API.

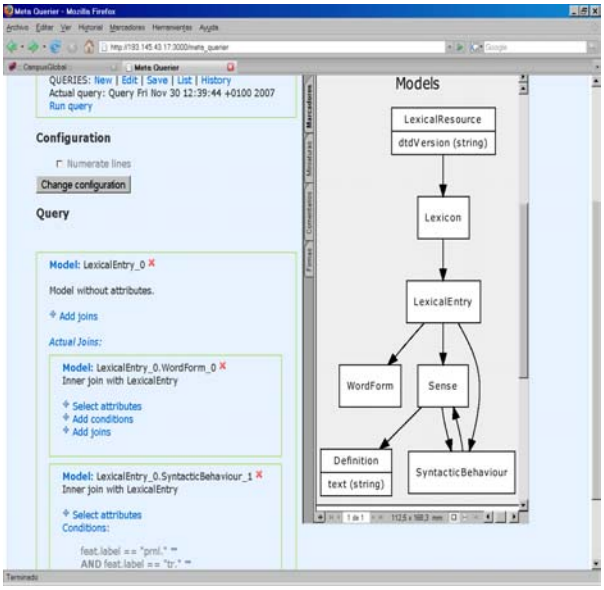


Figure 2: *Query Builder, main window. The instantiation of the model according to user data is ploted to support the construction of the quey.*

4. Technical features and availability

COLDIC has been developed in Ruby (htpp://www.ruby-lang.org) and uses the Ruby on Rails framework (http://www.rubyonrails.org). Ruby is a dynamic, open source programming language with a focus on simplicity and productivity. Ruby on Rails is an open source web framework that favors convention over configuration in order to get less and more understandable code.

The core of COLDIC, the automatic building of the platform out of a DTD, is based in what we have called MetaRails plugins. MetaRails plugins are open source Ruby on Rails plugins that handles the following modules of the platform (Figure 1): Automatic database generation, web services generation, the *querier* and the forms editor. The open source project MetaRails, created for COLDIC development, can be found at http://meta-rails.rubyforge.org and is distributed under the GPL License. COLDIC is also released as a open source project that can be found at http://COLDIC.sourceforge.net distributed under the GPL license.

5. References

[1] Baldwin, T., E. M. Bender, D. Flickinger, et al. (2004). Road-testing the English Resource Grammar over the British National Corpus. In Proceedings of the FourthInternational Conference on Language Resources and Evaluation (LREC 2004), Lisbon.

[2] Briscoe, T. and J. Carroll. 1997. 'Automatic extraction of subcategorization from corpora'. In Proceedings of the Fifth Conference on Applied Natural Processing, Washington.

[3] Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 Lexical Markup Framework (LMF), Proceedings of LREC Genoa.

[4] Kemps-Snijders et al. (2007). Data Category Registry API, Lirics Delivery 5.1. http://lirics.loria.fr.

[5] Kemps-Snijders and J. Nioche. (2007). API for Lexica, Lirics Delivery 5.1. http://lirics.loria.fr.

[6] Kemps-Snijders (2007). Data Category Usage Platform, Lirics Delivery 5.4. http://lirics.loria.fr.

[7] Lenci A., Bel N., Busa F., Calzolari N., Gola E., Monachini M., Ogonowski A., Peters I., Peters W., Ruimy N., Villegas M., Zampolli A. 2000. SIMPLE: A General Framework for the Development of Multilingual Lexicons, International Journal of Lexicography 13(4). Oxford University.

[8] Villegas, M.; Bel, N. (2002). "From DTDs to relational dBs. An automatic generation of a lexicographical station out off ISLE guidelines" dins LREC 2002 Third International Conference on Language Resources and Evaluation Proceedings. Las Palmas de Gran Canaria. Pp.. 694-700.

Fast and easy development of pronunciation lexicons for names[1]

Henk van den Heuvel[*], *Jean-Pierre Martens*[#], *Nanneke Konings* [*]

[*] CLST, Radboud University Nijmegen, The Netherlands
[#] ELIS, Ghent University, Belgium
H.vandenHeuvel@let.ru.nl

Abstract

We show that a good approach for the grapheme-to-phoneme conversion of Dutch proper names (e.g. person names, toponyms, etc), is to use a cascade of a general purpose grapheme-to-phoneme (G2P) converter and a special purpose phoneme-to-phoneme (P2P) converter. The G2P produces an initial transcription that is then transformed by the P2P. The P2P is automatically trained on reference transcriptions of names belonging to the envisaged name category (e.g. toponyms). The P2P learning process is conceived in such a way that it can take account of high order determinants of pronunciation, such as specific syllables, name prefixes and name suffixes. The proposed methodology was successfully tested on person names and toponyms, but we believe that it will also offer substantial reductions of the cost for building pronunciation lexicons of other name categories.

Index Terms: G2P conversion, names, pronunciation lexicon, Dutch, machine based learning

1. Introduction

Our objective is to reduce manual effort in producing Dutch pronunciation lexicons with correct phonetic transcription of names, using a minimum amount of training material for maximum performance.

Correct phonetic transcriptions are of paramount importance for both automatic speech-to-text (STT) conversion and text-to-speech (TTS) conversion. Good name transcription is imperative for success of many speech-based services: directory assistance, car navigation, etc.

The manual generation of these transcriptions is very time-consuming and subject to a great deal of inconsistencies. For this reason, automatic grapheme-to-phoneme (G2P) converters have been developed [1, 2, 3, 6, 9, 11]. However, general purpose converters often perform poorly when it comes to the transcription of names. Names typically do not adhere to the standard spelling conventions of a language due to their fossilized orthographic forms and/or their foreign origin. Therefore, they need special treatment [3, 7, 12].

In the AUTONOMATA[2] project we have chosen for an approach in which an initial transcription, emerging from a state-of-the-art general-purpose G2P converter, is 'corrected' by a special-purpose phoneme-to-phoneme (P2P) converter. This is an attractive option because it permits the P2P converter to profit from the knowledge of the general-purpose G2P converter, and it can focus on pronunciation rules that are typical for the envisaged name category. As such, it can be compact (few rules), and trainable on a relatively small pronunciation dictionary comprising only a few thousand names with their manually verified transcriptions.

Autonomata has delivered G2P-P2P tandems for person names and toponyms, as well as a *methodology* for the creation of good P2P's for other word classes for which a standard G2P is known to perform poorly (e.g. brand names, Points-Of-Interests, etc.). We have explored two different approaches. The first one is called the standard *decision tree (DT)* approach. A DT rule learning algorithm retrieves stochastic rules (attached to the leaf nodes of decision trees) that can explain many of the discrepancies between the correct transcription and the transcription produced by the G2P (in terms of phonemes and their immediate contexts). In the second approach, human expert knowledge is used in a top-down fashion to relate errors to higher order (viz. morphology, affix splitting, and language origin). A human expert analyzes the errors that the G2P makes, and formulates rules to correct these. For example, in street names we can identify morphemic entities such as 'erf' (yard), 'kamp' (parcel), 'straat' (street) which must be taken into account for the production of a proper syllabification, stress assignment and phonemization. We qualify this second approach as *HOK* (application of Higher Order Knowledge).

In this study we present experiments that were conducted in order to compare the DT and HOK approaches. In a pilot on toponyms only we asked ourselves the following questions:

1. What are the performances of the HOK and DT approaches when used independently?
2. What is the range of improvements that the synergy of both approaches can bring when compared to the DT approach alone?
3. What are the most relevant features the HOK approach can add to the DT approach?

In the main experiments we focused on the potential synergy of both approaches. The main research questions were:

1. Is it possible to incorporate the findings of the HOK approach into the P2P learning software?
2. Are additional improvements achievable by extracting and adding further HOK rules?
3. What is the effect of the training set size?
4. Can the G2P-P2P cascade perform as well as a dedicated G2P that was trained to generate a phoneme transcription of the names directly from the orthography?

We will only briefly report on the results of the pilot to leave more space for the main experiments.

2. Materials and tools

The reference lexicons that were available in the Autonomata project included first names, last names and toponyms (place names and, mostly, street names) that are encountered in the Netherlands and in Flanders. In the present study however, only names from the Netherlands were considered. Each lexicon was divided in a training, a development and an evaluation set [10]. Table 1 shows the sizes of these sets for the various name types.

Table 1. *Sizes of the manually verified reference lexicons*

Lexicon	Train set	Dev. set	Eval. set
First names	15,655	3,913	2,987
Last names	64,048	16,011	8,382
Toponyms	92,838	23,209	12,483

For each name, one or more manual phoneme transcriptions were available as reference transcriptions during training and evaluation. The manual transcriptions are broad phonetic transcriptions, enriched by syllable and stress markers.
The general-purpose G2P was provided by Nuance. The DT learning toolbox was developed by ELIS and is described in [12]. However, some important extensions were implemented since then. They are described in Section 4.
For the HOK approach we used internal tools of CLST. Based on different contexts of one, two or three phonemes to the left and/or right of the target phonemes, we first identified (in the training set) the most frequent input patterns that yielded incorrect pronunciations, and we computed rule application rates for these input-output pairs. The application rate is 1 if the input always leads to the same erroneous output. Next, we tried to relate the observed errors to higher order phenomena, more specifically to the morphological composition of the name. Correction rules were then formulated in the FONPARS format [8] and applied to the output of the G2P.
To evaluate the performance of the DT and HOK rules (or the combination thereof) the G2P-P2P output transcription is aligned with the reference transcription and the number of mismatches is counted at the phoneme level and the word (=name) level. Based on this comparison the evaluation yields a symbol error rate (SER) and a word error rate (WER). The WER is the percentage of names with one or more errors in their transcriptions. The evaluation tool also counts the mismatches between the original G2P transcription and the reference transcription. This way we can compute the word improvement rate (WIR) as the percentage of words for which the P2P offers an improvement minus the percentage of words for which it causes a degradation of the transcription. A transcription is called better if it has less symbols that deviate from the reference transcription (even if it is still wrong, and thus leaving the WER unchanged). In our experiments we will report SER, WER, and WIR. Further, we will report on the number of rules that were needed to arrive at the result, the idea being that given the same accuracy, a small rule set is preferred over a large one.
The output of the G2P-P2P cascade is compared to the best matching one of the manual pronunciations available in the evaluation lexicon. Person names and toponyms were treated separately in the experiments.
Although language origin may be an important factor, this information is not presumed to be available. In fact, the positive effect of distinguishing the language origin is not so straightforward as it might seem at first glance. E.g. the name Johnny, which is clearly of an English origin, has an accepted 'dutchified' pronunciation /ʃɔni/.

3. Pilot Experiment

As stated in the introduction we just briefly discuss the results of the pilot. We found that, for toponyms,

1. Both the DT and HOK approach yield substantial improvements over the general-purpose G2P. The DT approach performed better than the HOK approach (WIR of 27.4% vs. 20.6%).
2. Cascading HOK rules after the DT P2P did yield a further improvement. This synergetic approach raised the WIR from 27.4% to 31.3% and lowered WER from 38.1% to 32.5%.
3. The most relevant findings of the HOK approach were that it helps to take the identity of syllables, prefixes and suffixes into account for computing a transcription. For toponyms, these syllables, prefixes and suffixes were:

- **syllables**: *be de ge he ber der het ver kor laan sint sint-straat toor van*
- **suffixes**: *de el ne se ter baan dijk dreef hof jitte laan meester pad plein singel straat weg destraat sestraat seweg steenweg*
- **prefixes***: be de ge he ber ver bo ca ha ho ka ma mo ou pa ro burge sint-*

For person names, the syllables, prefixes and suffixes were:

- **syllables**: *de den der het te ten ter ver van a*
- **suffixes**: *je ske sje ke kje the na ne nus se ta us berg burg re ren sen de den en er ga gen ijer ker len man mans meijer ter veld ven zen*
- **prefixes***: be ca ge ka ma de van vande vander ber ger*

4. Main Experiment

Next, we investigated the synergetic potential of the HOK and DT approach.

4.1. Extension of the standard learning software

The original standard learning software was extended so that it can take into account the phenomena discovered by the HOK approach in the pilot. In the linguistic context of a pattern that is considered for modification (the rule focus), one can now include the syllable identity as well as the identities of the name prefix and suffix. These identities are either a syllable, prefix or suffix belonging to a predefined set, or *unknown*. In a final step the discovery of appropriate syllable, prefix and suffix sets is also automated. To that end, the learning software records all syllables, prefixes and suffixes it encounters, and it records how many times such an item occurs and how many times it co-occurs with a transcription error: a phonemization error in the syllable or an arbitrary error in a name with the given prefix or suffix. By just retaining the items with a sufficiently high co-occurrence rate, one can construct the syllable, prefix and suffix sets to supply to the learning software as part of the data file that also specifies the other linguistic features. One can even change the sets without having to change anything in the software. In this way, it is straightforward to apply the methodology on other name categories requiring other sets of prefixes, suffixes and syllables. In order to constrain the computation time and memory, the search for suitable prefixes and suffixes during the training stage is limited to syllables and syllable pairs as they are output by the general-purpose G2P

4.2. Experimental setup

To prove our case, we have compared our G2P-P2P approach against a direct approach in which a special purpose G2P is trained on the same data the P2P was trained on. We have used TiMBL (Tilburg Machine Based Learning) [5] to train such G2P's. For TiMBL the number of rules is simply the number of training instances (=names). In all tests, we developed two systems per approach: one for converting person names and one for converting toponyms. During evaluation we separately tested on first names and last names. This leads us to the following experimental conditions:

1. **G2P** proper: baseline general-purpose G2P
2. **TiMBL** (trained on full train set): special purpose G2P trained on the full training set
3. **G2P + P2P(DT):** standard DT rules only, but now with syllable, prefix and suffix features included. To measure the effect of the training set size, we distinguished between systems trained on the full training set (**A**), and on the smaller development set (**B**).
4. **G2P + P2P(DT) + P2P(HOK):** a combination of the previous system 3B and extra HOK rules obtained after

human inspection (using the development set) of the remaining errors made by this system.

5. **G2P + P2P(New):** As system 3B but with fully automatic detection of relevant syllables, prefixes and suffixes.
6. **G2P + P2P(New)2outputs:** As system 5 but with 2 name outputs, and selecting the best one for evaluation.

The same experimental conditions were applied to both person names and toponyms for Dutch.

4.3. Results

The results of our experiments are presented in Table 2. The 95% confidence intervals of the WER scores were added to show the statistical significance of the differences. We can summarize our findings as follows:

- System 6 with two outputs has by far the best performance.
- For first names none of the approaches with systems 2-5 lead to significant improvements, whereas for the two other categories all of these approaches result in significant improvements across all three measures (SER, WER and WIR).
- For toponyms for instance, one observes an extra gain of 3.5% absolute in the WIR, when compared with the pilot.
- For all name categories, only small insignificant gains in performance are obtained by adding further HOK rules.
- The relevant HOK features can be detected automatically (compare systems 3 and 5), yielding even further improvement for family names.
- The special purpose G2P's, developed with TIMBL, can only compete (in terms of accuracy) with the G2P-P2P systems for the case of family names.
- Comparing systems 3A and 3B for all three name types shows that the development set suffices for the training of a nearly optimal P2P.

The results for family names are much better than for first names, probably because the training set for person names consisted of 75% last names and only 25% first names. Also the length of first names is considerably shorter, giving the P2P learning tool less context to fine-tune its adjustments.

Table 2: *Experimental results for transcriptions containing both segmental and supra-segmental (syllabic, stress) information. Results for Dutch names only. Number of extra rules refers to the extra rules per component on top of the G2P.*

Name type	System	# extra rules	SER	WER	WIR	95% CI on WER
First names	**1. G2P**	N/A	11.9	39.9	N/A	[38.1 – 41.7]
	2. TiMBL (Full training set)	(79.711)	12.4	52.5	-10.5	
	3A. G2P + P2P(DT) (Full training set)	**2434**	**10.5**	**40.4**	**5.4**	
	3B. G2P + P2P(DT) (development set)	2064	10.9	41.6	3.5	[39.8 – 43.4]
	4B. G2P + P2P(DT)+P2P(HOK)	2064 + 28	10.7	40.9	4.3	
	5B. G2P + P2P(New)	1964	10.9	41.7	3.6	
	6B. G2P + P2P(New)2outputs	1964	8.7	35.1	14.7	
Family names	**1. G2P**	N/A	9.5	44.6	N/A	[43.5 – 45.7]
	2. TiMBL (Full training set)	(79.711)	5.7	32.4	17.2	
	3A. G2P + P2P(DT) (Full training set)	2434	6.5	35.5	20.3	
	3B. G2P + P2P(DT) (development set)	2064	6.5	35.2	20.3	[34.2 – 36.2]
	4B. G2P + P2P(DT)+P2P(HOK)	2064 + 28	6.3	34.4	21.3	
	5B. G2P + P2P(New)	**1964**	**6.0**	**31.9**	**21.5**	
	6B. G2P + P2P(New)2outputs	1964	5.0	28.3	27.7	
Toponyms	**1. G2P**	N/A	6.8	51.2	N/A	[50.3 – 52.1]
	2. TiMBL (Full training set)	(92.845)	5.4	37.8	23.5	
	3A. G2P + P2P(DT) (Full training set)	**1037**	**3.6**	**32.8**	**30.9**	
	3B. G2P + P2P(DT) (development set)	1358	3.6	32.9	30.4	[32.1 – 33.7]
	4B. G2P + P2P(DT)+P2P(HOK)	1358 + 68	3.6	32.3	29.9	
	5B. G2P + P2P(New)	1343	3.6	32.7	30.6	
	6B. G2P + P2P(New)2outputs	1343	2.5	22.6	42.1	

5. Discussion & conclusions

Returning to our initial research questions as formulated in the Introduction, we can draw the following conclusions when reviewing systems 1-5:

1. Taking syllabic context and morphological context (affixes) discovered by a HOK approach into account during DT learning, leads to an improved performance (as compared to the WIRs obtained in the pilot (section 3));
2. The obtained improvements must be close-to-optimal since the addition of more HOK elements does not yield any significant improvement anymore.
3. A relatively small development set suffices to train a nearly optimal P2P.
4. The comparison with TiMBL shows that the G2P-P2P approach is more effective in terms of required training data, and transcription accuracy than a restart from scratch approach in which a special purpose G2P like TiMBL is trained from scratch on the same training data.

Fully automatic detection of HOK elements (system 5) leads to further improvements (as compared to system 4) for person names. This shows that the learning software is able to take

this type of higher order phenomena more effectively into account than the human expert.
Except for first names, the G2P-P2P tandem yields substantial improvements over the G2P alone. Nonetheless, more than 30% of the name transcriptions still contain one or more errors. This makes a manual post-hoc correction necessary, be it less time consuming than before.
If the stress marks are excluded from the evaluation (not shown in Table 2), the WER scores of the best G2P-P2P systems are about 30% for the first names, 24% for the family names and 14% for the toponyms. The corresponding WER scores for the general-purpose G2P were 34%, 37% and 33%. Comparing these figures with those in Table 2 demonstrates that the remaining errors are mainly in the stress assignment for toponyms and in the phonemization for first names; for family names the situation is somewhere in between.
Turning to system 6, if the P2P converters are allowed to produce two outputs per name, and if one sets the probability threshold to 0.2 times the probability of the best transcription (according to the P2P that is), one gets about 1.6 transcriptions per name on average. If one then counts as errors the percentage of names for which none of the generated transcriptions is correct, the WERs drop significantly. Especially the WIR for first names is multiplied by a factor 4. In about 50% of the first name cases where the general-purpose G2P makes an error, one of the two transcriptions produced by the G2P-P2P tandem is already better than the general-purpose transcription. Making these two transcriptions available may seriously speed up the semi-automatic construction of a first name lexicon.
For comparison, the WER achieved by the G2P on infrequent but normal words (non-names) is typically below 25%. This shows that names are indeed much more difficult to convert than words; it also shows that the cascaded P2P is not able to bring down the WER to what is typical for normal words.
A further examination of the remaining errors for the three name categories (in the development sets) shows that almost no systematic errors were overlooked by the P2P's.

At a methodological level, the results of our experiments give interesting guidelines for developing a good P2P:
1. Select a development set of about 2500 names.
2. Use the G2P to obtain an initial transcription of 1000 of these names and correct them manually
3. Train a P2P on this corrected set, and transcribe the remaining names in the development set with the resulting G2P-P2P. Correct the output manually.
4. Retrain the P2P on the full development set and transcribe this set using the resulting G2P-P2P tandem.
5. Analyze remaining errors from a HOK perspective
6. Identify features related to higher order phenomena and incorporate them in the learning process.
7. Train the new P2P for optimal results

Higher order phenomena such as syllable and affix identities are automatically detected by the learning software, meaning that the HOK analysis in steps 5-7 can probably be skipped for many name types, thus saving a lot of manual efforts.
In terms of time effort, the HOK method is, not very surprisingly, much more costly than the DT method. The compilation of a HOK rule set for toponyms took at least one day for the full training set of 100k entries and at least half a day for the development set of 20k entries, using the flexible development tools for the HOK approach that CLST has at its disposal. In contrast, the P2P rule set was generated in less than an hour (full training set).

Future work in a follow-up project (called Autonomata Too) will involve the goodness of fit of the computed canonical pronunciations to real pronunciations as they are encountered in everyday speech in the context of a car navigation application involving street names, place names and POI's (Points of Interest). These type of comparisons will show the actual validity of our work for Automatic Speech Recognition.

6. Acknowledgements

The presented work was carried out in the Autonomata project, granted under the Dutch-Flemish STEVIN[3] program. The project partners are the universities of Ghent, Nijmegen and Utrecht and the companies Nuance and TeleAtlas.

7. References

[1] Bisani M., Ney H. (2003) "Multigram-Based Grapheme-to-Phoneme Conversion for LVCSR", Procs. Interspeech, 933-936.
[2] Black, A., Lenzo, K., Pagel, V. (1998). "Issues in building general letter to sound rules", Procs. ESCA/ COCOSDA workshop on Speech Synthesis (Jenolan Caves), pp. 77-81.
[3] Boula de Mareüil, P.; d'Alessandro, C.: Bailly, G.; Béchet, F.; Garcia, M.; Morel, M.; Prudon, R. and Véronis, J (2005). "Evaluating the pronunciation of proper names by four French grapheme-to-phoneme converters", Procs. Interspeech, Lisbon, 1521-1524.
[4] Bouma, G. (2000). "A finite state and data-oriented method for grapheme to phoneme conversion", Procs. ACL, 303-310
[5] W. Daelemans, J. Zavrel, K. van der Sloot, A.van den Bosch (2004). TiMBL: Tilburg Memory Based Learner, 5.1, Reference Guide. ILK Technical Report 04-02, http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf
[6] Damper, R.I, Marchand, Y., Adamson M.J. and Gustafson, K. (1998), "A comparison of letter-to-sound conversion techniques for English Text-to-speech synthesis", Procs. Institute of Acoustics, 20(6).
[7] Font Llitjós, A. and Black, A.W., "Evaluation and collection of proper name pronunciations online", Procs. LREC2002, Gran Canaria, 2002, 247-254.
[8] Kerkhoff, J., Rietveld, T. (1994). "Prosody in Niros with Fonpars and Alfeios." In: de Haan and Oostdijk (Eds.), Procs. Dept. of Language & Speech, Univ. of Nijmegen, Vol.18, 107-119.
[9] Polyakova, T., and Bonafonte, A. (2006) "Using error-driven approach to improve automatic g2p conversion accuracy", Procs. TC-STAR workshop on SLT, Barcelona.
[10] Stevens, G., and Bloothooft, G. (2006), "Autonomata namencorpora en p2p-evaluatie. http://speech.elis.ugent.be/autonomata
[11] Taylor, P. (2005), "Hidden Markov Models for grapheme to phoneme conversion" , Procs Interspeech 2005, Lisbon, 1973-1976.
[12] Yang, Q., Martens, J.P., Konings, N., Van den Heuvel, H. (2006). "Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names", Procs. LREC, Genua, 287-292.

[1] This paper is an extended and improved revision of:
Van den Heuvel, H., Martens, J.P., Konings, N., (2007). "G2P conversion of names. What can we do (better)?", Procs. Interspeech2007, Antwerp, 1773-1776.
[2] http://speech.elis.ugent.be/autonomata
[3] http://taalunieversum.org/taal/technologie/stevin/

Grammar Systems as Interfaces

Gemma Bel-Enguix, M. Dolores Jiménez-López

Research Group on Mathematical Linguistics
Rovira i Virgili University,
43005 Tarragona, Spain
{gemma.bel,mariadolores.jimenez}@urv.cat

Abstract

Basing human-computer interaction on human conversation can provide flexible and effective user interfaces. Conversational interfaces would provide the opportunity for the user to interact with the computer just as they would do to person. In this paper, we introduce a gammar systems model that may contribute to the building of better human-computer interfaces through a simulation of human language use.

1. Introduction

According to [1], human language technology plays a central role in providing an interface that will drastically change the human-machine communication paradigm from *programming* to *conversation*, enabling users to efficiently access, process, manipulate and absorb a vast amount of information. Effective conversational interfaces must incorporate extensive and complex dialogue modelling. In this paper, we introduce a Grammar System interface that may contribute to the building of more effective and efficient human-computer interaction tools through the simulation of human-human conversations.

Human-computer interfaces require models of dialogue structure that capture the variability and unpredictability within dialogue. The study of human-human conversation and the application of its features to human-machine interaction can provide valuable insights, as has been recognized by many authors (cf. [1], [2], [3], [4]). To be truly conversation-like, a human-computer dialogue has to have much of the freedom and flexibility of human-human conversations.

Even though many researchers agree that a complete simulation of human conversation is very difficult (maybe impossible) to be reached, it seems clear that knowledge of human language use can help in the design of efficient human-computer dialogues. It can be argued that users could feel more comfortable with an interface that has some of the features of a human agent. Therefore, our model is based on human-human interactions. The result is a highly formalized dialogue-based interface.

Throughout the paper, we assume that the reader is familiar with the basics of formal language theory, for more information see [5].

2. A GS Interface

In this section, we introduce a Grammar System Interface (GSI) that may contribute to the building of better human-computer dialogues through a simulation of human language use. Note that we do not intend to fully simulate human language use, but only to take advantage from research on human language in order to improve interfaces. Our formal model tries to capture human-conversation main features and is based on Eco-Grammar Systems (EGS).

Eco-grammar systems theory is a subfield of grammar systems theory, a consolidated and active branch in the field of formal languages [6]. Eco-grammar systems have been introduced in [7] and provide a syntactical framework for eco-systems, this is, for communities of evolving agents and their interrelated environment. Taking as starting point the notion of eco-grammar system, we introduce the notion of *Grammar Systems Interface* (GSI). GSI present several advantages useful for human computer interaction, we emphasize here only several aspects like the following: *a)* it is highly modularized by a distributed system of contributing agents; *b)* it is contextualized permitting to linguistic agents to re-define their capabilities according to context conditions given by mappings; *c)* it is emergent, from current competence of the collection of active agents emerges a more complex behaviour.

Definition 1 *A Grammar System Interface (GSI) of degree* n, $n \geq 2$, *is an* $(n+1)$*-tuple:*

$$\Sigma = (E, A_1, ..., A_n),$$

where:

- $E = (V_E, P_E)$,
  - V_E *is an alphabet;*
  - P_E *is a finite set of rewriting rules over* V_E.
- $A_i = (V_i, P_i, R_i, \varphi_i, \psi_i, \pi_i, \rho_i)$, $1 \leq i \leq n$,
  - V_i *is an alphabet;*
  - P_i *is a finite set of rewriting rules over* V_i;
  - R_i *is a finite set of rewriting rules over* V_E;
  - φ_i: $V_E^* \rightarrow 2^{P_i}$;
  - ψ_i: $V_E^* \times V_i^+ \rightarrow 2^{R_i}$;
  - π_i *is the start condition;*
  - ρ_i *is the stop condition;*
  - π_i *and* ρ_i *are predicates on* V_E^*.

The items of the above definition have been interpreted as follows: a) E represents the environment described at any moment of time by a string w_E, over alphabet V_E, called the *state of the environment*. The state of the environment is changed both by its own evolution rules P_E and by the actions of the agents of the system, A_i, $1 \leq i \leq n$. b) A_i, $1 \leq i \leq n$, represents an

agent. It is identified at any moment by a string of symbols w_i, over alphabet V_i, which represents its current state. This state can be changed by applying evolution rules from P_i, which are selected according to mapping φ_i and depend on the state of the environment. A_i can modify the state of the environment by applying some of its action rules from R_i, which are selected by mapping ψ_i and depend both on the state of the environment and on the state of the agent itself. Start/Stop conditions of A_i are determined by π_i and ρ_i, respectively. A_i starts/stops its actions if context matches π_i and ρ_i.

GSIs intend to describe dialogue as a sequence of *context-change-actions* allowed by the current environment and performed by two or more *agents.*

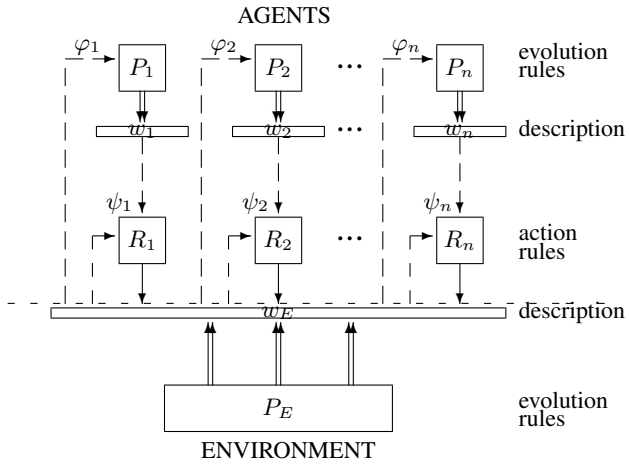


Figure 1: Grammar Systems Interface based on EGS

Definition 2 *By an action of an active agent A_i in state $\sigma = (w_E; w_1, w_2, \ldots, w_n)$ we mean a direct derivation step performed on the environmental state w_E by the current action rule set $\psi_i(w_E, w_i)$ of A_i.*

Definition 3 *A state of a GSI $\Sigma = (E, A_1, \ldots, A_n)$, $n \geq 2$, is an $n+1$-tuple:*

$$\sigma = (w_E; w_1, \ldots, w_n),$$

where $w_E \in V_E^*$ is the state of the environment, and $w_i \in V_i^*$, $1 \leq i \leq n$, is the state of agent A_i.

This rule is applied *by an active agent* and it is a rule selected by $\psi_i(w_E, w_i)$. We define an *active agent* in relation to the allowable actions it has at a given moment. That is, an agent can participate in conversation –being, thus, active– only if its set of allowable actions at that moment is nonempty:

Definition 4 *An agent A_i is said to be active in state $\sigma = (w_E; w_1, w_2, \ldots, w_n)$ if the set of its current action rules, that is, $\psi_i(w_E, w_i)$, is a nonempty set.*

Since conversation in GSI is understood in terms of *context changes*, we have to define how the environment passes from one state to another as a result of agents' actions:

Definition 5 *Let $\sigma = (w_E; w_1, \ldots, w_n)$ and $\sigma' = (w'_E; w'_1, \ldots, w'_n)$ be two states of a GSI $\Sigma = (E, A_1, \ldots, A_n)$. We say that σ' arises from σ by a simultaneous action of active agents $A_{i_1}, \ldots, A_{i_r}$, where $\{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}$, $i_j \neq i_k$, for $j \neq k$, $1 \leq j, k \leq r$, onto the state of the environment w_E, denoted by $\sigma \overset{a}{\Longrightarrow}_\Sigma \sigma'$, iff:*

- *$w_E = x_1 x_2 \ldots x_r$ and $w'_E = y_1 y_2 \ldots y_r$, where x_j directly derives y_j by using current rule set $\psi_i(w_E, w_{i_j})$ of agent A_{i_j}, $1 \leq j \leq r$;*
- *there is a derivation:*
  $w_E = w_0 \overset{a}{\Longrightarrow}^*_{A_{i_1}} w_1 \overset{a}{\Longrightarrow}^*_{A_{i_2}} w_2 \overset{a}{\Longrightarrow}^*_{A_{i_3}} \ldots \overset{a}{\Longrightarrow}^*_{A_{i_r}} w_r = w'_E$

  such that, for $1 \leq j \leq r$, $\pi_{i_j}(w_{j-1}) =$ true and $\rho_{i_j}(w_j) =$ true. And for $f \in \{t, \leq k, \geq k\}$ the derivation is:
  $w_E = w_0 \overset{a}{\Longrightarrow}^f_{A_{i_1}} w_1 \overset{a}{\Longrightarrow}^f_{A_{i_2}} w_2 \overset{a}{\Longrightarrow}^f_{A_{i_3}} \ldots \overset{a}{\Longrightarrow}^f_{A_{i_r}} w_r = w'_E$

  such that, for $1 \leq j \leq r$, $\pi_{i_j}(w_{j-1}) =$ true[1], and
- *$w'_i = w_i$, $1 \leq i \leq n$.*

Definition 6 *Let $\Sigma = (E, A_1, ..., A_n)$ be a GSI. And let $w_E = x_1 x_2 ... x_r$ and $w'_E = y_1 y_2 ... y_r$ be two states of the environment. Let us consider that w'_E directly derives from w_E by action of active agent A_i, $1 \leq i \leq n$, as shown in definition 5. We write that:*

$w_E \overset{a}{\Longrightarrow}^{\leq k}_{A_i} w'_E$ *iff* $w_E \overset{a}{\Longrightarrow}^{\leq k'}_{A_i} w'_E$, *for some $k' \leq k$;*

$w_E \overset{a}{\Longrightarrow}^{\geq k}_{A_i} w'_E$ *iff* $w_E \overset{a}{\Longrightarrow}^{\leq k'}_{A_i} w'_E$, *for some $k' \geq k$;*

$w_E \overset{a}{\Longrightarrow}^{*}_{A_i} w'_E$ *iff* $w_E \overset{a}{\Longrightarrow}^{k}_{A_i} w'_E$, *for some k;*

$w_E \overset{a}{\Longrightarrow}^{t}_{A_i} w'_E$ *iff* $w_E \overset{a}{\Longrightarrow}^{*}_{A_i} w'_E$ *and there is no $z \neq y$ with $y \overset{a}{\Longrightarrow}^{*}_{A_i} z$.*

In words, $\leq k$-derivation mode represents a time limitation where A_i can perform at most k successive actions on the environmental string. $\geq k$-derivation mode refers to the situation in which A_i has to perform at least k actions whenever it participates in the derivation process. With $*$-mode, we refer to such situations in which agent A_i performs as many actions as it wants to. And finally, t-derivation mode represents such cases in which A_i has to act on the environmental string as long as it can.

However, in the course of a dialogue, agents' states are also modified and the environmental string is subject to changes due to reasons different from agents' actions. So, in order to complete our formalization of dialogue development, we add the following definition:

Definition 7 *Let $\sigma = (w_E; w_1, \ldots, w_n)$ and $\sigma' = (w'_E; w'_1, \ldots, w'_n)$ be two states of a GSI $\Sigma = (E, A_1, \ldots, A_n)$. We say that σ' arises from σ by an evolution step, denoted by $\sigma \overset{e}{\Longrightarrow}_\Sigma \sigma'$, iff the following conditions hold:*

- *w'_E can be directly derived from w_E by applying rewriting rule set P_E;*
- *w'_i can be directly derived from w_i by applying rewriting rule set $\varphi_i(w_E)$, $1 \leq i \leq n$.*

[1] In this latter case the stop condition $\rho_i(w_j) =$ true is replaced by the stop condition given the f-mode.

Definition 8 *Let $\Sigma = (E, A_1, ..., A_n)$ be a GSI as in definition 1. Derivation in Σ terminates in:*

- *Style (ex) iff for $A_1, ..., A_n$, $\exists A_i : w_i \in T_i, 1 \leq i \leq n$;*
- *Style (all) iff for $A_1, ..., A_n$, $\forall A_i : w_i \in T_i, 1 \leq i \leq n$;*
- *Style (one) iff for $A_1, ..., A_n$, $A_i : w_i \in T_i, 1 \leq i \leq n$.*

According to the above definition, a derivation process ends in style *(ex)* if there is *some* agent A_i that has reached a terminal string. It ends in style *(all)* if *every* agent in the system has a terminal string as state. And it finishes in style *(one)* if there is *one* distinguished agent whose state contains a terminal string. Styles *(all)*, *(ex)* and *(one)* might account for three different ways of closing a dialogue.

In GSI, the development of dialogue implies that both the *state of the environment* and *state of agents* change. Such changes take place thanks to two different types of processes: *action steps* and *evolution steps*. At the end, what we have is a *sequence of states* reachable from the initial state by performing, alternatively, action and evolution derivation steps:

Definition 9 *Let $\Sigma = (E, A_1, \ldots, A_n)$ be a GSI and let σ_0 be a state of Σ. By a state sequence (a derivation) starting from an initial state σ_0 of Σ we mean a sequence of states $\{\sigma_i\}_{i=0}^{\infty}$, where:*

- $\sigma_i \stackrel{a}{\Longrightarrow}_\Sigma \sigma_{i+1}$, *for* $i = 2j, j \geq 0$; *and*
- $\sigma_i \stackrel{e}{\Longrightarrow}_\Sigma \sigma_{i+1}$, *for* $i = 2j + 1, j \geq 0$.

Definition 10 *For a given GSI Σ and an initial state σ_0 of Σ, we denote the set of state sequences of Σ starting from σ_0 by* $Seq(\Sigma, \sigma_0)$.

The set of environmental state sequences is:

$$Seq_E(\Sigma, \sigma_0) = \{\{w_{Ei}\}_{i=1}^{\infty} \mid \{\sigma_i\}_{i=0}^{\infty} \in Seq(\Sigma, \sigma_0), \sigma_i = (w_{Ei}; w_{1i}, \ldots, w_{ni})\}.$$

The set of state sequences of the j-th agent is defined by:

$$Seq_j(\Sigma, \sigma_0) = \{\{w_{ji}\}_{i=1}^{\infty} \mid \{\sigma_i\}_{i=0}^{\infty} \in Seq(\Sigma, \sigma_0), \sigma_i = (w_{Ei}; w_{1i}, \ldots, w_{ji}, \ldots, w_{ni})\}.$$

3. An Example

The following fragment of a dialogue illustrates how the dialogue-based interaction system we have introduced above works. Dialogues of this type can be handled by our system in a simple way. Let us consider the following dialogue that can take place between a customer (the user) who wants to buy a table and the shop assistant (the system) that guides the customer in his choice. The interaction is collaborative, with neither the system nor the user being in control of the whole interaction. Each of them contributes to the interaction when they can.

User: Good morning.

System: Good morning. Can I help you?

User: Yes, I would like to buy a table.

System: Which type of table, for which room?

User: I need a living room table.

System: We have different shapes, which do you prefer?

User: I would like a rectangular table.

System: What about the material?

User: I would like a glass table.

System: We have different glass colours, any preference?

User: I would like a transparent glass table.

System: I guess this is the table you are looking for.

User: Yes, it is wonderful. Thank you very much.

System: Your are welcome.

Starting from this example, we develop the following GSI with two agents.

The alphabet of the environment is:

- I is the initial state,
- h stands for "Hello",
- Q, for "Can I help you?",
- t, for "I need a table",
- R for "What kind of room?",
- ℓ, for "A living room",
- S for "What shape?",
- r for "Rectangular.",
- M for "What material?",
- g for "Glass",
- C for "Colour?",
- w for "White",
- K for "OK, something else?",
- e for "End of my request, bye",
- B for "Bye".

The environment is in the initial state I. For this dialogue we do not need special functions φ, they just copy the environment to the agents' status.

The function ψ_1 adds a single rule R_1 to its set of rules, as follows:

- $R_1 = \{I \rightarrow h\}$ for $\omega_E = xI, x \in V_E^*$,
- $R_1 = \{Q \rightarrow Qt\}$ for $\omega_E = xQ, x \in V_E^*$,
- $R_1 = \{R \rightarrow R\ell\}$ for $\omega_E = xR, x \in V_E^*$,
- $R_1 = \{S \rightarrow Sr\}$ for $\omega_E = xS, x \in V_E^*$,
- $R_1 = \{M \rightarrow Mg\}$ for $\omega_E = xM, x \in V_E^*$,
- $R_1 = \{C \rightarrow Cw\}$ for $\omega_E = xC, x \in V_E^*$,
- $R_1 = \{K \rightarrow Ke\}$ for $\omega_E = xK, x \in V_E^*$,
- $R_1 = \{b \rightarrow bB\}$ for $\omega_E = xb, x \in V_E^*$.

The function ψ_2 is defined in a similar way as follows:

- $R_2 = \{h \rightarrow hhQ\}$ for $\omega_E = xh, x \in V_E^*$,
- $R_2 = \{t \rightarrow tR\}$ for $\omega_E = xt, x \in V_E^*$,
- $R_2 = \{\ell \rightarrow \ell S\}$ for $\omega_E = x\ell, x \in V_E^*$,
- $R_2 = \{r \rightarrow rM\}$ for $\omega_E = xr, x \in V_E^*$,
- $R_2 = \{g \rightarrow gC\}$ for $\omega_E = xg, x \in V_E^*$,
- $R_2 = \{w \rightarrow wK\}$ for $\omega_E = xw, x \in V_E^*$,
- $R_2 = \{e \rightarrow eb\}$ for $\omega_E = xe, x \in V_E^*$.

This implementation gives only one possible dialogue that is: $I \rightarrow hhQtRlSrMgCwKebB$. If we want to get more diverse dialogues on the same topic, we can imagine that after asking the about the table, the dialogue continues with the specifications of details, colour, shape, material, room in an arbitrary order like: $I \rightarrow hhQtRlCwMgSrKebB$ or $I \rightarrow hhQtSrRlMgCwKebB$. To do this, we keep the definition of the function ψ_1 and we should modify a little bit the function ψ_2 in this way. $R_2 = \{a \rightarrow aB\}$ for $\omega_E = xa, x \in V_E^*$, $a \in \{t, l, r, g, m, w\}$ and B randomly selected from the set $\{R, S, M, C\}$.

4. Final Remarks

Many researchers believe that natural language interfaces can provide the most useful and efficient way for people to interact with computers. According to [1], "for information to be truly accessible to all anytime, anywhere, one must seriously address the problem of user interfaces. A promising solution to this problem is to impart human-like capabilities onto machines, so that they can speak and hear, just like the users with whom they need to interact."

The Grammar System Interface we have introduced can be considered a mixed-initiative interaction model in which there is a dynamic exchange of control of the dialogue flow. GSI are able to model interfaces with a high degree of flexibility, what means that they are able to accept new concepts and modify rules, protocols and settings during the computation. The main characteristic of the model is the use of simple grammars in order to generate a dialogue structure. It should not to be seen as a psychologically realistic cognitive model, but as a model that might successfully emulate human linguistic behaviour in some specific situations such as natural language interfaces.

Up to now, we have developed the formal model and we have worked on the comparison and interplay between GSI and other multi-agent dialogue systems. The formal properties of GSI, obtained by the mathematical analysis of the model, show the flexibility required for modelling dialogue. Moreover, since the definition of the framework is based on features that characterize human-human conversation, GSI capture and simulates (in a very abstract and formal way) human language use. Currently, we are working on the implementation of the model. This phase of our research will provide us a definite answer about the GSI adequacy for building a conversation-like human-computer interface.

5. References

[1] Zue, V., "Conversational Interfaces: Advances and Challenges", Kokkinakis, G., Fakotakis, N. and Dermatas, E. (eds.), Eurospeech'97. 5th European Conference on Speech Communication and Technology, vol. 1: 9-18, 1997.

[2] Bunt, H., "Non-problems and social obligations in human-computer conversation", Proceedings of the 3rd International Workshop on Human-Computer Conversation, 2000.

[3] Larsson, S., "Dialogue systems: Simulations or interfaces?", Gardent, C. and Gaiffe, B. (eds), Proceedings of the Ninth Workshop on the Semantics and Pragmatics of dialogue. DIALOR'05, Loria-Inria, Nancy pages, 2005, 45–52.

[4] Ogden, W.C., Bernick, Ph., Using Natural Language Interfaces, Helander, M.G., Landauer, T.K., Prabhu, P.V. (eds.), Handbook of Human-Computer Interaction, Elsevier, Amsterdam, 1997, 137–161.

[5] Rozenberg, G. and Salomaa, A., Handbook of Formal Languages, Springer, Berlin, 1997.

[6] Csuhaj-Varjú, E., Dassow, J., Kelemen, J. and Păun, Gh. Grammar Systems: A Grammatical Approach to Distribution and Cooperation, Gordon and Breach, London, 1994.

[7] Csuhaj-Varjú, E., Kelemen, J., Kelemenová, A. and Păun, Gh., "Eco-Grammar Systems: A Grammatical Framework for Studying Lifelike Interactions", Artificial Life, 3: 1-28, 1997.

[8] Allen, J.F., Byron, D.K., Dzikovska, M., Ferguson, G., Galescu, L. and Stent, A., "Towards Conversational Human-Computer Interaction", AI Magazine, 22(4): 27-37, 2001.

HLT and communicative disabilities: The need for co-operation between government, industry and academia

Catia Cucchiarini [1], *Dirk Lembrechts* [2] *and Helmer Strik* [3]

[1] Nederlandse Taalunie (Dutch Language Union), The Netherlands

[2] MODEM, Consultancy Centre on Communicative Disabilities, Wilrijk, Belgium

[3] Department of Language and Speech, Radboud University Nijmegen, The Netherlands

ccucchiarini@taalunie.org, dirk.lembrechts@vzwkinsbergen.be, strik@let.ru.nl

Abstract

To improve the position of people with communication disabilities, it is first essential to identify the tools they require to improve their communicative capabilities. HLT can be instrumental in restoring functions and compensating for impairments by providing solutions that integrate knowledge of speech and language into automatic processes. Against this background, an initiative was taken of analysing the specific needs of communicatively disabled people in terms of applications and related HLT resources so as to identify a minimum common set of HLT resources that would be useful for developing applications for a number of communicative disabilities. The priorities set in this survey could be used to inform policy, research and development and eventually stimulate take-up by industry. In this paper we describe this approach.

Index Terms: communicative disabilities, HLT applications, health telematics.

1. Introduction

The Dutch Language Union is a Dutch-Flemish intergovernmental organization that has the aim of promoting the Dutch language. In the last decade, the DLU has taken a serious interest in human language technologies because these can play a vital role in strengthening the position of a language in the information society. Together with the relevant ministries in the Netherlands and Flanders, the DLU has set up a number of initiatives aimed at promoting the development of digital language resources and language and speech technology for the Dutch language. Governmental support was considered to be mandatory because since Dutch is a so-called mid-sized language [1, 2], companies are not always willing to invest in developing such technology for a language with a relatively small market. On the other hand, the development of language and speech technology is considered to be crucial for a language to be able to survive in the information society.

To promote the use of Dutch, the DLU tries first of all to create the right conditions for making it easier for Dutch speakers to get by with their language in as many different situations as possible. The DLU wants to achieve this for all speakers of Dutch, hence also for those who have communicative disabilities. Given that information and communication technology are gradually but steadily pervading our lives, creating the right conditions for ensuring the use of Dutch in daily life partly entails supporting the development of HLT applications. For the group of communicatively disabled speakers of Dutch, HLT can be instrumental in restoring functions and compensating for impairments by providing solutions that integrate knowledge of speech and language into automatic processes. In the field of communicative disabilities, HLT can be used for diagnosis, therapy, training and monitoring, compensation and augmentative and alternative communication. This policy is in line with European policy aimed at realising an inclusive information society where accessibility, universality and user-friendliness are considered to be essential to ensure full participation and to enhance the quality of life for all individuals.

2. HLT and communicative disabilities

To improve the position of people with communication disorders in the Netherlands and Flanders, it is first essential to identify the tools they require to improve their communicative capabilities: tools that assist verbal dialogue, reading and writing, and the use of communication devices. Against this background, the DLU was first interested in finding out whether people with communicative impairments need specific HLT products and services that are currently unavailable or insufficiently so. In addition, it was important to find out what role the business sector can play in providing these products and services.

To answer these questions, Rietveld and Stolte [3] carried out a survey in which recent research and initiatives in the Netherlands and Flanders were examined and experts, care providers, providers of communicative tools were interviewed as well as professionals from consultancy and knowledge centres such as MODEM [11].

Employing the World Health Organisation's ICF classification system, the researchers identified target groups from the viewpoint of the person as an organism. They identified four body functions and related impairments:

- mental (aphasia, dyslexia, mental disabilities)
- sensory (blindness and partial sight, deafness and partial hearing, deafblindness)
- voice and speech (dysarthria / anarthria, mutism, stuttering)
- movement and mobility (RSI / UEMSD, dyspraxia / apraxia).

It is also important to distinguish between congenital disorders and those acquired in life, because congenitality or acquiredeness can determine the suitability of a tool. Comorbidity (a combination of disorders) also imposes specific requirements on tools.

From the viewpoint of how disabilities limit human behaviour, the researchers examined the HLT applications that could help disabled people understand (read, hear, or understand sign language) and express (write or speak) messages and operate communications tools (telephones, fax machines, mobile phones).

For each category, the researchers tried to answer four questions:

- Who has this disability (which impairments cause it)?
- What HLT tools are available to compensate for this disability?
- What experiences have users had with tools (in terms of user friendliness, knowledge of the product, quality, and applicability)?
- How can the requirements relating to this disability be met in the short, mid, and long term?

The study that resulted form this investigation, Human language technologies and communicative disabilities [3] showed a world of very diverse desires, requirements, and possibilities – which helps explain why communicative disabilities arouse so little interest in the business sector. The diversity of disorders and requirements makes it impossible to develop products that everyone can use. Furthermore, the development of HLT-based products requires considerable investments in basic language resources, and the majority of HLT companies are not in a position to make such investments, especially for languages with a relatively limited market, like Dutch.

The researchers concluded that the Netherlands and Flanders have a wide range of requirements for HLT product and services. They also indicated which ones should be realized in the short term (i.e. synthetic whisper voice), in the mid-term (i.e. high-quality speech training for the deaf and hard of hearing) and in the long term (i.e. automatic lexical simplification of texts). Furthermore, the researchers noted a number of ways in which the business sector could help meet these requirements, such as by localising products, by making them more accessible and flexible, and by offering information and product support. Finally, they pointed out the need for a more extensive programme for action in the mid and long term.

In spite of the difficulties highlighted by Rietveld and Stolte [3], it seems nevertheless that HLT could play an important role in the development of solutions for communicative disabilities. It seems therefore worthwhile to consider how the problems signalled in the above-mentioned report could be surmounted or at least alleviated to try and make it easier for companies to develop HLT-based products.

A viable solution would seem to be an approach similar to the one that was adopted some years ago in the Dutch language area for strengthening the digital language infrastructure [4]. This approach is shortly described in the next section.

3. Basic Language Resources Kit

At the end of the previous century an initiative was launched by the DLU to strengthen the Dutch-Flemish HLT infrastructure. An important element of this initiative was a survey of existing HLT resources at that time, which was carried out by a working group of researchers who in turn were supervised by a steering committee of HLT experts. This committee first defined what the BLARK (Basic LAnguage Resources Kit) should be and then the working group carried out the survey on the availability and the quality of the existing resources.

In defining the BLARK several matrices were used. A distinction was made between applications, modules, and data [5, 6].

Applications: refers to classes of applications that make use of HLT. The following classes were defined: CALL (Computer Assisted Language Learning), access control, speech input, speech output, dialogue systems, document production, information access, and multilingual applications or translation modules.

Modules: refers to the basic software components that are essential for developing HLT applications. A distinction was made between 'Language Technology' modules (such as Morphological analysis, Parsers and grammars, Shallow parsing, Constituent recognition, Semantic analysis, Referent resolution, etc.), and 'Speech Technology' modules (such as Pronunciation lexicon, Speaker identification, Speaker tracking, Utterance verification, Language identification, etc.) [see 5, 6].

Data: refers to data sets and electronic descriptions that are used to build, improve, or evaluate modules. The following types of data were defined: monolingual lexica, multilingual lexica, thesauri, annotated corpora, unannotated corpora, speech corpora, multilingual corpora, multimodal corpora, and multimedia corpora.

In order to guarantee that the survey would be complete, unbiased and uniform, matrices were drawn up by the steering committee describing (1) which modules are required for which applications, (2) which data are required for which modules, and (3) what the relative importance is of the modules and data.

These matrices served as the basis for defining the BLARK. For instance it indicated that monolingual lexicons and annotated corpora are required for the development of a wide range of modules; these were therefore included in the BLARK. Furthermore, semantic analysis, syntactic analysis, and text pre-processing (for language technology) and speech recognition, speech synthesis, and prosody prediction (for speech technology) serve a large number of applications and were therefore included in the BLARK, as well. Note that only language specific modules and data were considered in this survey.

By defining a general BLARK, by identifying which elements were missing in the BLARK, and by analyzing the availability and the quality of the various resources, priority could be assigned to the development of those parts of the BLARK that were considered to be crucial and appeared to be missing. One list of priorities for speech technology and one for language technology were drawn up and were subsequently submitted to representatives from the whole HLT field (about 2,000 people). Definitive priority lists were then produced [5; 7] and submitted to various policy institutions such as the Dutch Ministry of Economic Affairs, the Netherlands Organisation for Scientific Research (NWO), the Dutch Ministry of Education, Culture and Science, the Flemish Institute for the Promotion of Innovation by Science and Technology (IWT), the Department of Economy, Science and Innovation (EWI) of the Ministry of the Flemish Community, and the Flemish Fund for Scientific Research (FWO).

These institutions acknowledged the importance of developing the resources mentioned in the priority lists. Besides, the Dutch Ministry of Economic Affairs decided to carry out an additional study aimed at determining whether some other form of economic support, in combination with the BLARK proposal, would stimulate the HLT sector even more. The results of this study indeed provided interesting insights into how effective action in the HLT sector should be shaped. The view developed [8] appeared to be shared by the other Dutch and Flemish financing institutions, which decided to combine their HLT subsidies in one common research and industry stimulation programme called, STEVIN, which started in 2004 under the auspices of the DLU [9].

4. A new initiative to stimulate the development of HLT applications for communicative disabilities

The success of the STEVIN programme [10] has convinced us that a similar approach with respect to HLT for communicative disabilities is worth investigating. This is also in line with the recommendations by Rietveld and Stolte [3] concerning the more extensive programme for action in the mid and long term.

With this aim in mind the DLU took the initiative of organising a round table conference with experts from the HLT sector and experts from various disciplines covering different communicative disabilities. New target groups are people suffering from dementia and cognitively impaired adults and children. The aim of the conference was to discuss whether an approach similar to 'BLARK / STEVIN' would be feasible for this sector. The initiative DLU had in mind was a survey aimed at identifying a minimum common set of HLT resources that would be useful for developing applications for a number of communicative disabilities, some sort of minimum common denominator. The rationale behind this initiative is that if indeed it is possible to identify such a core of resources that could be employed for developing applications for a wide range of disabilities, then it would be easier to convince policy institutions to finance the development of such HLT resources.

The similarity between this approach and the BLARK / STEVIN one lies in the fact that in both cases priorities are set as to the resources to be developed and that important criteria in setting priorities are multipurposedness and reusability.

4.1. Matrices

At the round table conference mentioned above, a working group of experts was formed, who has the task of investigating whether matrices can be defined that can later be used for drawing up an inventory of HLT resources that would be required for developing applications for people with communication disabilities, as was done for the BLARK for HLT in general. In other words, what we aim at is an overview of the specific needs in terms of applications and related HLT resources to guide policy, research and development and stimulate take-up by industry.

The MATRIX working group discussed the various possible dimensions of these matrices. Below some preliminary results are presented. However, the dimensions presented below may be subject to change, as s result of feedback from the field. This was also the case with the BLARK: during the survey we often had to adjust the taxonomy. Nevertheless, the dimensions presented below indicate that the matrices in this case clearly differ from those used for defining the BLARK. Regarding the applications we discern the following three dimensions:

- the purpose of the applications: diagnosis, monitoring, training, therapy, compensation, and augmentative and alternative communication (AAC).
- the function it concerns: reading, writing, listening, and speaking
- the target group, the disability: visual, hearing, mental / cognitive, motor, neurological, and oncological.

The combination of these three dimensions indicates what kind of applications should be considered. Clearly, some combinations make more sense than others. Nevertheless, these dimensions can be employed for defining (classes of) applications. Concrete examples of applications that appear to be needed are: training systems using speech recognition for dysarthria patients, diagnostic instruments and training material capable of self-adaptation for patients with cochlear implants, automatic recognition of sign language, automatic conversion of speech to symbols, and speech-based applications for cognitively heavily impaired people.

On the one hand we thus have the applications. On the other hand, we should consider which types of HLT are needed for these applications. It turned out that a useful way of looking at the technologies is to consider them as conversions between the following five modalities:

1. Auditive – spoken language
2. Visual 1 – written language
3. Visual 2 – images, animations: symbols, gestures, agents
4. Tactile – Braille, 3D-images (with relief)
5. Cognitive – concepts

To make clear what the relation is between conversions of modalities and (classes of) technologies, some examples are given here:

- 5 → 2 : text (language) generation: e.g. writing tools
- 2 → 2 : text modification, summarizing, indexing, etc.
- 1 → 2 : speech recognition
- 2 → 1 : speech synthesis
- 2 → 4 : text to Braille conversion
- 1 → 1 : speech manipulation, e.g. delayed or frequency-altered auditory feedback (DAF & FAF)
- 2 → 3 : from text to virtual talking heads, agents, gestures, etc.

Also in this case some combinations make more sense than others, but again considering these combinations is useful for defining the possible (classes of) technologies.

A possible fourth dimension is age, since some of the demands will differ depending on the age of the users. For instance, the interface will often differ between age groups, and automatic speech recognition of children is different from speech recognition of adults. However, with age it is difficult to specify the classes: in some cases distinguishing between children, adults, and elderly people could be a useful classification, but in other cases a more detailed classification is needed, and in other cases age is not an important factor at

all. The conclusion is that in making an inventory and priority list one should always keep in mind that age could be a factor.

4.2. Future steps

Once the matrices have been defined, important questions about relevance, availability and assessment have to be answered such as:

- What is the relevance of technologies to applications?
- What is already available, and what isn't?
- What is the quality level, is it suitable?

These results together could then be used to define a priority list that indicates what is most needed and which technologies have the highest priority.

Also in this case it is important that the majority of the actors in the field of communicative disabilities subscribe to the priorities and recommendations identified by the working group. To this end, a provisional report containing the inventory, the priority lists and the recommendations will be submitted to a large number of people active in the field of communicative disabilities, ranging from healthcare professionals, care providers, research centres, universities, health, social service organizations, patient and user organizations, and other stakeholders in this domain. In addition, a serious attempt will be made to involve HLT companies and potential product manufacturers to make them aware of the opportunities that emanate from HLT. The relevant comments will then be incorporated in the report and the same group of people will be invited to participate in a workshop in which the results (priority lists and recommendations) will be officially presented to the public.

On this occasion some people will be given the opportunity to publicly present their views on the results of the survey. The workshop will be concluded with a general discussion between the audience and the panel of experts that were responsible for the survey.

As was the case with the BLARK, the next step will consist in finding funds to finance the development of the resources that have been prioritized. Beside the institutions that are now involved in the STEVIN programme, other policy institutions may be interested in stimulating the development of HLT for communicative disabilities. For instance, many of the resources could be used to develop telecare systems, many of the possible products and services may be relevant for health telematics and could contribute to cutting down health expenditure while maintaining the same level of quality in health care. In addition, since a number of communicative disabilities are related to growing older, the number of people with communicative disabilities is likely to increase as a result of the ageing population. HLT applications will then play an important role in ensuring non-invasive personal assistance and independent living, in improving or maintaining functional abilities, in enhancing productivity and in improving the quality of life. Therefore, it is to be expected that health institutions may also be interested in supporting a stimulation programme in the domain of HLT and communicative disabilities in which government, industry and academia cooperate.

5. Conclusions

The Dutch Flemish STEVIN programme for HLT constitutes a good example of co-operation between governmental bodies, academia and industry to stimulate resource development, strategic research, knowledge transfer and eventually innovation and take-up by the HLT industry in the Netherlands and Flanders. In the field of HLT and communicative disabilities there is even more need for such a programme because in this case the market is even smaller, the needs and abilities are more heterogeneous and the user groups are smaller. A comprehensive stimulation programme supported by government, industry and academia would contribute to reducing the barrier of fragmentation that clearly characterizes this field and to making industry, politicians and users aware of the new market opportunities offered by this sector.

6. Acknowledgements

The authors would like to thank the other members of the MATRIX working group: Lilian Beijer, Vincent de Jong, Hugo Van hamme, and Emiel Krahmer, and also Toni Rietveld for their contribution.

7. References

[1] Pogson, G. (2005). Language Technology for a Mid-Sized Language, Part I. Multilingual Computing & Technology, 16(6), 43-48.

[2] Pogson, G. (2005). Language Technology for a Mid-Sized Language, Part II. Multilingual Computing & Technology, 16(7), 29-34.

[3] Rietveld, A. and Stolte, I. (2005) Taal- en spraaktechnologie en communicatieve beperkingen. Nederlandse Taalunie, Den Haag. http://taalunieversum.org/taal/technologie

[4] Cucchiarini, C. and D'Halleweyn, E. (2002) How to HLT-enable a Language: the Dutch-Flemish experience, http://www.hltcentral.org/htmlengine.shtml?id=996

[5] Strik, H., Daelemans, W., Binnenpoorte, D., Sturm, J., De Vriend, F., Cucchiarini, C. (2002) Dutch HLT resources: From blark to priority lists. Proc. of ICSLP-2002, Denver, USA, pp. 1549-1552.

[6] Binnenpoorte, D., De Vriend, F., Sturm, J., Daelemans, W., Strik, H., Cucchiarini, C. (2002) A Field Survey for Establishing Priorities in the Development of HLT Resources for Dutch, Proceedings of LREC2002, Gran Canaria de Las Palmas, Spain.

[7] Daelemans, W. and Strik, H. (2002) Het Nederlands in taal- en spraaktechnologie: prioriteiten voor basistaalvoorzieningen; een rapport in opdracht van de Nederlandse Taalunie..

[8] Akkermans, J., van Berkel, B. Frowein, C., van Groos, L., Van Compernolle, D. (2004) Technologieverkenning Nederlandstalige Taal- en Spraaktechnologie, een rapport in opdracht van het Ministerie van Economische Zaken.

[9] Cucchiarini C. & D'Halleweyn E., (2004), The new Dutch-Flemish HLT programme: a concerted effort to stimulate the HLT sector, Proceedings LREC 2004.

[10] D'Halleweyn E., Odijk, J., Teunissen, L. and Cucchiarini, C., (2006), The Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources , Proceedings LREC 2006, pp. 761-766.

[11] http://www.modemadvies.be/

HUMAN LANGUAGE TECHNOLOGIES FOR SPEECH THERAPY IN SPANISH LANGUAGE

Carlos Vaquero, Oscar Saz, W.-Ricardo Rodríguez, Eduardo Lleida

Communications Technology Group (GTC),
I3A, University of Zaragoza, Zaragoza, Spain
{cvaquero, oskarsaz, wricardo, lleida}@unizar.es

Abstract

This paper introduces Vocaliza, an application for computer-aided speech therapy in Spanish language based on the use of Human Language Technologies (HLT). The objective of this application is to help the daily work of the speech therapists that train the linguistic skills of Spanish speakers with different speech impairments, working at three levels of language: phonological, semantic and syntactic. Furthermore, Vocaliza is designed to enable those who suffer speech disorders to train their communication capabilities in an easy and entertaining way, with little or no supervision once a speech therapist has configured the application for the impairment of the user.

The HLT systems used in the application are Automatic Speech Recognition (ASR), speech synthesis, speaker adaptation and utterance verification. The ability of these technologies, namely ASR and speaker adaptation, to actually help users to improve their language is shown by means of the accuracy of the ASR system to detect correct and incorrect utterances according to a manual labeling of a recently acquired database containing impaired speech. The results show that accuracy reaches 87.66% when using speaker adaptation, due to its ability to model the inter speaker variability of every speaker but not their pronunciation errors.

Index Terms: Human Language Technology, speech therapy, Spanish language

1. Introduction

Recently, the demand for computer-aided speech therapy software has increased as computer technologies were getting more reliable and affordable to speech therapists and people suffering speech impairments. The most popular of these systems has been SpeechViewer by IBM, but the non existence of a version for the Spanish language and its lack of modularity made it very uncomfortable for speech therapists in Spain to use on a regular basis.

In terms of research work, during the last decade many European projects related to Human Language Technology (HLT) and speech therapy such as Orto Logo-Paedia [1], SPECO [2], ISAEUS [3] and HARP [4] have been carried out, some of them resulting in the development of software applications for speech therapy at the end of the research process. However, there are no versions of these softwares available in Spanish language, so the applications developed in these projects can not be used by speakers and speech therapists to train communication skills in this language. Due to that, the Aragon Institute for Engineering Research (I3A) with the collaboration of experts in pedagogy and speech therapy from the Public School for Special Education Alborada has developed a research work which aims to provide speech technologies as a tool to aid speech impaired and handicapped people, obtaining as a result a software application for speech therapy in Spanish language, which is free to distribute. This article, which explains the work carried out to obtain the application, is organized as follows: section 2 describes the objectives that are set to the development of a computer-aided speech therapy software in Spanish language. In section 3, there is a wide description of the application architecture while section 4 explains the experiments and results carried out to validate the application. Finally the conclusions to this work are explained in section 5.

This work has been supported by the national project TIN-2005-08660-C04-01 from MEC of the Spanish government.

2. Objectives and Requirements

The objective of this work was the development of a free distribution software application for speech therapy in Spanish language.

For this purpose, the collaboration of experts in speech therapy and pedagogy is strongly necessary. This work has counted on the assistance of the staff of the Public School for Special Education Alborada, located in Zaragoza, Spain, which is a Reference Center for Technical Aids and Communication appointed by the Regional Government of Aragon. Their knowledge in different fields of work with disabled children was essential for setting the application requirements prior to the start of the work and for reaching the objectives of this work, as they had been tracking the whole application development process.

The requirements set for the application can be separated from four points of view:

- In terms of linguistic levels, the application should train several levels of language, from phonological level to semantic and syntactic levels, in order to work on a wide range of speech imparements.
- Regarding application usability, the application should provide enough flexibility for speech therapists to work on different speech impairments, while methods used to treat these impairments should be amusing to attract end users (mainly children).
- The application should have a modular way of dealing with the users, this is, information about every user speech impairments and most suitable methods to train user speech should be stored in order to enable speech therapist to work with different users in an easy way.
- The application should be easy to use, as speech therapists and speech impaired people may not be used to work with computers.

All these requirements were taken into account for the final development of the application, whose given name was Vocaliza.

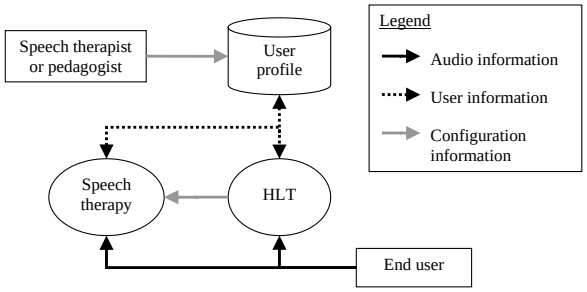


Figure 1: *Block diagram of Vocaliza.*

3. Vocaliza Architecture

Vocaliza architecture can be summed up in a block diagram as shown in Fig. 1. Blocks exchange audio information (solid black arrow), user information (dotted black arrow) and configuration information (solid grey arrow). As shown in Figure 1, the application must be configured previously by a speech therapist, to obtain the desired operation, and after that, the end user, which will be a speech impaired person, will be able to use the application with little or without supervision. Every block functionality is explained next.

3.1. Speech Therapy

The main purpose of Vocaliza is to provide methods for improving user communication skills. The application trains three levels of language, namely phonological, syntactic and semantic levels. Each level is trained by a different method which is shown as a game, in order to attract young users.

Phonological level is trained forcing the user to utter a set of words previously selected by a speech therapist during a configuration procedure. These words are selected to focus on every user specific speech impairment. The application evaluates every utterance and displays a mark with an animated motion on the screen, that the user will be able to understand easily.

Syntactic level is trained forcing the user to utter a set of sentences, previously selected by a speech therapist. Again, the application will evaluate user utterances to display a mark, showing user improvement.

Semantic level is trained by means of a set of riddles, previously defined by a speech therapist. The application ask a question to the user and gives three possible answers. The user must utter the correct answer to go on with the next riddle. The application will show again a mark depending on the user ability to solve the riddle.

All games are based on Automatic Speech Recnognition (ASR), which will decide if the word or sentence uttered by the user is the one the application was expecting.

Fig. 2 shows a screen shot of the main window of Vocaliza. In this window, every game is represented as a picture in order to enable the user to access the desired game easily.

3.2. User Profile

User profile stores all information regarding user configuration, including all words, riddles and sentences selected by a speech therapist to train user speech, as well as all utterances recorded by the user or all speaker dependent acoustic models. This provides flexibility and modularity so that speech therapists will be able to work with different patients fast and easily, merely loading the user profile in the application.

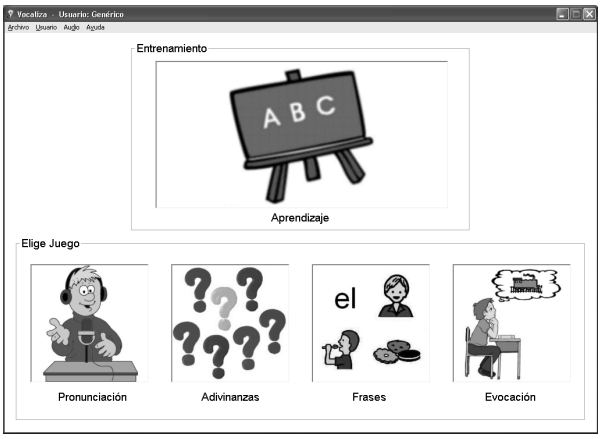


Figure 2: *Main window of Vocaliza.*

3.3. HLT in Vocaliza

Most of Vocaliza functionalities are provided by different HLTs, which are explained next.

ASR constitutes the core of the application. Speech therapy games need ASR to decode user utterances, and to decide which word sequence has been pronounced so that the application will be able to let the user know if the game has been completed successfully.

The ASR system used is based on Hidden Markov Models (HMM). Speech signals are acquired with a sampling frequency of 16 kHz and a bit depth of 16 bits. Signals are windowed with a Hamming window of 25 ms length, with an overlap of 15 ms, and the features used for the ASR are 37 MFCC (Mel Frequency Cepstral Coefícients), consisting on 12 static parameters, 12 delta parameters, 12 delta-delta parameters and the delta-logenergy. The acoustic model is composed by a set of 822 context dependent units plus a silence model and an interword model for a total set of 824 units. Every unit is modeled with 1 state per model and a 16-Gaussian mixture for every state.

Speech synthesis provides a way to show the user how a word or sentence should be pronounced, which is useful in speech therapy games. As soon as a speech therapist adds a new word, sentence or riddle to the application, it is able to synthesize a correct Spanish utterance of the corresponding word, sentence or question. However, speech synthesis may be a very strict method to teach the user how to pronounce a word or a sentence, thus, to provide flexibility, Vocaliza allows speech therapists to record word, riddle, and sentence utterances, which the application will use instead of speech synthesis, in order to show different utterances depending on user age, speech impairments and other requirements of the user.

Speaker adaptation enables the application to estimate speaker dependent acoustic models adapted to each user. Vocaliza uses Maximum A Posteriori (MAP) estimation [5] which, given a speaker independent acoustic model and a set of user utterances, can estimate a speaker dependent acoustic model, adapted to the user. MAP is a well known and reliable estimation method which does not require a great number of utterances to retrieve a reliable acoustic model adapted to the user. This is a very interesting feature since the application will estimate acoustic models from a set of utterances recorded by the user, which in most cases will consist of a small number of utter-

ances due to two factors: speech therapists can not spend long time recording speech of every user, and users with speech impairments will find very hard and tiring to record a great amount of speech utterances. Moreover, MAP estimation convergence make this method a very interesting one when the number of utterances is a priori unknown.

Speaker adaptation is strongly necessary in this application since impaired speech can reduce dramatically ASR systems performance, so that users suffering severe speech impairments would not be able to train their speech with this application using speaker independent models.

Utterance Verification (UV) is a technique embedded in the application to provide a mechanism to evaluate the improvement of user communication skills. Vocaliza uses a Likelihood Ratio (LR) based UV [6] procedure to assign a measure of confidence to each hypothesized word in an utterance. This procedure gives the confidence measure as the ratio of the target hypothesis acoustic model likelihood with respect to an alternate hypothesis acoustic model likelihood. Choosing suitable acoustic models as target and alternate hypothesis can provide a measure of confidence which quantifies improvement in user speech. To achieve this, the application uses a speaker independent acoustic model, which is assumed to model correct speech, as target hypothesis, and a speaker dependent acoustic model, which is assumed to be adapted to impaired speech, as alternate hypothesis. Therefore, this measure of confidence involves a relative evaluation method to quantify improvement of user communication skills.

4. Evaluation and Results

To evaluate the performance of the HLT used in this application and how it can be used to improve the language abilities of the user, a set of recordings in an actual environment of use were made for testing.

4.1. Database acquisition

These recordings form a database of impaired speech recorded from 14 young speakers ranging from 11 to 21 years old, 7 boys and 7 girls. The recordings were made in the same school facilities they are attending currently, and acquired via a wireless close-talk microphone connected to a laptop where the audio capture feature of Vocaliza was used to store the signals. The Signal-to-Noise Ratio (SNR) in the signals is 26.35 dBs, an optimal value for the correct operation of the application and for the evaluation of the system. The 14 speakers suffer from different physical and mental disabilities that affect in several ways their speech and language abilities.

The vocabulary used for the recordings was a set of 57 words gathered in the "Registro Fonológico Inducido" (RFI) [7]. This set of words is a common tool in the community of speech therapy in Spain for the diagnosis of speech disorders as it contains all the phonemes and most of the allophones in Spanish language as well as different combinations of them. The average length of the words is 5,22 phonemes per word. Every speaker recorded four sessions and uttered these 57 words once in every session. Hence, the total number of utterances of isolated words in the database is finally 3,192 utterances. Sessions were recorded on different days to be more realistic with the presence of intra speaker variability among sessions.

During the recordings of the impaired speech database, another database containing speech from children in the range from 11 to 18 years old without any kind of disability was recorded. This database stores speech from 168 speakers to be used as a reference of the speech in the age range of the impaired speakers. Every speaker recorded one session of the 57 words leading to a total number of utterances of 9,576. The same recording process was used for this database and the average SNR is 25.59 dBs.

Training	ML	MAPref	MAPspk
WER	52.22%	31.45%	16.07%

Table 1: *WER results for different acoustic models. ML stands for the adult acoustic model. MAPref stands for the infants adapted model. MAPspk stands for the speaker dependent models.*

For evaluation purposes, an annotation of the corpus was made to know which utterances contain pronunciation errors. This manual annotation has been carried out by a group of independent labelers. Every phoneme in every word of the corpus has been labeled as correct or incorrect by three different judges, and has been finally labeled as correct or incorrect considering the majority of votes of the three judges. In this annotation, a 17.41% of the phonemes in the impaired speech corpus have been labeled as incorrectly uttered.

4.2. Results in ASR

A set of experiments in ASR were carried out over this database with the same specifications of the ASR system and HMM acoustic modeling that are used in the speech therapy application as shown in Section 3.3. Results are shown in Table 1.

The first acoustic model was obtained via the Maximum Likelihood (ML) algorithm from the utterances contained in the databases SpeechDat-Car and Albayzin containing adult Spanish speech and it is the same model used in the Vocaliza application. The baseline results for the 14 speakers give an average Word Error Rate (WER) of 52.22%. Variability in the results among speakers is high, as the speaker with the worst results obtains a 89.47% in WER, while the speaker with the best results obtains a 13.16% in WER. This variability is related to the deep variability in their kinds and degrees of impairment.

A second acoustic model adapted to infants speech was trained over the non-impaired speech database via MAP adaptation [5] and tests were carried out over this model with the impaired speech database. The results obtained show a decrease in WER to an average 31.45% among the speakers. Although not initially in the application, this model could be easily included in the application as a way to reduce recognition errors without the use of speaker adaptation when enough data is not available.

Finally, a set of experiments with the MAP algorithm for speaker adaptation implemented in the application was carried out. In this case, a strategy of leave-one-out was taken; this is, every one of the four sessions of every speaker was used for testing a model trained with the speech in the three remaining sessions of the speaker. Average WER among all the speakers in this experiment drops to 16.07%.

4.3. Accuracy detection results

At this moment, a evaluation of the ability of the ASR system has been made. A reduction of the WER has been proved by the use of speaker dependent models. Regarding the speech therapy application, the improvement in the performance of the ASR system avoids the user getting frustrated of been rejected even when he utters perfectly the word. But the objective of the application is to help people with disabilities to correct their pronunciation errors. Because of this, a WER of 0% would be of no use if the speaker is making a great number of mistakes, as the system would not be helping him to correct them.

# of Incorrect phonemes	1	2	3
Words labeled as incorrect	47.72%	21.87%	10.59%
Accuracy (ML)	69.96%	63.66%	56.64 %
Accuracy (MAPref)	71.31%	77.91%	75.47%
Accuracy (MAPspk)	65.41%	87.66%	86.94%

Table 2: *Correct/incorrect detection accuracy for different acoustic models. ML stands for the adult acoustic model. MAPref stands for the infants adapted model. MAPspk stands for the speaker dependent models.*

Thus, a new measure is needed to know the real performance of the speech therapy system. This measure has to tell the accuracy of the system to discriminate correct pronunciations from incorrect pronunciations. For this purpose a traditional way to measure accuracy in acceptation/rejection systems was taken. In this case, the accuracy (Acc) was considered as the number of mispronunciations not recognized (TN) (this is, the system correctly does not accept them as pronunciations of the given word) plus the number of correct utterances recognized (TP) (the system accepts them as correct) divided by the total number of utterances (U).

$$Acc = \frac{TP + TN}{U} \quad (1)$$

Although it is clear to define when a phoneme is mispronounced, it is not so clear the definition of a mispronounced word, specially in the case of ASR. Considering a word as incorrectly uttered when at least one phoneme is mispronounced would make 47.72% of the words in the speech impaired database fit that definition. If at least two mispronounced phonemes are required to consider a whole word as incorrect, that would lead to a 21.87% of incorrect words. Furthermore, taking at least three mispronounced phonemes to consider a word as incorrect would mean a 10.59% of mispronounced words. Going further in this classification (at least four incorrect phonemes in a incorrect word) is a not significant case as less than 6% of the words would fit as incorrect and there are some words with only three phonemes who would never be mispronounced.

Table 2 shows the number of incorrect words according to every possible definition of what a mispronounced word is. Also, it shows the results in the accuracy of the ASR system to recognize or not the uttered words according to their condition as correctly or incorrectly pronounced words. Results show that considering that a mispronounced word contains at least two incorrect phonemes the accuracy rises to 87% when using speaker dependent models.

4.4. Discussion

Relevant discussion can be made out of the results achieved. The most relevant is the fact of how accuracy rises from 60% to 76% just by changing the adult speaker independent model to an infants speaker independent model; and then to a 87% by using speaker dependent models. This shows that the use of speaker dependent models really is important for a better operation of the system. This is due to the fact that the speaker adaptation process obtains a speaker dependent model that eliminates the errors due to the inter speaker variability in the recognition phase, but speaker adaptation by itself is unable to learn the strong mispronunciations of the speaker, so those errors stay in the system.

However, these results achieved for the cases in which a minimum of two or three incorrect phonemes are set for a word to be considered mispronounced do not follow the same trend for the case of a minimum of one incorrect phoneme. This is due to the chosen vocabulary, the 57 words of the RFI, that does not provide words with a high confusability. That is to say, every pair of words in the vocabulary differs in at least two phonemes (and usually in more than three phonemes). Because of this, when only a phoneme is mispronounced in a given word the system keeps recognizing that word as it is still the closest word in the vocabulary.

5. Conclusions

As a result of this work, a totally functional application which aims to help the work of the speech therapists in three levels of the language (phonological, semantic and syntactic) has been developed. The software is ready to be distributed at this moment, and is free to use for every speech therapist who may require it.

All the requirements set at the beginning of the work have been completely fulfilled. Furthermore, AAC methods embedded in Vocaliza provide added value to the application, making possible to use it as a educational software, not only for people with communications disorders but for people with cognitive disorders or even for young people without disorders.

Also, it has been shown how the HLT included in the application really can help the user to effectively detect pronunciation errors and correct them. Particularly, the use of speaker adaptation rises the accuracy rate of the system according to a manual labeling from 60% to 87%.

6. Acknowledgements

The authors want to thank Prof. Richard C. Rose from McGill University in Montréal (Canada) for his fruitful discussion and ideas that have lead to some of the results in this work.

7. References

[1] Protopapas A. Öster A-M, House D. and Hatzis A., "Presentation of a new EU project for speech therapy: Olp (ortho-logo-paedia)", TMH-QPSR vol. 44, Fonetik, 2002.

[2] Öster A. Kacic Z. Barczikay Z. Vicsi K., Roach P. and Sinka I., "Speco — a multimedia multilingual teaching and training system for speech handicapped children", Eurospeech, 6th Conference on Speech Communication and Technology, Interspeech, 1999.

[3] García Gómez et al., "Isaeus speech training for deaf and hearing-impaired people," Tech. Rep., Eurospeech, 6th Conference on Speech Communication and Technology, Interspeech, 1999.

[4] "Harp — an autonomous speech rehabilitation system for hearing impaired people," Final report, HARP (TIDE project 1060), May 1996.

[5] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains", IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291–298, 1994.

[6] E. Lleida and R.C. Rose, "Utterance verification in continuous speech recognition: Decoding and training procedures", IEEE Transactions on Speech and Audio Processing, vol. 8, no. 2, pp. 126–139, 2000.

[7] Monfort M., Juárez-Sánchez A., "Registro Fonológico Inducido (Trajetas Gráficas)", Ed. Cepe, Madrid, 1989.

Legal Taxonomy Syllabus: Handling Multilevel Legal Ontologies

Gianmaria Ajani[1], *Guido Boella*[2], *Leonardo Lesmo*[2],
Alessandro Mazzei[2], *Daniele P. Radicioni*[2], *Piercarlo Rossi*[3]

[1]Dipartimento di Scienze Giuridiche, Università di Torino - Italy
[2]Dipartimento di Informatica, Università di Torino - Italy
[3]Dipartimento di Studi per l'Impresa e il Territorio, Università del Piemonte Orientale - Italy
gianmaria.ajani@unito.it, {boella,lesmo,mazzei,radicion}@di.unito.it,
piercarlo.rossi@eco.unipmn.it

Abstract

The Legal Taxonomy Syllabus methodology has been used to represent legal information at different levels such, e.g., European Directives, and their transpositions into national legislations. In this paper we point out the main issues of this approach, and extend it to account for a further level, the Acquis Principles level.
Index Terms: Formal Ontologies, European Directives, Legal Drafting Support.

1. Introduction

The European Union produces each year a large number of Union Directives (EUD), which are translated into each of the Member States' languages. The EUD are sets of norms that have to be implemented by the national legislations. The problem of multilingualism in European legislation has recently been addressed by using linguistic and ontological tools, e.g. [1, 2, 3]. The management of EUD is particularly complex since the implementation of a EUD does not correspond to the straight transposition into a national law. An EUD is subject to further interpretation, and this process can lead to unexpected results. Comparative Law has studied in details the problems concerning EUD and their complexities. On the other hand managing with appropriate tools this kind of complexity can facilitate the comparison and harmonization of national legislation [1]. Based on this research, in this paper we describe a tool for building multilingual conceptual dictionaries that we developed for representing and analysing terminologies and concepts used in EUD. We also point out some recent advances of the Legal Taxonomy Syllabus[1] (LTS) that have been designed to generalize our methodology to a broader set of legal contexts.

The main assumptions of our methodology are motivated by studies in comparative law [4] and ontologies engineering [5] and they can be listed as follows:

1. Terms and concepts must be distinguished; for this purpose, we use lightweight ontologies, i.e. simple taxonomic structures of primitive or composite terms together with associated definitions. They are hardly axiomatized as the intended meaning of the terms used by the community is more or less known in advance by all members, and the ontology can be limited to those structural relationships among terms that are considered as relevant [6][2].
2. We distinguish the ontology implicitly defined by EUD, the *EU level*, from the various national ontologies. Each one of these "particular" ontologies belongs to the *national level*: i.e., each national legislation refers to a distinct national legal ontology. We do not assume that the transposition of an EUD automatically introduces in a national ontology the same concepts that are present at the EU level.
3. Corresponding concepts at the EU level and at the national level can be denoted by different terms in the same national language.

Another feature of European law has to be taken into account in knowledge engineering: the *Acquis communautaire*, the existing body of EU primary and secondary legislation as well as the ECJ decisions. Nowadays the *acquis* comprising 80,000 pages around. However, notwithstanding the importance of this existing body of settled laws, the Acquis is also a far wider notion, encompassing an impressive set of principles and obligations, going far beyond the internal market and including areas, such as, agriculture, the environment, energy and transport. Today, some areas of Acquis are in the way to be consolidated in order to ensure greater coherence in their implementation in the Member States and their interpretation by courts. In February 2003, the European Commission published an Action Plan aimed at achieving greater coherence in European contract law by adopting a non-binding Common Frame of Reference (CFR) [7].

In this paper we show how LTS is used to build a dictionary of Consumer Law within the broader scope of the Uniform Terminology Project[3] [4]. The paper is structured as follows. In Section 2 we stress two main problems raised in comparative law as regards as EUD and their transpositions. In Section 3 we describe how the methodology of the LTS allows to cope with these problems. In section 4 we describe the *Acquis level* and illustrate how the LTS can be enriched to account for the Acquis Principles [8], as well. Finally in Section 5 we provide some conclusions and elaborate about future works.

2. Terminological and conceptual misalignment

Comparative Law has identified two key points in dealing with EUD, which make more difficult dealing with the polysemy of

[1] The current version of our system can be found at the address: www.eulawtaxonomy.org.

[2] See http://cos.ontoware.org/

[3] http://www.uniformterminology.unito.it

legal terms: we call them the *terminological* and *conceptual misalignments*.

In the case of EUD (usually adopted for harmonising the laws in the Member States), the terminological matter is complicated by the need to implement them in the national legislations. In order to have a precise transposition in a national law, a Directive may be subject to further interpretation. Thus, a *legal concept* can be expressed in different ways in a Directive and in its implementing national law. A single concept in a particular language can be expressed in a number of different ways in a EUD and in the national law implementing it. As a consequence we have a terminological misalignment. For example, the concept corresponding to the word *reasonably* in English, is translated into Italian as *ragionevolmente* in the EUD, and as *con ordinaria diligenza* into the transposition law.

In the EUD transposition laws a further problem arises from the different national *legal doctrines*. A legal concept expressed in an EUD may not be present in a national legal system. In this case we can talk about a conceptual misalignment. To make sense for the national lawyers' expectancies, the European legal terms have not only to be translated into a sound national terminology, but they also need to be correctly detected when their meanings are to refer to EU legal concepts or when their meanings are similar to concepts which are known in the Member states. Consequently, the transposition of European law in the parochial legal framework of each Member state can lead to a set of distinct national legal doctrines, that are all different from the European one. In case of consumer contracts (like those concluded by the means of distance communication as in Directive 97/7/EC, Art. 4.2), the notion to provide in a *clear and comprehensible manner* some elements of the contract by the professionals to the consumers represents a specification of the information duties which are a pivotal principle of EU law. Despite the pairs of translation in the language versions of EU Directives (i.e., *klar und verständlich* in German - *clear and comprehensible* in English - *chiaro e comprensibile* in Italian), each legal term, when transposed in the national legal orders, is influenced by the conceptual filters of the lawyers' domestic legal thinking. So, *klar und verständlich* in the German system is considered by the German commentators referring to three different legal concepts: 1) the print or the writing of the information must be clear and legible (*gestaltung der information*), 2) the information must be intelligible by the consumer (*formulierung der information*), 3) the language of the information must be the national of consumer (*sprache der information*). In Italy, the judiciary tend to control more the formal features of the concepts 1 and 3, and less concept 2, while in England the main role has been played by the concept 2, though considered as plain style of language (not legal technical jargon) thanks to the historical influences of plain English movement in that country.

Note that this kind of problems identified in comparative law has a direct correspondence in the ontology theory. In particular Klein [5] has remarked that two particular forms of ontology mismatch are *terminological* and *conceptualization* ontological mismatch which straightforwardly correspond to our definitions of misalignments.

3. The methodology of the Legal Taxonomy Syllabus

A standard way to properly manage large multilingual lexical databases is to do a clear distinction among terms and their interlingual acceptions (or *axies*) [9, 10]. In our system to properly manage terminological and conceptual misalignment we distinguish in the LTS project the notion of legal term from the notion of legal concept and we build a systematic classification based on this distinction. The basic idea in our system is that the conceptual backbone consists in a taxonomy of concepts (ontology) to which the terms can refer to express their meaning. One of the main points to keep in mind is that we do not assume the existence of a single taxonomy covering all languages. In fact, it has been convincingly argued that the different national systems may organize the concepts in different ways. For instance, the term *contract* corresponds to different concepts in common law and civil law, where it has the meaning of *bargain* and *agreement*, respectively [11]. In most complex instances, there are no homologous between terms-concepts such as *frutto civile* (legal fruit) and *income*, but respectively civil law and common law systems can achieve functionally same operational rules thanks to the functioning of the entire taxonomy of national legal concepts [12]. Consequently, the LTS includes different ontologies, one for each involved national language plus one for the language of EU documents. Each language-specific ontology is related via a set of *association* links to the EU concepts, as shown in Fig. 1.

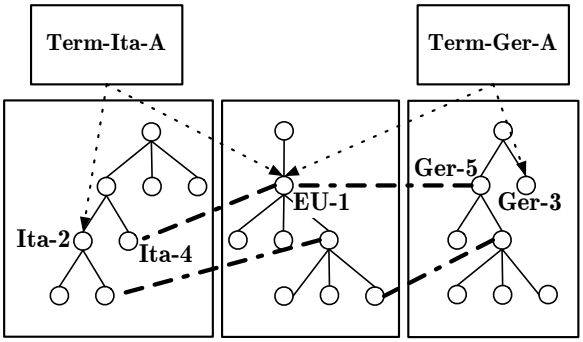


Figure 1: Relationship between ontologies and terms. The thick arcs represent the inter-ontology "association" link.

Although this picture is conform to intuition, in LTS it had to be enhanced in two directions. First, it must be observed that the various national ontologies have a reference language. This is not the case for the EU ontology. For instance, a given term in English could refer either to a concept in the UK ontology or to a concept in the EU ontology. In the first case, the term is used for referring to a concept in the national UK legal system, whilst in the second one, it is used to refer to a concept used in the European directives. This is one of the main advantages of LTS. For example *klar und verständlich* could refer both to concept `Ger-379` (a concept in the German Ontology) and to concept `EU-882` (a concept in the European ontology). This is the LTS solution for facing the possibility of a correspondence only partial between the meaning of a term has in the national system and the meaning of the same term in the translation of a EU directive. This feature enables the LTS to be more precise about what "translation" means. It puts at disposal a way for asserting that two terms are the translation of each other, but just in case those terms have been used in the translation of an EU directive: within LTS, we can talk about direct EU-national translations of terms, but only about *implicit* national-system translations of terms. In other words, we distinguish between

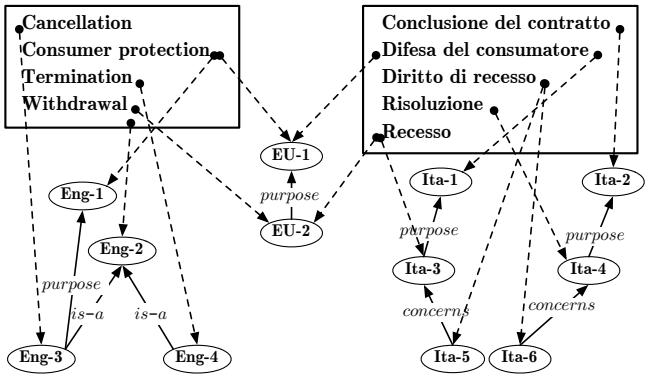


Figure 2: An example of interconnections among terms.

explicit and *implicit* associations among concepts belonging to different levels. The former ones are direct links that are explicitly used by legal experts to mark a relation between concepts. The latter ones are indirect links: if we start from a concept at a given national level, by following a direct link we reach another concept at European level. Then, we will be able to see how that concept is mapped to further concepts at the various national levels.

The situation enforced in LTS is depicted in Fig. 1, where it is represented that, the Italian term *Term-Ita-A* and the German term *Term-Ger-A* have been used as corresponding terms in the translation of an EU directive, as shown by the fact that both of them refer to the same EU-concept `EU-1`. In the Italian legal system, *Term-Ita-A* has the meaning `Ita-2`. In the German legal system, *Term-Ger-A* has the meaning `Ger-3`. The EU translations of the directive is correct insofar no terms exist in Italian and German that characterize precisely the concept `EU-1` in the two languages (i.e., the "associated" concepts `Ita-4` and `Ger-5` have no corresponding legal terms). A practical example of such a situation is reported in Fig. 2, where we can see that the ontologies include different types of arcs. Beyond the usual *is-a* (linking a category to its supercategory), there are also a *purpose* arc, which relates a concept to the legal principle motivating it, and *concerns*, which refers to a general relatedness. The dotted arcs represent the reference from terms to concepts. Some terms have links both to a National ontology and to the EU Ontology (in particular, *withdrawal* vs. *recesso* and *difesa del consumatore* vs. *consumer protection*).

The last item above is especially relevant: note that this configuration of arcs specifies that: 1) *withdrawal* and *recesso* have been used as equivalent terms (concept `EU-2`) in some European Directives (e.g., Directive 90/314/EEC). 2) In that context, the term involved an act having as purpose the some kind of protection of the consumer. 3) The terms used for referring to the latter are *consumer protection* in English and *difesa del consumatore* in Italian. 4) In the British legal system, however, not all *withdrawals* have this goal, but only a subtype of them, to which the code refers to as *cancellation* (concept `Eng-3`). 5) In the Italian legal system, the term *diritto di recesso* is ambiguous, since it can be used with reference either to something concerning the *risoluzione* (concept `Ita-4`), or to something concerning the *recesso* proper (concept `Ita-3`).

Finally, it is possible to use the LTS to translate terms into different national systems via the transposed concepts at the European level, i.e. by using the implicit associations. For instance suppose that we want to translate the legal term *credito al consumo* from Italian to German. In the LTS *credito al consumo* is associated to the national umeaning `Ita-175`. We find that `Ita-175` is the transposition of the European umeaning `EU-26` (*contratto di credito*). `EU-26` is associated to the German legal term *Kreditvertrag* at European level. Again, we find that the national German transposition of `EU-26` corresponds to the national umeaning `Ger-32` that is associated with the national legal term *Darlehensvertrag*. Then, by using the European ontology, we can translate the Italian legal term *credito al consumo* into the German legal term *Darlehensvertrag*.

4. Work in progress: adding further levels

One major feature of the LTS approach relies on distinguishing legal information as belonging to different levels. At the current stage of development, the system manages terms and meanings at both EU and national levels. The former one is an ontology of legal concepts derived from the EUD; the latter one includes national legal ontologies coming from the various national legal systems. It is worth emphasizing that not only the current approach is general enough to account for heterogeneous legal sources (like, e.g., EUD and "Decreti Legislativi" for European and Italian national levels respectively), but also it be generalised by adding further levels.

To add a level into the system, we link the new legal ontology to that in one of the existing levels. Linking a new ontology means that we define *explicit* associations between concepts in the new ontology and concepts in an ontology from an existing level.

Moreover, the EC Commission on Common Frame of Reference should provide common principles, terminology, and rules for contract law to address gaps, conflicts, and ambiguities emerging from the application of European contract law. In drafting the Action Plan the Commission emphasized that the CFR would eliminate market inefficiencies arising from the diverse implementation of European directives, providing a solution to the non-uniform interpretation of European contract law due to vague terms and rules, now present in the existing Acquis. In particular, two issues arise out from the vague terminology of EUD. First, directives adopt broadly defined legal concepts, therefore leaving too much discretion in their implementation to national legislators or judges. Second, directives introduce legal concepts that are different from national legal concepts. Thus, when judges face vague terms, they can either interpret them by referring to the broad principles of the acquis communautaire, or they can refer to the particular goals of the directive in question.

To respond to the Action Plan, in the last few years, within the general framework of a "Network of Excellence" EU Project, a research group aiming at consolidating the existing EC law is working on the "Principles of the Existing EC Private Law" or "Acquis Principles" (ACQP). These Principles will be discussed and compared with other outcomes from different European research groups, such as Von Bar or Lando group, and, during a complex process of consultation with stakeholders under the direction of EC Commission, the CFR will be set up. The Acquis Principles should provide a common terminology as well as common principles to constitute a guideline for uniform implementation and interpretation of European law [13, 14].

One outcome of such project is the Acquis Principles glossary, i.e., a set of interconnected terms and concepts. We introduce the *Acquis level* into the LTS by defining explicit associations between Acquis Principles concepts and EU-level con-

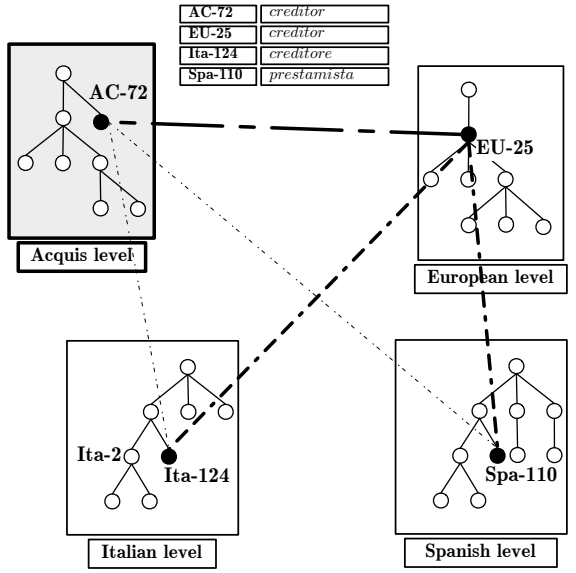


Figure 3: LTS augmented with the Acquis level. Thick lines indicate *explicit* associations; thin lines indicate *implicit* associations.

cepts.

For example, in Figure 3 we have that the concept `EU-25` (corresponding to the English legal term *creditor*) present in a EUD is explicitly associated to the national legal concepts `Ita-124` (*finanziatore*) and `Spa-110` (*prestamista*) for Italian and Spanish, respectively. We now add the term *creditor* from the Acquis Level by inserting an explicit association between the Acquis legal concept `AC-72`. As a consequence, the concept `AC-72` is implicitly associated to the legal concepts `Ita-124` and `Spa-110`.

This fact has deep consequence on the way one can build systems for reasoning, that are allowed to make paths passing through more than two levels, thereby offering new insights (and ready-to-use associations between terms) to scholars in comparative law.

5. Conclusions

In this paper we discuss some features of the LTS, a tool for building multilingual conceptual dictionaries for the EU law. The tool is based on lightweight ontologies to allow a distinction of concepts from terms. Distinct ontologies are built at the EU level and for each national language, to deal with polysemy and terminological and conceptual misalignment.

Many attempts have been done to use ontologies in legal field, e.g. [15, 3] and LOIS project (that is based on EuroWordNet project [16], `http://www.loisproject.org`), but to our knowledge the LTS is the first attempt which starts from fine grained legal expertise on the EUD domain.

The present work illustrates how further levels can be added to the EU and national levels. In particular, we introduced how a novel set of principles (along with a terminology) can be added to the LTSThis work has two main virtues: firstly, legal experts will be allowed to recover information from diverse kinds of data. Secondly, legal reasoning systems will benefit of a framework enriched by new explicit and implicit associations between Acquis and European and national levels.

Future work will address how the LTS can be used as a thesaurus for general EUD, even if the currently implemented version of the LTS knowledge base is limited to EUD concerning consumer law.

6. References

[1] A. Boer, T. van Engers, and R. Winkels, "Using ontologies for comparing and harmonizing legislation." in *ICAIL*, 2003, pp. 60–69.

[2] E. Giguet and P.-S. Luquet, "Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia: Association for Computational Linguistics, July 2006, pp. 271–278.

[3] S. Després and S. Szulman, "Merging of legal micro-ontologies from european directives," *Journal of Artificial Intelligence and Law*, February 2007.

[4] P. Rossi and C. Vogel, "Terms and concepts; towards a syllabus for european private law," *European Review of Private Law (ERPL)*, vol. 12, no. 2, pp. 293–300, 2004.

[5] M. Klein, "Combining and relating ontologies: an analysis of problems and solutions," in *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA, 2001.

[6] D. Oberle, Ed., *Semantic Management of Middleware*. Springer Science+Business and Media, 2005.

[7] European Commission, "Communication from the Commission to the European Parliament and the Council - A More Coherent European Contract Law - An Action Plan," COM, 2003.

[8] R. G. on the Existing EC Private Law, *Principles of the Existing EC Contract Law (Acquis Principles) Contract I*. Sellier. European Law Publishers, 2007.

[9] G. Sérasset, "Interlingual lexical organization for multilingual lexical databases in NADIA," in *Proc. COLING94*, 1994, pp. 278–282.

[10] V. Lyding, E. Chiocchetti, G. Sérasset, and F. Brunet-Manquat, "The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated," in *Proc. of the Wokshop on Multilingual Language Resources and Interoperability, ACL06*, 2006, pp. 25–31.

[11] R. Sacco, "Contract," *European Review of Private Law*, vol. 2, pp. 237–240, 1999.

[12] M. Graziadei, "Tuttifrutti," in *Themes in Comparative Law*, P. Birks and A. Pretto, Eds. Oxford University Press, 2004, pp. –.

[13] G. Ajani and H. Schulte-Nölke, "The Action Plan on a More Coherent European Contract Law: Response on Behalf of the Acquis Group," 2003. [Online]. Available: http://www.acquis-group.org

[14] R. Schulze, "European Private Law and Existing EC Law," *European Review of Private Law (ERPL)*, 2005.

[15] P. Casanovas, N. Casellas, C. Tempich, D. Vrandecic, and R. Benjamins, "OPJK modeling methodology," in *Proceedings of the ICAIL Workshop: LOAIT 2005*, 2005.

[16] P. Vossen, W. Peters, and J. Gonzalo, "Towards a universal index of meaning," in *Proc. ACL-99 Siglex Workshop*, 1999.

META-MultilanguagE Text Analyzer

Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile,
Leo Iaquinta, Pasquale Lops, Giovanni Semeraro

[1]Department of Computer Science, University of Bari, Italy
{basilepp,degemmis,gentile,iaquinta,lops,semeraro}@di.uniba.it

Abstract

Natural Language Processing *(NLP)* has a significant impact on many relevant Web-based and Semantic Web applications, such as information filtering and retrieval. Tools supporting the development of NLP applications are playing a key role in text-based information access on the Web.

In this paper, we present META *(MultilanguagE Text Analyzer)*, a tool for text analysis, designed with the aim of providing a general framework for NLP tasks over different languages. The system implements both basic and advanced NLP functionalities, such as Word Sense Disambiguation. After describing the main ideas behind the architecture of META, we discuss some results about the processing of different corpora in English and Italian. Finally, we show how META has been integrated in a recommender system for content-based information filtering.

Index Terms: Natural Language Processing, Information Filtering, Document Indexing

1. Introduction

A vast portion of the Web consists of text documents, thus methods for automatically analyzing text have great importance in the context of the Web.

Several techniques have been developed within the fields of Information Retrieval (IR) and Information Filtering (IF), and include indexing, scoring, and categorization of textual documents. Filtering and retrieval systems deal with the ranking of textual documents in order of relevance. Retrieval refers to the selection of documents from a fixed set, whereas filtering typically refers to selection of relevant documents from a stream of incoming data. Retrieval systems are generally concerned with satisfying a users one-off information need (query); filtering systems are usually applied to attaining information for a users long term interests (profiles). Categorization or classification of documents is another useful technique, somewhat related to IR and IF, that consists of assigning a document to one or more predefined categories. A classifier can be used, for example, to distinguish between relevant and irrelevant documents (where the relevance can be personalized for a particular user or group of users), or to help in the semiautomatic construction of large Webbased knowledge bases or hierarchical directories of topics like the Open Directory[1].

In this scenario, the development of robust tools for both basic and more complex NLP tasks is becoming crucial. This paper describes META (*MultilanguagE Text Analyzer*), an infrastructure for processing textual documents over different languages. The main features of the proposed tool are:

- The system is designed to clearly separate low-level tasks (such as data storage, location and loading of language resources) from data structures and algorithms.
- The tool provides a baseline set of NLP components (Tokenizer, POS-tagger, ...) that can be extended and modified by the user according to the tasks to be accomplished.
- The architecture was conceived so that language-independent components for both basic and more complex tasks, such as Word Sense Disambiguation, can be easily included.
- Indexing structures produced by the META can be exported in different formats, thus allowing an easy integration with both IR and IF systems.

The rest of the paper is structured as follows. We first describe the META architecture in Section 2, then we provide some detail about document representation in Section 3. Some application scenarios are reported in Section 4, while a brief description of related work is given in Section 5, while conclusions and future work close the paper.

2. System Architecture

The architecture of META is depicted in Fig. 2, in which the three main components of the system are showed:

1. COLLECTION MANAGER - This component provides the tools for the import of documents in different formats (HTML, PDF, DOC,...), allows the user to organize them in collections, and includes algorithms for the segmentation of documents, that is each document is logically viewed as structured in different sections (e.g., a scientific paper can be structured into: *title*, *abstract*, *authors*, *body* and *references*). The COLLECTION MANAGER allows also the annotation of sections with tags stored in a domain ontology;
2. NLP ENGINE - This engine is devoted to the management of different NLP annotators. An annotator is a component that performs a specific NLP task (e.g. tokenization, stop word elimination, POS-tagging). The NLP ENGINE schedules the annotators, loads the lexical resources required for each annotator, and runs the annotator over all the documents into a collection.
3. EXPORT MANAGER - This component is able to export the results carried out by the NLP ENGINE into different formats, according to the user's request (XML, RDF, specific DBMS, ...).

The whole process of document analysis performed by META is described in the following. The COLLECTION MANAGER imports the documents to be processed from the user's

[1]http://dmoz.org/

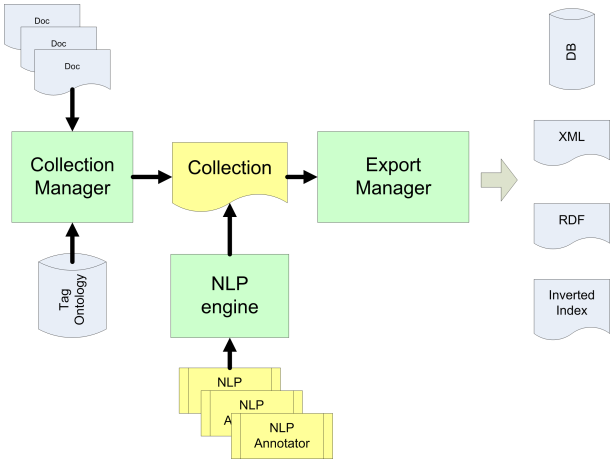


Figure 1: META conceptual architecture

file system (HTML, DOC, RTF, PDF) and groups them in a *collection*. Each document is assigned with a unique identifier (ID) in the collection, then segmentation is performed and the raw text is extracted from the original document. In this stage, it is also possible to associate both collections and single documents with tags stored in a domain ontology. After these preliminary steps, the documents are ready for the next stage performed by the NLP ENGINE.

First, the *NLP Engine* detects the document language; this step is strictly required in order to load the right lexical resources for each language. Then, the NLP engine normalizes (for example, all formatting characters are removed) and tokenizes the text. At this stage, each document is turned in a list of tokens. Each token can be associated with a set of *annotations*. An annotation is a pair *(annotation_name,value)*, which specifies the kind of annotation and the corresponding value (e.g., the position of the token in the text). Annotations are produced by different components called NLP ANNOTATORS, whose scheduling is managed by the NLP ENGINE.

Currently, the following annotators have been developed and included in META:

1. *Stop words elimination*: all commonly used words are deleted;
2. *Stemming*: it is the process of reducing inflected (or sometimes derived) words to their stem. In META, we adopt the *Snowball stemmer* [2];
3. *POS-tagging*: it is the process of assign a part-of-speech to each token. We develop a JAVA version of *ACOPOST tagger* [3] using Trigram Tagger T3 algorithm. It is based on Hidden Markov Models, in which the states are tag pairs that emit words;
4. *Lemmatization*: it is the process of determining the lemma for a given word. We use WordNet Default Morphological Processor (included in the WordNet distribution) for English. For the Italian language, we have built a different lemmatizer that exploits the *Morph-it!* morphological resource [4];
5. *Entity Recognition Driven by Ontologies*: it is the process of finding ontology instances into the text;
6. *Word Sense Disambiguation* (WSD): it is the problem of selecting a sense for a word from a set of predefined possibilities, by exploiting a sense inventory that usually comes from a electronic dictionary or thesaurus. We have implemented a WSD algorithm, called JIGSAW [1], able to disambiguate both English and Italian text.

At the end of the pipeline ran by the NLP ENGINE, the output could be exported in different formats by the EXPORT MANAGER. This component is devoted to turn the internal output produced by META into different formats such as XML or RDF.

3. Document representation

The internal representation of META is a collection that contains a list of documents. Each document is subdivided into segments, each one corresponding to a specific part of the document. Documents are composed by one segment at least. Each segment contains a list of token, each one associated with one annotation at least. An annotation represents a particular feature extracted during text processing (e.g. token, stemming, lemma, entity, sense, ...). The logical structure of a document in depicted in Fig. 3.

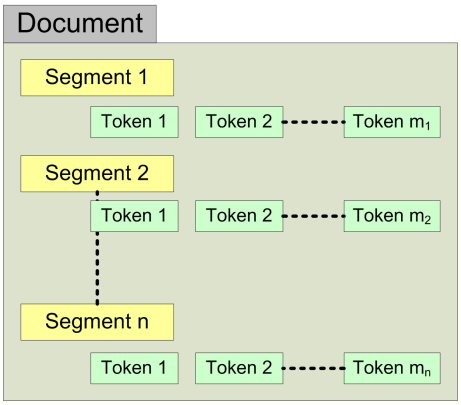


Figure 2: Conceptual document structure

For example, if the user want to analyze the text '*In this paper we present META (MultilanguagE Text Analyzer), it's a tool for text analysis which implements some NLP functionalities.*'. The NLP ENGINE executes the following operations: tokenization, stemming, pos-tagging, lemmatization and WSD. The output of the system is a list of tokens and corresponding annotations. Fig. 3 shows the logical structure for the token *paper*. In particular, the token has a *sense annotation* produced by the WSD annotator, whose value is *n12660433*, the number which identifies the WordNet synset assigned by JIGSAW.

The snapshot of the META GUI that represents the output of the above example is showed in fig. 3. The GUI of the system allows the visualization of the output by using a table format: tokens are represented in rows and annotations in columns. Also, from the GUI it is possible to access EXPORT MANAGER functionalities.

[2] http://snowball.tartarus.org/
[3] http://acopost.sourceforge.net/
[4] http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

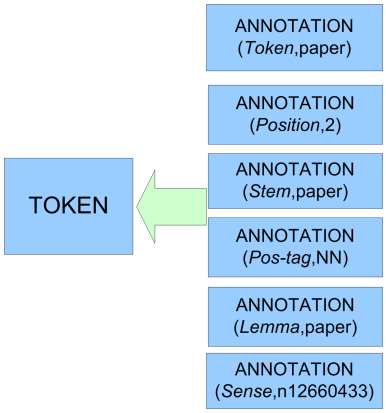


Figure 3: Conceptual token structure

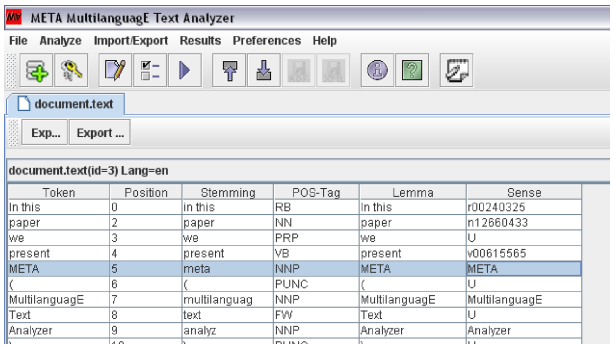


Figure 4: META GUI snapshot

4. META @ Work

META has been employed for the processing collection of documents in different scenarios, in order to evaluate its performance:

1. Disambiguation of a whole collection of document in English;
2. Disambiguation of a whole collection of document in Italian;
3. Indexing of a whole collection of scientific papers for personalized filtering;

In the following, we describe each one of the scenario in which we system was tested.

4.1. WSD on English

JIGSAW, the WSD algorithm included in META, has been tested in the context of SemEval 1-Task 1 competition [2]. This task is an application-driven one, where the application is a fixed Cross-Lingual Information Retrieval (CLIR) system. Participants must disambiguate text by assigning WordNet synsets, then the CLIR system must perform both the expansion to other languages and the indexing of the expanded documents; the final step is the retrieval (in batch) for all the languages. The retrieved results are taken as a measure of the disambiguation accuracy. The dataset consisted of 29,681 documents, including 300 topics (short text). Results are reported in Table 1. Besides the two systems (JIGSAW and PART-B) that participated to SEMEVAL-1 Task 1 competition, a third system (ORGANIZERS), developed by the organizers themselves, was included in the competition. The systems were scored according to standard IR/CLIR measures as implemented in the TREC evaluation package[5].

system	*IR documents*	*IR topics*	*CLIR*
no expansion	0.3599		0.1446
full expansion	0.1610	0.1410	0.2676
1st sense	0.2862	0.1172	0.2637
ORGANIZERS	0.2886	0.1587	0.2664
JIGSAW	0.3030	0.1521	0.1373
PART-B	0.3036	0.1482	0.1734

Table 1: SEMEVAL-1 Task1 Results

All systems showed similar results in IR tasks, while their behaviour was extremely different on CLIR task. Probably, the negative results of JIGSAW in CLIR task depends on complex interaction of WSD, expansion and indexing. Contrarily to other tasks, the task organizers do not plan to provide a ranking of systems on SEMEVAL-1 Task 1. As a consequence, the goal of this task - what is the best WSD system in the context of a CLIR system? - is still open.

4.2. WSD on Italian

An important applications scenario is EVALITA [6], that is an initiative devoted to the evaluation of Natural Language Processing tools for Italian. In this context, we have evaluated META for Italian language. Experiments were performed by using the instructions for EVALITA WSD All-Word-Task. The dataset consisted of about 5000 words. Precision and Recall are reported in Table 2.

SYSTEM	*P*	*R*	*attempted*
UniBa_Basile (JIGSAW)	0.560	0.414	73.95%
1st sense (baseline)	0.669	0.669	100%

Table 2: JIGSAW results on EVALITA All-Words Task

The results are encouraging as regards precision, considering that our system exploits only ItalWordNet as knowledge base. JIGSAW was compared only with the *baseline* (for all words, the first sense in ItalWordNet is selected), which achieves very high results. In Table 3 the precision for each POS-tag is showed. It is possible to notice that the precision is quite acceptable for nouns, and very high for proper nouns because generally they have only a sense. The results show that the verb disambiguation is very hard due to high polysemy. High precision is achieved for adjectives and adverbs, but recall is lower due to POS-tagger errors. The process of WSD requires lemmatization and POS-tagging, which introduce errors, thus influencing the recall. We estimated lemmatization and POS tagging precision respectively to 77,66% and 76,23%. More details are reported in [3].

4.3. META in an Information Filtering Scenario

META has been used as Content Analyzer into a content-based recommender system [4]. The recommender automatically infers the user profile, a structured model of the user interests,

[5] http://trec.nist.gov/
[6] http://evalita.itc.it/

$POS-tag$	P	R	$attempted$
$NOUN$	0,556	0,444	79,96%
$VERB$	0,375	0,283	75,60%
$OTHERS$	0,676	0,321	47,55%
$PROPERNOUN$	0,913	0,724	79,25%

Table 3: JIGSAW results for each POS-tag on EVALITA All-Words Task

from documents that were already deemed relevant by the user. The profile is used to filter new documents and to produce personalized suggestions. We used META in the indexing phase for the extraction of both lexical and semantic features from documents. The learning algorithms embedded in the recommender are able to infer user profiles from the feature produced by META. The system produced both a classical Bag-Of-Word (BOW) document representation and a new representation that we call Bag-of-Synset (BOS). In this model, a document is represented by a vector of WordNet synsets recognized by the WSD procedure.

5. Related Work

The design of META was strongly inspired by GATE-General Architecture for Text Engineering [7] developed by the NLP group at Sheffield University. GATE [5] is an infrastructure for developing and deploying software components that process human language. GATE helps scientists and developers in three ways: by specifying an architecture for language processing software; by providing a framework, or class library, that implements the architecture and can be used to embed language processing capabilities in various applications; by providing a development environment built on top of the framework made up of convenient graphical tools for developing components. The goal of GATE is to enable users to develop and deploy language engineering components and resources in a robust fashion. On the other hand, META is a tool for the management of documents collections, the organization of multi-lingual NLP pipelines, and the storage of processed documents.

The main differences between META and GATE are:

1. META provides powerful tools for both the management of collections and document segmentation and annotation;
2. META provides an NLP pipeline that allows the development of NLP annotators not only for English;
3. META is oriented toward semantic indexing of documents, by making easier the integration of WSD algorithms;
4. META was not designed specifically for information extraction or text mining as GATE, but it is possible to convert the output produced by META in several formats. Therefore, META is prepared for the export also in formats required by external mining tools like WEKA[8].

UIMA [9] is a framework for NLP developed by IBM. The UIMA framework is an open, industrial-strength, scalable and extensible platform for building analytic applications or search solutions that process text or other unstructured information to find the latent meaning, relationships and relevant facts buried within. It enables developers to build analytic modules and to compose analytic applications from multiple analytic providers, encouraging collaboration and facilitating value extraction for unstructured information. UIMA is able to deal with both text and other media format such as videos and images.

In conclusion, META is more useful for NLP tasks that require the indexing of documents and the extraction of semantic features from text.

[7] http://gate.ac.uk
[8] http://www.cs.waikato.ac.nz/ml/weka/
[9] http://uima-framework.sourceforge.net/

6. Conclusion and Future Work

NLP tools has a crucial role in the success of Semantic Web technologies because they provide an automated way to extract semantic features from text. In this paper, we described META, a tool that support the development of NLP applications. This component allows the management of collection of documents and provide an engine able to run different NLP operations on documents. The output of this operations could be exported in different way or could be used in Information Retrieval, Information Filtering or Information Extraction tasks.

An ongoing work in which META is involved is the adoption of the system as indexer for a semantic search engine designed and developed in our lab. This search engine provides different document representations that we call *levels*. Each level has a local scoring function, then a global ranking function is defined in order to merge the results produced by local scoring functions. META is adopted to build the different levels of document representation. at the moment, we have three levels: keyword, synset and entity.

As future work, we plan to develop new components able to carry out statistical report on the extracted features. We are working also in order to provide tools for the evaluation of NLP algorithms included in the META pipeline.

7. Acknowledgements

We would like to thank Franco Grieco for his help in the design and development of the first release of META, and all the students who contributed to improve META.

8. References

[1] P. Basile, M. de Gemmis, A. Gentile, P. Lops, and G. Semeraro, "Jigsaw algorithm for word sense disambiguation," in *SemEval-2007: 4th Int. Workshop on Semantic Evaluations*. ACL press, 2007, pp. 398–401.

[2] E. Agirre, B. Magnini, o. Lopez de Lacalle, A. Otegi, G. Rigau, and Vossen, "Semeval-2007 task 1: Evaluating wsd on cross-language information retrieval," in *SemEval-2007: 4th Int. Workshop on Semantic Evaluations*. ACL press, 2007, pp. 1–6.

[3] P. Basile and G. Semeraro, "Jigsaw: an algorithm for word sense disambiguation," *Rivista dell'Associazione Italiana per l'Intelligenza Artificiale*, vol. IV(2), pp. 53–54, 2007.

[4] G. Semeraro, M. Degemmis, P. Lops, and P. Basile, "Combining learning and word sense disambiguation for intelligent user profiling." in *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence IJCAI-07*, 2007, pp. 2856–2861, m. Kaufmann, San Francisco, California. ISBN: 978-I-57735-298-3.

[5] H. Cunningham, "Information Extraction, Automatic," *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.

Mining the News with Semantic Press

Eugenio Picchi, Eva Sassolini, Sebastiana Cucurullo, Francesca Bertagna
Istituto di Linguistica Computazionale (CNR-ILC)
Consiglio Nazionale delle Ricerche, Pisa, Italy
{eugenio.picchi|eva.sassolini|nella.cucurullo|francesca.bertagna}@ilc.cnr.it

Abstract

In this paper, we present Semantic Press, a tool for automatic press review based on text mining technologies and tailored to meet the requirements of eGovernment and eParticipation. The paper first provides a general description of the applicative exigencies that emerge from the eParticipation and eGovernment sectors. Then, an introduction of the general framework (the so called Linguistic Miner) for the automatic analysis and classification of textual content is provided. The core of the paper is the description of the tool for the analysis and presentation of newspapers content, its underlying technologies and final functionalities.
Index Terms: text mining, press review

1. Introduction

eParticipation is the extension and transformation of participation in political deliberation and decision-making processes through information and communication technologies (ICT). The notion is complementary the eGovernment one, which concerns more the use of ICT to improve and to innovate the quality of services offered by Public Administration to citizens.
Language plays a fundamental role in eParticipation, since it is the medium through which all the communication takes place: it is the language we find in institutional sites to explain to citizens how to obtain a particular service, it is the language of political discourse, the language of people expressing political opinions on a non official forum. NLP can be an instrument to deal with all these types of messages in an automatic or semiautomatic way.
A very sensitive issue for eParticipation and eGovernment is the necessity, for citizens but also for professionals of politics and consensus formation, to know salient facts and features, hidden in very large quantity of data, which standout for their frequency: this allows deriving interesting and constantly updated information about trends, tendencies and most important topics in a given period. For this kind of exigencies, ignited by the availability of huge quantity of information, constantly changing and dislocated on a great number of web sites, Text Mining techniques are very useful and promising.
In this paper we present Semantic Press, a tool for automatic press review based on text mining technologies and tailored to meet the requirements of eGovernment and eParticipation. The paper first provides a general description of the applicative exigencies that emerge from the eParticipation and eGovernment sectors. Then, an introduction of the general framework (the so called Linguistic Miner) for the automatic analysis and classification of textual content is provided. The core of the paper is the description of the tool for the analysis and presentation of newspapers content, its underlying technologies and final functionalities.

2. Linguistic Miner

Semantic Press is one of the applications derived by the so-called Linguistic Miner [1], a project started in 2003 with the aim of developing a framework for the automatic extraction of linguistic knowledge from very large amounts of texts (from different sources and in different formats) to be exploited in didactic, editorial and cultural products.
Building the Linguistic Miner involves two fundamental steps: first of all, the data are gathered, then they are linguistically analysed to be further processed and classified. The first step produced a repository (a “mine”) of around 200 millions words, together with an automatic topic classification of texts. This was achieved by exploiting procedures for the upgrade and augmentation of textual data in the “mine” and for the automatic acquisition from the Web through spider technology, both with periodic updating and also by means of user-defined paths. The Mine is thus constantly augmented in size.
The second step consists in the automatic linguistic processing of the textual material, by using modules of the PiSystem [2], an integrated framework for the treatment of textual and lexical material, where the most important module is the DBT (Data Base Testuale, Textual Data Base). The most effective procedures for further analysis of texts are POS tagging and lemmatization, which have been performed over 90% of the whole repository.
Many are the frameworks in which text mining techniques are applied and exploited (such as Inxight's LinguistX [3], IBM's Intelligent Miner [4], TextWise etc.). In this scenario, the Linguistic Miner stands out for its being based on tools and basic technologies developed to obtain good linguistic analysis as support of the entire application. Linguistic Miner is specifically tailored for analyzing Italian, but it is obviously open to other languages.
In the last year, Linguistic Miner has been addressed to meet the exigencies of political and institutional bodies, such as Regione Toscana, which have expressed their interest to use and exploit a tool for the intelligent access to the flow of news and information provided by Italian newspapers available on the Internet. This aim is in line with the current interest of CNR-ILC to the themes of eParticipation and eGovernment, testified also by the participation to the DEMO-net project (http://www.demo-net.org/demo).

3. Semantic Press

Semantic Press specializes some of the functionalities of Linguistic Miner towards the analysis of information available in Italian on-line newspapers. Semantic Press can be reached at ULR http://serverdbt.ilc.cnr.it/edicola/. Semantic Press is different from other tools for automatic press review (such as Press Today, see http://test.presstoday.com/): as a matter of fact, it is not only a way to present and to incrementally store news and articles pertinent to different

sectors, but also a powerful tool, based on NLP and text mining techniques, for highlight emerging subjects, issues and words. Fig. 1 provides an overview of the entire system.
Every morning, Semantic Press downloads all the articles of the most important and most read Italian newspapers: Sole 24h, La Stampa, La Repubblica, Il Corriere della Sera, Il Messaggero, La Nazione, Reuters and La Voce. The acquisition procedure downloads not only the text and the title of the article, but also visits and saves all the textual material available in the web pages linked to the article. Some filters are activated, in order to avoid downloading of dossiers, tables and other not interesting sources. During the day, Semantic Press performs the updating of the articles, adding new information if available and handling cases of similarity between different versions of the same article. In this way, about 1200 new articles are stored every day.
The acquired textual material is saved and converted in an internal format based on the DBT specifications. When the article is saved, it is also classified according to a pre-defined set of ten topics (politics, finance, sport etc.). The classification is based on the classifying tags already present in the sources, which are normalized and mapped onto a shared and common classification scheme.

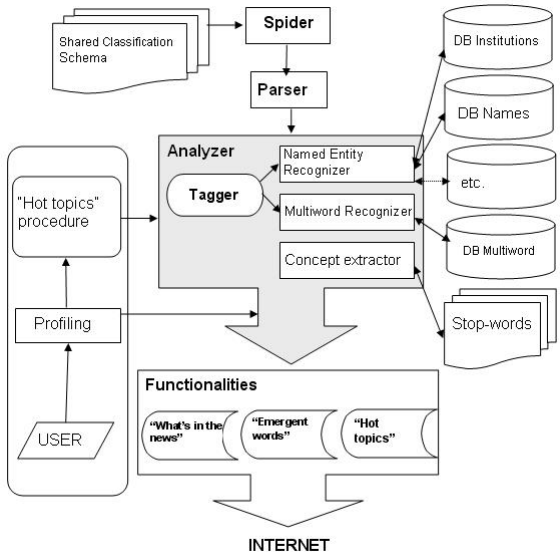


Fig. 1: "Semantic Press" system overview

Three basic technologies are used in the analysis phase: named entity recognition, multiword recognition and concept extraction. Different experiments have been carried out to evaluate the impact and performance of the basic technologies exploited, as well as an evaluation of the various strategies adopted.

The basic technologies are exploited by the application modules which provide results interesting for the users: the "What's in the news.." and the "Emergent Words.." functionalities. A particular step is represented by the so-called "Hot Topics" functionality, which performs classification on the basis of user-specific and user-defined exigencies.

4. System basic technologies

4.1. Named Entities Recognition

An important module of the system performs Named Entity Recognition on the bulk of knowledge acquired every day, extracting names of person, locations, addresses and organizations. These classes, which belong to the larger set of entities recognized within the context of international evaluation campaigns, such as MUC and EVALITA, are those more in line with the aim of a press review.
The approach is hybrid and exploits both external ad-hoc resources and consolidated techniques based on Support Vector Machine. The NERec module consists of a set of finite-state automata which benefit of ad-hoc databases containing more than 200.000 pre-classified named entities derived by on-line repositories (such as gazetteers, lists of names and surnames, lists of locations etc.). These databases are also the base of the dynamic recognition of new entities, resulting from the combination of only partially known entities and specific heuristics which provide a feedback on the original databases and constantly enrich them. In this way, the system is able to recognize the unknown Italian surname Andreoli by exploiting the co-occurrence with a known name, Alberto. Then, after the recognition, the "new" surname is added to the database. The NERec module also performs term disambiguation in case of ambiguous classes (typical the case geo-political entities) by exploiting ad-hoc statistical strategies. An important functionality is represented by the possibility to use lists of synonyms and variants, able to recognize, as a same entity, different forms: this is done, in particular, for the names of organizations, which are often alternatively present in form of acronym or whole name.

4.2. Multiword Recognition

This module is based on the analysis of very large corpora, both belonging to the Linguistic Miner and the repository of press articles; multiwords are extracted by exploiting pattern-matching techniques (the typical N-Adj and N-preposition-N patterns of Italian constructions) and filtered on the basis of the frequency distributions on the various sectors with respect to the reference corpora of the Linguistic Miner.

4.3. Concept Extraction

This basic functionality extracts all the terms which are above a given frequency threshold and are recognized as "semantically salient" terms. The module uses list of stop-words and other heuristics to determine the relevance of the term. In the specific case of the Hot Topics functionality (see section 5.3), the term extraction is carried out by means of the following process:

1. starting from a small set of pivot terms, a vocabulary is obtained based on mutual information;
2. the terms of the vocabulary are used to "weight" each single article;
3. the vocabulary is then enriched by exploiting the ranking of the archive of the news. In this way, the weight of each term and of each article is compared with the weight they have in the reference generic corpus.

Then, the articles are linguistically analysed in order to obtain texts annotated at lemma and PoS level.

5. System Functionalities

The textual analysis performed on the daily press allows the system to provide users with three functionalities: "Emergent Words", "What's in the News" and "Hot Topics". They are conceived as alternative views and access to the same content. Fig. 2 is a screenshot of the Semantic Press home page: the frame menu presents the set of available functionalities, while in the principal frame the various journalistic sources, the different themes and the relevant news are presented.

5.1. What's in the News

Semantic Press presents to the user the most important themes emerged from the automatic analysis of the daily news. User can know, each day, which are the most discussed themes, in form of name of persons (often politicians, but also people of the show business, important scientific or artistic personalities, protagonists of crime news etc.), events and facts, locations where important things are happening etc. Examples of possible topics, as they emerged from the Italian news in the current period, are *immigrants*, *tesoretto* (a revenue which exceeds the expectation), *welfare*, *Padoa Schioppa* (the Italian Minister of finance) etc.

Themes are selected, on the basis of statistical evidences, among the concepts highlighted during the textual analysis phase.

5.2. Emergent Words

Emergent words are automatically obtained by exploiting information on relative frequency of a term and its belonging to specific sectors. The aim is making emerge all the different words in a newspaper article, by providing a general overview of the vocabulary predominant in a particular day.

5.3. Hot Topics

A particular form of content classification is the one implemented in the functionality called Hot Topics, which allows the user to retrieve all the articles and news concerning a specific argument. First, by exploiting techniques based on mutual information and words co-occurrence, specific dictionaries are extracted by the bulk of information starting from a selection of "pivot terms". This allows, for example, deriving a dictionary of terms concerning sport starting from pivot terms like "Tour de France" and "cycling". Then, these argument-specific dictionaries are projected on the news repository and the system provides a ranking of the relevancy of each article to the given argument. The selection of the arguments is driven by user-specific interests: if a user wants to investigate a particular aspect present in the news in a given period, it can ask the service to report its evolution and behavior. The functionality is provided of a module for user profiling, which can direct the search and the extraction to specific interests of the user.

5.4. Mining the local news

Users are often interested in information of very local nature (see paragraph 7 for a generic use-case). For this reason, a specific functionality performs text mining on local news, offering a customized service to citizens living in different Italian cities.

This kind of "transversal" classification requires a specific treatment of the textual materials: first of all, the frequency thresholds established to select salient terms in the national news have to be modified. Moreover, it is not possible anymore to exploit the domains that classify the content in the national press: as a matter of fact, the on-line version of newspapers does not contain a reliable classification of the local news. Thus, the only classification provided in this case is the one supported by the "Hot Topics" functionality.

5.5. Web Alert

A specific functionality of the system is the one that sends an e-mail to the user each time news of his/her interest is published on the press. This functionality allows tracking the evolution of specific information in the news.

A specific feature of our Web Alert, with respect to similar available web solutions, is that it works by exploiting all the mining solutions implemented in the system, in particular the innovative classification strategies created as support of the "Hot Topics" functionality.

This means that the information will be found and announced not only in case of a simple keyword-based matching but also if the article is selected by projecting, on the incoming news, the dynamic, ad-hoc vocabulary extracted by the archive.

For example, if the user is interested on the situation in the Middle East, he/she will be "alerted" not only in case of an incoming article containing the string "Medio Oriente" (Middle East), but also if the news contain the words "Afghanistan" and "guerra" (war) or "Abu Mazen" and "Palestina".

6. Accessing normative texts

A further access modality is foreseen for normative text. In Semantic Press the aim is offering the wider range of access modalities and of prompt announcement of news. We want to obtain similar results also in the Legal domain, which is represented by a very particular type of text (the normative text). Specific modules for multiword recognition and concept extraction, more tailored to the analysis of this particular type of text, are used to access and browse laws and regulations. This new application is called Edicola Juris and works on an archive of legal texts derived by the "Gazzetta Ufficiale" (the official, periodic publication that collects all the new national Italian laws). In Edicola Juris, the extracted terminology (composed by single terms and multiwords) is used to help the user to restrict the scope of his/her search. For example (see Fig. 3) the user may ask all the articles of law concerning the "corte dei conti" (the Italian state audit court): the articles will be returned, together with the terminology calculated on the articles themselves. Each single term of the terminology may be used, as a sort of *facet*, to restrict the search and to obtain more precise results, for example articles that concern the "corte dei conti" and "contrasto all'evasione" (fight against tax evasion).

7. Use Case

"Profiling" is one of the most interesting features in Semantic Press: it allows the identification of specific needs of very different users.

Every levels of access allow the specialization of the informative offer, not only by means of the selection of the most suitable sources of information and of the sectors closer to the user exigencies but also by exploiting the classificatory capabilities of the "Hot Topics" functionality, which offers in this sense a high flexibility for user-specific needs.

An exemplifying user-scenario is the one of a service company that can offer to its users (often not "normal" citizens but rather other companies, typically public utilities companies) the possibility of obtaining very precise information on very particular sectors of interest. Public utilities companies rarely have the appropriate size and the personnel to be constantly "on the news": this is why they often ask an external service to provide them with the information of their interest. Often, the interesting information for them is of very "local" nature and very specific relevance. This type of information can hardly be restricted to a sector defined in aprioristic way. If we look at a public utilities company for refuse collection, for example, we see that salient information usually concern calls for bids and notice of modification in regulations of the specific sector. How can such specificity be successfully handled by an automatic system? Which are the words that can help us to circumscribe the sector we are trying to deal with? In a situation like this, the technology underlying the "Hot Topics" functionality may be of great help.

As a matter of fact, by using the "Hot Topics" option, the search is not restricted to single pivot words typical of the sector, but it is projected on an entire customized vocabulary, automatically created on the basis of the written material pertaining to the specific sector. The more specific is the sector, the more detailed and particular will be the vocabulary. In this way, users can find the information they are looking for, with a highly customized service.

8. References

[1] Picchi E., Ceccotti, M. G., Cucurullo, S., Sassi, M., Sassolini, E. (2004). Linguistic Miner: an Italian Linguistic Knowledge system. In Proceedings of LREC 2004

[2] Picchi E., Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian, in Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex '94, Amsterdam, The Netherlands, 1994.

[3] Effective Information Discovery, Supporting the Analytical Mission through Entity-Based, Semantic Discovery. An Inxight Federal Systems Group White Paper, November 2006.

[4] Installing the Intelligent Miner products: Modeling, Scoring, Visualization V8.2. IBM Manuals for DB2 Intelligent Miner Modeling V8.2.

.

Fig. 2: Semantic Press home page

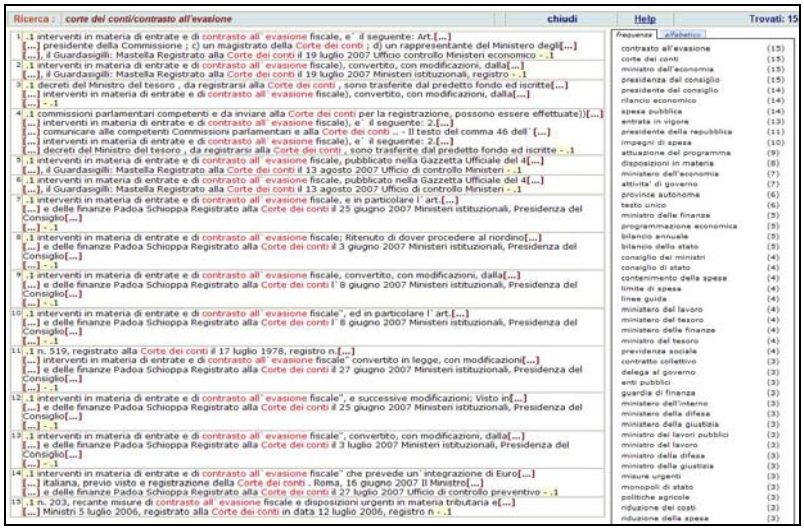


Fig. 3: Edicola Juris: analyzing normative text

Text Processing Tools and Services from iLexIR Ltd

Ted Briscoe, Paula Buttery, John Carroll
Ben Medlock, Rebecca Watson

iLexIR Ltd, Cambridge, UK
www.ilexir.com, ejb@ilexir.co.uk

Abstract

We describe the text processing tools that iLexIR has developed in collaboration with the Universities of Cambridge and Sussex. iLexIR has sole commercial rights to an extensive toolkit for English text processing applications and undertakes additional software development as well as tool tuning and porting for SMEs marketing applications with a text processing component.

To date, we have worked with clients to develop sentiment classification systems, mobile phone based question-answering services, and text mining tools for use in ESOL examination design and biomedical information extraction. Our toolkit has been extensively deployed for non-commercial research and proven its utility in projects on ontology and lexicon construction, anonymisation, anaphora resolution, word sense disambiguation, and many forms of text classification at the document, passage and sentence levels.

Index Terms: text analytics, text mining, text classification, question-answering, information extraction, ontology construction

1. Introduction

The RASP (robust accurate statistical parsing) toolkit is being developed by research groups based at the universities of Cambridge and Sussex (Briscoe & Carroll, 2002; Briscoe, Carroll & Watson, 2006). iLexIR was incorporated in 2003 as the sole commercial agent and owner of the intellectual property rights in RASP. We have deployed this toolkit, in conjunction with a range of open-source tools such as machine learning classifiers (e.g. Mallet, mallet.cs.umass.edu), information retrieval engines (e.g. Lucene, www.lucene.sourceforge.net) and XML-based document metadata handling systems (e.g. UIMA, uima-framework.sourceforge.net), to solve a diverse range of real-world text processing tasks. As a consequence of this activity, the RASP system is now also available embedded in UIMA (Andersen *et al.*, submitted; www.digitalpebble.com/resources.html), and iLexIR also offers tight integration of the toolkit with its own timed aggregate perceptron classifier, an innovative machine learning classifier with the accuracy comparable to support vector machines but training time closer to a naive bayes classifier (Medlock, forthcoming).

The resulting suite of tools, and expertise in using them, allows us to tackle almost any English text processing problem rapidly and effectively, yielding systems with state-of-the-art performance. In this paper, we describe the functionality of the toolkit and briefly discuss and reference some of the applications we have developed using the toolkit.

2. The RASP Toolkit

RASP is implemented as a series of modules written in C and Common Lisp, which are pipelined, working as a series of Unix-style filters. RASP runs on Unix-based platforms and is compatible with most C compilers and Common Lisp implementations. The public release includes Lisp and C executables for common 32- and 64-bit architectures, shell scripts for running and parameterising the system, documentation, and so forth. Potential commercial users may download the freely-distributed system under the non-commercial licence to conduct their own evaluation of its suitability for their application – see www.informatics.susx.ac.uk/research/nlp/rasp/ for licence and download details. An overview of the system is given in Figure 1.

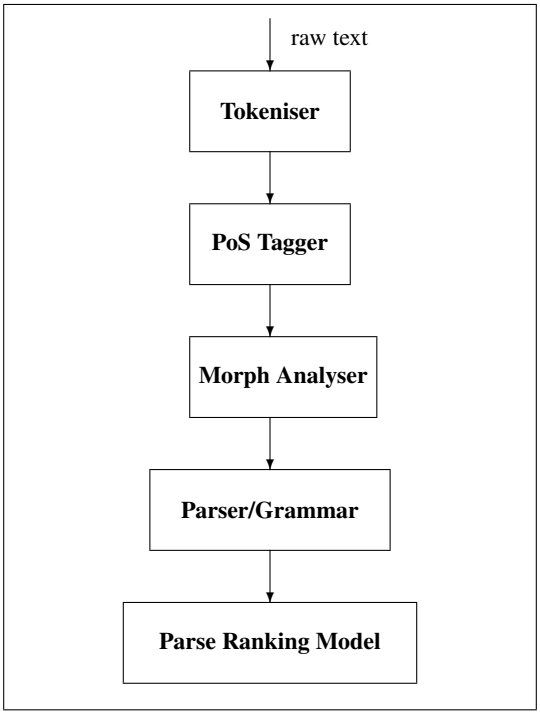


Figure 1: RASP Pipeline

2.1. Sentence Boundary Detection and Tokenisation

The system is designed to take unannotated text or transcribed (and punctuated) speech as input, and not simply to run on pre-tokenised input. Sentence boundary detection and tokenisation modules, implemented as a set of deterministic finite-state rules

in Flex (an open source re-implementation of the original Unix Lex utility) and compiled into C, convert raw ASCII (or Unicode in UTF-8) data into a sequence of sentences in which, for example punctuation tokens are separated from words by spaces, and so forth. Users are able to modify the rules used and recompile the modules. All RASP modules now accept XML mark up (with certain hard-coded assumptions) so that data can be pre-annotated – for example to identify named entities – before being passed to the tokeniser, allowing for more domain-dependent, potentially multiword tokenisation and classification prior to parsing if desired (e.g. Vlachos et al., 2006), as well as, for example, handling of text with sentence boundaries already determined, and retention of any document metadata encoded as XML.

2.2. PoS and Punctuation Tagging

The tokenised text is tagged with one of 150 part-of-speech (PoS) and punctuation labels (derived from the CLAWS tagset). This is done using a first-order ('bigram') hidden markov model (HMM) tagger implemented in C (Elworthy, 1994) and trained on the manually-corrected tagged versions of the Susanne, LOB and BNC corpora. The tagger has been augmented with an unknown word model which performs well under most circumstances as well as an extended lexicon better able to assign appropriate tags to rare words. The new tagger has an accuracy of just over 97% on the DepBank part of section 23 of the Wall Street Journal (WSJ), suggesting that this modification has resulted in competitive performance on text types at some remove from the original training data. The tagger implements the Forward-Backward algorithm as well as the Viterbi algorithm, so users can opt for tag thresholding rather than forced-choice tagging (giving >99% tag recall on DepBank, at some cost to overall system speed).

2.3. Morphological Analysis

The morphological analyser is also implemented in Flex, with about 1400 finite-state rules incorporating a great deal of lexically exceptional data. These rules are compiled into an efficient C program encoding a deterministic finite state transducer. The analyser takes a word form and CLAWS tag and returns a lemma plus any inflectional affixes. The type and token error rate of the current system is less than 0.07% (Minnen, Carroll and Pearce, 2001). The primary system internal value of morphological analysis is to enable later modules to use lexical information associated with lemmas, and to facilitate further acquisition of such information from lemmas in parses.

2.4. PoS and Punctuation Sequence Parsing

The manually-developed wide-coverage tag sequence grammar utilised in this version of the parser consists of around 700 unification-based phrase structure rules. The preterminals to this grammar are the PoS and punctuation tags. The terminals are featural descriptions of the preterminals, and the nonterminals project information up the tree using an X-bar scheme with 41 attributes with a maximum of 33 atomic values. The current version of the grammar finds at least one parse rooted in S for about 85% of the Susanne corpus (used for grammar development), and most of the remainder consists of phrasal fragments marked as independent text sentences in passages of dialogue. The coverage of our WSJ/DepBank test data is 84%. In cases where there is no parse rooted in S, the parser returns a connected sequence of partial parses covering the input. The criteria are partial parse probability and a preference for longer but non-lexical combinations (Kiefer *et al.*, 1999).

2.5. Probabilistic Generalised LR Parser

A non-deterministic LALR(1) table is constructed automatically from a CF 'backbone' compiled from the feature-based grammar. The parser builds a packed parse forest using this table to guide the actions it performs. Probabilities are associated with subanalyses in the forest via those associated with specific actions in cells of the LR table (Inui *et al.*, 1997). The n-best (i.e. most probable) parses can be efficiently extracted by unpacking subanalyses, following pointers to contained subanalyses, and choosing alternatives in order of probabilistic ranking. The probabilities of actions in the LR table are computed using bootstrapping methods which utilise an unlabelled bracketing of the Susanne Treebank (Watson *et al.*, 2007). This makes the system more easily retrainable after changes in the grammar and opens up the possibility of quicker tuning to in-domain data. In addition, the structural ranking induced by the parser can be reranked using (in-domain) lexical data which provides conditional probability distributions for the SUBCATegorisation attributes of the major lexical categories.

2.6. Grammatical Relations Output

The resulting set of ranked parses can be displayed, or passed on for further processing, in a variety of formats which retain varying degrees of information from the full derivations. The most common output format is a set of named grammatical relations (GRs), illustrated as a subsumption hierarchy in Figure 2. Factoring rooted, directed graphs of GRs into a set of bilexical dependencies makes it possible to compute the transderivational support for a particular relation and thus compute a weighting which takes account both of the probability of derivations yielding a specific relation and of the proportion of such derivations in the forest produced by the parser. A weighted set of GRs from the parse forest is computed efficiently using a variant of the inside-outside algorithm (Watson *et al.*, 2005).

2.7. Evaluation

The system has been evaluated using our reannotation of the PARC dependency bank (DepBank; King *et al.*, 2003) – consisting of 560 sentences chosen randomly from section 23 of the WSJ – with GRs compatible with our system. Relations take the following form: (**relation subtype head dependent initial**) where **relation** specifies the type of relationship between the **head** and **dependent**. The remaining **subtype** and **initial** slots encode additional specifications of the relation type for some relations and the initial or underlying logical relation of the grammatical subject in constructions such as passive. We determine for each sentence the relations in the test set which are correct at each level of the relational hierarchy. A relation is correct if the head and dependent slots are equal and if the other slots are equal (if specified). If a relation is incorrect at a given level in the hierarchy it may still match for a subsuming relation (if the remaining slots all match). Thus, the evaluation scheme calculates unlabelled dependency accuracy at the most general level in the hierarchy. The micro-averaged precision, recall and F_1 score are calculated from the counts for all relations in the hierarchy. Table 1 gives the microaveraged F_1 score for RASP, the Collins Model 2 parser, the Parc XLE parser, and the CCG parser. Only the CCG parser which is trained on in-domain data and includes many lexical parameters derived from

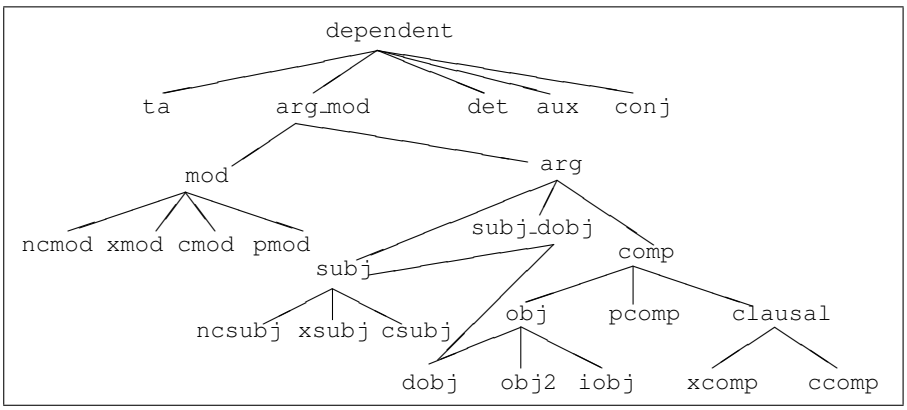


Figure 2: The GR hierarchy

the WSJ treebank outperforms the unlexicalised RASP parse ranking system (see Briscoe & Carroll, 2006 and Clark & Curran, 2007 for detailed discussion of the evaluation and results).

System	Precision	Recall	F_1
Collins	78.3	71.2	74.6
XLE	79.4	79.8	79.6
RASP	81.5	78.1	79.7
CCG	82.4	81.3	81.9

Table 1: Overall Microaveraged Scores

3. Text Classification with RASP

Standard text classification adopts the 'bag of words' (BoW) model in which a document is treated as an unstructured multiset of terms and information about word position or syntactic structure is ignored. This approach works well for document topic classification but less well for sentiment or genre classification, or for (sub)sentential classification tasks such as named entity recognition, anonymisation, or (non)-speculative assertion identification (e.g. Medlock, 2006).

The RASP toolkit makes available a range of features beyond BoW, based on morphological analysis (lemmas, stems), part of speech tags, and word cooccurrences mediated by grammatical relations rather than by adjacency or windowing. These additional feature types can be made available to machine learning classifiers, and feature instances from these types that are effective for a given classification task can be selected during the training phase by the classifier for run-time application.

The timed aggregate perceptron (TAP) classifier (Medlock, forthcoming) is a highly scalable linear classifier which has been shown to outperform SVMs and Bayesian logistic regression (BLR) on topic and other text classification tasks. The TAP classifier achieved better classification accuracy than either popular alternative, but trained in near linear time. This means that a classifier trained on the entire Reuters Rcv1 corpus of around 800K news stories (Lewis *et al.*, 2004) divided into 103 classes could be built in around 3.5hrs CPU time (as opposed to around 20hrs for the SVM or 50hrs for BLR). This is a significant advantage for real world applications where reductions in training time allow vital experimentation into enhancing feature generation and selection as well as frequent retraining as data is accumulated.

The TAP classifier has been tightly integrated with the RASP toolkit so that it is easy to undertake experiments to find the optimal set of feature types and instances for a particular classification task, whether this be at the document, passage, sentence or (sub)sentence level. However, in many real world applications it is not possible to train a classifier in a fully supervised fashion because data is only partially or noisily labelled. A significant element of the research undertaken with the toolkit has been to explore the use of bootstrapping and other semi-supervised techniques to circumvent the need for large quantities of well-annotated training data. In areas such as anonymisation (Medlock, 2006) and biomedical named entity recognition (Vlachos *et al.*, 2006) we have been successful in bootstrapping accurate classifiers from text automatically annotated with RASP.

4. Text Information Retrieval/Extraction with RASP

To date, RASP has been applied to around one billion words of English text drawn from genres as diverse as biomedical scientific papers through to second language learners' examination scripts. The additional annotations produced by RASP, possibly in conjunction with text classifiers, can form the basis for enhanced information retrieval at the document, passage or sentence level, based on going beyond keyword (BoW based) search for documents to search for named entitities in specific relations or contexts.

The fact that RASP and our text classifiers' produce XML annotations on text that may already be annotated with metadata allows us to efficiently exploit the new generation of XML aware open-source indexing engines such as Lucene, Indri (www.lemurproject.org/indri) and Xapian (www.xapian.org). These provide flexible search interfaces that allow Boolean combination of constraints based on XML path specifications and, thus, support seamless extension from information retrieval to information extraction, only limited by the extent of annotation in the indexed text. If a free text question interface is used, the approach can be extended to parsing the query, extracting the GRs in the query, and using the resulting annotation to find matches in the annotated document database. As an illustration of one application, Figure 3 shows a screenshot of the FlyBase curator interface to an article annotated with gene names.

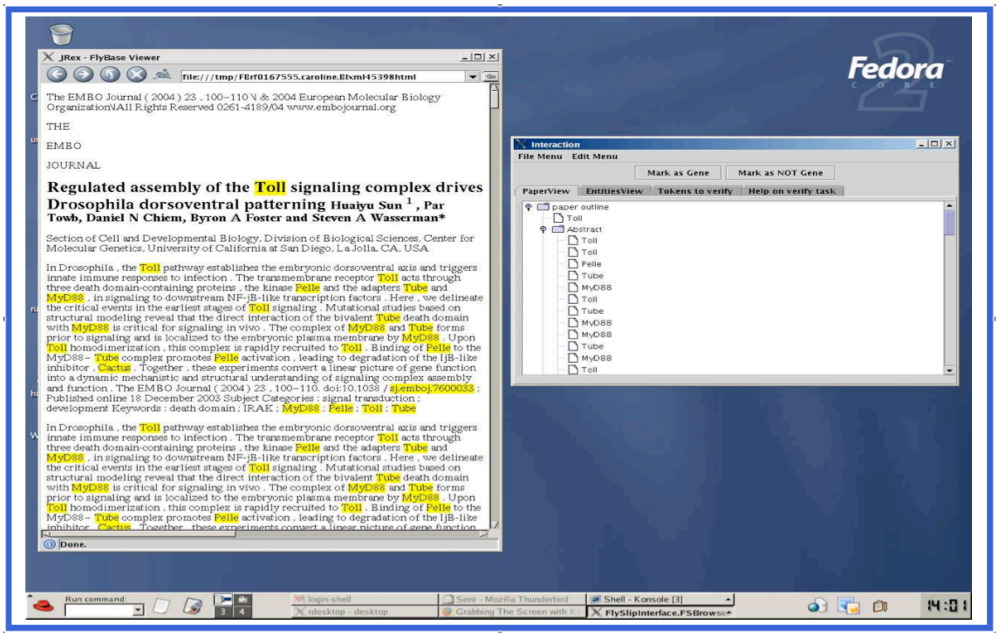


Figure 3: The FlyBase Curation Tool

5. Conclusions

Through integration of state-of-the-art tools from natural language processing, information retrieval and machine learning, we have been able to build a flexible toolkit with which it is possible to efficiently develop an optimal solution to most text processing tasks. The toolkit together with the know-how gained from research with its precursors has allowed us to rapidly develop components for commercial applications involving text mining, classification and question-answering. As a small research-led company, we expect to continue to develop the toolkit informed by the latest research in all three fields, whilst adding functionality in-house which will enhance robustness and scalability and decrease the resources required for tuning and adaptation to new applications.

6. References

Andersen, O., J. Nioche, E.J. Briscoe and J. Carroll (submitted) 'The BNC parsed with RASP4UIMA', *Proceedings of the 6th Int. Conf. on Language Resources and Evaluation (LREC),* Marrakech, Morocco.

Briscoe, E.J. and J. Carroll (2002) 'Robust accurate statistical annotation of general text', *Proceedings of the 3rd LREC,* Las Palmas, Gran Canaria, pp. 1499–1504.

Briscoe, E.J. and J. Carroll (2006) 'Evaluating the Accuracy of an Unlexicalized Statistical Parser on the PARC DepBank', *Proceedings of the 44th Assoc. Computational Linguistics(ACL)-Coling, Main Conf. Poster Sesssion,* Sydney, Australia, pp. 41–48.

Briscoe, E.J., J. Carroll and R. Watson (2006) 'The Second Release of the RASP System', *Proceedings of the 44th ACL-Coling, Interactive Presentation Session,* Sydney, Australia, pp. 77–80.

Clark, S. and J. Curran (2007) 'Formalism independent parser evaluation with CCG and DepBank', *Proceedings of the 45th ACL,* Prague, Czech Republic, pp. 248–255.

Elworthy, D. (1994) 'Does Baum-Welch re-estimation help taggers?', *Proceedings of the 4th ACL Conf. on Applied NLP,* Stuttgart, Germany, pp. 53–58.

Inui, K., V. Sornlertlamvanich, H. Tanaka and T. Tokunaga (1997) 'A new formalization of probabilistic GLR parsing', *Proceedings of the 5th Int. Workshop on Parsing Technologies (IWPT),* MIT, pp. 123–134.

Kiefer, B., H-U. Krieger, J. Carroll and R. Malouf (1999) 'A bag of useful techniques for efficient and robust parsing', *Proceedings of the 37th ACL,* University of Maryland, pp. 473–480.

King, T.H., R. Crouch, S. Riezler, M. Dalrymple and R. Kapl an (2003) 'The PARC700 Dependency Bank', *Proceedings of the 4th Int. Workshop on Linguistically Interpreted Corpora (LINC-03),* Budapest.

Lewis, D., Y. Yang, T. Rose, F. Li (2004) 'Rcv1: a new benchmark collection for text categorization research', *J. Machine Learning Research, vol.5,* 361–397.

Medlock, B. (2006) 'An introduction to NLP-based textual anonymisation', *Proceedings of the 5th LREC,* Genoa, Italy.

Medlock, B. (forthcoming) *Scalability for text categorization and the timed aggregate perceptron,* m.s..

Minnen, G., J. Carroll and D. Pearce (2001) 'Applied morphological processing of English', *Natural Language Engineering, vol.7.3,* 225–250.

Watson, R., J. Carroll and E.J. Briscoe (2005) 'Efficient extraction of grammatical relations', *Proceedings of the 9th IWPT,* Vancouver, Ca..

Watson, R., E.J. Briscoe and J. Carroll (2007) 'Semi-supervised Training of a Statistical Parser from Unlabeled Partially-bracketed Data', *Proceedings of the 10th IWPT,* Prague, Czech Republic.

Vlachos, A., Gasperin, C., Lewin, I., Briscoe, E. J. (2006) 'Bootstrapping the recognition and anaphoric linking of named entities in Drosophila articles', *Proceedings of the Pacific Symposium in Biocomputing,* Hawaii.

THE IMPACT OF STANDARDS ON TODAY'S SPEECH APPLICATIONS

Paolo Baggia
Loquendo SpA

ABSTRACT

At the end of the last century, the landscape of speech applications was abruptly changed due to the convergence of several factors: the maturity of speech technologies and the creation of standards to promote the development of speech applications.
The intention of this paper is to give a clear picture of this evolution, to summarize the major standards and to highlight future evolution paths.

Index Terms— Automatic Speech Recognition, Text-To-Speech, Spoken Dialog Systems, Voice Browsers, Speech Standards.

1. INTRODUCTION

At the end of the last century, the landscape of speech applications abruptly changed, not only because speech applications became common in many areas (e.g. customer care, self-service applications, voice portals), but also because of a shift from proprietary applications to standards based ones. A convergence of different factors drove this change, certainly speech technologies had reached a level of maturity to allow their use in many applications; however, the ongoing development of the Web promoted the adoption of a novel architecture called Voice Browsing. The development of standards for speech applications was another important driver, which was able to allow the creation of powerful building blocks. The intention of this paper is to give a clear picture of this evolution, to summarize the major standards and to highlight evolution paths.
All the major speech technologies were heavily studied during the second half of the last century and very important research results were found in every field. For instance:

- Text-To-Speech (TTS) reached the first goal of high intelligibility during the '80s [1], after decades of research, then during the late '90s the *Concatenative Unit Selection* [2] provided the more natural sounding voices in use today.
- Automatic Speech Recognition (ASR) research provided many results in the '70s-'80s, when the statistical approaches were developed, e.g. dynamic programming, hidden Markov models and statistical language models (cfr [3-4]). Performance was pushed by competition on large corpora like DARPA and EU funded projects.
- Natural Language Understanding (NLU) is another area that moved from pure recognition of words to meaning representation, for a survey see [5-6].
- Spoken Dialog Systems (SDS) [7] were created in the late '90s after successful projects like EU SUNDIAL and COMMUNICATOR in US.

If the technology was ready to allow the creation of a speech industry, the architectures were either results of research projects or proprietary IVR systems. In this context a standard approach mainly promoted by World Wide Web Consortium (W3C) was begun and it forced an abrupt change in the industry.
Section 2 will describe the Voice Browsing approach to speech applications, with a short description of the major standards, then Section 3 will briefly introduce the architectural changes related to the Voice Browsing Platforms in use today. Section 4 will talk of other areas of standardization and finally Section 5 will draw conclusions and discuss future evolutions of this area.

2. VOICE BROWSER STANDARDS

From an historical point of view the time was right to change the landscape of speech applications. The first signal was a W3C Workshop in 1999 promoted by Dave Raggett (W3C Lead) and James Larson, who soon became the co-chair of W3C Voice Browser working group (VBWG) [8]. This group catalyzed efforts from different companies to create a new framework for developing speech applications, which was called "Voice Browsing".
The group received the VoiceXML 1.0 proposal, which had been recently developed by the VoiceXML Forum. This became the basis of subsequent standardization activities. The seminal insights were captured by Jim Larson in the "W3C Speech Interface Framework" [9], whose illuminating chart is depicted in Fig. 1.

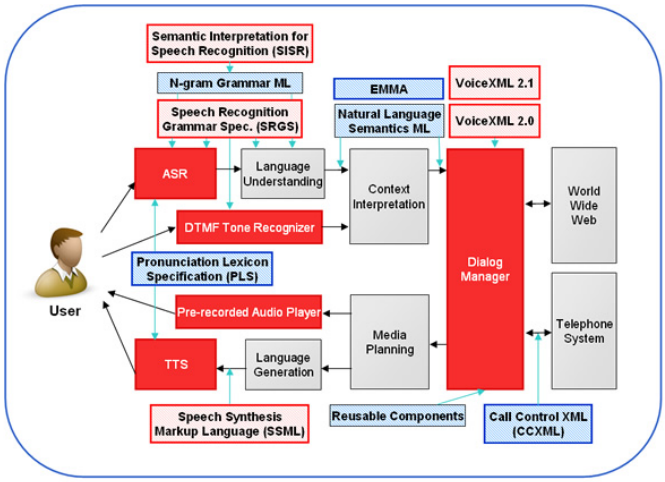


Fig. 1. Speech Interface Framework revisited

Figure 1 shows the processing modules needed to accomplish a speech (or DTMF) interaction - ASR, Language Understanding (for input processing), Context Interpretation, Dialog Manager (for dialog interaction), Media Planning, Language Generation and TTS (for output processing). The boxes outlined in red/blue are the standards to be created to support this framework; many of which are now W3C Recommendation (the ones outlined in red.) Moreover, if data are based on standards, the modules of the framework can also be completely standardized today (shown in red boxes).
The industry very soon adopted this change and interest grew significantly, which forced even the big IVR vendors to take this picture very seriously. The following sections give a brief introduction to the major standards shown in Figure 1.

2.1 Spoken Dialog (VoiceXML)

The VoiceXML 2.0 [10] standard was the key factor in the innovation. Its key features are:

- It is an XML declarative language;
- It is easy to author, the motto was: “Simple things must be easy and complex things must be possible!”
- It assumes the existence of the Web architecture.

All these features carry clear advantages; to be an XML language allows: a clean syntax checked by DTD/Schema, extensibility by Namespaces, and encodings are available from XML open source processors. Because of the second feature, VoiceXML 2.0 can be either edited by a normal text editor (then uploaded as a static page in a Web Server) or it can be dynamically generated by sophisticated Web applications, like the majority of Web pages today. Finally, to be within the Web architecture today means to share an enormous background of tools and techniques and to be placed in the mainstream of technology evolution.

From a functional point of view VoiceXML 2.0 allows the replacement of DTMF and pre-recorded applications (i.e. customer care today still offers this kind of application). This was the reason why all the IVR platforms migrated, in the last five years, to VoiceXML applications, instead of the exclusively proprietary applications developed previously. However, VoiceXML was designed to allow applications to take advantage of ASR to recognize user speech input, and of TTS to do prompting, perhaps mixed with pre-recorded audio files for speech or audio jingles. VoiceXML 2.1 [11] added eight additional features to extend VoiceXML 2.0.

```
<?xml version="1.0" encoding="UTF-8"?>
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml"
  xml:lang="en-GB">

<form id="dep_arr_airports">

 <grammar src="dep_arr.grxml"
          type="application/srgs+xml"/>

 <initial name="start">
  <prompt>
     What are the arrival and departure airports?
  </prompt>
 </initial>

 <field name="fromcity">
  <prompt>Tell me the departure airport.</prompt>
 </field>

 <field name="tocity">
  <prompt>Tell me the arrival airport.</prompt>
 </field>

 <field name="go_ahead" type=”boolean” modal="true">
  <prompt>Do you want to leave from
    <value expr="fromcity"/> and arrive
    to <value expr="tocity"/>?
  </prompt>
  <filled>
     <if cond="go_ahead">
       <submit next="/servlet/dep_date"
                namelist="fromcity tocity"/>
     </if>
     <clear namelist="fromcity tocity go_ahead"/>
  </filled>
 </field>
</form>
</vxml>
```

Fig. 2. A simplified VoiceXML document

Figure 2 shows a simplified VoiceXML 2.0 document that implements a dialog to obtain departure and arrival airports. The dialog tries first to recognize both the locations, if that fails, they are asked in sequence. A final confirmation is given before transitioning to another page of the application. For a detailed introduction see [12].

2.2 Automatic Speech Recognition (SRGS and SISR)

Two standards were created to define knowledge sources for ASR, which specify prior knowledge about the language to be recognized by the ASR. The syntax for speech grammars is in SRGS 1.0 - “Speech Recognition Grammar Specification Version 1.0” [13], which has been largely adopted by the speech industry and fully supported by all ASR engines. SRGS allows the definition of grammars for speech as well as for DTMF inputs.

Moreover, *semantic interpretation* (SI), which is the part of a speech grammar devoted to the generation of a semantic result, is given in SISR 1.0 - “Semantic Interpretation for Speech Recognition Specification Version 1.0” [14]. SISR can be a simple transliteration (e.g. “coke” for “coca cola”), called *literal semantics*, or a *script semantics*, which is based on ECMAScript/JavaScript.

```
<?xml version="1.0" encoding="UTF-8"?>
<grammar version="1.0" xml:lang="en-GB"
  xmlns="http://www.w3.org/2001/06/grammar"
  tag-format="semantics/1.0" root="fromto">
  <rule id="fromto" scope="public">
    from <ruleref uri="#city"/>
         <tag>out.fromcity=rules.latest();</tag>
    to <ruleref uri="#city"/>
       <tag>out.tocity= rules.latest();</tag>
  </rule>
  <rule id="city">
    <one-of>
      <item>London<tag>out="LHR"</tag></item>
      <item>Paris<tag>out="CDG"</tag></item>
      <item>Rome<tag>out="FCO"</tag></item>
    </one-of>
  </rule>
</grammar>
```

Fig. 3. Simple SRGS grammar with SISR script

The SRGS grammar in Figure 3 includes SISR script semantics inserted in `<tag>` elements, while the syntax is organized into rules with sequences and alternatives for words/phrases. For example, for the utterance “from Rome to Paris” the result is the ECMAScript object `{fromcity: "FCO", tocity: "CDG"}`.

2.3 Text-To-Speech (SSML)

The SSML 1.0 - “Speech Synthesis Markup Language Version 1.0” [15] is a standard to control and improve TTS rendering.

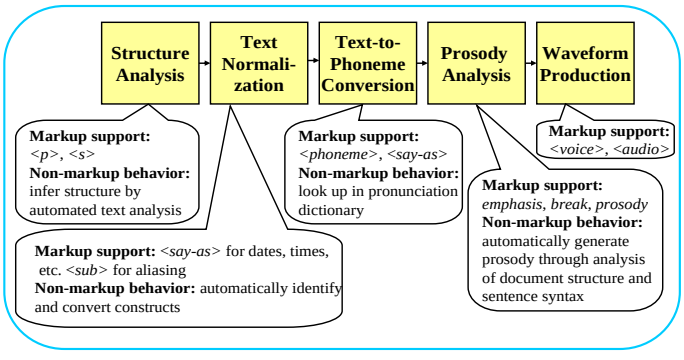


Fig. 4. Steps for TTS rendering and SSML markup

Figure 4 shows the five major processing steps present in all TTS engines. For each of them the engine already offers a normal behaviour, called “non-markup behaviour” in the picture. If needed, SSML allows the engine to improve the default rendering by means of elements of the language. Each element is related to one specific processing step and it is interpreted as a request by the author to perform an action. It

is then up to the processor to determine whether and in what way to realize the command.

```
<?xml version="1.0" encoding="UTF-8"?>
<speak version="1.0"
  xmlns="http://www.w3.org/2001/10/synthesis"
  xml:lang="en-GB">
  <p>The requested flight leaving from
        <s xml:lang="it-IT">
          <sub alias="Roma Fiumicino">FCO</sub></s>
        airport
        <emphasis>with destination
          <sub alias="London Heatrow">LHR</sub>
          </emphasis>
      are: <break time="1s"/>
      <s>
        <sub alias="British Airways 0 3 0 2">BA0302</sub>
        <break time="0.5s"/>
        leaving at
            <say-as interpret-as="time">3:45pm</say-as>
        <break time="0.5s"/>
        from gate number A63.
      </s>
      <!-- Other flight options -->
  </p>
</speak>
```

Fig. 5. A simple SSML document

The SSML example in Figure 5 shows a prompt for a flight information system. The prompt is organized into a single paragraph (<*p*>) and two sentences (<*s*>). Acronyms are substituted (<*sub*>) into expanded versions, pauses are added (<*break*>) and a time expression is explicitly labelled (<*say-as*>) to select the correct way of reading it.

2.4 Pronunciation Lexicon (PLS)

A complimentary standard of the previous ones, but still under development[1], is focused on improving pronunciation for both ASR and TTS processing. This W3C specification is PLS 1.0 – "Pronunciation Lexicon Specification Version 1.0" [16].

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon version="1.0"
  xmlns="http://www.w3.org/2005/01/pronunciation-lexicon"
  alphabet="ipa" xml:lang="en-GB">
  <lexeme>
    <grapheme>Alitalia</grapheme>
    <phoneme>ˌæl.ɪ.ˈtæi‿ə</phoneme>
    <phoneme prefer="true">a.li.ˈta:.lja</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>Lufthansa</grapheme>
    <phoneme>ˈluft.hænzə</phoneme>
    <phoneme prefer="true">ˈluftˌhan.za</phoneme>
  </lexeme>
  <lexeme>
    <grapheme>AF</grapheme>
    <alias>Air France</alias>
  </lexeme>
  <lexeme>
    <grapheme>BA</grapheme>
    <alias>British Airways</alias>
  </lexeme>
</lexicon>
```

Fig. 6. PLS 1.0 document for flight applications

A PLS document stores words, or tokens, (<*grapheme*>) with the corresponding pronunciations, which may be expressed either by textual substitutions (<*alias*>) or phonetic transcriptions (<*phoneme*>). The pronunciations might be multiple to accommodate different ways of saying a word/token, or different spelling for the same pronunciation.

[1] PLS 1.0 became a Candidate Recommendation on 21 Dec. 2007 and it is in the final stage of specification, for details see: http://www.w3.org/TR/pronunciation-lexicon/

A simple PLS 1.0 document is given in Figure 6 to be used in a flight application. For "Alitalia" and "Lufthansa" the pronunciations inside the <*phoneme*> element are given in IPA (International Phonetic Alphabet) [17] - a standard way of expressing the pronunciations for all spoken human languages. Moreover, the two lexemes have a double pronunciation; the first is the normal English one while the second is more similar to the original language (Italian and German respectively). The *prefer* attribute indicates which pronunciation has to be selected for TTS rendering.

2.5 Call Control (CCXML)

If a speech interaction is a mostly synchronous activity of collecting prompts, then speaking them and waiting to listen for some input from the user, telephony by contrast is completely based on asynchronous events. Therefore, the VBWG decided to start a new separate but interlaced standard. This was CCXML 1.0 – "Call Control XML" [18].

```
<?xml version="1.0" encoding="UTF-8"?>
<ccxml version="1.0"
  xmlns="http://www.w3.org/2002/09/ccxml">
  <var name="currentState"/>
  <var name="myDialogId"/>
  <var name="myConnId"/>
  <eventprocessor statevariable="currentState">
    <transition event="connection.alerting">
      <assign name="myConnId" expr="event$.connectionid"/>
      <accept connectionid="event$.connectionid"/>
    </transition>
    <transition event="connection.connected">
      <dialogstart
        src="'http://www.example.com/flight.vxml'"
        connectionid="myConnId" dialogid="myDialogId"/>
    </transition>
    <transition event="dialog.started">
      <log expr="'VoiceXML appl is running now'"/>
    </transition>
    <transition event="connection.disconnected">
      <dialogterminate dialogid="myDialogId"/>
    </transition>
    <transition event="dialog.exit">
      <disconnect connectionid="myConnId"/>
    </transition>
    <transition event="*">
      <log expr="'Closing, unexpected:' + event$.name"/>
      <exit/>
    </transition>
  <eventprocessor/>
</ccxml>
```

Fig. 7. Basic handling of incoming calls with CCXML

This new specification is still under development, but it is on the final stage of specification. CCXML 1.0 has the potential to be a second shake up in the IVR field. CCXML 1.0 addresses simple tasks of call handling (see Figure 7), as well as complex tasks, i.e. conditional call handling, conferencing, coaching, etc.

Each CCXML document describes transitions to handle specific events. In Figure 7 a "connection.alerting" event (incoming call) is accepted, a VoiceXML dialog is started when the "connection.connected" event is received, and then the CCXML processor waits until either the caller disconnects ("connection.disconnect") or the VoiceXML dialog exits ("dialog.exit"). These are simple actions performed by telephony calls, both TDM and VoIP.

3. VOICE BROWSER PLATFORM

Besides the standards described in the previous section, another interesting advance was made on architectures to speech applications. In the past an application was developed on proprietary SDK and then deployed inside a proprietary IVR platform. After the advent of standards the architecture changed: the platform become independent from the service

to be deployed, where the application is generated by a Web Application and accessed through HTTP, like in a Web browser.

Figure 8 shows a diagram of a voice browser platform (VoxNauta from Loquendo). On the left side it interfaces to either traditional TDM telephony (fixed or mobile) or Voice-over-IP, which is today the mainstream of evolution. On right side is shown all the knowledge to be uploaded from a Web Server, which includes CCXML and VoiceXML documents, but also the SRGS grammars and audio files. For all of them caching policies can be applied to allow efficient multi-channel applications.

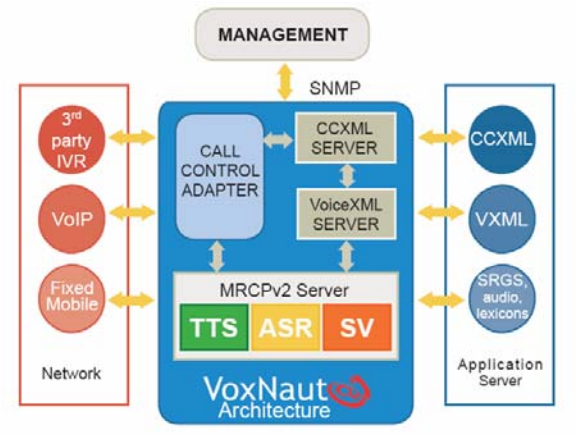


Fig. 8. VoxNauta - a Voice Browser platform

An example of a complete interaction can be the following. The platform receives an incoming call to start a new CCXML session to be controlled by a CCXML document. If the application requires a voice interaction, a VoiceXML session is started on a VoiceXML document, which in turn will provide the prompts and audio to be played by a TTS engine and SRGS grammars to be fed into the ASR engine. In VoxNauta platform the speech engines are integrated by means of the IETF MRCPv2 protocol [19].

4. RELATED STANDARDS

Other very interesting standards under development inside the W3C VBWG are SCXML [20], a generic language to represent and synchronize interactions based on State Charts that will allow the integration of multiple input modalities, and the next version of VoiceXML 3.0, which will be more modular and even more integrated with the Web.

Standards under development from the W3C Multimodal Interaction WG (MMIWG) [21] include: EMMA [22], an XML annotation of input results, InkML [23], an XML representation of digital ink to represent gesture and handwriting in a standard way, and the Multimodal Architecture and Interfaces specification [24].

Many developments have been, and are currently, done by IETF, for instance, the Media Resource Control Protocol (MRCP) [19] is the best way to integrate server speech technology into a platform.

5. CONCLUSIONS/FUTURE ADVANCES

The "Speech Interface Framework" [9], designed by W3C VBWG in 2002, is near to being completely achieved. This means that the current industry around speech technology has very solid roots and it is well placed in the midst of the advanced development of the Web.

Exciting new applications may arise from the adoption of multimodal applications that integrate speech and other modalities on small, portable devices. From an architectural point of view both multimodal and voice browsing will take advantage of the current work of the VBWG with SCXML and new features for VoiceXML 3.0, i.e. biometrics for Speaker Verification and Identification.

Speech technologies and applications are a reality today and they will certainly expand their presence in our everyday lives.

7. ACKNOWLEDGEMENTS

I would like to thank Dr. James Larson (co-chair of W3C Voice Browser WG) and Dr. Deborah Dahl (chairman of Multimodal Interaction WG) for introducing me to this world and for spending precious time reading this paper; Simon Parr for fighting with my imperfect English and many other Loquendo people, a great team to work with.

8. REFERENCES

[1] D. Klatt, "Review of text-to-speech conversion for English", JASA, 82(3), Sept. 1987.

[2] M. Balestri, A. Pacchiotti, S. Quazza, P. L. Salza and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system", *Proc. of EUROSPEECH-99*, Vol. 5, pp. 2291-2294, 1999.

[3] X. Huang, A. Acero, H.W. Hon, *Spoken Language Processing*, Prentice Hall, 2001.

[4] R. De Mori, "Spoken Dialogues with Computers: Signal Processing and Its Applications", *Academic Press*, 1998.

[5] D. Jurafsky, and J. Martin, *Speech and Language Processing*, Prentice Hall, 2000.

[6] J. Allen, *Natural Language Understanding*; Addison-Wesley, 1995.

[7] D. Dahl, *Practical Spoken Dialog Systems*; Springer; 2005.

[8] W3C Voice Browser WG: http://www.w3.org/voice/

[9] J. Larson, "Speech Interface Framework", *W3C note*, 2000, http://www.w3.org/TR/voice-intro/

[10] S. McGlashan, et al., "Voice Extensible Markup Language (VoiceXML) Version 2.0", *W3C Recommendation*, Mar. 2004.

[11] M. Oshry, et al., "Voice Extensible Markup Language (VoiceXML) 2.1", *W3C Recommendation*, Jun. 2007.

[12] J. Larson, *VoiceXML: Introduction to Developing Speech Applications*, Prentice Hall, 2003.

[13] A. Hunt, and S. McGlashan, "Speech Recognition Grammar Specification Version 1.0", *W3C Recommendation*, Mar. 2004.

[14] L. van Tichelen, and D. Burke, "Semantic Interpretation for Speech Recognition (SISR) Version 1.0", *W3C Recommendation*, Apr. 2007.

[15] D. Burnett, et al., "Speech Synthesis Markup Language (SSML) Version 1.0", *W3C Recommendation*, Sep. 2007.

[16] P. Baggia, "Pronunciation Lexicon Specification (PLS) Version 1.0", *W3C Candidate Recommendation*, Dec. 2007.

[17] IPA, *Handbook of the International Phonetic Association*, Cambridge University Press, 1999.

[18] RJ Auburn, "Voice Browser Call Control: CCXML version 1.0", W3C Working Draft, Jan. 2007.

[19] S. Shanmugham, and D. Burnett, "Media Resource Control Protocol Version 2 (MRCPv2)", IETF, Jan. 2008.

[20] J. Barnett, et al., "State Chart XML (SCXML): State Machine Notation for Control Abstraction", *W3C Working Draft*, Feb. 2007.

[21] W3C Multimodal Interaction: http://www.w3.org/2002/mmi

[22] M. Johnston, "EMMA: Extensible MultiModal Annotation markup language", *W3C Candidate Recommendation*, Dec. 2007.

[23] Y.-M. Chee, et al.,, "Ink Markup Language (InkML)", *W3C Working Draft*, Oct. 2006.

[24] J. Barnett, et al., "Multimodal Architecture and Interfaces", *W3C Working Draft*, Dec. 2006.

Using LMF to Shape a Lexicon for the Biomedical Domain

Monica Monachini, Valeria Quochi, Riccardo Del Gratta, Nicoletta Calzolari

Istituto di Linguistica Computazionale – CNR, Pisa, Italy

name.surname@ilc.cnr.it

Abstract

This paper describes the design, implementation and population of the BioLexicon in the framework of BootStrep, an FP6 project. The BioLexicon (BL) is a lexical resource designed for text mining in the bio-domain. It has been conceived to meet both domain requirements and upcoming ISO standards for lexical representation. The data model and data categories are compliant to the ISO Lexical Markup Framework and the Data Category Registry. The BioLexicon integrates features of lexicons and terminologies: term entries (and variants) derived from existing resources are enriched with linguistic features, including subcategorization and predicate-argument information, extracted from texts. Thus, it is an extendable resource. Furthermore, the lexical entries will be aligned to concepts in the BioOntology, the ontological resource of the project. The BL implementation is an extensible relational database with automatic population procedures. Population relies on a dedicated input data structure allowing to upload terms and their linguistic properties and "pull-and-push" them in the database. The BioLexicon teaches that the state-of-the-art is mature enough to aim at setting up a standard in this domain. Being conformant to lexical standards, the BioLexicon is interoperable and portable to other areas.

Index Terms: domain terminologies, computational lexicons, lexical standards, lexical architectures

1. Motivation and background

Bio-literature is continuously being produced and new knowledge is continuously being developed and it is of paramount importance to share and disseminate knowledge in the biomedical domain especially for boosting and supporting discoveries of new illnesses, treatments, medicaments, and similar. The reuse of information however requires time and efforts because it needs to integrate often redundant and partial pieces of information, which are often stored in different formats.

Intensive research has being carried out to develop language technologies that provide intelligent access to such knowledge and build lexical and ontological resources targeted to fulfill special demands for the biologist community: i.e. normalized nomenclatures (see Kors et al. 2005), extensible databases for storing terminological information like Termino (Harkema et al. 2004), lexical and ontological resources like the SPECIALIST lexicon. Still, access and interoperability of biological databases is hampered, due to persistent lack of structuring and uniformity of formats. Moreover, available bio-terminologies lack information relevant to knowledge extraction, such as predicate argument structures and syntactic complementation patterns. A comprehensive and continuously growing resource where bio-terms from different sources are integrated, encoded on the basis of the most accredited standards, enriched with relevant linguist description and linked to concepts in the ontology would significantly improve text analysis and knowledge capture systems (Hahn and Markó 2001). One of the main resources of BOOTStrep knowledge core is the BioLexicon: an expected state-of-the-art lexical resource that meets both bio-domain requirements and the most recent standards for lexical representation. The BioLexicon is an integrated resource in that it is semi-automatically populated with data collected from different available biomedical sources (e.g. UniProt/ Swiss-Prot, ChEBI, BioThesaurus, NCBI taxonomy) and is further integrated with morphological, syntactic and lexical semantic features either extracted from texts and or from domain ontologies.

2. The BioLexicon

The BioLexicon is a computational lexicon for the biology domain, designed to be reusable and flexible enough to adapt to different application needs: e.g. text mining, information extraction, information retrieval. The BioLexicon accounts for (English) lemmas and terms related to the bio-domain and contain morphological, syntactic and lexical semantic properties of them.

Since one of our main aims is to foster semantic interoperability in the community, the ISO Lexical Markup Framework (Francopoulo et al. 2006a) was chosen as the reference meta-model for the structure of the BioLexicon. The Lexical Markup Framework provides a common and shared representation of lexical objects that allows for the encoding of rich linguistic information. The BioLexicon is modeled in an XML DTD according to the LMF DTD: it implements the core model plus objects taken from the NLP extensions for the representation of morphological, syntactic and lexical semantics aspects of words and terms. The model consists of a number of independent lexical objects (or classes) and a set of Data Categories (DCs), i.e. attribute-value pairs which represent the main building blocks of lexical representation. In conformity to the ISO philosophy, the Data Category Selection for the BioLexicon is partially drawn from the ISO 12620 Data Category Registry (Francopoulo et al. 2006b,c, Wright 2004), and partially defined for the specific purposes of the project and the special domain. Furthermore, in order to be able to automatically constrain and check the consistency of the DCs on each specific object most DCs have been typed.

A key innovation is that the DB comes equipped with automatic loading procedures for its population with data coming from partners. Also, the BioLexicon will be linked to the BioOntology, and the two will serve as the terminological backbone for harvesting information from documents.

3. The BioLexicon data model

The core lexical objects of the BioLexicon are: *LexicalEntry, Lemma, Sense*, and *Syntactic Behaviour*.

The *Lexical Entry* class represents the abstract units of vocabulary at three levels of description: morphology, syntax and semantics. To ensure modularity and extendibility the three levels of description are accounted for in separate lexical objects, independently linked to the *LexicalEntry*, which, thus, functions as a bridge among the *Lemma* – and their forms – its related *Sense(s)*, and *Syntactic Behavior(s)*. *Lexical Entry* bears a Part-Of-Speech DC, plus additional non mandatory attributes.

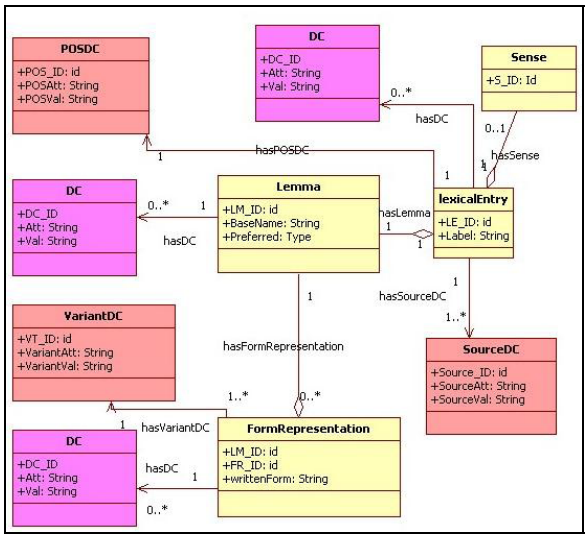


Figure 1: *The BioLexicon morphology extension.*

A specific requirement coming from the biology community is that the resource should keep track of the ids of the terms in other well known reference databases and ontology (see Harkema et al. 2004). External references in the BioLexicon are thus represented as typed data categories that are added as attributes to the *Lexical Entry* object. *Lemma* is used to represent the base form of lexemes plus additional grammatical properties; because it is in a one-to-one relation with the *Lexical Entry*, homonyms in the BioLexicon are represented as separate entries.

Syntactic Behaviour is dedicated to the representation of how lexical items and terms are used in context. One *Syntactic Behaviour* describes specific syntactic properties of an item related to one of the possible contextual behaviors of a lexical entry.

Finally, the basic information units at the semantic level are senses. *Sense* is therefore the class used for the representation of the lexical meanings of a word/term, and it is inspired by the SIMPLE Semantic Unit (Ruimy *et al.* 2003). Each *Sense* instance represents and describes one meaning of a given *Lexical Entry*, contains information on the specific (sub)domain to which the sense applies, and contains a link to the Bio-ontology.

3.1. The morphology extension

In a terminological lexicon for biology a key requirement is the representation of the different types of term variants. Variants in fact are extremely frequent and common in the biology literature (Nenadic et al. 2004). Given that linguistic information are automatically extracted from texts, in the BioLexcon we chose to distinguish only between two types of variants: variants of form and semantic variants. The morphology extension therefore has been implemented mainly to allow for a rich and extensible representation of variants of form. The *FormRepresentation* object has in fact the function of representing multiple orthographies. The basic DC specifying the *FormRepresentation* is the *writtenform*, i.e. the string identifying the form in question. Each variant is then adorned with properties represented by specific DCs: the type of variants ("orthographic", for variants and "preferred" for baseforms), and a confidence score that the automatic extraction techniques assigned to each variant (for details on the treatment of variants see Quochi et al 2007). The *InflectedForm* class is used in the BioLexicon to represent the automatically generated inflected forms of domain-relevant verbs.

3.2. The syntactic extension

As mentioned above, *Syntactic Behavior* represents one of the possible behaviors that a lexical entry shows in context. A detailed description of the syntactic behavior of a lexical entry is further defined by the *Subcategorisation Frame* object, which is the "*hearth*" of the syntax module.

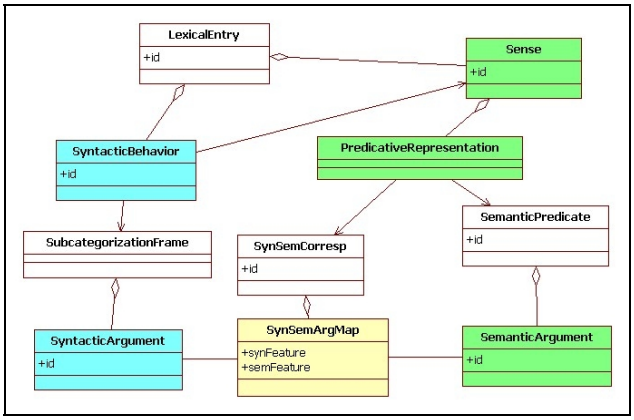


Figure 2: *The BioLexicon syntactic extension.*

Subcategorisation Frame is used to represent one syntactic configuration and does not depend on individual syntactic units; rather it may be shared by different units. The LMF syntax extension is adapted in view of accommodating the subcategorisation behaviors of terminological verbs automatically extracted from texts by appropriate NLP algorithms, and thus a probability score will be recorded as a property of the *Syntactic Behavior* belonging to a give *SubcategorisationFrame*.

3.3. The semantic extension

The semantic module of the lexicon is made of lexical objects related to the *Sense* class. As said above, *Sense* represents lexical items as lexical semantic units. Semantic relatedness among terms is expressed through the *SenseRelation* class, which encodes (lexical) semantic relationships among instances of the Sense class. The BioLexicon *Semantic Relations* build on the 60 *Extended Qualia relations* of the SIMPLE model and are represented as Data Categories drawn from the Data Category Selection specifically defined to meet the needs of the bio-domain and of the BOOTStrep project (for details on bio-relations and the semantic extension in general see Monachini et al. 2007).

The *Semantic Predicate* class, instead, is independent from specific entries and represents an abstract meaning

together with its associated semantic "arguments". It represents a meaning that may be shares by more senses that are not necessarily considered as synonyms. It is referred to by the *Predicative Representation* class, which represents the semantic behavior of lexical entries and senses in context, i.e. it describes the complete semantic argument structure of a predicative lexical item.

4. The BioLexicon data base

The software implementation of the BioLexicon consists of two modules: a relational database MySQL and a java-based loading software for the automatic population of the database. External to the DB, but fundamental for its automatic population, is an XML Interchange Format (XIF hereafter) specifically tailored to the BioLexicon structure.

4.1. Database Architecture

The database is structured into three logically distinct but strongly interconnected layers (see Figure 3). The TARGET FRAME layer contains the actual BioLexicon tables, i.e. tables that directly instantiate the lexical objects and relations designed in the conceptual model presented in the sections above and defined in a corresponding DTD.

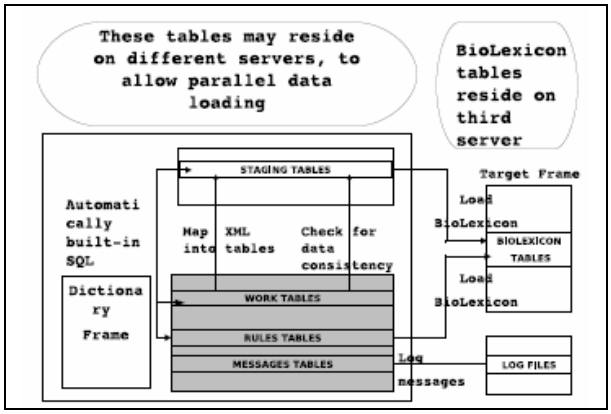


Figure 3: *The BioLexicon database architecture*

Each module of the BioLexicon structure (syntax, semantics, morphology) is independently accessible by queries and represents a self-consistent set of data. Each table has its internal identifier, but native identifiers, i.e. original source identifiers, are maintained as attributes of a given lexical entry (i.e. as DCs). The other two layers, DICTIONARY and STAGING, may be considered as operational layers: the DICTIONARY contains rules to populate target tables, whereas STAGING is an intermediate set of hybrid tables for the storage of volatile data: staging table columns consist of attributes of the XML Interchange Format (see below) and attributes of target tables. In addition, the staging frame is the level dedicated to data-cleaning and normalization. This neat separation between target tables (the BioLexicon proper) and operational tables allows for the optimization of the uploading of data into the BioLexicon DB and ensures extensibility both of the database and of the uploading procedures. Faced with the need to add a new table (i.e. a new lexical object) or a new attribute to an existing table, for example, it is sufficient to add only the definition of the new tables or attributes to the DB and to modify accordingly only the relevant portion of the Dictionary Frame layer to handle the novelties.

4.2. The XIF and Automatic Population

The XML Interchange Format (XIF) is designed with the main purpose of automatically populating the BioLexicon with data provided by domain experts and by lexical acquisition systems. Within the project, data are extracted and gathered through automatic procedures both from existing resources and by research papers in biology. The XIF DTD is to be considered a simplified version of the BioLexicon DTD which accommodates the needs of data providers and facilitates the automatic uploading of the DB. By means of the XIF, we therefore allow for a standardization of the data extracted from the different terminological resources and from texts. Differently from other similar lexical systems (like Termino), the XIF allows for the independency of the uploading procedures from native data formats. This way, any system/group wishing to feed new data into the BioLexicon would only need to encode this in an XML file according to the XIF DTD. The XIF DTD partially mirrors the way biological data are stored into domain knowledge databases and also accommodates the way these data are extracted from those resources. The XIF is organized in clusters of terms, i.e. in sets of coherent types of information. A cluster contains one or more synonymous entries with information related to their lemmas, parts-of-speech, inflected forms, semantic relations and external references. Such an organization, furthermore, permits the splitting of the input file by clusters, which in turn allows for a parallel uploading of the data into the DB. The XIF, therefore, has been conceived as a link between existing resources and the BioLexicon. From the implementation perspective, the XIF may be considered as the physical counterpart of the Dictionary Frame: that is to say, the loading software uses rules contained in the dictionary tables to correctly interpret the input file. Faced with the need to add a new table or to alter an existing one, it is sufficient to add new elements or new attributes to the XIF DTD and to add relevant instructions to the dictionary frame. The loading software interprets the XIF as usual and applies the new Dictionary instructions automatically inserting the new table or attribute. This property of the XIF together with the neat separation of the three layers of the DB mentioned above allows any agent (human or machine) to easily populate, update and create new objects and attributes. The data that we are currently uploading are terms gathered by Project BOOTStrep partners from existing databases with information relevant to their external source references, variants, lexical category and, to some degree, semantic relations.

5. Statistics and Validation

Since we are dealing with the building of a lexical resource within an ongoing project, no evaluation is available yet. However, some kind of content validation can be made, taking into account the input resources and the documentation so far produced. For validation of the resource we readapt templates from the ELRA Validation Manual for Lexica 2.0 (Fersøe, 2004).

The BioLexicon is a monolingual English lexical resource that represents Bio-terms as well as general lexical items relevant to the bio-medical domain. Both nouns and verbs are represented: nouns cover a wide portion of the existing biological terminologies and come from the most used databases in the sector. Mainly they represents terms denoting enzymes, chemical, species, genes and/or proteins, especially those relevant for the gene regulation topic. Verbs are also

represented, but are very limited in number: for the time being only verbs relevant to the E.Coli species have been included (see Table 1).

Table 1: *The BioLexicon coverage*

POS	Sem.Type	Lexical Entries	Variants
N	Enzymes	4,016	9,379
N	Genes/Proteins	841,164	1,547856
N	Species	367,565	439,336
N	Chemicals	13437	51332
POS	**Sem.Type**	**Lexical Entries**	**Infl. Forms**
V		489	2435

For each entry the part-of-speech is encoded together with the written form of both its lemma and its variants. Also some semantic relations are instantiated: synonymy, part-of, and a few other biological relations (see Table 2).

Table 2: *Instantiated semantic relations*

Relation Type	#
is_a	464,078
is_synonym_of	847,873
is_part_of	189
is_conjugate_base_of	637

The BioLexicon resource is implemented as a MySQL relational database that runs both under Linux and Windows systems. The database is shaped according to the model XML DTD, and therefore easily allows for XML outputs. So far, the DB has automatically uploaded all the input files provided by the bio-experts within the BOOTStrep project (EBI-Cambridge and MIB Manchester), which gathered and systematized biological terminology from the major online databases. Altogether the DB contains 25 million records and occupies ca. 1.7G of memory space. It consists of 1,309,496 Lexical Entries, Lemmas and Senses; 2,047,903 orthographic variants and 1,697,804 semantic relations.

6. Conclusions

The biological literature is continuously developing, which leads to the need for large-scale terminological lexicons that can support text mining and information extraction applications, which would make the life of biologists much easier. The BioLexicon, described in this paper, is designed to integrate both typical information provided by domain ontologies and typical linguistic information generally available in open-domain computational lexicons. The DB, as well as the BioLexicon model, is modular, extensible and, by means of the protocol defined through the XIF, can easily and automatically upload new data, and provide outputs by means of web services. A brief hint at internal validations of the resource has been added, based on the first data coming from our project partners. A suitable evaluation of the resource is not feasible for the moment, and will only be possible in the future, when the DB will actually be integrated in the BOOTStrep UIMA infrastructure; that is when it starts to dynamically interact with applications.

7. Acknowledgements

The work presented here has been funded by the EC´s 6th Framework Programme (4th call), conducted within the BOOTStrep consortium under grant FP6-028099.

8. References

[1] Calzolari N., Bertagna F., Lenci A., Monachini M. (eds) 2003. Standards and best Practice for Multilingual Computational Lexicons. MILE (The Multilingual ISLE Lexical Entry). ISLE CLWG Deliverable D2.2 & 3.2 Pisa.

[2] Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006a. Lexical Markup Frame-work (LMF). Proceedings of the LREC 2006, Genova, Italy.

[3] Francopoulo G. Monachini M., Declerck T., Romary L.. 2006b. The relevance of standards for research infra-structure. Proceeding of the LREC 2006, Genoa, Italy

[4] Francopoulo G., Monachini M., Declerck T., Romary L., 2006c. Morpho-syntactic Profile in the ISO-TC37/SC4 Data Category Registry. In Proceedings of the LREC2006, Genova, Italy.

[5] Hahn U., Markó K. 2001. Joint Knowledge Capture for Grammars and Ontologies. Proceedings of the 1st in-ternational conference on Knowledge capture Victoria, British Columbia, Canada.

[6] Harkema H., Gaizauskas R., Hepple M., Angus R., Roberts I., Davis N., Guo Y. et al. 2004. A Large Scale Teminology Resource for Biomedical Text Processing. HLT-NAACL 2004 Workshop: Bio-LINK 2004, Linking Biological Literature, Ontologies and Database, Boston, Massachusetts, USA.

[7] Ide N. and Romary L. 2004. A registry of standard data categories for linguistic annotation. In Proceeding of the LREC04, Lisbon, Portugal.

[8] ISO-12620 2006. Terminology and other content language resources – Data Categories – Specifications of data categories and management of a Data Category Registry for language resources. ISO/TC37/SC3/WG4.

[9] Kors J. A. et al. 2005. Combination of Genetic Databases for Improving Identification of Gens and Proteins in Text, Rotterdam, Netherlands

[10] ISO-12620. 2006. "Terminology and other content language resources – Data Categories – Specifications of data categories and management of a Data Cate-gory Registry for language resources. ISO/TC37/SC3/WG4.

[11] Monachini M., Quochi V., Ruimy N., Calzolari N. 2007. Lexical Relations and Domain Knowledge: The BioLexicon Meets the Qualia Structure. In GL2007: 4th International Workshop on Generative Approaches to the Lexicon, 10-11 May 2007, Paris. CD-ROM.

[12] Quochi V., Del Gratta R., Sassolini E., Monachini M., Calzolari N. 2007. Toward a Standard Lexical resource in the Bio Domain. In Vetulani (ed.), Proceedings of 3rd Language & Technology Conference. Fundacja Uniwersytetu im A. Mickiewicza, Poznań. 295-299.

[13] Ruimy N., Monachini M., Gola E., Calzolari N., Del Fiorentino M. C., Ulivieri M., Rossi Sergio. 2002. In Linguistica Computazionale, Vol.XVIII-XIX, I.L.C. and Computational Linguistics, special issue, A. Zampolli, N. Calzolari, L. Cignoni, (Eds.), I.E.P.I., Pisa-Roma.

[14] Wright S.E. 2004. A global data category registry for interoperable language resources. In Proceedings of LREC04 Lisbon, Portugal.

ORAL PRESENTATIONS

Welcome to the participants

Antonio Sassano
Director General
Fondazione Ugo Bordoni

It is my very pleasant duty to welcome all participants to the third Langtech meeting.
As many of you may know, I represent the Fondazione Ugo Bordoni. Fondazione Ugo Bordoni (FUB), incorporated on October 13th, 2000 upon closing the former Institution with the same name, is recognised by the Law as an Institution of high culture that elaborates and proposes strategies on the development of the communications sector that it can support in the national and international competent centres, and assists the Ministry of Communications to tackle and solve technical, economical, financial, managerial, normative, and regulatory problems encountered in its statutory activities.
FUB carries on research, study and consultancy activities in the area of Information and Communications Technologies.
FUB has a sound experience, recognised at international level, in several areas including multimedia communications, and others. At the international level, it cooperates with several institutions by participating to relevant standardisation for European research programmes.
As part of research into communications technologies and more specifically of speech processing, field in which the Foundation has consolidated experience, we are proud to organise a Langtech meeting hoping that this opportunity could favour all possible synergies among the research groups, the developers and the users.
With my warm thanks to dr. Lönnroth, the director general of the translation section of European Communities and to the Organising Committee I wish a fruitful workshop to all participants.

General Chairman's message

Giordano Bruno Guerri
Conference Chair
Fondazione Ugo Bordoni

I am extremely privileged to have the honour and pleasure to welcome you, on the behalf of the Organizing Committee, to Langtech 2008, the third Langtech forum.
It is not necessary for me to explain to you the importance of international meetings that facilitate communication and cooperation among communities and organizations which work in the field of the development, deployment, and exploitation of TAL (Trattamento Automatico della Lingua) technology, the automatic processing of spoken and written language.
The first Langtech forum was held in Berlin in 2002 while the second one was held in Paris in 2003. Subsequently, due to the completion of the Euromap project, the meetings were discontinued, and therefore it is a special honour for us to revive this initiative in Italy, where many companies and research institutions are working actively on the development of language technologies.
After our conference there is the concrete prospect of a new Langtech meeting next year in another European country. The aim is to make these meetings more regular, and in this way to strengthen the exchange of experiences in this area which is crucial not only for technological development but also for the maintenance of multilingualism and multiculturalism.
The event focuses on the exhibition area, which is located in the adjacent room and which contributes concretely to demonstrating even to the most sceptical that linguistic technologies can contribute in important ways to the quality of everyday life. Particularly useful, from this point of view, are automatic translation systems that are increasingly effective and precise and that can now also translate speech, and not only written texts.
I also want to mention, in a world which is submerged by increasing quantities of information, the new systems that are capable of extracting from enormous amounts of unstructured data the information that we request.

The meeting will also provide several working sessions in which qualified experts will review various issues related to language and language technologies. Among these working sessions, for the sake of brevity, I will only mention the session dedicated to the analysis of the market, which certainly will be of interest to all participants.
Finally, in this edition of Langtech the Organizing Committee has reserved some space for the presentation of more academic research work, with the aim to encourage greater interaction between research and development. In order to promote synergies between ideas and their implementation, there will also be talks by economic organisations, such as a speech concerning the role of "venture capitalism" in this area.
To underline the importance of language technology for Europe, I am particularly happy that dr. Lönnroth, the director general of the translation section, will participate to this session on the behalf of Dr. Orban, the commissioner on multilingualism.
Many thanks to all those who, in various ways, have contributed to the event: those who have sent their work to the Scientific Committee, which has evaluated them, to the organizing committee and the local committee and to everyone else who has contributed to the preparation of this Langtech. In particular many thanks to the Ugo Bordoni Foundation, represented here by its Director Antonio Sassano, to all our sponsors, from Loquendo, which first believed in the success of the event, to the newly created Pervoice and to all other sponsors: IBM, CELCT, FBK, Rai, Speech Technology, AVIOS and others.
Once again, most cordially welcoming all of you, participants and guests of Langtech, I wish you an enjoyable and fruitful conference and a very pleasant stay in this so prestigious building and in this beautiful city.

Language Technologies and the European Commission

Karl-Johan Lönnroth
Director General
Directorate-General for Translation, European Commission

The activities of the European Commission depend heavily on language technologies. The huge mass of texts that are published every day in the Official Journal of the European Union, on the Internet and in thousands of brochures, leaflets and reports in 23 different languages could not be managed without powerful electronic tools and communication networks.

The linguistic diversity of the European Union keeps growing, both at institutional level and within its Member States, and so does demand for information and participation in European policy-making. Human language technologies are therefore vital to ensure the sustainability of multilingualism policy.

The European Commission finances research in the field of language technologies through its Research Framework Programmes. The Directorate-General for Translation (DGT) is one of the largest laboratories for testing, designing, refining and implementing new linguistic tools.

The Commission is also a power user of language technologies, from content management to computer-assisted translation, from multi-engine and multilingual searches to terminological and documentary databases, from dictation software to applications for the management and sharing of glossaries and to sophisticated authoring tools. It has for over 30 years invested in such tools, thereby contributing to the development of the market.

The combined effects of an unrivalled degree of multilingualism with the unique nature of the texts translated at the Commission – for the most part, legislative texts, demanding extreme accuracy and absolute concordance – makes its expertise in this field a precious asset for language research and for the whole translation industry.

Access to the technologies developed at the Commission or to the data made available through such technologies – machine translation, translation memories, terminological and documentary databases, multilingual news

gathering and aggregation – is offered to the public, to national authorities or to the research community.

As the technological environment evolves, the professional profile of the translator will also have to develop. A huge effort will be needed in terms of training translators and assistants to make the best use of available resources in a constantly and rapidly evolving market.

While developing new services, including summarising, linguistic editing, web translation and editing, and localisation, DGT emphasises the quality of the Commission's written communication, thereby improving its legitimacy, transparency and efficiency.

This holistic approach requires a constant search for the best technologies available, to be developed in cooperation with the language industry and adapted to the Commission's very special needs. It also requires some realism, given that technology is a necessary aid but not a complete recipe for meeting all translation challenges.

Using frames in Spoken Language Understanding

Renato De Mori

Laboratoire d'Informatique – Université d'Avignon - France
renato.demori@univ-avignon.fr

Abstract

This paper reviews basic concepts for natural spoken language interpretation by computers. Frame structures are described as suitable computer representations of semantic compositions. A process is introduced for obtaining basic semantic constituents by translating word sequences into basic semantic constituents and for composing constituent hypotheses into frame structures. Experimental results with the French telephone corpora are reported. They show that Finite State conceptual language models are useful for translating word hypotheses into states representing progressive semantic compositions and the use of Conditional Random Fields (CRF) improves the accuracy of constituent hypothesization.

Index Terms: Spoken Language Understanding, computer meaning representation, meaning representation languages, Frames, finite-state conceptual language models.

1. Introduction

Epistemology, the science of knowledge, considers a datum as basic unit. A datum can be an object, an action or an event in the world and can have time and space coordinates, multiple aspects and qualities that make it different from others. A datum can be represented by a word or it can be abstract and be represented by a concept. There may be relations among data.

Computer epistemology deals with observable facts and their representation in a computer. Knowledge about the structure of a domain represents a datum by an object and groups objects into *classes* by their properties. Classes are organized into *hierarchies*. An object is an *instance* of a class. Judgment is expressed by *predicates* which describe. Predicates have arguments which are variables whose values have to respect some constraints.

Natural language refers to data in the world and their relations. Sentences of a natural language are sequences of words. Groups of words have associated conceptualizations also called *meanings* which can be selected and composed to form the meaning of the sentence.

Semantics deals with the organization of meanings and the relations between signs or symbols and what they denote or mean (Woods, 1975). Human conceptualization of the world is not well understood. Nevertheless, good models for this organization assume that basic *semantic constituents* expressed by a language are organized into *conceptual structures*.

In (Jackendoff, 2002, p. 124) it is suggested that semantics is an independent generative system correlated with syntax through an interface. Computer semantics performs a conceptualization of the world using well defined elements of programming languages. Programming languages have their own syntax and semantic. The former defines legal programming statements, the latter specifies the operations a machine performs when an instruction is executed. Specifications are defined in terms of the procedures the machine has to carry out. Semantic analysis of a computer program is essential for understanding the behavior of a program and its coherence with the design concepts and goals.

Natural language interpretation by computers generate concept hypotheses represented in a *semantic language*. The definition of a semantic language can be based on a formal grammar but has to include procedures for obtaining interpretations from sentences. Procedures are executed by computational processes belonging to an *interpretation strategy*.

Computer programs conceived for interpreting natural language differ from the human process they model. They can be considered as approximate models for developing useful applications, interesting research experiments and demonstrations. Semantic representations in computers usually treat data as *objects* respecting logical *adequacy* in order to formally represent any particular interpretation of a sentence. Even if utterances, in general, convey meanings which may not have relations which can be expressed in formal logic (Jackendoff, 2002, p. 287), formal logic has been considered adequate for representing natural language semantics in many application domains.

In many applications, computer systems interpret natural language for performing actions such as a data base access and display of the results and may require the use of knowledge which is not coded into the sentence but can be inferred from the system knowledge stored in long or short term memories.

It is argued in (Woods, 1975) that a specification for natural language semantics requires more than the transformation of a sentence into a representation. In fact, computer representations should permit, among other things, legitimate conclusions to be drawn from data (Mc Carty and Hayes, 1969).

Spoken Language Understanding (SLU) is the interpretation of signs conveyed by a speech signal. This is a difficult task because meaning is mixed with other information like speaker identity and environment. Natural language sentences are often difficult to parse and spoken messages are often ungrammatical. The knowledge used is often imperfect and the transcription of user utterances in terms of word hypotheses is performed by an Automatic Speech Recognition (ASR) system which makes errors.

Some important challenges in SLU are:

- meaning representation,

- definition and representation of signs,
- conception of relations between signs and meaning and between instances of meaning,
- processes for sign extraction, generation of hypotheses about units of meaning and constituent composition into semantic structures,
- robustness and evaluation of confidence for semantic hypotheses,
- automatic learning of relations from annotated corpora,
- collection and semantic annotation of corpora.

This paper describes a process for SLU. Reviews on SLU research can be found in in (De Mori, 1998, Wang 2006 and Mc Tear 2006).

2. Computer representations of meaning using frames

Computer representation of meaning is described by a Meaning Representation Language (MRL). It is preferable that MRL is conceived with reference to a representation model coherent with a theory of epistemology. As such, it should take into account, *intension* and *extension*, relations, reasoning, composition of semantic constituents into structures, procedures for relating them with signs.

The semantic knowledge of an application is a *knowledge base (KB).* A convenient way for reasoning about semantic knowledge is to represent it as a set of logic formulas. Formulas contain variables which are bound by constants and may be typed. An object is built by binding all the variables of a formula or by composing existing objects.

Semantic compositions and decisions about composition actions are the result of an inference process. Basic *inference problem* is to determine whether $KB \models F$ which means that KB *entails* a formula F, meaning that F is true in all possible variable assignments (worlds) for which KB is true.

The formulas in a KB describe concepts and their relations which can be represented in a network called *semantic network*. A semantic network is made of nodes corresponding to entities and links corresponding to relations. This model combines the ability to store factual knowledge and to model associative connections between entities (Woods, 1975).

The structure of semantic networks can be defined by a graph grammar. Computer programming classes and objects called *frames* can be defined to represent entities and relations in semantic networks. Frame representation can be derived from semantic networks They are computational structures (Kifer et al., 1995) and also cognitive structuring devices in a semantic construction theory (Fillmore, 1968).

Part of a frame is a data structure which represents a concept by associating to the concept name a set of roles which are represented by *slots*. Finding values for roles corresponds to fill the frame slots. A *slot filler* can be the instance of another frame. There may be *necessary* and *optional* slots. *Fillers* can be obtained by *attachment* of procedures or detectors (of e.g. noun groups), *inheritance*, default.

A *facets* can be associated to a slot. Constraints on the values that can fill a slot can be stored into a slot facet. Constraints can be expressed by probability distributions on the possible filler values (Koller, 1998).

Descriptions are attached to slots to specify constraints. Descriptions may have connectives, coreferential (descriptions attached to a slot are attached to another and vice-versa), declarative conditions.

Verbs are fundamental components of natural language sentences. They represent actions for which different entities play different roles. Actions reveal how sentence phrases and clauses are semantically related to verbs by expressing cases for verbs. A *case* is the name of a particular *role* that a noun phrase or other component takes in the state or activity expressed by the *verb* in a sentence. There is a case structure for each main verb. Attempts were made for mapping specific *surface cases* into a deep semantic representation expressing a sort of semantic invariant. Many deep semantic representations are based on *deep case n-ary relations* between concepts as proposed by Fillmore (Fillmore, 1968). *Deep case* systems have very few cases each one representing a basic semantic constraint.

Early frame representations were used to represent facts about an object with a property list. For example, a specific address can be represented by the following frame:

{a0001

instance_of	address
loc	Avignon
area	Vaucluse
country	France
street	1, avenue Pascal
zip	84000*}*

Here a0001 is a handle that represents an instance of a class which is specified by the value of the first slot. The other slots, made of a property name and a value, define the property list of this particular instance of the class “address”.

The above frame can be derived (Nilsson, 1981), after skolemization from the following logic formula:

$$(\exists x)\left\{\begin{array}{l}\text{instance\_of}(x,\text{address}) \wedge \text{loc}(x,\text{Avignon}) \wedge \\ \wedge\ \text{area}(x,\text{Vaucluse}) \wedge \text{country}(x,\text{France}) \wedge \\ \wedge\ \text{street}(x,1\ \text{avenue Pascal}) \wedge \text{zip}(x,84000)\end{array}\right\}$$

A definition, with a similar syntax, but with a different semantic is provided for the address class which defines the structure of any address:

{address

loc	TOWN
area	DEPARTMENT OR PROVINCE OR STATE
country	NATION
street	NUMBER AND NAME
zip	ORDINAL NUMBER*}*

The syntactic analysis of a parsable sentence can be used for establishing relations between syntactic structures and

meaning. Concerning the relation between syntax and semantics, in (Jackendoff, 1990), it is observed that:

- Each major syntactic constituent of a sentence maps into a conceptual constituent, but the inverse is not true.
- Each conceptual constituent supports the encoding of units (linguistic, visual,…).
- Many of the categories support *type|token* distinction.
- Many of the categories support quantification.
- Some realizations of conceptual categories in conceptual structures can be decomposed into a *function|argument* structure.

For certain types of applications, domain-dependent semantic knowledge has been integrated into stochastic semantic grammar. A survey on these grammars and their use can be found in (Wang, 2005)

3. Conceptual language models for a modular SLU architecture

Generation of hypotheses about semantic constituents and semantic composition are different operations in nature and can be performed by different techniques implemented in different modules. Each module can integrate different models in order to improve robustness. Specific *conceptual language models* can be used in ASR decoding to obtain constituent hypotheses directly from the signal or from word hypotheses. Other types of knowledge are used in shallow parsers(Pradhan, 2004). In order to avoid the complexity of context-free and context-sensitive grammars, finite-state approximations of context-free grammars are proposed in (Pereira, 1990). Approximations of TAG grammars are described in (Rambow et al., 2002). A review of these approximations is provided in (Erdogan et al., 2005).

In both cases, a generic n-gram LM can be used with specific stochastic finite-state machines (FSM), one for each semantic constituent c_j.An example of LMs based on stochastic FSMs can be found in (Prieto et al., 1994). Stochastic Automata and their use for hypothesizing semantic constituents are proposed in (Gorin 1997, Nasr., 1999). Finite-state Hidden Markov Models (HMM) for SLU are proposed in (Pieraccini, 1991).

In (Kawahara et al., 1999), an automaton extracts key phrases from continuous speech and converts them to commands for a multi-modal interaction with a virtual fitting room. Interpolation of generic n-gram models and specific concept models is performed by maximizing the divergence between a linear interpolation of the two models and the generic n-gram model. A greedy algorithm is proposed (Riccardi and Gorin, 2000).

In (Drenth and Ruber, 1997), it is proposed to obtain a semantic interpretation of a dialog "turn" (one or more sentences) by extracting concept hypotheses from a word lattice. Each concept hypothesis is extracted with a *conceptual semantic context-free grammar*.

Finite state models can be made more robust by modifying the original topology to take into account possible insertions, deletions and substitutions. Insertion of words not essential for characterizing a semantic constituent can be modeled by groups of syllables.

Recent advances in research on stochastic FSM made it possible to generate a probabilistic lattice of conceptual constituent hypotheses from a probabilistic lattice of word hypotheses.

The solution proposed in (Raymond et al., 2006) is now introduced. A stochastic finite-state conceptual language model CLM_j is conceived for every semantic constituent c_j. An initial ASR activity uses a generic LM, indicated as GENLM, for generating a graph of word hypotheses. Let WG be the stochastic FSM representing the lattice of word hypotheses generated by an ASR system. A knowledge source, is built by connecting all the CLM_j in parallel as shown in Figure 1. Such a knowledge source is composed with WG leading to an automaton SEMG in which concept tags representing semantic constituents are added to arcs in WG:

$$SEMG = WG \circ \left(\bigcup_{c=0}^{C} CLM_c \right)$$

operator $\circ$ indicates composition.

CLM_0 is a generic model for sequences of words which do not express concepts in the application domain.

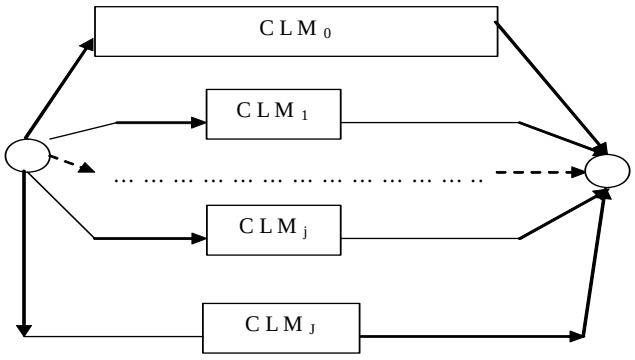


Figure 1 – Composition of conceptual language models

In order to obtain the concept tags representing hypotheses that are more likely to be expressed by the analyzed utterance, SEMG is projected on its outputs leading to a weighted Finite State Machine (FSM) with only indicators of beginning and end words of semantic tags. The resulting FSM is then made deterministic and minimized leading to an FSM SWG given by:

SWG=OUTPROJ(SEMG)

where OUTPROJ represents the operation of projection on the outputs followed by determinization and minimization.

A sequential interpretation strategy for a dialogue service in the France Telecom 3000 system (Minescu, 2007) using confusion networks (Hakkani-Tur, 2006) on relevant messages only. Results are reported in Table 1.

Conditional Random Fields (CRF) have been used for generating hypotheses about semantic constituents for the

MEDIA French corpus. Results and comparisons with other methods obtained by (Raymond, personal communication, 2007) on predicate/attribute pairs using the 1 best ASR hypothesis are provided in Table 2. Further 10% error reduction have been observed by method combination.

Table 1 – Interpretation results using conceptual LMs

	Baseline (1-best)	*sequential strategy*
Insertion rate	17.2 %	8.8 %
Substitution rate	6.1 %	5.6 %
Deletion rate	2.7 %	5.2 %
Interpretation error rate (IER)	**26.0 %**	**19.6 %**

Table 2 – Comparison of interpretation results obtained in the MEDIA corpus

	concept error rate (CER)
Conditional Random Fields	25.2 %
Finite State Transducers	29.5 %
Support Vector Machines	29.6 %

4. Probabilistic logic and inference for slu

In practical applications, SLU is part of a dialogue system whose objective is the execution of actions to satisfy a user goal. Actions can be executed only if some pre-conditions are asserted true and their results are represented by post-conditions. Preconditions for actions can be formulated is in formal logic. Preconditions for actions depend on instances of semantic structures.

The system knowledge is made of general knowledge, e.g. knowledge about dates and time, and specific domain knowledge, e.g. the details of a telephone service. Let us call the resulting knowledge *in-domain knowledge*.

As a dialogue progresses, part of the domain knowledge is instantiated. The purpose of the dialogue is to interpret the user beliefs and goals and represent them with the MRL. Eventually, system actions like accessing a data base, are performed to satisfy a user request. If MRL contains frames, then user sentences should cause the instantiation of some frames, the assignment of values to some frame roles and functions to describe them. Instantiation is based on what the user says, but also on what can be inferred about the implicit meaning of each sentence.

Control strategies for interpretation determine how semantic structures are built, how expectations are defined and how knowledge structures are matched with input data in the presence of constraints and imprecision.

There are two basic types of strategy. One is based on path extraction from a semantic or a frame network. The other adopts a constructionist approach that can use one or more of the following methods: inference, parsing, abduction, agenda-based formation and scoring of interpretation hypotheses called *theories*.

In the constructionist approach, the meaning of a complex phrase is considered to be a function of the meanings of its constituent parts and the way in which these parts are syntactically combined. Reasoning is performed by programs that activate memory structures by placing activation markers on them. Nodes of the structure are activated when the corresponding concepts are instantiated. Active nodes may spread activation markers to hypothesize or predict the activation of concepts which have not yet been instantiated. When two markers collide in the same node, a path is identified indicating a possible inference. Frame-activated inference is discussed in (Norvig 1987).

Early approaches to SLU used semantic representations in terms of partitioned semantic networks (Walker, 1975). Marker propagation was used for making predictions about concepts likely to appear in the natural language messages. Concept hypotheses were generated by templates matching word and partial parses (obtained with a best first parser) with semantic structures.

In the Hearsay II SLU architecture (Erman et al., 1980), a heterarchical architecture was used for applying rules for matching and inference. An agenda based control strategy selects a rule whose precondition matches the content of a blackboard. If matching is successful, then actions are performed which modify the content of the blackboard.

The weakness of these approaches was that they did not contain an effective method for evaluating the confidence of the generated hypotheses.

If instances of semantic constituents are structured into probabilistic frames, it is possible to have a probability model for the values that can fill a slot (Koller, 1998). It is also possible to inherit probability models from classes to subclasses, to use probability models in multiple instances and to have probability distributions representing structural uncertainty about a set of entities.

It is shown that it possible to construct a Bayesian Network (BN) for a specific instance-based query and then perform standard BN inference if the graph obtained from a list of statistical dependencies between slot values is acyclic. Otherwise, Markov Logic Networks (MLNs) can be used (Richardson, 2006).

Probabilities obtained with these models can be combined with probabilities computed by SLU components in a way that is now introduced.

Let us consider an instance $\Gamma_{i,j}$ of a frame F_i.

Let us indicate by $\Gamma_{i,j} : \lfloor \gamma_{i,j,1}, ..., \gamma_{i,j,k}, ..., \gamma_{i,j,K} \rfloor$ the set of roles (slots) of $\Gamma_{i,j}$ that are instantiated and possibly filled by a value.

The instantiation of each slot is based on a casual relation graphically represented as follows:

$$Y_k \rightarrow W_k \rightarrow C_k \rightarrow \gamma_{i,j,k}$$

Y_k is a sequence of acoustic feature vectors from which a word sequence W_k has been hypothersized. W_k contains the support for a semantic constituent C_k expressed by a predicate in first-order logic. If C_k has been expressed in relation to $\Gamma_{i,j}$, then it becomes the slot hypothesis $\gamma_{i,j,k}$.

There may be other dependences between slot values, represented by lings in the following feature:

Each slot hypothesis can be evaluated by the following probability:

$$P[W_k, C_k, \gamma_{i,j,k} | Y_k] \approx P[C_k, \gamma_{i,j,k} | W_k] P[W_k | Y_k] \approx$$
$$\approx \frac{P[W_k | C_k, R(\gamma_{i,j,k})]}{P[W_k]} P(\gamma_{i,j,k}) P[W_k | Y_k]$$

since $P[C_k | \gamma_{i,j,k}] = 1$

The ratio $\frac{P[W_k | C_k, R(\gamma_{i,j,k})]}{P[W_k]}$ can be obtained with two different language models (LMs) a generic LM for the denominator and an LM estimated on dialog turns expressing C_k and a relation $R(\gamma_{i,j,k})$ to an instance of F_i. Notice that C_k is hypothesized in a dialog turn using a specific concept LM and $P[W_k | C_k, R(\gamma_{i,j,k})]$ could also be approximated by $P[W_k | C_k)]$ obtained directly with this concept LM. Notice also that if the LMs are estimated on entire turns rather than concept supports, the ratio of probabilities will be mostly determined by the n-grams of the words characterizing the supports, especially if unigram LMs are considered. The LM used for computing the numerator can also be obtained by interpolating a generic LM with a relation specific one.

As an evidence indicator for the entire instantiation $\Gamma_{i,j}$, let us define the following vectors
$C_{i,j} : [C_1, ..., C_k, ..., C_K]$,
$W_{i,j} : [W_1, ..., W_k, ..., W_K]$,
$Y_{i,j} : [Y_1, ..., Y_k, ..., Y_K]$. Assuming also that each concept has a support that is somehow different from the supports of other concepts and assuming independence among supports, one gets:

$$P\{\Gamma_{i,j}, C_{i,j}, W_{i,j} | Y_{i,j}\} =$$
$$= P[\Gamma_{i,j}] \left\{ \prod_{k=1}^{K} \frac{P[W_k | C_k, R(\gamma_{i,j,k})]}{P(W_k)} \right\} P[W_{i,j} | Y_{i,j}]$$

Confidence indicators can be introduced to replace some probabilities. Let $\Phi_{i,j} = [\varphi_{i,j,1}, \quad , \varphi_{i,j,k}, ..., \varphi_{i,j,K}]$ be a vector of confidence indicators, one for each slot. In this case, the following computation can be performed:

$$P\{\Gamma_{i,j} | \Phi_{i,j}\} = \frac{P\{\Phi_{i,j} | \Gamma_{i,j}\} P[\Gamma_{i,j}]}{P\{\Phi_{i,j}\}}$$

Vector quantization can be introduced for $\Phi_{i,j} = [\varphi_{i,j,1}, \quad , \varphi_{i,j,k}, ..., \varphi_{i,j,K}]$.

User goals can be represented by frames. A plan for achieving each goal can be represented by a sequence of states. If different goals are hypothesized in a dialog control agenda, then the set of the corresponding plans are represented by a finite state machine. This corresponds to represent by a state a cluster of instances $\Gamma_{i,j}, C_{i,j}, W_{i,j}$ corresponding to successive slot filling of a frame instance.

As different states can be reached with different probabilities, a set of states can be active at a turn k of a dialogue A system was proposed in (Damnati, 2007) which interprets a dialogue turn message in two phases. In the first phase, a word-to-constituent transducer translates a word lattice into a constituent lattice. In the second phase, a set of precondition-action rules encoded as a transducer transforms concept hypotheses into state transitions. A lattice of words is thus translated into a set of states with attached probabilities p(S|Y) where S is a dialogue state and Y is the acoustic description of a spoken message.

The results reported in Table 3 are obtained with system 3000 data, using this approach (strategy 2) and are compared with the results obtained with a pure sequential solution (strategy1) consisting in taking the 1-best word sequence and mapping it into the 1-best concept sequence. The abbreviations are defined in tables 1 and 2, WER stays for Word Error rate.

Table 3 – Performance on goal detection using a two different strategies

	WER	IER
strategy 1	40.1	15.0
strategy 2	38.2	14.5

5. Conclusions

A modular SLU architecture has been introduced. It uses CRFs, classifiers and stochastic FSMs, which are approximations of more complex grammars, for generating semantic constituent hypotheses and probabilistic logic for performing semantic compositions.
Annotating corpora for these tasks is time consuming suggesting that it is suitable to use a combination of knowledge acquired by a machine learning procedure and human knowledge (Riccardi, 2005). Finding the best combination of these approaches is still a research issue. Other challenging problems concern the use of probabilistic logic, the introduction of suitable confidence indicators (as in Sarikaya, 2005), the design of interpretation strategies and their integration with dialog management.

6. Acknowledgements

The work described in this paper is performed in the European project LUNA, IST contract no 33549. (www.ist-luna.eu). Warm thanks to F. Béchet, G. Damnati, F. Duvert, M..J. Meurs, C. Raymond, C. Servan (LIA Avignon) and H. Ney, S. Hann (RWTH, Aachen).

7. References

G. Damnati, F. Bechet, R. de Mori, (2007) Spoken language understanding strategies on the France Telecom 3000 voice agency corpus, IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, Hawai

R. De Mori, *Spoken dialogues with computers* Academic Press, 1998.

E.W. Drenth and B.Ruber (1997) Context-dependent probability adaptation in speech understanding *Computer Speech and Language*, 11(3):225-252.

H. Erdogan, , R., Sarikaya, S.F Chen, Y.Gao and M Picheny (2005) Using Semantic Analysis to improve Speech Understanding Performance. *Computer Speech and Language*, 19(3):321-344.

L. D. Erman, F. Hayes-Roth, V. R. Lesser et R. D. Reddy. The Hearsay-II Speech Understanding System : Integrating Knowledge to Resolve Uncertainty. *ACM Computing Surveys*, 12(2):213-253, 1980.

C. J. Fillmore (1968). The case for case. in E. Bach and R. Harms eds. *Universals in linguistic theory*, Holt, Rinehart and Winston, New York, 1968.

A. L. Gorin, G. Riccardi, and J. H. Wright, 1997 How may I help you? *Speech Commununication* vol. 23, no. 1–2, pp. 113–127, October.

D. Hakkani-Tur, F. Bechet, G. Riccardi and Gokhan Tur (2006) Beyond ASR 1-Best: Using Word Confusion Networks for Spoken Language Understanding, *Computer Speech and Language* 20(4):495-514.

R.Jackendoff (1990). *Semantic Structures*, The MIT Press, Cambridge Mass.

R.Jackendoff (2002). *Foundations of language*, Oxford University Press, Oxford UK.

T. Kawahara, K. Tanaka and S. Doshira (1999) Virtual fitting room with spoken dialogue interface. *Proc. ESCA Workshop on Interactive Dialog in Multi-Modal Systems*, Kloster Irsee Germany, pp.5-8, 1999

M. Kifer, G. Lausen and J. Wu (1995) Logical Foundations of Object-Oriented and Frame-Based Languages *Journal of the Assocmtmn for Computmg Machlncry*. 42 (4), July1995. pp. 741-843

D. Koller and A. Pfeffer (1998) Probabilistic frame-based systems. *Proc. AAAI98*, pages 580–587, Madison, Wisc.,

J. McCarty and P.J. Hayes. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, Ed. by B. Meltzer and D. Michie, Edimburg University Press.

M. McTear (2006) Spoken language understanding for conversational dialog systems, IEEE/ACL Workshop on Spoken Language Technology Aruba, December 2006

B. Minescu, G Damnati, F. Béchet, R. De Mori (2007) Conditional use of Word Lattices, Confusion Networks and 1-best string hypotheses in a Sequential Interpretation Strategy Proc. European Conference on Speech Communication and Technology, Interspeech 07, Antwerpen, Belgium

A. Nasr, Y. Estéve, F. Béchet, T. Spriet, R. de Mori (1999) A language model combining n-grams and stochastic finite state automata. Proc. Eurospeech-99, Budapest, Hungary, pp :2175-2178

N. Nilsson., (1981) *Principles of Artificial Intelligence* Tioga Press, 1981

P. Norvig (1987) Inference in Text Understanding. Proceedings Sixth National Conference on Artificial Intelligence, July 13-17 Seattle, WA. AAAI Press 561-565

F. Pereira (1990) Finite-state approximations of grammars Proc. DARPA, Speech and Natural language workshop, Hidden Valley, PA, June 1990, pp. 12-19
http://dblp.uni-trier.de/rec/bibtex/conf/aaai/Norvig87
R. Pieraccini, E. Levin and C.H. Lee (1991). Stochastic Representation of Conceptual Structure in the ATIS Task. Proceedings of the, 1991 *Speech and Natural Language Workshop*, 121-124, Morgan Kaufmann publ, Los Altos, CA.

S.S. Pradhan, W. Ward, . K. Hacioglu, J.H. Martin and D. Jurafsky, .(2004) Shallow Semantic Parsing using Support Vector Machines Proc HLT-NAACL Conference, Boston, Massachusetts, USA, May 04, pp.233-240.

N. Prieto, E. Sanchis and L. Palmero (1994) Continuous speech understanding based on automatic learning of acoustic and semantic models. Proc ICSLP 1994, Yokohama

O. Rambow, S. Bangalore, T. Butt, A. Nasr, R. Sproat (2002) Creating a Finite State Parser with Application Semantics , 19th International Conference on Computational Linguistics (Coling 2002) Taipei

C. Raymond, F. Béchet R. de Mori and G. Damnati, (2006) On the use of finite state transducers for semantic interpretation, *Speech Communication*, 48(3): 288-304, March 2006.

G. Riccardi and A.L. Gorin (2000) Stochastic language adaptation over time and state in natural spoken dialog *systems IEEE Transactions on Speech and Audio Processing*, SAP-8(1):3-10.

G. Riccardi and D. Hakkani-Tur (2005) Active Learning: Theory and Applications to Automatic Speech Recognition *IEEE Transactions on Speech and Audio Processing*,SAP-13 (4) : 534-545

M. Richardson and P. Domingos (2006) Markov Logic Networks, *Machine Learning*, 62:107-136.

R. Sarikaya, Y., Gao, M. Picheny and H. Erdogan (2005) Semantic Confidence Measurement for Spoken Dialog Systems *IEEE Transactions on Speech and Audio Processing*,SAP-13 (4) : 534-545

D. Walker, (1975) The SRI speech understanding system *IEEE Transactions On Acoustics, Speech, And Signal Processing*, ASSP-23, NO- 5, pp. 397-416

Ye-Yi Wang; Li Deng; A. Acero, "Spoken language understanding." *IEEE Signal Processing Magazine.* 22 (5) pp. 16- 31, 2005.

W. A. Woods(1975). What's in a link? in D.G. Bobrow and A. Collins Eds, *Representation and understanding* , Academic Press, New York.

Voice Search on Mobile Devices

Geoffrey Zweig
Microsoft Research

Cellphones are among the most widely used technological devices in the world today, with over two billion cellphone subscribers worldwide, and approximately one billion new sales per year - a significant fraction of the human population has a cellphone. The use of these devices is coming at the same time that web-based services can offer an incredible richness of information to those who are able to access it, which has traditionally been done with a large-screen computer. This talk explores the convergence of these trends in voice search on mobile devices, which offers the potential to get people the information they need even when they are on-the-go and away from a traditional computer. The talk explores the new application areas that utilize mobile voice search; advertising models to support these services; and the technological challenges of voice and multi-modal interfaces for search on cellphones. Microsoft's recently released "Live Search for Windows Mobile" application will be used to illustrate the technology.

DATAMONITOR

Understanding the Market Movements in Network Speech: Aligning Business and Technology

Daniel Hong
Lead Analyst, Customer Interaction Technologies
Langtech 2008
2/8/08
dhong@datamonitor.com

quality data expert analysis innovative delivery

the home of **Business Intelligence**

Company overview

DATAMONITOR

- **Global premium business information firm**
  - Offices in New York, San Francisco, London, Frankfurt, Sydney, Shanghai and Tokyo
  - More than 500 analysts across seven vertical industries
  - Recently acquired by Informa, a leading international provider of specialist information and services for the academic and scientific, professional and commercial business communities
- **Highlights for technology team**
  - Recently acquired Ovum research to improve telecom coverage
  - Acquired Butler Group to improve end-user analysis and consulting
  - Purchased Computerwire and Computer Business Review (CBRonline) in the past
- **Trusted reputation**
  - Trusted provider of premium research to over 5,000 institutional & corporate clients
  - 600+ media mentions per month

Market drivers and inhibitors

DATAMONITOR

Drivers

- Cost reduction
- Increasing call volumes
- Customer service improvement
- Revenue generation
- Branding and differentiation
- Long-term automation and customer care strategy

Inhibitors

- Viewed as expensive and complex
- Lack of market awareness and education
- Poor implementations in the past
- Fear of customer backlash
- Proprietary solutions caused 'vendor lock-in'

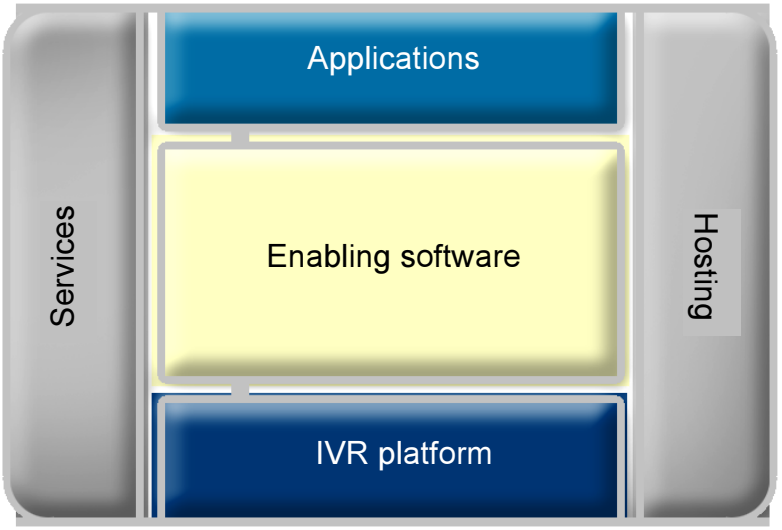
Global market trends

DATAMONITOR

- The emergence of Voice-XML is changing the IVR and speech landscape…
- Speech market is shifting from package to build environments…
- Simplifying application development is becoming even more crucial…
- Sharpening focus on Service-Oriented Architecture (SOA)…
- Increasing importance of Eclipse…
- Growing demand for speech as a hoste… service…
- Speech-enabled mobile search has arrived…

Understanding the Voice Business Value Chain

DATAMONITOR

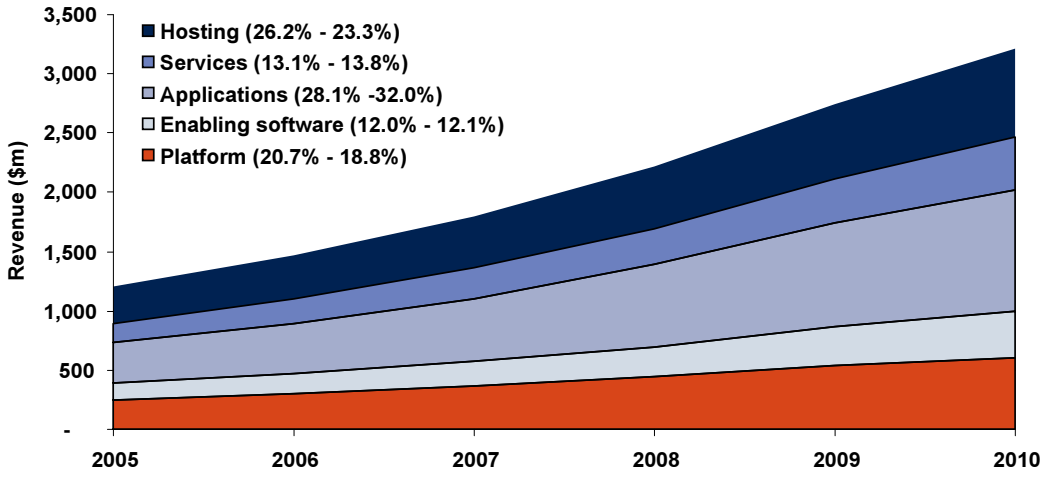


- IVR platform
  - Traditional (proprietary) IVR
  - Open standards-based platforms and browsers (Voice-XML)
- Enabling software
  - Tools for application development and management
  - Advanced Speech Recognition (ASR) engine
  - Text-to-speech (TTS) engine
  - Voice authentication engine
- Applications
  - Pre-built or packaged
  - Custom developed
- Services
  - Requirements/discovery
  - Project management
  - Systems integration
  - Implementation
  - Modifications
- Hosting
  - Fully dedicated hosted services
  - Premise-based managed services

Market sizing and segmentation

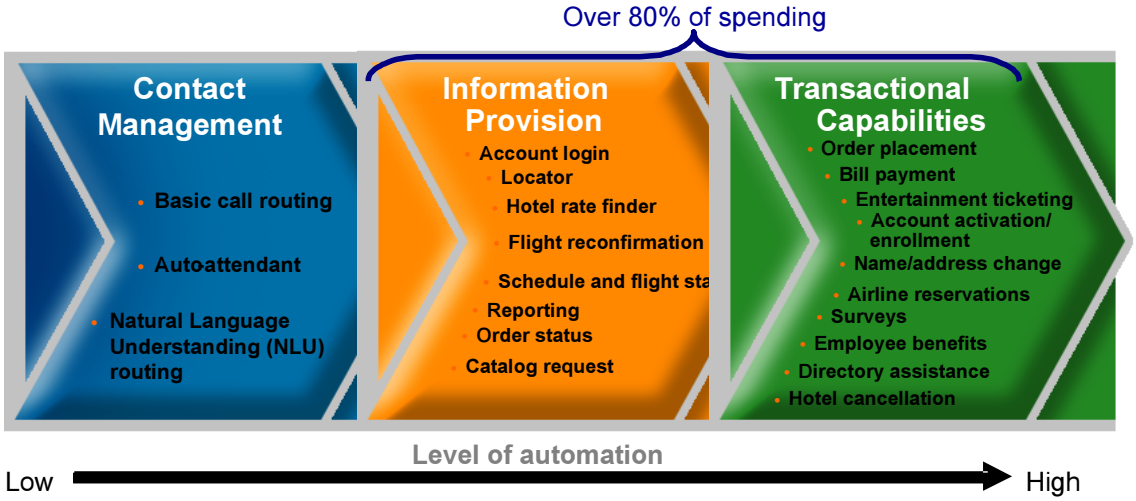
DATAMONITOR

- **Global spending on speech recognition across the value chain:**
  - $1.2bn in 2005
  - $3.2bn by 2010
  - CAGR of 21.7%



Speech application types and uses

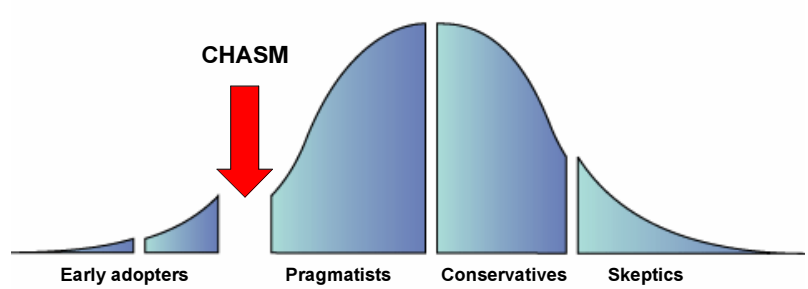
DATAMONITOR



- **Contact Management** – Solution that enables routing and dialing.
- **Information Provision** – Solution that enables the delivery of information.
- **Transactional Capabilities** – Solution that enables fully automated transactions.

Adoption curve for speech

DATAMONITOR



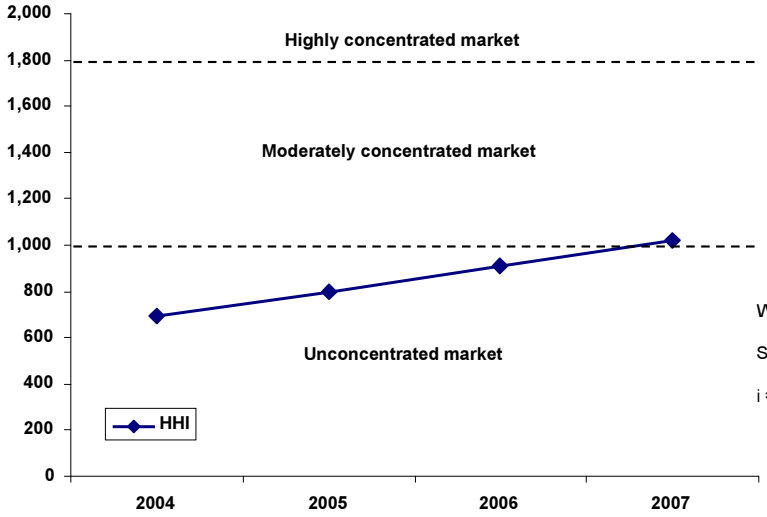
- Financial services
- Communications
- Travel & Tourism
- Healthcare
- Retail
- Technology
- Outsourcing
- Utilities
- Public Sector
- Distribution and wholesale
- Entertainment, media, leisure
- Manufacturing
- Other

- The adoption curve for speech is moving beyond early adopter markets and penetrating pragmatist and conservative markets.
- Verticals will vary by region, but early adopters are common across all geographic regions.
- Uptake will continue across all markets through the next five years.

Plotting the speech industry

DATAMONITOR

- HHI measure < 1,000 is an unconcentrated market;
- 1,000 < HHI measure < 1,800 is a moderately concentrated market;
- 1,800 < HHI measure is a highly concentrated market.



[redacted]

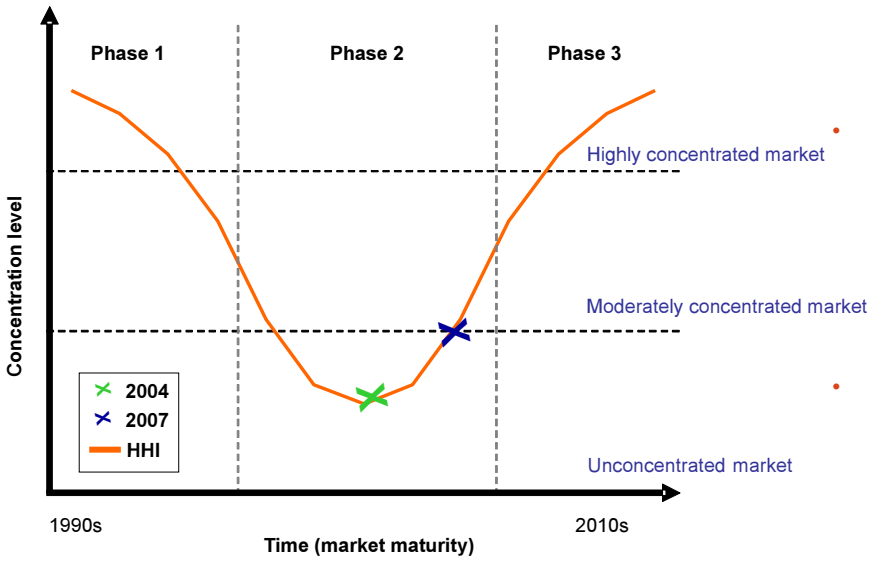
Where f = number of firms participating in the industry

Si = each firm's market share

i = firm in a given industry

Plotting the speech industry

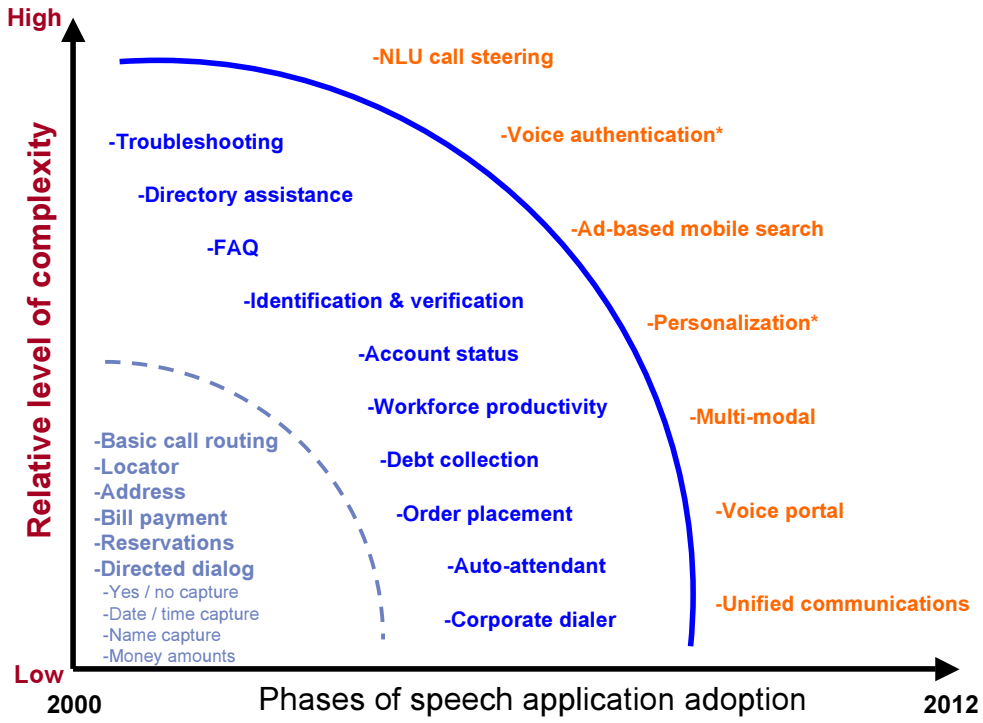
DATAMONITOR



- **Phase 1** – A nascent market with very few vendors that have majority market share.
- **Phase 2** – An emerging and competitive market. Barriers to entry have come down and new entrants enter the market causing fragmentation, downward pricing pressure and increased competition.
- **Phase 3** – A mature market. Saturation has been reached in the addressable market. Heavy M&A activity has occurred and the majority of market share is held by a handful of large enterprise software vendors and systems integrators.

Phases of speech application adoption

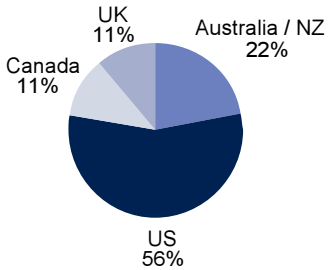
DATAMONITOR



Voice Industry Index

DATAMONITOR

Countries where speech solutions have been deployed

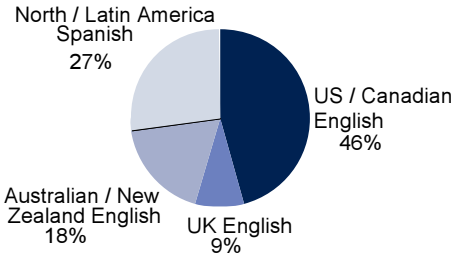


Annual revenues ($)

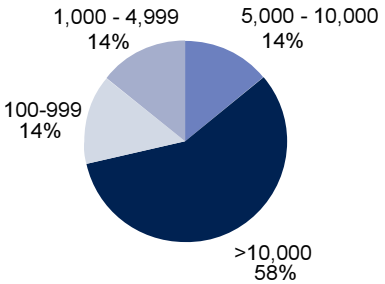
Range: $50m – $3bn

Average: $500m – $1bn

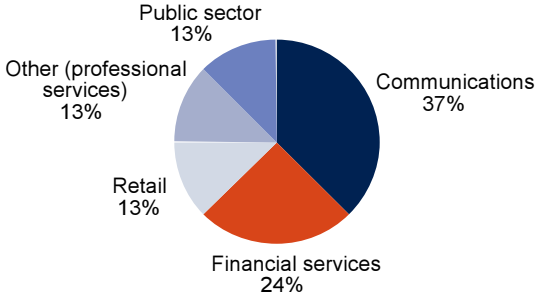
Languages supported in speech solutions



Number of employees



Vertical markets served

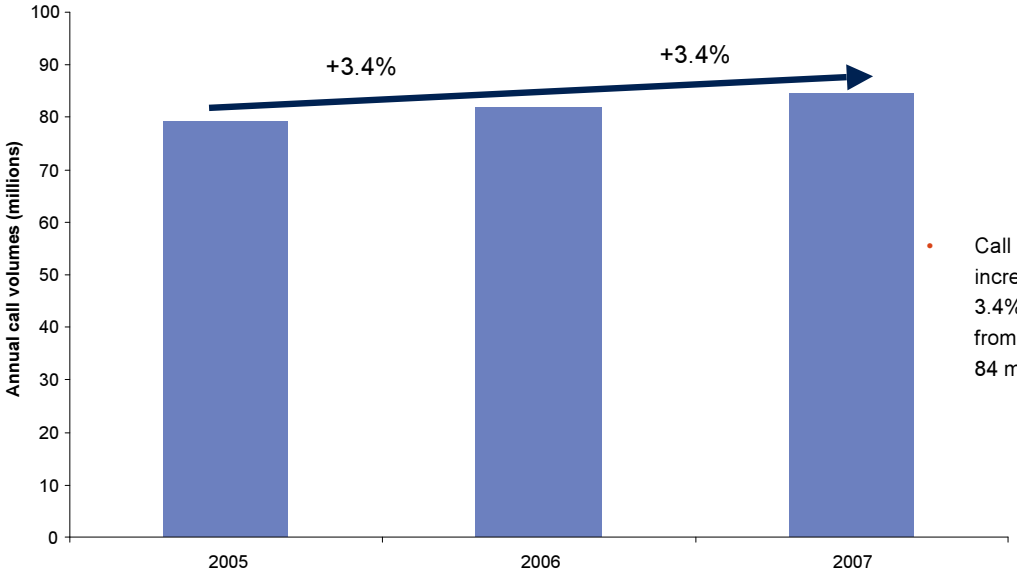


Voice Industry Index

DATAMONITOR

What were your annual call volumes (inbound calls that entered the ACD / PBX or IVR if this is placed first) for 2005, 2006 and 2007?

- Annual call volumes continue to increase year-over-year as consumer population grows and new strains of calls enter the contact center.

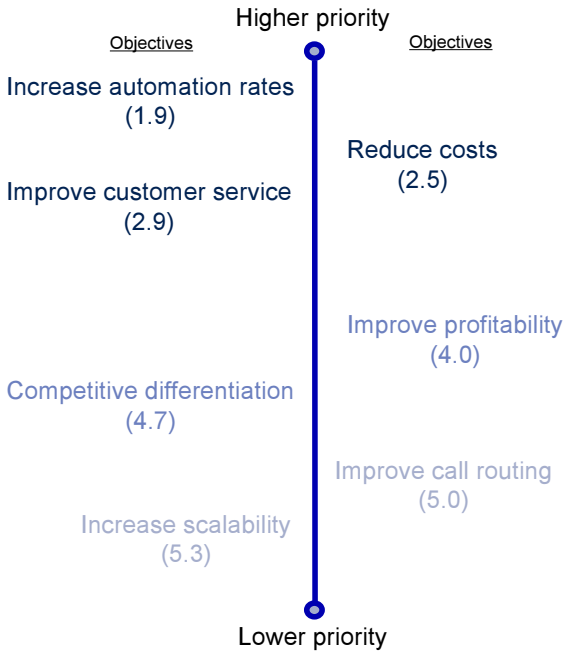


- Call volumes are increasing at a rate of 3.4% per year, growing from 79 million in 2005 to 84 million in 2007.

Voice Industry Index

DATAMONITOR

What were the initial objectives you set out to achieve with a speech solution - can you rank in order these objectives in terms of priority, 1 being the highest priority?



Did you complete these objectives?

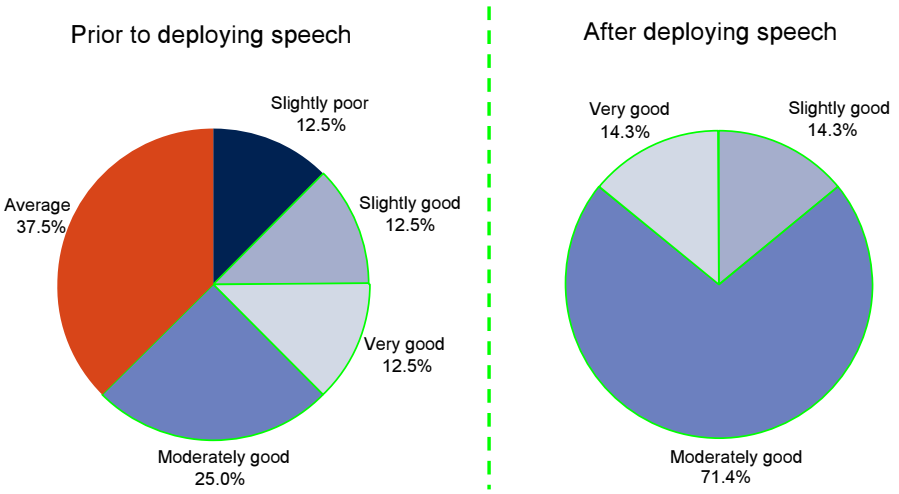
- **37.5%** - We have completed some of the objectives
- **25.0%** - We have successfully completed all of the objectives
- **25.0%** - We have completed some of these objectives, and will achieve the rest in the next 12 months.
- **12.5%** - We have not completed any these objectives, but will do so in the next 12 months.
- **0.0%** - We have not completed any of the objectives

- Most respondents have achieved some or all of the objectives they originally set out when deploying speech.
- Respondents that have achieved some or none of the objectives believe they will achieve them in the next 12 months.

Voice Industry Index

DATAMONITOR

How would you rate your (phone-based) customer satisfaction score prior to deploying speech and after deploying speech?



- Phone-based customer satisfaction scores showed strong improvement with speech. All respondents indicated slightly good to very good customer satisfaction rates after deploying speech.
- The major improvements in customer satisfaction came from those respondents that viewed their phone-based customer service as average or slightly poor.

Majority of customer satisfaction scores were obtained through phone-based customer surveys

Key takeaways

DATAMONITOR

- Speech recognition has transitioned from a ‘cool technology’ to an applied business solution…
  - Emergence of Voice-XML (open-standards)
  - Crossed the chasm into new markets
- Speech is an accepted UI and companies are establishing long-term positions using the technology…
  - Network, PC and embedded deployments
- Companies have benchmarked success…
  - Continue to optimize application performance
  - Leverage speech and DTMF to innovate in other lines of business (upsell, cross-sell, outbound)

Venture Capital and language technology business

Carlo Paris
Paris & Partners

New European Infrastructural and Networking Initiatives

Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR
Pisa, Italy
glottolo@ilc.cnr.it

I'll point at two new infrastructural initiatives – launched by the European Commission – in the area of Language Resources (LR) and Language Technologies (LT), which will influence how we shape the future of the field: CLARIN and FLaReNet.

CLARIN

CLARIN (*Common Language Resource and Technology Infrastructure*) is an ESFRI project (http://www.mpi.nl/clarin/) whose mission is to create an infrastructure that makes LRs (annotated texts, lexicons, ontologies, etc.) and technologies (speech recognisers, lemmatisers, parsers, information extractors, etc.) available and readily usable to scholars of all disciplines, in particular of the humanities and social sciences (HSS), to make them ready for an eScience scenario. The purpose of the infrastructure is to offer persistent services that are secure and provide easy access to language processing resources. Without the proper infrastructure, the technologies to make these tasks possible will only be available to a few specialists.

The CLARIN vision is that the resources for processing language, the data to be processed as well as appropriate guidance, advice and training can be made available and can be accessed over a distributed network from the user's desktop. CLARIN proposes to make this vision a reality: the user will have access to guidance and advice through distributed knowledge centres, and to repositories of data with standardised descriptions, processing tools ready to operate on standardised data. All of this will be available on the internet using a service oriented architecture based on secure grid technologies.

The nature of the project is therefore primarily to turn existing, fragmented technology and resources into accessible and stable services that any user can share or adapt and repurpose. CLARIN will build upon the rich history of national and European initiatives in this domain, and will ensure that Europe maintains a leading position in HSS research in the current highly competitive era. Infrastructure building is a time-consuming activity and only robustness and persistency of the offered solutions will convince researchers to step over to new and more efficient ways to carry out leading research. The preparatory phase, starting in 2008, aims at bringing the project to the level of legal, organisational and financial maturity required to implement a shared distributed infrastructure that aims at making language resources and technology available to the HSS research communities at large. This necessitates an approach along various dimensions in order to pave the way for implementation.

FLaReNet

International cooperation and re-creation of a community are among the most important drivers for a coherent evolution of the LR area in the next years. The Thematic Network *FLaReNet (Fostering Language Resources Network),* proposed in the context of an eContent*plus* call, will be a European forum to facilitate interaction among LR stakeholders. Its structure considers that LRs present various dimensions and must be approached from many perspectives: technical, but also organisational, economic, legal, political. The Network addresses also multicultural and multilingual aspects, essential when facing access and use of digital content in today's Europe.

FLaReNet, organised into thematic Working Groups, each focusing on specific objectives, will bring together, in a layered structure, leading experts and groups (national and European institutions, SMEs, large companies) for all relevant LR areas, in close collaboration with CLARIN, to ensure coherence of LR-related efforts in Europe. FLaReNet will consolidate existing knowledge, presenting it analytically and visibly, and will contribute to structuring the area of LRs of the future by discussing new strategies to: convert existing and experimental technologies related to LRs into useful economic and societal benefits; integrate so far partial solutions into broader infrastructures; consolidate areas mature enough for recommendation of best practices; anticipate the needs of new types of LRs.

The outcomes of FLaReNet will be of a directive nature, to help the EC, and national funding agencies, identifying those priority areas of LRs of major interest for the public that need public funding to develop or improve. A blueprint of actions will constitute input to policy development both at EU and national level for identifying new language policies that support linguistic diversity in Europe, in combination with strengthening the language product market, e.g. for new products and innovative services, especially for less technologically advanced languages.

These initiatives call for international cooperation also outside Europe, and will be relevant for setting up a global worldwide Forum of Language Resources and Language Technologies.

ForumTAL initiative

Andrea Paoloni

Fondazione Ugo Bordoni, Roma, Italia

apaoloni@fub.it

www.fub.it

What's TAL

The acronym TAL (Automatic Processing of Language) denotes technologies of automatic processing of spoken and written language. This field can be subdivided into two broad areas: "Speech Processing" and "Natural Language Processing". The first one is related to the reproduction of human faculty to communicate through words, and the second concerns the reproduction ability of human faculty to understand spoken and written language.

The applications for the first area are *voice coding*, *automatic voice dictation systems*, *screen readers for the blind;* while for the second area are *spelling and syntax*, *machine translation*, *assisted management* of search engines on the Internet or on large archives of knowledge.

Our society is based on understanding and communication of knowledge, i.e. the ability and the opportunity to exchange information through a shared language. The pen or even the chisel with which our ancient predecessors left signs of their culture are now replaced by computers that can save words, sounds and images. The TAL intends to make the communication easier and at the same time most complete and immediate.

What's Forum

TAL is a research area of particular interest to the Ministry of Communications because language is the primary communication tool and because of the important influence that language, Italian in our case, has on culture. In fact this research area presupposes a close interaction and collaboration between humanistic scholars and researchers with a scientific and technical training.

Forum members are convinced that the potential of TAL, yet largely ignored, can be disclosed in order to facilitate the dissemination of these techniques in the different fields of communications.

The forumTAL has the following objectives:

1. Monitor the activities of institutions involved in TAL to promote synergies and stimulate new interest;
2. Promoting research and development in the field of linguistic tools;

3. Studying the initiatives that can lead to an enlargement of the market and to the development of national competitiveness in this sector;
4. Promoting public and private investment in the area, including the preservation of the Italian language and its dissemination in the world;
5. Studying research and tools in the area of TAL with particular attention to European initiatives;
6. Promoting the use of Italian technology in other countries.

Initiatives of the Forum

Established in 2003 the Forum has produced a White Paper on TAL that identifies areas of development, the size of current and future investments, the characteristics of the market, and the state of education and training. The following organisations and their institutional representatives take part in the TAL Forum, providing a balance of diverse cultural and industrial domains:

CNIPA, CRUI, Expert System, Bordoni Foundation, ILC-CNR, ISTC-CNR, Loquendo, Dante Alighieri Society, and the Ministries of Culture, Environment, Production, Communications, Public Administration, University and Research, and Justice.

The Forum has promoted numerous meetings in the area of language and has created a website (http://www.forumtal.it) providing many types of information concerning this area.

The last TAL Conference “Men and machines, a possible conversation”, hold in Rome on March 2006, discussed all the typical TAL themes from research issues to industrial applications. The Conference success, with nearly 800 participants, 19 exhibitors, and 45 well-known speakers, convinced the Forum members to organize this international event.

Captioning - Accessibility to Education for Hearing Impaired

Fausto Ramondelli
Senato della Repubblica

Captioning (real time subtitling) is the reporting service that more than others demonstartes the social and economic utility of fast captioning techniques; all based on the rational treatment of language.
The effectivness of computerized machine shorthand in the Italian language makes it easy the training. Captioning is requested more an more to ease the access of hearing impaired people to education. At the University, an inceasing number of students chose this kind of assistance.
The more relevant experiences are in the universities of Rome and Padova, but captioning is speading in many other cities in Italy.
Difficulties occur in training qualified captioners: excetpional skills both in technique and knowledge are requested. Experience and suggestions of students allow to improve the service.
We are late in informig on how and how much this service is useful; captioning is not yet known as an alternative way of access compared with Sign Language.
Due to the features of the service and to the limited extension of the captioning market costs are still high but technologies allow to develop new ad more flexible ways of producing subtitles, thus reducing prices for users.
Webcaptioning enables to provide remote subtitles as far as the place is provided with DSL; thus shorthand reporters located in distant areas. A webcaptioning experience was made at the University "Roma Tre".
Increasing spreading of this service lead also to provide on-demand captioning, not only to associations and institutions.
A further edge, where better results can be achieved, is the use of SR for subtitling: the SR performance are lower than shorthand machine, which allow a more analitical input.
The feature of SR engines imply further developments and higher accuracy of the captioner in order to ensure better performance; this path was passed also by shorthand reporters who use a phonetic shorthand method based on matching input strokes and job dictionary definitions.

Realtime Speech to Text: A Means to an End

Mark J. Golden, CAE
Executive Director and Chief Executive Officer

National Court Reporters Association

LangTech 2008
February 28, 2008

Speech Capture

- Three variables
  - The human element
    - Skill and other human judgments that go in to creating the input to the text processing system
  - Technology
  - The functional requirements of the party for whom the text is being produced
    - What will the text be used for?

Functionality

- Two parameters
  - Speed
    - How quickly is text required?
  - Accuracy
    - How exact must the translation be?
- Interrelationship of these two parameters
  - E.g.; if some delay in production of the final text is acceptable, there is an opportunity to review and revise in order to correct for any deficiencies in the initial capture and translation.

Realtime

- There are many speech capture/text production applications where you need text ***as soon as possible*** but not immediately
- There are some applications where a comprehensible text is needed immediately, but a rough translation is adequate
- There are numerous applications, however, that rely upon true realtime
  - Instant text production, with no opportunity to revise or correct
  - High degree of accuracy

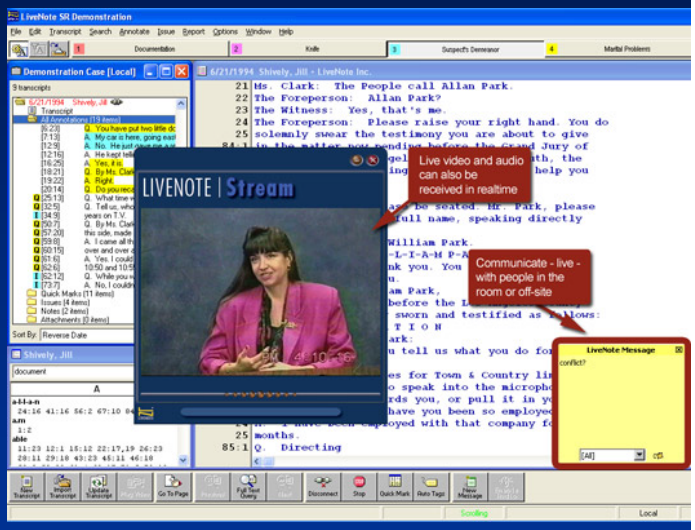
High-Tech Courtroom

- Reduce costs
- Improve efficiency
- Better customer service
- Secure data

NCRA Guardians of the Record

Litigation Support

- Immediate access to the record
  - Annotation
  - Highlights
  - Search
  - Instant message
- Complete multimedia record



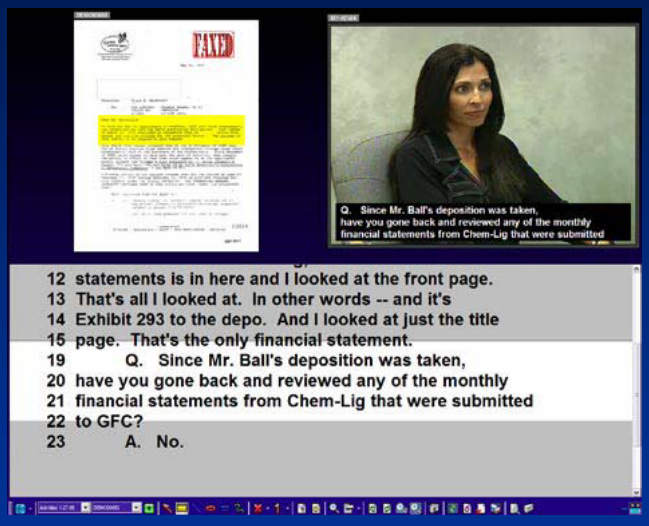
NCRA Guardians of the Record

Evidence Presentation

- Realtime text hyperlink to multimedia record
- Immediate display of evidence and other documentation

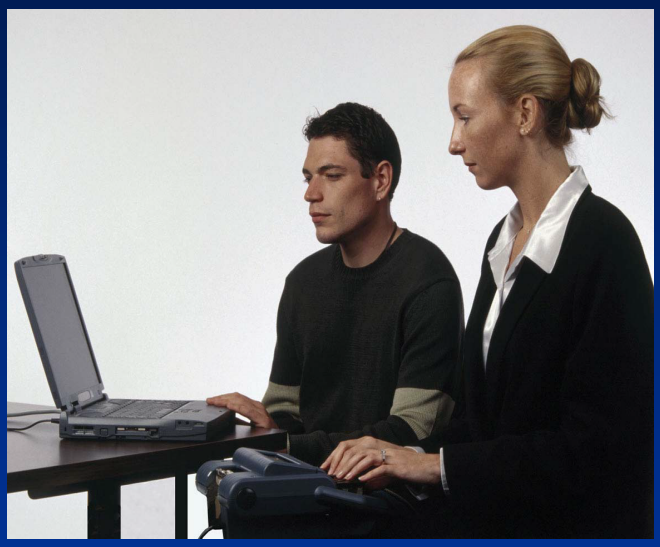
Virtual Justice

- Complete access to multimedia record
- Remote video participation by parties or witnesses – from anywhere in the world



Communication Access

- Full and effective participation
- Complete understanding



Stenographic Realtime

- Only current method for high accuracy, immediate voice-to-text translation
  - Multiple speakers
  - Near perfect accuracy at high speeds
- Serves as foundation to the applications previously described

The Question of Accuracy

Realtime Accuracy: 100 Percent

The hurricane, with winds exceeding 150 miles per hour, is expected to hit the coast by 10 in the morning. Local government officials have ordered all coastal residents to leave their homes and move farther inland. More to come later.

Realtime Accuracy: 90 Percent

The hurricane, with winds exceeding 150 miles per hour, is expected to hit the coast by TEPB in the morning. Local government officials have ORD/-D all coastal residents to HRAOEFB their homes and PHOFB farther inland. More to come later.

Realtime Accuracy: 80 Percent

The hurricane, with winds exceeding 150 miles per hour, is EBGS/PEBGT/-D to HEUT the coast by TEPB in the morning. Local government officials have ORD/-D all coastal RES/TKEPBTS to HRAOEFB their homes and PHOFB farther EUPB/HRAPBD. More to come later.

Mistranslates: Steno and Voice

- How do you justify stopping the defendant's vehicle?
  - How do you just if I stopping the defendant's vehicle? (voice)
  - How do you just if he stopping the defendant's vehicle? (steno)
- Do you recognize this document I've just marked as Exhibit 1?
  - Do you recognize the stockman I've just marked as it's a bit 1? (voice)
  - Do you recognize this doctor I've just marked as Kent 1? (steno)

Voice Capture ≠ Instant Translation

Appropriate Technology +

Competent Reporter

↓

High Accuracy and Immediate Translation

↓

Full, Complete and Instant Understanding

Thank You!

Questions and Answers

Stentor, a new Computer-Aided Transcription software for French language

Thierry Spriet

SténoMédia,
20 Bd Bastille, 75012 Paris, France
http://www.stenomedia.com

thierry.spriet@stenomedia.com

Abstract

We are presenting in this paper the technology used in STENTOR, the new software for Computer-Aided Transcription (CAT) in French.
The stenotypy domain is between speech recognition and text analysis.
In the most used stenotypy method in France, words are described by their sounds using a syllabic approach.
This implies some difficulties because of the high number of homophones in French: this problem is similar with a classical speech recognition problem. Homophonic rate in the French language is about 1.8 and in the most widely used method of stenotypy in France is not adapted to reduce it.
On the other hand, most of time, we have at our disposal punctuation information and the use a comprehensive dictionary, which is more like texts analysis.
In STENTOR, in order to avoid homophone ambiguity, we are proposing a classical treatment based on the linear interpolation of a 3-class and a 3-gram statistical language models. Some adjustments were proposed, as a word-class factorization to reduce the linguistic model size.
The speeches which have to be processed are various and often highly specialized. Even with the use of a comprehensive dictionary, very often new words have to be introduced during the transcription process.
We use a specific training corpus about 4.5 million words issued from more than several hundreds hours of transcripts.
The rate of error of STENTOR was tested on a corpus of only 5 000 words, but the comparison test have shown that we are very competitive indeed.

Index Terms: stenotypy transcription, french homophony

1.Introduction

Stenotypy is a very good lab to apply and to develop researches in linguistic engineering. Between speech recognition and texts analysis, stenotypy transcription has to deal with words described by their pronunciation, but without acoustic treatment problems.

Applied to the French language, we have the same homophony problematic than in speech recognition. On the other hand, as in a text analysis approach, we have to work with very large vocabulary, with the possibility of managing different dictionaries and adding news words during the transcription process.

In this paper, we first briefly compare stenotypy and speech recognition system. Then we explain some specific problem encountered in French language and their effect in CAT system. We present the technology used in Stentor, a new CAT system developed in order to process French language.
At last, we present some results from experiment made in order to evaluate the performance of this new system.

2.Stenotypy vs speech recognition

Computer-Aided Transcription (CAT) and automatic speech recognition (ASR) can be compared in two ways:

- the first one is about their use: why to use CAT systems instead of ASR system?
- the second one is about the similarities between the linguistic technologies used in both systems.

2.1.Using stenotypy

Why to continue to use stenotypy while speech recognition has improved a lot and has now a very small error rate? In fact, stenotypy is used in situations where some constraints can be resolved by speech recognition.

A great difference in stenotypy is the human interpretation made by the verbatim reporter. Even if this extra intervention has to be minimized, it allows to delete stammering and hesitation.
The stenotypist can also make a very powerful speaker identification even if two speakers speak together.

The stenotypist can also add some extra speech events like "*Mr. Smith leaves the room*" or "*Mr. Brown approves*".

If we want to know what happens exactly in a meeting, during the examination of a witness or in a court room, we need this human intervention.

At last, ASR technology is not mature enough to efficiently process speech in a noisy environment, or overlapped speech: stenotypist's brain is still the best cognitive system to process such phenomena.

2.2.Stenotypy and speech recognition similarities

As in speech recognition, words in the most widely used French stenotypy method are described by their pronunciation.
The problematic is quite similar to speech recognition in French language, such as:

- acoustic variations due to typing errors,
- acoustic similarities due to ambiguities of the French stenotypy method used in France,
- high rate of homophones, plus ambiguities provided by the French stenotypy method,
- high rate of homographs, which cannot be efficiently reduced by a n-gram model;
- long span syntactic constraints, which are not well modeled by n-class model but need special models like in [1]

3.Text analysis context

Stenotypy is used in a lot of specific areas such as court reporting, technical meetings, boards of directors, conventions, arbitrations, conciliation boards and so on. Each time we need a specific vocabulary depending on the firm or on the agenda of the meeting. Even when using a very large vocabulary, we have to manage with new words. To do that, we take in account the *unknown word* in the linguistic model and offer the possibility to the stenotypist to add words while working realtime.

Something also very interesting in computer-aided transcription is the knowledge of the sentence boundaries. A linguistic model including this information is much more efficient.

4.Professional and historic context

For several years now, French stenotypists use the same method in France. Step by step, some of them adapt this method to avoid ambiguities which generate errors when using a CAT system. To develop the Stentor software, we had to take this into account.

The problem is that these adaptations are personal and lead to specific dictionaries for each user.

Each change needs a more or less long period of adaptation for the stenotypist. It is impossible to ignore these modifications, and we have to deal with that.

When the modification is limited to the orthography a word to drop an ambiguity between this word and another one, it is easy to include this to a user dictionary.
But when the adaptation is about the grammatical word description, this can generate some problems with our linguistic models.

We have decided to be independent of the stenotypy method and to accept all user variations. But we recommend to use the real syntactic class of the words, in order to be coherent with the linguistic models used in the Stentor CAT system.

5.linguistic models

The figure 1 gives a good example of what it just be presented above. In this example, we have only 4 keystrokes:

POUL
L
F*E
ST*OUD

As we can observe, we have between 4 and 10 candidates for each reference word. In this case, the solution is very easy for a human analysis, but the computer can find at least 3 good paths in this graph.
Except if we are in a meeting about avian influenza or something about chickens, the right path is "*pour lever ce doute*".
The problem can be more complex if in this meeting we have someone named *Mister Pourlever*. In this case, we need the context of the sentence to decide.

Without the use of language model, any path in this graph could be proposed by a CAT system: this shows the need, in French language, of the use of language models.

The software STENTOR actually use a mixed approach, using statistics and knowledge rules as presented in [2]. We use a linear combination of a 3-gram and a 3-class language models. This technology is classical for automatic speech recognition.
In order to reduce the size of the model, we apply some factorizations on the 3-class model.

5.1.3-gram model

The 3-gram model is in fact a combination of 3-gram, 2-gram en 1-gram models. We use a specific training

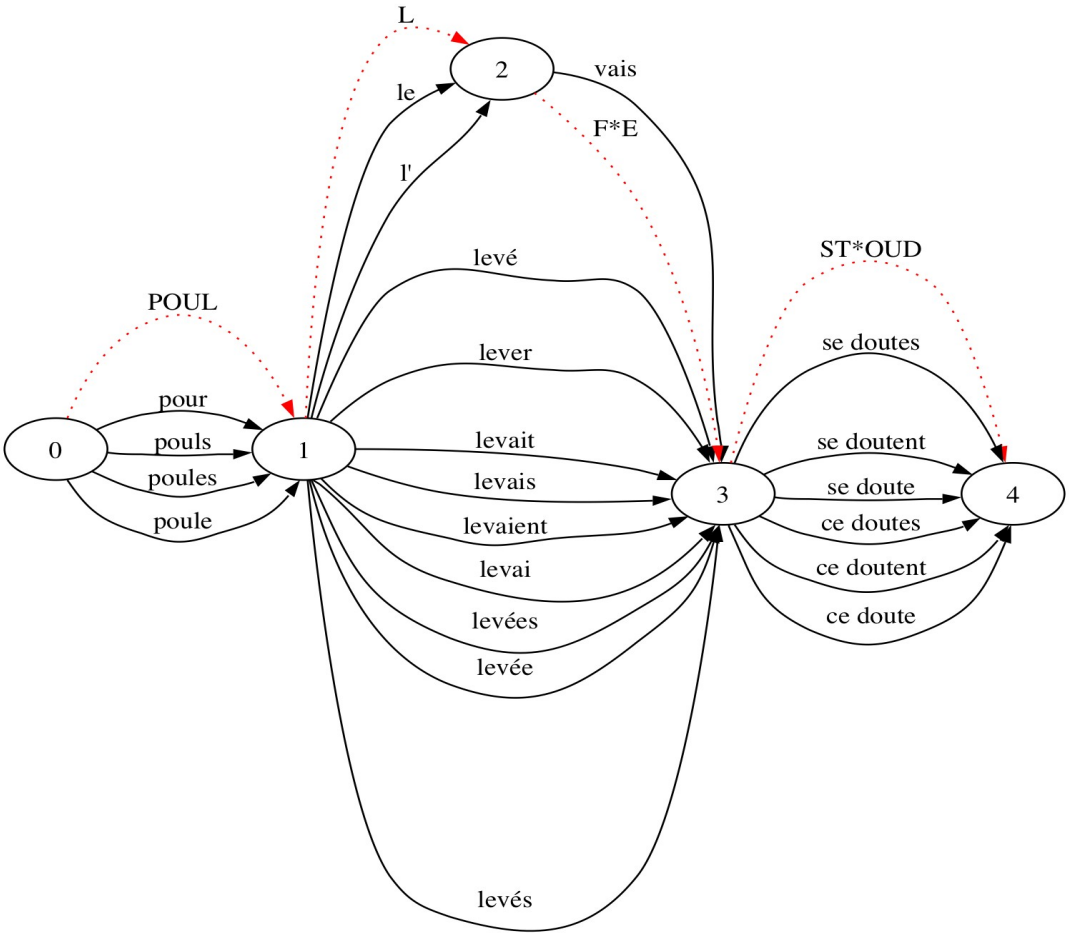


Figure 1: example of ambiguity graph in french stenotypy

corpus of 4.5 millions of words and a lexicon of about 150K most used words.
The training corpus is a corrected transcription of some 200 hours of meetings, boards of directors or town councils.
We have extracted the 150K most used words of this corpus to build the lexicon.
We have a special token for unknown words and the n-gram models are training with this token. So even if these models cannot guess an unknown word, they don't eliminate this possibility.

5.2. 3-class model

This model is based on statistics association of Part Of Speech (POS).
The training corpus was annotated with POS. This tagging was performed with a stochastic part of speech tagger [3] using a modified tag set of 105 POS. We add 2 special POS, one for foreign words and one for acronyms.

To reduce the size of the model, we have merged some classes which had the same behavior.

6.Experiment

To estimate the quality of the transcription we decided to use the word error rate as used to evaluate speech recognition systems. To this end we use the NIST Scoring Toolkit (SCTK) as used in NIST international evaluation campaigns [4]. The word error rate takes into account word substitutions, word deletions and word insertions.

Stentor is based on statistical language modeling. So, it is necessary to estimate the language models on a training corpus. As seen above, the training corpus contains 4.5 million words and is composed by a collection of real data provided by stenotypists.

The test corpus is distinct from the training corpus and was manually corrected. There is no longer typing

errors: they were manually corrected too in order to evaluate only the CAT system, not the stenotypist.
We had not time to set up a large test corpus, so we have only a 5K words corpus. This test corpus is extracted from real data built by a stenotypist.
In this corpus we have a word error rate of 6%.

We also made a comparative test with the French computed-aided transcription TASF+ [5]. On the same test corpus, our software had 10% less word error than the TASF+ software using the same user dictionaries.

7.Conclusions

The first version of STENTOR is now out and can be used for computer-aided transcription in realtime reporting or post treatments.
The word error rate is already competitive and we are planning in reducing it further. For this, we are going to use stochastic Finite State Automata in order to take in account long span dependencies as in [6]. Of course, we first have to integrate a larger corpus in the training step. We hope that the French stenotypy community is ready to help us in this task, giving us a lot of public and corrected transcriptions.
Even if STENTOR is a good lab to apply our researches, it is also a professional software which offers a lot of functionalities for a more efficient work :

- audio-sync, for post correction, when the user selects a word or a place in the stenotypy, he can hear directly this passage;
- dictionaries builder, a set of tools which make easy dictionary management like insertion of new words, importation of a former dictionary, merging of dictionaries and so on;
- realtime word insertion, during a realtime reporting, the user can add very easily a new word (a proper noun for example). This new word will be taken in account immediately by the system.
- computer assisted correction: for each word of the transcription, the software proposes all the alternatives ordered by decreasing pertinence.
- short cuts, the most frequent short cuts used in word processing softwares are implemented in STENTOR.

8.References

[1] Frédéric Béchet and Alexis Nasr and Thierry Spriet and Renato de Mori, Large Span Statistical Language Models: Application to Homophone Disambiguation in Large Vocabulary Speech Recognition in French, Eurospeech 99, p 1763-1766, Budapest, Hongrie, 1999

[2] Thierry Spriet and Marc El-Bèze, Introduction of Rules into a Stochastic Approach for Language Modelling, Computational Models of Speech Pattern Processing, NATO ASI Series F, vol. 169, ed. Keith Ponting, pp. 350-355,1998

[3] Thierry Spriet, Marc El-bèze *Etiquetage probabiliste et contraintes syntaxiques*. *TALN 95,* 1995

[4] Jonathan G. Fiscus Nicolas Radde John S. Garofolo Audrey Le Jerome Ajot Christophe Laprun, *The Rich Transcription 2005*, Spring Meeting Recognition Evaluation, 2005

[5] http://www.stenotype-grandjean.com/

[6] Alexis Nasr and Yannick Estève and Frédéric Béchet and Thierry Spriet and Renato de Mori., A Language Model Combining N-grams and Stochastic Finite State Automata, Eurospeech99, Vol 5 p 2175-2178, Budapest, Hungary, 1999

Directorate-General for Translation

Machine translation in the European Commission

Josep Bonet
Rome, 28/02/2008

DG Translation

Directorate-General for Translation

The beginning of it

- ✓ Machine translation (MT) introduced in the mid-70s
- ✓ DG XIII, Telecommunications, Information Market, and Exploitation of Research
- ✓ Aim: overcome language barriers
- ✓ Aim: help advance enabling technology
- ✓ Acquisition of a system for EN=>FR and FR=>EN
- ✓ System based on rules

DG Translation

Directorate-General for Translation

The continuation of it

- ✓ Improvement of the initial systems
- ✓ More and better rules
- ✓ Enlarged vocabularies
- ✓ Development of new systems

DG Translation

Directorate-General for Translation

Two parallel ways

- ✓ The Eurotra project
- ✓ Big research project
- ✓ Involvement of teams in most Member States
- ✓ Innovative approach to MT
- ✓ But too heavily dependent on computing power
- ✓ No working MT system delivered
- ✓ But big boost to language industries in Europe

DG Translation

Directorate-General for Translation

New orientation in the 90s

- ✓ EC finances rather than develops research projects
- ✓ Aim: develop the industry, let new products come to live
- ✓ Stress on pre-competitive research => no support for end-user products

DG Translation

Directorate-General for Translation

Translation in the EC

- ✓ The 90s: translation becomes digital
- ✓ Mid-90s: all translators use a PC, all translations are electronic
- ✓ The 80s: terminology and documentary databases
- ✓ The 90s: electronic translation tools
- ✓ MT as productivity tool
- ✓ MT as consistency tool

DG Translation

MT in EC's DGT

Directorate-General for Translation

- ✓ Development of MT transferred to DGT mid-90s
- ✓ Improvement of dictionaries and lexical routines
- ✓ Total number of language pairs available: 26
- ✓ English and French as source or target language
- ✓ Coupled with German, Italian, Spanish, Dutch, Portuguese, Greek, Swedish and Danish
- ✓ Varying quality
  - ✓ From rather acceptable
  - ✓ To rather unacceptable (prototype)
- ✓ Quick or light post-editing of best output

DG Translation

How to get an MT?

Directorate-General for Translation

- ✓ 1) E-mail to a dedicated mailbox (one per language pair and direction): phased out
- ✓ 2) Euramis: integrator of language services; MT is one of them
- ✓ 3) Web interface
- ✓ 4) All files to be translated automatically preprocessed, including MT when available.

- ✓ Methods 2 and 4 available only to translators
- ✓ Methods 1 and 3 also available for all EC staff

DG Translation

Directorate-General for Translation

The issue of quality

- ✓ Quality is in the eye of the reader
- ✓ The case of the Tchernobyl documentation
- ✓ Fully Automatic High Quality Translation is out
- ✓ Fully Automated Usable Translation is in
- ✓ MT against TM as the translation tool of choice
- ✓ The day corpus availability beat MT

DG Translation

Directorate-General for Translation

Caring about the user

- ✓ Professional or non professional
- ✓ Translator prefers quality, at the cost of partial translation
- ✓ End-user prefers full information, at the cost of lesser readability
- ✓ Different tolerance thresholds. Important to keep in mind
- ✓ Language is extremely personal
- ✓ It involves deep feelings: sensitive issue
- ✓ MT should be seen as an added service, not a denial of service

DG Translation

Some trends

Directorate-General for Translation

- ✓ Hype about systems based on corpora
- ✓ Great availabilities of corpora enable developments
  - ✓ DGT has released its corpus of legislation in force in 22 languages
- ✓ Exponential growth of demand for quick and cheap translation
- ✓ Global markets as triggers for this demand
- ✓ User satisfaction with MT solutions
  - ✓ Resignation? Better something imperfect than nothing
  - ✓ Satisfaction survey on Microsoft Knowledge Base
- ✓ But concentration on major languages
- ✓ Most European languages have few speakers
- ✓ MT remains mostly intelligence-related

DG Translation

Open Source Tools for Statistical Machine Translation

Philipp Koehn
University of Edinburgh

The EuroMatrix project is aimed at the creation of the infrastructure for the development of machine translation technology for all European languages. In particular, statistical machine translation has emerged as a promising new direction to machine translation. Within the EuroMatrix project, the open source statistical machine translation toolkit Moses is being developed that follows this direction.

Moses is a complete toolkit that, given a corpus of translated text, allows the creation of a statistical machine translation system that is ready to use. The software was mainly developed for fostering research, but it has already attracted much interest in the commercial machine translation developer community. This talk will cover aspects of the underlying technology and its application.

LangTech 2008 February 28 - 29, 2008
San Michele a Ripa
Rome - Italy

NEC
Machine Translation Service and Technology for Mobile Phones

Akitoshi OKUMURA
NEC Corporation, JAPAN
February 28, 2008

Background

In a ubiquitous network society,
new forms of communication and new values will emerge.





Pleasure of connecting
Generating mutual understanding
Creating new ideas
Sharing feelings and ideas
Building fellowship and consensus

Machine translation technology will break through language barriers to create the ubiquitous network society.

NEC Concept: C&C

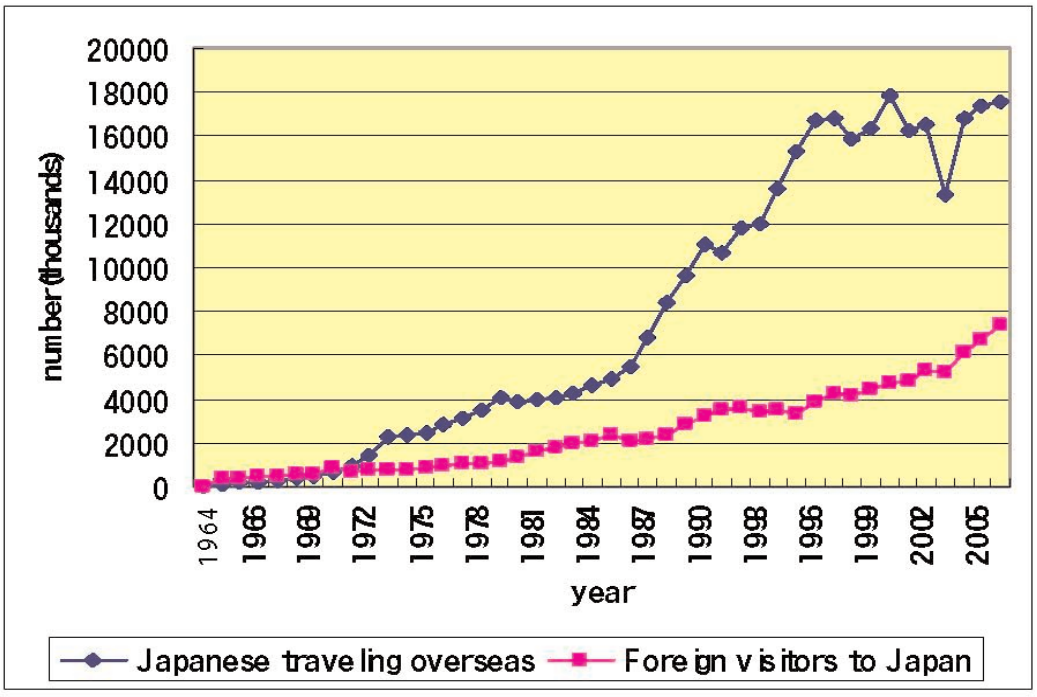
- Dr. Koji Kobayashi
  – former CEO of NEC, 1907-1996
- Presented C&C, integration of Computers and Communications at INTELCOM 1977 in Atlanta
- Proposed the concept of speech translation telephone at Telecom 1983 in Geneva

INTELCOM'77

Telecom'83

http://www.nec.co.jp/techrep/en/journal/g07/n02/070223.htm#c1
http://www.nec.co.jp/profile/empower/history/1977_1.html
(in Japanese)

Social Needs: Travelers to and from Japan



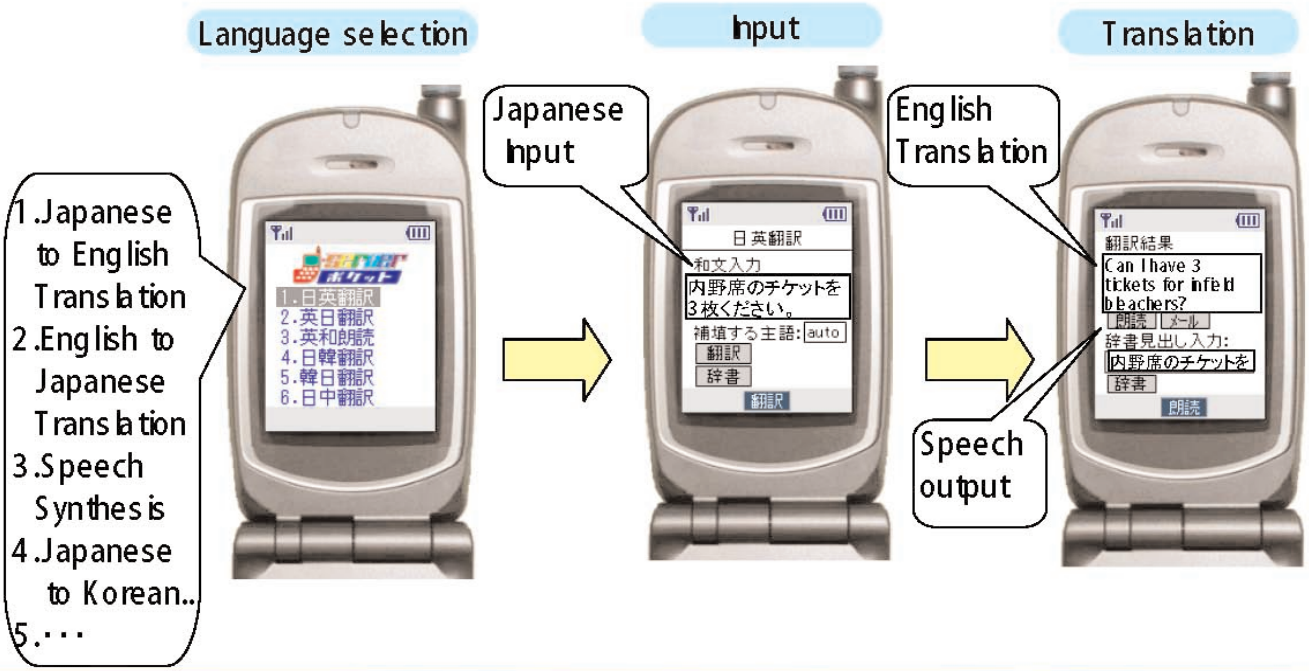
From White Paper on travel by Ministry of Land, Infrastructure and Transport

I. NEC Text Translation Service for Mobile Phones

- J-SERVER Pocket provided by NEC and Kodensha offers mobile phone users bi-directional text translation between Japanese and English, Japanese and Chinese, and Japanese and Korean.
- http://www.nec.co.jp/press/ja/0501/1101.html (in Japanese)
- The service is available via four major portals of mobile-phone carriers in Japan.

J-SERVER Pocket

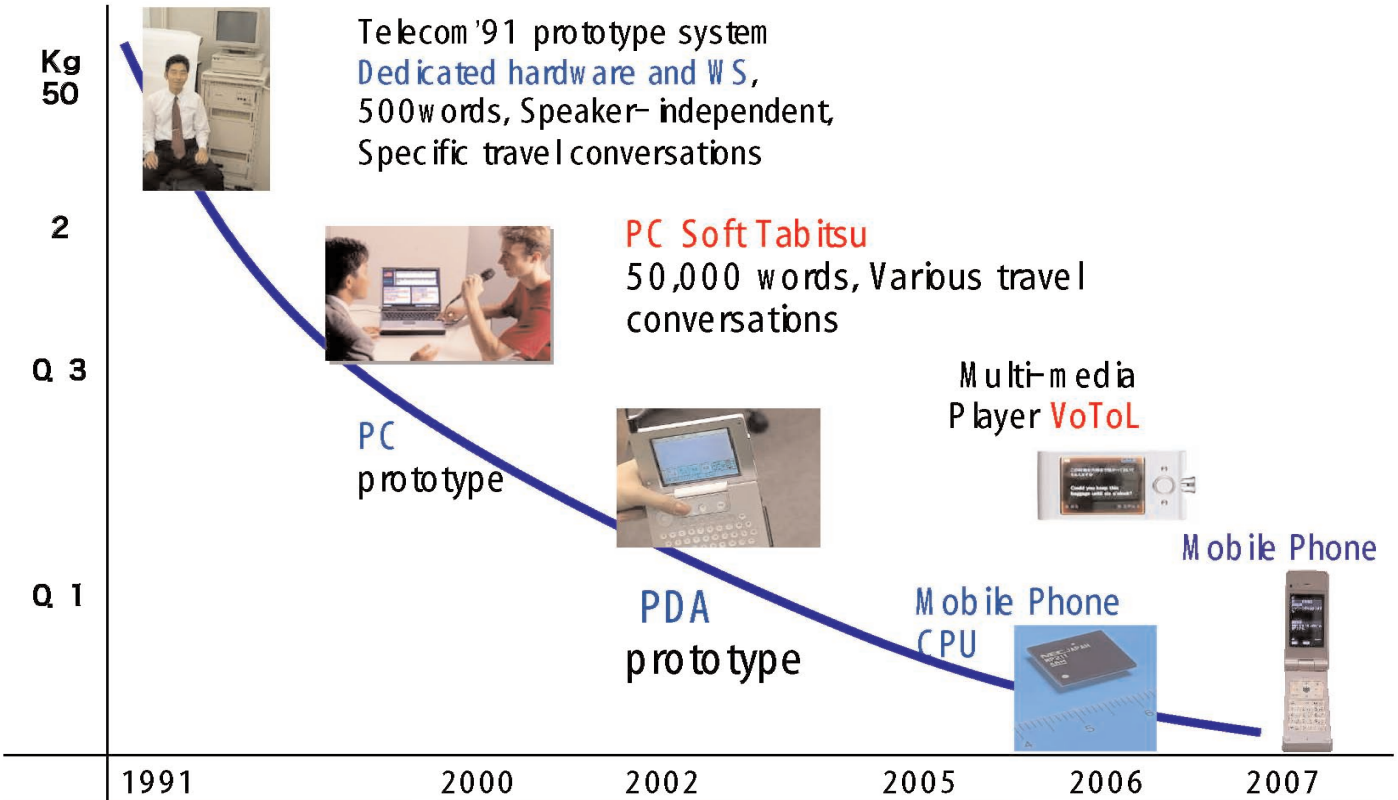
Multi-lingual Translation Service through Internet



II. NEC Speech Translation Software

- Speech translation software for embedded devices with low power consumption
- Related URL list
  - http://www.nec.co.jp/rd/Eng/innovative/E5/top.html
  - http://www.nec.co.jp/rd/innovative/E5/top.html (in Japanese)
  - http://www.nec.co.jp/rd/Overview/soshiki/media/natural-language.html (in Japanese)
  - http://www.nec.co.jp/techrep/ja/journal/g05/n05/t050511.pdf (in Japanese)
- Press Release
  - http://www.nec.co.jp/press/en/0601/0401.html
  - http://www.neceurope.com/news_and_events/news_archive_2005/24_october_2005.html
  - http://www.nec.co.jp/press/ja/0501/1101.html (in Japanese)
  - http://www.nec.co.jp/press/ja/0711/3002.html (in Japanese)

Progress of Speech Translation Software and Devices



Compact Implementation on PDA and Mobile Phone

- Speech recognition
  - decreased memory size of acoustic model and decoder by reducing total number of Gaussian mixtures and by improving dictionary structure
- Language translation
  - decreased memory size by using external storage effectively and improving internal data structure
- Japanese speech synthesis
  - deceased memory size by improving pronunciation dictionary structure and speech synthesis units

cf: Isotani et al., "An Automatic Speech Translation System on PDAs for Travel Conversation", Proc. ICMI'02, pp.211-216, Oct. 2002.

Mobile Multimedia Player *VoToL* featuring Speech Translator

- Provide users with music and video playing functions as well as speech translation between English and Japanese

http://www.nec.co.jp/press/ja/0602/1401.html (in Japanese)

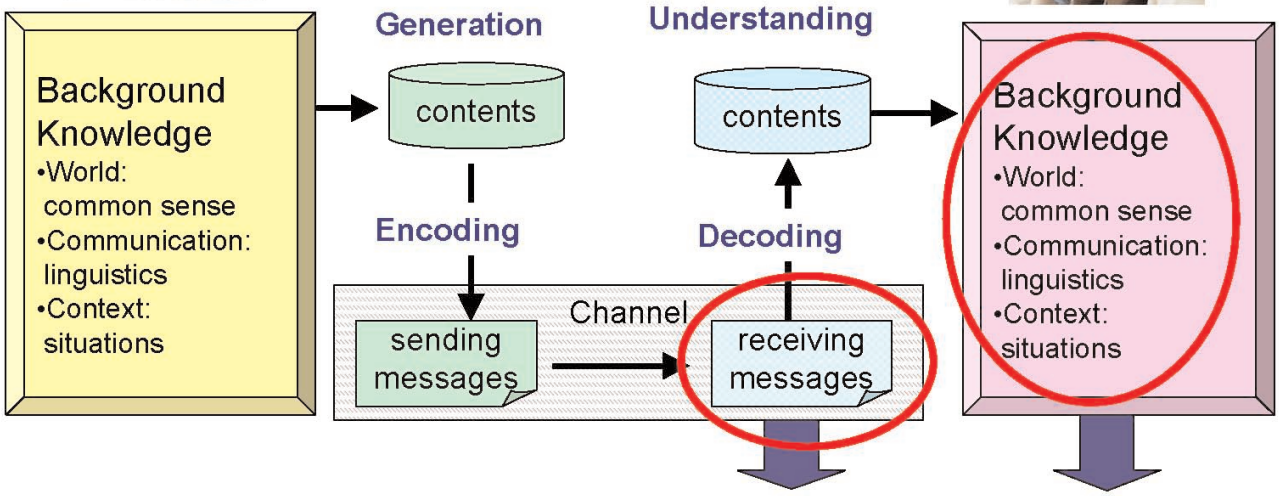
III. Future Work: Communication Agent

- Communication Agent can help mutual understanding by filling in gaps between messages and background knowledge.
  - Rich-media message creation robot is the first step.
- Related URL list
  - http://www.nec.co.jp/press/en/0703/0501.html
  - http://www.incx.nec.co.jp/robot/english/robotcenter_e.html
  - http://www.nec.co.jp/press/ja/0703/0501.html (in Japanese)
- Related Papers
  - Okumura et all, "Multimedia Blog Creation System using Dialogue with Intelligent Robot", Proceedings of the ACL 2007 Demo and Poster Sessions, pages 9–12, Prague, June 2007.
  - Okumura et all, "Evaluation of Multimedia Blog Creation System using Dialogue with Intelligent Robot", FIT2007 (The 6th Forum on Information Technology), LE-009,pp.135-138, September 2007 (in Japanese)

Understanding Model using Background Knowledge

Sender

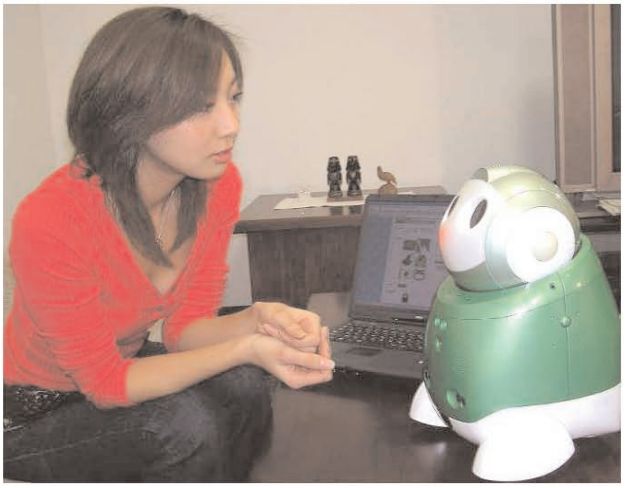
Receiver



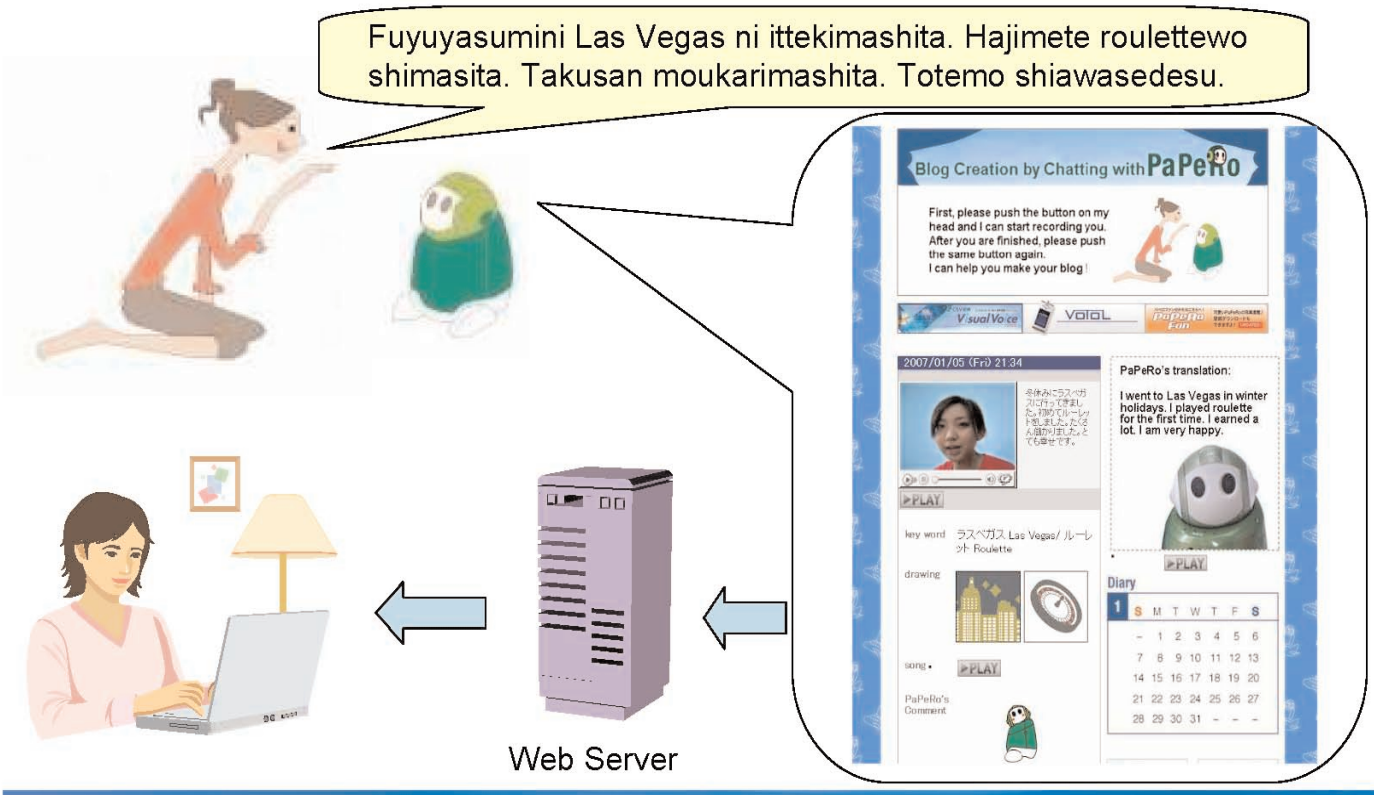
Filling in gaps between his messages and her background knowledge can help a receiver understand!

Rich-media Message Creation Robot

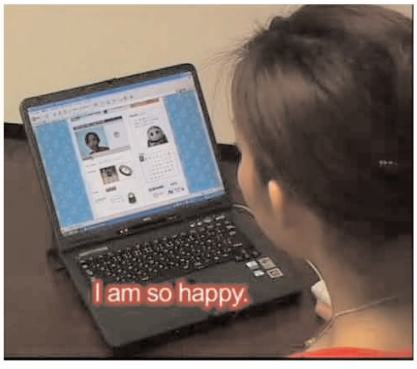
- Purpose
  - Enhancement of blog messages for easy understanding
  - Facilitation of blog creation
- Process
  - Recording a video message through dialogue
  - Translating the message
  - Searching related information
  - Creating rich-media message
- Platform
  - Personal robot, PaPeRo



Outline of Rich-media Message Creation



Video Demo



Thank you very much.

Empowered by Innovation NEC

Language Technology and Intelligence

Giuseppe Fabbrocino
General Technical Coordination Office (UGCT)

The General Directorate of "Telecommunications, IT and Advanced Technologies (TELEDIFE)" is involved in research and development, standardization, type testing, procurement, transformation, support of radar and electronic systems (non integral parts of weapon system's naval, air and ground), Command, Control, Communication, Calculation and Information (C4I) systems, space observation and communication, IT and network systems of all types (economic-financial, meteorological, geo-topographic, medical, logistic, secure, message handling and intelligence).
In conclusion, the competences of TELEDIFE cover all technologies included in modern C4ISTAR term.
In this vary wide range of technologies, the modern language technologies are assuming an always more importance on several applications of commercial, industrial and military applications where are demonstrating:

- in commercial-industrial applications, more cost-effective of traditional tools-methodologies for extract and correlate information from various type of database and text;
- in military and intelligence applications, more effective of extract, correlate and identify the useful operational information in very wide quantity of communications of various technologies (traditional radio and TV communications, satellite communications, open sources spoken or written, etc), also in various languages.

I hope that following presentations can explain characteristics and possible capabilities of this technologies

Speaker recognition for surveillance scenario against terrorism and organised crime

Pasquale Angelosanto
ROS Carabinieri

The presentation will focus on the use of Speaker Recognition SW technology in a surveillance scenario against terrorism and organised crime. Individuals who are likely to be the object of LEA (Law Enforcement Agency) investigation are growingly aware of the risk of being intercepted whenever executing a phone call, may that be through the fixed or wireless network. Hence the growing efforts to disguise their phone calls as a countermeasure. When listening to one person's voice we try to recognize that person identity from his/her voice; to do that our cognitive process follows a series of logic steps, which interpret the signals received by the ear at different levels. From all these characteristics of sounds and voices the human brain is capable to recognise the identity of the person when his/her voice is already known from other previous listening. This process has been automated by complex SW algorithms making possible therefore the use of voice as a biometric datum (like fingerprints, iris recognition, etc.). The speaker recognition systems verify the people identity by comparing the speaker's voice against a relevant statistical model, created and stored in advance from a "certified" voice sample, belonging to the same speaker. That model is named voiceprint. In scenarios having the following characteristics:

- Huge volume of telephone intercepts;
- Hundreds of target speakers;
- Different languages spoken;
- Spotting of targets as calls come in;
- Multiple investigation scenarios.

The Speaker Recognition technology allows to identify rapidly the calls made by specific targets. The effect of such a system can be that of enhancing the investigative process efficiency and efficacy thus assuring a high return on investment for the LEA agency.Given the maturity of the underlyng technology, it is time for Speaker Recognition potentiality to be leveraged in counter terrorism and against organised crime, by Law Enforcement Agencies operating within the appropriate legal boundaries.

Multilingual and multimedia information for Business Intelligence

Christian Fluhr
Director of research
CEA/LIST
Christian.fluhr@new-phenix.com

1 Introduction:

To maintain a good view of the their ever changing world environment each company, state or public organization need to organize the data collection from the open sources, analysis and synthesis of this information to help deciders in their managing activity. This is Business Intelligence.

This concerns an observation of the competitors, possible partners, change in legislations, evolution of technologies and patents, evolution of markets in various countries, activity of disinformation from competitors, etc. The world wide economy necessitates a world wide observation.

Open sources are of various types : web sites, news wires, news papers, scientific papers and congresses, radio, television, groups of discussions, blogs, reports on contacts by people going to congresses or visiting companies or public organizations.

This shows that information must be processed coming from various media (text, images, video, radio) and in various languages. The assertion that all valuable information can be found in English is false. English is used by non English speaking companies to let others know what they want them to believe. The remaining information is found in their native language. News papers, television, radio is generally only in the national language.

This shows the importance of multilingual language engineering in the Business intelligence activity. These technologies are said dual because the same tools can be used both by companies on open sources and by security and intelligence services on a more diversified source of information. The only notable difference is the set of languages to process that can be different. This convergence gives more financial support for the development of tools.

2 Language engineering for worldwide business intelligence:

Various tools are uses like spoken or written language identification, multilingual speech to text conversion, crosslanguage interrogation, crosslanguage information filtering, crosslanguage clustering, machine translation.

Perhaps it is necessary to clarify the distinction between multilingual and crosslingual systems. A multilingual system is a system that can separately process various languages. Generally the software is common but it uses different languages resources to process different languages.

Crosslingual systems are multilingual systems but in addition they can consider a set of documents or video in various languages as a whole that can be interrogated, filtered or clustered like a monolingual one.

For example, a crosslanguage search engine can index documents in various languages and a query in one language can retrieve relevant documents in any language in the indexed database.

About multimedia in this paper we will limit our presentation to text and speech which is the domain of the Langtech conference but there is also a lot of research to extract semantics from still or moving images and to combine these results with the semantic interpretation of language.

3 Role of names entities in worldwide business intelligence:

In Business intelligence, observation of competitors or partners, activities of persons working in the field, presentation of new products from competitors, geographical repartition of activities in the field, links between persons and companies, between persons, shows the importance of named entities recognition. This means recognition of persons, organizations, places, product names and associated numerical information like dates, amount of dollars, Kg, %, etc.

In addition, action involving names entities are very important in the process of watching environment. For example, nomination of managers, control taken by a company on another, new product presentation, etc.

In a multilingual environment, the problem of the recognition of names entities is much more difficult. Names of places can change when changing language (Paris (FR) Pariggi (IT)), they can change with the time (Petrograd, Leningrad, Saint-Petersbourg).

Identification of the same entity along the documents can be also a challenge like “George W Bush, George Bush, Bush, President Bush, the president, he (anaphora). It is also the case for dates (13/02/2008, yesterday, a week ago) that can represent the same information.

But, the economy is no longer controlled by the Latin character set. Crossing the character set barrier is sometimes difficult.

For example between Latin and Cyrillic :

Transliteration is generally linked with the phonetics of the target language

Чернобыль (RU), Чорнобіль (UA), Tchernobyl (Fr), Chernobyl (ISO US)

For Arabic, the lack of vowels gives a lot of possible transliterations.

فيصل , Faical, Faicel, Faisal, Faissal, Faycal, Faycel, Faysal, Fayssal
فيروز , Fairouz, Fayrouz, Ferouze, Ferouz
Jean, جان , جون

In Chinese : the representation in Latin character (Pinyin) can give thousand of possible ideogram strings. The best way to find the right spelling is to filter using the web.

wen jiabao温 家宝	7,210 pages ← the right spelling has a larger number of occurrences	
wen jiabao文 家宝	219	pages
wen jiabao闻 家宝	32 pages	

4 Crosslanguage interrogation:

The crosslanguage research activity has been launch in Europe in the beginning of the 90, with the EMIR project (1990-1994), followed by others like MOULINEX.

In the middle of the 90s This theme has been considered as an important theme of research especially after a report from DARPA considering the language barrier “This barrier puts the United States at a distinct STRATEGIC DISADVANTAGE in the international competition for Information Dominance, because technical and military personnel in other nations have far better skills in English than we have in their languages”

The problem of crosslanguage interrogation has been treated in various ways.

- use of machine translation tools for query translation of database translation
- use of statistical models for database containing a subset of translated documents. Models like LSI (Latent Semantic Indexing) can be applied to use implicit links between words and/or between documents to access relevant documents that are not in the subset of translated ones.
- Use of bilingual dictionaries (approach of the EMIR European project). This approach tries all possible translations of each query word and eliminates the wrong translations using the database (or web) as a semantic filter).
- Use of a limited set of concepts which can be recognised in each language. This approach is possible only for general interrogation but not for very precise queries.

Example of crosslingual interrogation :

This example is taken from the ALMA European project. Languages are French, English, Spanish and Arabic. The technology by NewPhenix uses the bilingual dictionary approach

5 Crosslanguage filtering

Filtering and interrogation are slightly different. Interrogation is performed to ask generally a precise query on an existing set of data corresponding to the problem to be solved at this moment. Filtering is a continuous process of observation of the information entering into a system. The query (named in this case profile) has a permanent value. So, it can be more elaborated, it can be modified along the time to have more relevant results or modified because the user changes slightly his strategy according to the first results.

Elaboration of filtering profiles can use more manual/intellectual work than instant query for interrogation.

But it is the only big difference. At the end the same process must compare a query (short or long) with a document text which could be in a different language.

6 Crosslanguage clustering :

In Business intelligence it is important to identify new important event as early as possible. Some events that have very few occurrences in one language can more easily be seen if observed across languages. Crosslanguage clustering is able to show that if an event is not significant in one language it can be identified as important if various countries show interest in this event.

Building clusters across languages can be done using the crosslanguage interrogation systems. A matrix of distance can be computed using the interrogation of the database by each

document. The system gives a distance between theis document-query and the closer documents in the database.

When the matrix is built, a clustering algorithm can be used like SNN.

7 Speech to text translation

Speech to text translation is compulsory to extract information from broadcasted or on the web radio and video. A lot of research teams or companies are in this field speech to text processing. Most of the activity is not for radio or video processing but for speech dialog to interrogate databases.

The maximum of effort has been spent on news and a little bit less for telephone conversations. The funding of research by intelligence agencies (directly or through research foundations) can explain this limitation. But for the purpose of Business intelligence, the news are good sources of information so business intelligence can really use the tools developed by these research programs.

Limsi (CNRS) and VECSYS (its industrialisation body) have very good results in the NIST competitions. Outputs of such programs give a character string divided into words. For each word a time code gives the position of its pronunciation in the radio or video signal. This can be used by crosslingual interrogation systems to locate relevant information and show the only relevant part of TV or radio news.

8 Machine translation:

Crosslanguage interrogation can show relevant parts of documents even if the user doesn't understand the language. In this case a rough translation is very interesting to confirm the relevance of the document. That is the reason why a link between crosslanguage interrogation and machine translation is very useful. The high quality of translation is not compulsory. So most of existing machine translation systems can be used.

Machine translation has a long history in computer sciences. Historically systems are based on a linguistic processing to analyse the source language, a transfer that is specific to a couple of source-target languages and a generation in the target language. Systems like SYSTRAN (Systran SA) or Reverso (Softissimo) are of this kind.

Few years ago, translation memories have been designed to give a better productivity to human translators especially in case of translation of technical documentation concerning consecutive versions of a same product. This technology is based on existing translations. Bilingual texts are aligned on the sentence level. When a new sentence is to be translated, a comparison with stored sentences in the same language is done and if a sentence with a maximum of intersection is found, its existing translation can be used.

The increasing number of bilingual texts that can be found in the web has given an opportunity to develop statistical technologies to translate new texts by analogy with the existing translations found in bilingual texts. The technologies are similar to those used for speech to text recognition. Correlation between n-gramms (succession of n words) in the source language and n-gramms in the target language are used to propose translation. The

University of Aachen has been a pioneer in this domain. Google today uses this kind of technology for some couple of languages like Arabic-English.
This technology based on existing bilingual texts and n-gramm model of languages brings some limitation. Evolution to systems that use cross language interrogation based on large bilingual dictionaries and an indexation of the full target languages can eliminate the limitations. It can also solve the problem of translation from a rare language to a largely represented one because the source language necessitates only a syntactic analysis. For statistic translation no large bilingual corpus can be found if the source language is rare.

9 Conclusion :

Technologies for multilingual and multimedia processing can already be used in the framework of business intelligence. Of course, a lot of work remains to be done to ameliorate the quality of the results, the number of languages available, the quality of speech recognition out of the news domain and the semantic interpretation of images. But the state of the art today can already give a greet help to the business intelligence activity.

University and Intelligence: an Italian point of view

Mario Caligiuri
University of Calabria
University of Rome "La Sapienza"

After having highlighted the relevance of Intelligence in the globalized world, we will first analyse the possible relation between Intelligence and University, completely absent in Italy at present.
The issues discussed in such contest will include the promotion of intelligence as an academic discipline, the selection of the best graduates and senior year students, the identification of the required skills, the analysis of the political and economical phenomena, the multidisciplinary integration of diverse scientific sectors, the transparent development of a technological and psychological research and the expansion of a culture of Intelligence.
The course will then examine those scientific disciplines more strictly linked to intelligence issues, such as communication, journalism, business organization (particularly decision-making skills), economy (specifically the business intelligence branch, overlapping more and more with the government type) history, law, sociology, psychology, technology and others.
A review will be carried out of how the relation between university and intelligence evolved in Italy and abroad (with particular reference to the United States), followed by an analysis that will draw attention to the need of developing further the culture of intelligence in our country, taking in consideration the historical, social and economical specificity of the Italian context.
The conclusions will argue that the intelligence agents have to be an highly specialized and trained elite at State service to safeguard its security and welfare.
The relation between University and Intelligence proves vital to reach this goal.

Language technology evaluation in Europe
Key achievements and the need for an infrastructure

Khalid Choukri

ELRA/ELDA
55-57 Rue Brillat-Savarin, 75013 Paris, France
Tel.: +33 1 43 13 33 33 – Fax: +33 1 43 13 33 30
Email: choukri@elda.org
URL: http://www.elda.org or http://www.elra.info

ABSTRACT

This abstract aims at describing briefly the evolution of language technology evaluation in Europe. This will be directly linked to the activities of the European Language Resources Association (ELRA) and its operational body, the Evaluation and Language resources Distribution Agency (ELDA). The rational behind the foundation of the European Language Resources Association (ELRA) and its Evaluation and Language Distribution Agency (ELDA) in 1995 will be elaborated upon and the HLT Evaluation activities carried out since then highlighted. We would like to focus on the issues to address for making language resources available to different sectors of the language engineering community and, in particular, on those needed to carry out evaluation activities. The presentation will introduce a number of Evaluation projects and Services established through a large number of European and nationally funded projects. In addition, ELRA carries out promotion tasks in the field of Human Language Technology (HLT), in order to advertise resources and any relevant activity (with the maintenance of catalogues, the edition of its Newsletter, the organization of the LREC Conference, which is now a major event for the HLT community, and the maintenance of the HLT Evaluation Portal, among others).

ELRA's Mission

ELRA's initial mission was to set up a centralized Not-for-profit organization for the collection, distribution, and validation of speech, text, terminology resources and tools. In order to play this role of a central repository, ELRA had to address issues of a different nature such as technical and logistic problems, commercial issues (prices, fees, royalties), legal issues (licensing, Intellectual Property Rights, …), information dissemination (to act as a clearing house). This mission is tuned from time to time to anticipate future requirements. As of today, this can be reflected by the following tasks:

- The identification of useful Language Resources.
- The handling of legal issues related to the availability of Language Resources.
- The Language Resource distribution activities and Pricing policy.
- The validation and Quality assessment of Language Resources.
- The commission of the production of needed Language Resources & Market watches.
- The information dissemination, Promotion and Awareness.
- **Last but not least, the supply of the Evaluation services to the HLT community.**

HLT Evaluation activities

For any HLT research effort to be successful, it is essential that it be assessed through rigorous evaluations of the developed technologies. This allows performance benchmarking and a better understanding of possible limitations and challenging conditions. Since 2000, ELDA had an objective to design and validate evaluation packages for several Human Language Technologies. An ELDA's evaluation package (or Evaluation Kit) comprises the following items:

1) An evaluation protocol specification, including specification of the task to be performed by the systems being evaluated, metrics, data representation formalisms, and the relevant documentation.
2) Development data representative of the task and in sufficient amount to enable a full validation of the evaluation protocol.
3) The test data that will be used to score systems' performance.
4) All the software tools required to run an evaluation campaign implementing the protocol defined in 1), i.e. format standardization and validation tools, measuring tools, result presentation tools, data server, storing and communication tools, etc.

These evaluation packages had to be made available for organizing large evaluation campaigns, involving all key players from laboratories developing technologies targeted for evaluation. The evaluation packages also had to be made commercially available upon request for government agencies or industries wishing to organize other evaluation campaigns. Finally, these packages are distributed for industrial or public research entities wishing to evaluate a technology (possibly the one they develop) and compare it to the state-of-the-art.

In many cases, ELDA also helps to provide data for system training and, since 2006, it provides an on-line evaluation platform for several technologies (web-service and/or UIMA-based platform) to avoid the installation of scoring tools at each site.

In order to achieve such goal, ELDA has established an evaluation Department that takes care of assessing and benchmarking Human Language Technologies both within R&D projects and for customers. In order to do so, ELDA has joined efforts with some of the largest consortia involved in HLT development and it has managed to ensure that the consortia capitalize on the evaluation work through the packaging of all needed pieces to carry out similar initiatives afterwards. A new paradigm refers to this task as the "project exit strategy". Such strategy ensures that the availability of the "evaluation package" as described above (the full documentation, definition and description of the evaluation methodologies, protocols and metrics, alongside the data sets and software scoring tools) is an essential part of each project. An evaluation package can be conditioned so that it can be distributed through ELRA's Catalogue. This allows any organization to reproduce one of the technology evaluations that were conducted during the project Evaluation Campaign. The exploitation of this outcome is one of the achievements of the project.

ELRA's Focus on Evaluation

The ELSE project (Evaluation in Language and Speech Engineering, 1999) conducted a study on the possible implementation of HLT evaluations in Europe and compared them to other initiatives conducted in the USA and Japan. Among the findings of this project we can quote the need for comparative evaluations conducted by a truly European infrastructure that would ensure long-term availability of expertise and resources so as to avoid the loss that occurs when projects are funded through R&D programs for a short period of time. The ELSE project also highlighted how important this was for the developers that benefit indirectly from evaluation through the acquisition of the complete evaluation toolkits and by-product data that become available afterwards but also through the knowledge sharing that takes place systematically in the post-campaign workshops during which experts compare approaches and techniques used by each system.

Within these evaluation activities, ELDA has participated in a large number of projects and initiatives which have helped reinforce its expertise in the area and have supported the development of HLT Evaluation in Europe. Among these projects we will mention just a few here to highlight the huge European investment and the crucial need to ensure a serious Return on investment through the exploitation of such packages but also through the support of the ELDA infrastructure to become self-sustainable.

A hot topic these days is Machine Translation technology, including Speech to Speech translation systems. Through its involvement in the FP6 project **TC-STAR**, ELDA has contributed to the evaluation of speech recognition systems, machine translation, and speech synthesis systems. In addition, ELDA conducted end-to-end evaluations and compared the achievements of TC-STAR systems with the work of human interpreters for English and Spanish. A review of such work will be described during the talk. One of ELDA's tasks was the collection and annotation of huge sets of spoken multilingual corpora and the corresponding written corpora that was used to train the systems. ELDA has also been in charge of elaborating the global evaluation plan for the 3 evaluation campaigns of the project. Today, all packages are being made available to assess system performance and a number of copies have already been distributed.

Another major project worth mentioning here is **CHIL** ("Computers in the Human Interaction Loop", an FP6 IP). The implication of ELDA made it easy to capitalize on the work conducted within the project and to ensure that all data sets and evaluation toolkits are made widely and immediately available under very fair conditions. CHIL has addressed the largest number of technology components ever done before with the goal to develop computer assistants that attend to human activities, interactions, and intentions. CHIL's thirteen technological components were evaluated, which focused on Vision technologies (Face Detection, Visual Person Tracking, Visual Speaker Identification, Head Pose Estimation, Hand Tracking), on Sound and Speech technologies (Close-Talking Automatic Speech Recognition, Far-Field Automatic Speech Recognition, Acoustic Person Tracking, Acoustic Speaker Identification, Speech Activity Detection, Acoustic Scene Analysis) and on Contents Processing technologies (Automatic Summarization and Question Answering on Spoken Transcriptions, conducted in partnership with CLEF). The corresponding evaluation packages are being made available through ELRA's Catalogue.

A further international achievement resulting from this project, and of great importance, is the establishment of the open international evaluation workshop **CLEAR** - "CLassification of Events, Activities, and Relationships", in partnership with NIST and other players. So far, two CLEAR evaluation campaigns were conducted, partly with CHIL packages.

Another major area being tackled by most of the HLT key players is the Multilingual/Cross-Lingual Information Access and Retrieval. Through some partial European funding, **CLEF** (Cross-Language Evaluation Forum) was launched in 2000 with the aim to develop an infrastructure for the evaluation, testing and tuning of information retrieval systems operating on European languages in both monolingual and cross-language contexts and, beyond this, to experiment the setting up of a European HLT evaluation infrastructure. The project managed to create test suites of reusable data which are part of the ELRA evaluation catalogue and which are extensively employed by system developers for benchmarking purposes. The exploitation of the methodologies implemented by CLEF for the testing and tuning of information retrieval systems is now part of ELDA's assets and it allows conducting the evaluation of commercial products and applications with a strong and reliable technical and scientific background. More than 11 copies of the CLEF packages have been distributed so far.

Another important initiative, funded by a national agency, is the French programme "**Technolangue/Evalda**": the Evalda projects that ELDA has coordinated consisted of 8 evaluation campaigns with a focus on the spoken and written language technologies for the French language: ARCADE II (evaluation of bilingual corpora alignment systems), CESART (evaluation of terminology extraction systems), CESTA (evaluation of machine translation systems), EASY (evaluation of parsers), ESTER (evaluation of broadcast news automatic transcribing systems), EQUER (evaluation of question answering systems), EVASY (evaluation of speech synthesis systems), and MEDIA (evaluation of in and out-of context dialog systems). As planned and achieved within the other evaluation projects, all Evalda evaluation resources have been packaged and made available and more than 15 copies have been distributed so far.

Further to its participation in such projects, ELDA has also run a number of initiatives, such as discussion and brainstorming events. One such example is ELRA's 10th Anniversary Workshop on Evaluation, celebrated in Malta in December 2005. The discussions initiated at this occasion have been taken further with the Evaluation Workshop celebrated during the MT Summit 2007 (Automatic Procedures in MT Evaluation), and a coming ELRA Evaluation Workshop (Looking into the Future of Evaluation: when automatic metrics meet task-based and performance-based approaches) to be celebrated jointly with the LREC 2008 Conference, this coming May-June 2008.

Some topics that will be addressed during the talk

During the talk, I will elaborate on the role of evaluation on the research progress, on the need for a truly European infrastructure for HLT evaluation and the potential role of ELRA within such structure, on the main reasons to promote an international dimension of the evaluation, insisting on the multilingual issues. I will introduce and describe some evaluation concepts (comparative evaluation versus competition, Technology evaluation versus Usage/usability evaluation). I will also describe the different types of evaluation and how to ensure that evaluation does not kill very innovative not-yet-mature approaches.

Conclusion

As stated above, ELRA has been entrusted with a crucial mission: to ensure that Language Resources needed by Language Engineering players are made available when they already exist or to produce them in a cost-effective frame. It is of paramount importance that regional organizations emerge and co-operate between themselves with respect to the issues described herein. The main common task would be to achieve, all together, a better streamlining of efforts in the development of new Language Resources that are of interest to "local" and "global" players. This role should be extended to Evaluation in particular in geographical areas that do not have a dedicated organization.

At the same time, the paradigm of evaluation should be reconsidered by the funding agencies and funded as a major part of their investments, as it allows both to measure if the money they have invested in technology development has led to significant progress and to identify areas where the technology needs further improvement.

Evaluation also allows application developers/integrators and end-users to understand where the technology is and how it can help them and provide them with new solutions to the problems they face.

ELRA also initiated the HLT Evaluation portal that is designed to be an online information resource about HLT evaluation and related topics of interest to the HLT community at large.

Putting HLT research and technology into action for European multilingualism

Kimmo Rossi

Unit E1 "Interaction and Interfaces", Directorate-General for Information Society and Media, European Commission

kimmo.rossi@ec.europa.eu

Abstract

These speech notes outline the state of play and orientations of language technology in the context of publicly co-funded European technology projects.

Index Terms: language technology, European projects, HLT, multilingualism

1. Background and policy context

Research and innovation in language technology have been intensively supported for over 20 years in the European technology funding framework. Since the 1990's, research Framework Programmes were complemented by specific programmes such as the Multilingual Information Society (MLIS) and eContent. These programmes provided direct financial support to projects for translation and terminology tools, multilingual information retrieval, machine translation, access to multilingual content etc. Research and development in language technology continue to be supported in the 6th and 7th framework programmes [1][2], although the trend is towards integration and mainstreaming of language understanding in interactive systems and knowledge management. Significant progress has been achieved in areas such as speech technology and data-driven machine translation. Large amounts of language resources have been collected, annotated and refined.

The i2010[3] is the EU policy framework for the information society and media. It promotes the positive contribution that information and communication technologies can make to the economy, society and quality of life. One of the main objectives of the i2010 framework is to create a *single European information space*. This single information space relies on rich and diverse online content and digital services.

2. The challenge of multilingual content

The provision of rich and diverse content relies on paying appropriate attention to linguistic diversity – how to enjoy the full potential and richness of European multilingualism and overcome the language barrier? How can global online services and content be offered to all citizens and businesses, irrespective of language? These are questions that require new insight on how language technologies are applied and integrated with associated technologies such as (web) content and knowledge management and search technologies. Language technology is above all an enabling technology for ICT systems and online solutions.

But much more is needed than simple machine translation of strings. Plug-in solutions have their limitations. Language diversity needs to be mainstreamed into all phases starting from content production, authoring, content management, presentation, IT architectures etc. Many of the enabling technologies may already exist, but the main challenge lies in the integration. We are still far from the ideal online use scenario where anyone can seamlessly access content, use services and purchase goods, solve problems and communicate across language boundaries.

3. Diversity means business

Business has migrated to the web, and customers require solutions and information in their own language. Multilingual websites with true multilingual access and services are likely to sell better in the global market than English-only websites. On the other hand, the efficient use of information and content generated by the public sector (including the European institutions) requires efficient solutions that address the language barrier. There are plenty of success stories of companies that have been able to expand and win new markets by successfully making use

of linguistic diversity. Translation and localization can be considered as an inevitable operating cost, but it can also be considered an investment in marketing and product development. If successful, the return on this investment can be considerable, and multilingualism can be a source of expansion and growth of business.

Recent advances in language technology already offer great potential to businesses, but a lot of work needs to be done. Barriers seem to exist to the efficient take-up and exploitation of language technologies (e.g. speech recognition, machine translation, cross-lingual search) as these are not as widely employed as they could be. Further efforts are needed to put the state-of-the-art language technologies into productive use. This will not only justify the heavy investments in language technology research but will also provide a competitive edge to companies and better services to the citizens. While the major languages are well equipped with language resources and tools, there are still gaps in the coverage of some less widely spoken languages. Efforts are still necessary to make languages more equal.

4. European technology actions promoting multilingualism

As mentioned, the emphasis of EU-funded technology actions has shifted from dedicated language technology research and development projects towards integrated and applicative approaches which have illustrated the power of language-enabled interaction and demonstrated how the operation of systems and appliances can be facilitated with natural language interfaces.

European projects, such as TC-STAR[4] and EUROMATRIX[5], have made a significant contribution towards setting up a comprehensive and comparable evaluation infrastructure for language resources and data-driven machine translation systems. Other projects such as TALK[6] have contributed to more natural and adaptive human-machine dialogues. A particularly large number of actions have addressed language resources, both spoken and written, some more language-oriented, others semantic and ontology-driven. The LIRICS[7] project contributed to the exchange and reuse of language resources and defined an ISO standardization framework to organize this work. While the continued importance of language resources is undisputable, the requirements of portability, scalability and cost-effectiveness require that more is done to promote standardization and automation whether for collection, annotation, processing or use of language resources.

While the European technology programmes continue to support efforts to overcome the language barrier as well as promote the use of language technologies and language resources, it is equally necessary that providers of language technologies and online service providers find new ways of satisfying the need for multilingual services. New partnerships are needed to put in place systems that will effectively reach out to customers in their own language. European projects and other collaborative actions have proven to be a well suited instrument to pursue these efforts.

5. References

[1] CORDIS: the European research web service http://cordis.europa.eu/
[2] A list of European projects (FP5 and FP6) on interaction, interfaces and language technology http://cordis.europa.eu/ist/ic/projects.htm.
[3] http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm
[4] http://www.tc-star.org/
[5] http://www.euromatrix.net/
[6] http://www.talk-project.org/
[7] http://lirics.loria.fr/

Crossing media for improved information access: the REVEAL THIS example

Stelios Piperidis
Head of Language Technology Applications Department
Institute for Language and Speech Processing – Athena R.C.
spip@ilsp.gr

Abstract

The explosion of multimedia digital content and the development of technologies that go beyond traditional broadcast and TV have rendered access to such content important for all end-users of these technologies. REVEAL THIS develops content processing technology able to capture, semantically index, categorise and cross-link multiplatform, multimedia and multilingual digital content, providing the system user with search, retrieval, summarisation and translation functionalities.

Introduction

The development of methods and tools for content-based organization and filtering of the large amount of multimedia information that reaches the user through heterogeneous channels is a key issue for its effective consumption. Despite recent technological progress in the new media and the Internet, the key issue remains "how digital technology adds value to information channels and systems".

The outcome of the REVEAL THIS project (www.reveal-this.org)addresses this issue by helping people keep up with the explosion of digital content scattered over different platforms (radio, TV, Web), different media (speech, text, image, video) and different languages. It provides users with search, retrieval, categorization, summarisation and translation functionalities for multimedia content, through the use of automatically created semantic indices and links across media.

In designing its technological solutions, REVEAL THIS had to tackle the following scientific and technological challenges:

- enrichment of multilingual multimedia content with semantic information like topics, speaker names, facts or events and their participating entities, keyframes and face names relevant to user profiles
- establishment of semantic links between pieces of information presented in different media and languages within the same as well as across multimedia documents
- development of cross-media categorization and summarization engines
- deployment of cross-language information retrieval and machine translation to allow users to search for and retrieve information according to their language preferences.

Users of this technology include a) **content providers** who want to add value to their content, restructure and re-purpose it and offer personalised content to their subscribers, and b) **end users** who wish to gather, filter and categorize information collected from a wide variety of sources in accordance with their preferences.

Web, TV and /or Radio content is fed into the REVEAL THIS prototype, it is analysed, indexed, categorized, summarized and stored in an archive. This content can be searched and/or pushed to a user according to his/her interests. Novice and advanced computer users are targeted; the former use mainly a simple mobile phone interface where information is pushed to, while the latter use a web interface for searching. EU politics, news and travel data are handled by the system in English and Greek.

The REVEAL THIS system comprises the following subsystems : (i) the Cross-media Content Analysis & Indexing (CAIS), (ii) the Cross-media Categorisation (CCS), (iii) the Cross-media Summarisation (CSS), (iv) the Cross-lingual Translation (CLTS), and (v) the Cross-media Content Access and Retrieval.

Cross-media Content Analysis and Indexing

The main REVEAL THIS subsystem, the Cross-media Content Analysis and Indexing Subsystem (CCAIS) caters for processing single media and automatically generating metadata for each medium such as:

- speaker turn and identity, transcriptions and stories for speech data,
- named entities (persons, places, organizations), topics and facts for web text and transcribed speech
- keyframes, shots, faces (detected and identified) and image categories for video and images

The metadata produced by the single-media processing components are aligned, synchronized and encoded in XML. The result is an XML/MPEG-7 document containing all information gathered and linked to the corresponding points of the source material (text, audio, video). Segmentation suggested by audio processing (speaker turns) and topic detection is taken into account to produce unified segments. Several consecutive segments are aggregated to form “stories”, i.e. sections of the document that deal with the same topics. The cross-media indexing component decides on the most appropriate indexing terms per story. Crossing media (*or cross-mediality*) in REVEAL THIS is conceived of as the process of intelinking evidence, in the form of indexical data, provided by the different media participating in the message formation process within the same multimedia document (e.g. a video file). Deploying this indexical information and representations, REVEAL THIS uses state-of-the-art technologies for categorizing, summarizing and translating multimedia documents.

Cross-media categorisation

The categorization subsystem operates along the cross-lingual and cross-media directions. In the cross-lingual dimension, a categorizer based on a pivot language category model (in English) is deployed. Documents in other languages go through a translation phase, thus enabling their categorisation. Such a strategy overcomes constraints at the level of the training set, which often are not sufficiently big for building models in all languages involved.

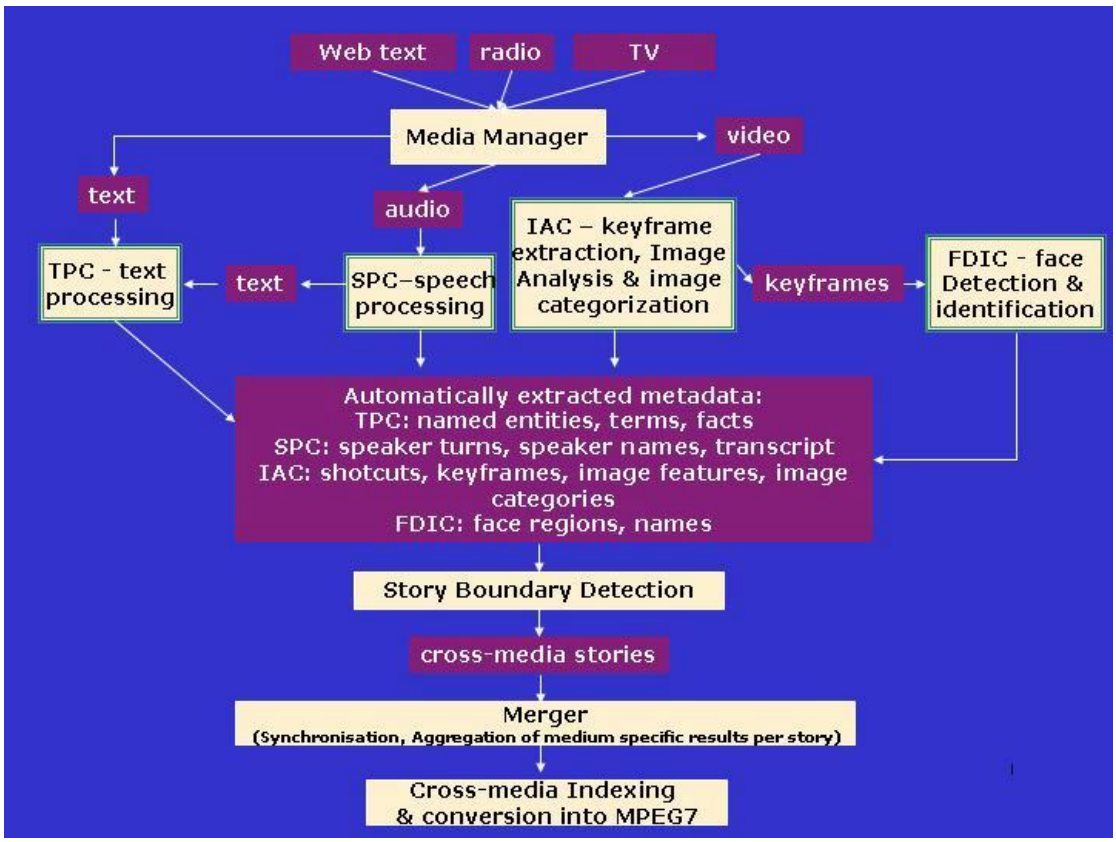


Figure 1: **REVEAL THIS Cross-media Content Analysis and Indexing**

In the cross-media dimension, documents containing not only text or images but a combination of different types of media (text, image, speech, video) are considered. The multiple-view fusion method adopted builds 'on top' of two single-media categorizers, a textual and an image categorizer, without the need to re-train them. Data annotated manually for both textual and image categories is used for training the cross-media categorizer. In that set dependencies between single-media category systems are exploited in order to refine the categorization decisions made.

Cross-media summarisation Subsystem

The task of the Cross-media Summarization Subsystem (CSS) is to determine and present the most salient parts according to users' profiles and interests by fusing video, audio and textual metadata. The CSS subsystem consists of three major components: the textual-based summarization component, the visual-based summarization component, and the cross-media summarization component aiming at the fusion of the two analyses and creating a self-contained object. Additionally, the Cross-media summarization subsystem provides the necessary visualization interfaces enabling the user to preview a specific multimedia object before downloading a file of interest. Cross-media summaries for politics, news and travel-related audiovisual files have been designed and implemented in the project; while the module interfaces also with a web-based translation service for creating "translated" versions of the summaries for English and Greek.

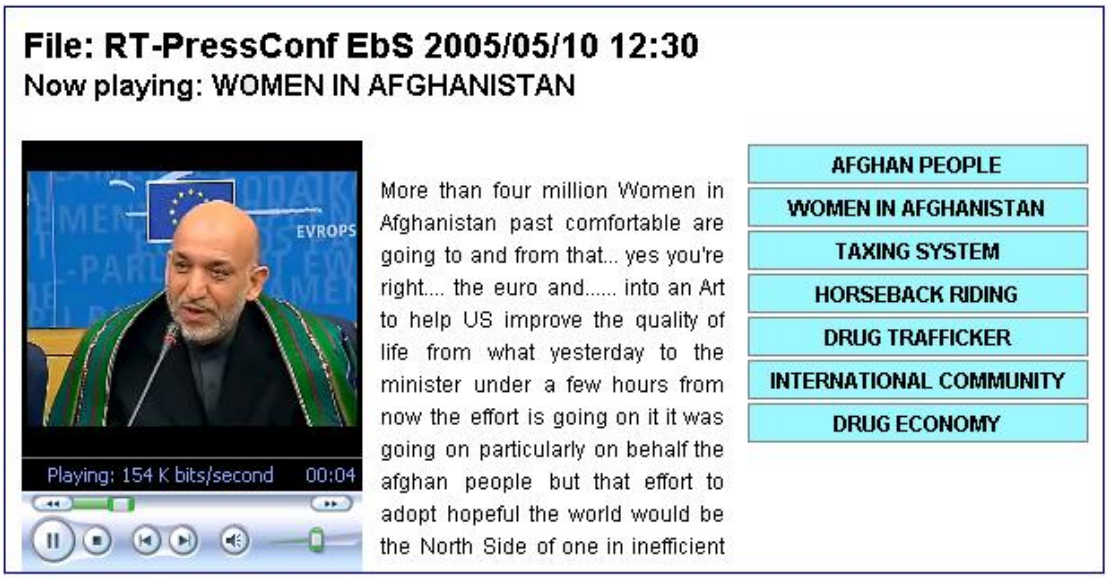


Figure 2: Multimedia summary (file summary)

Cross-lingual Translation

The Cross-lingual Translation subsystem (CLTS) allows users to query documents written in different languages, to categorise content expressed in different languages and to preview language specific summaries. A bilingual lexicon extraction module is used to generate lexical equivalences used for query translation purposes, but also to replace keywords in a target language, in case a document is linguistically not well formed (e.g. output from a speech recognizer) and thus not effectively translated. Last, a statistical machine translation module is responsible for providing translations of the textual part of the summaries produced by the Cross-Media Summarization Subsystem. The translation pair is English and Greek and translation takes place in both directions.

Integrated Application Prototype

The REVEAL THIS system consists of two major components: The *Multimedia Indexer* (based on the cross-media indexing component – CMIC) and the *Media Server*. All results of the Indexer together with the original media files are uploaded to the Media Server. Media files are uploaded in two formats: in Window Media format (.wmv/.wma) and in 3gp format (targeted at mobile devices). The server supports *searching for content* (pull scenario) or *filtering content* (push scenario) based on the metadata. It provides multi-lingual search, multi-media presentation of retrieved content, personalization, summarisation, and delivery to different devices and in particular to PCs/Laptops and mobile phones. The web-interface of the prototype provides full search and retrieval functionalities targeting professional users with demanding information needs, while the mobile-interface provides a light-weight access to the different functionalities of the prototype, suitable for mobiles and for laymen.

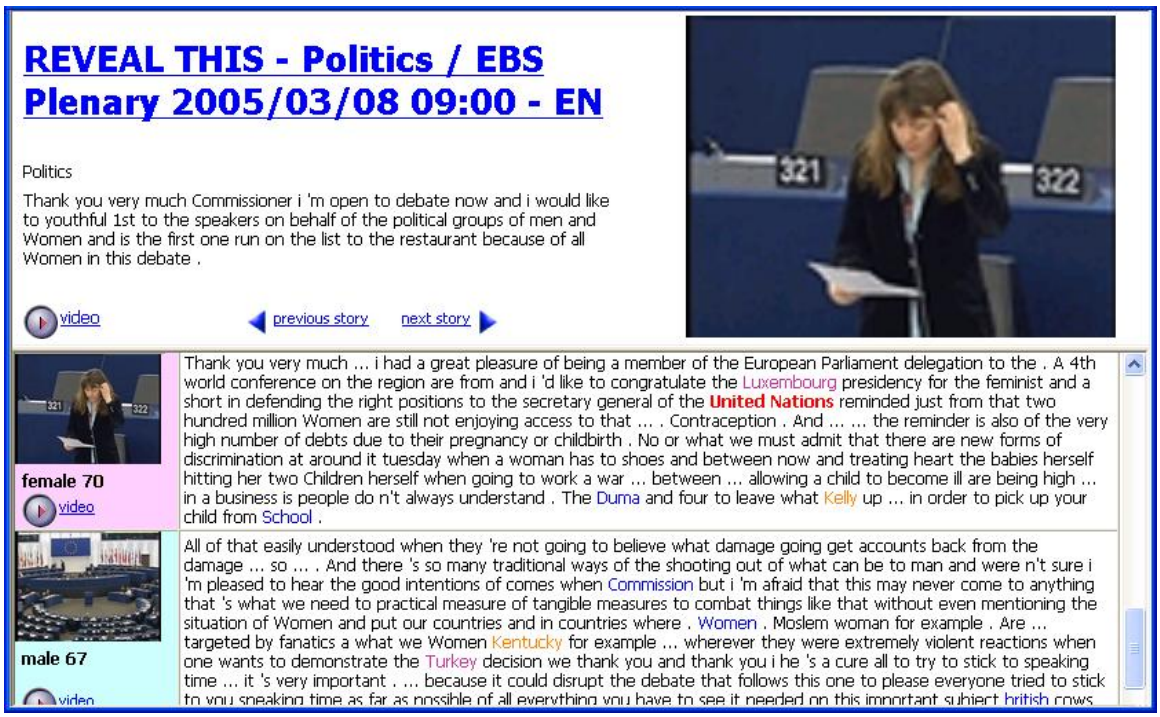


Figure 3 : Story View in the REVEAL THIS web-interface

The system was tested in the two domains of EU politics and travel information, and in two languages, English and Greek. Advanced users, experienced in multimedia search, used the prototype in its pull (retrieval) mode (the web interface), while novice users, with limited or no search experience used it in its push (filtering) mode through mobile phones. The results showed that the innovative REVEAL THIS functionalities were more than welcome by both user groups, the performance was generally satisfactory for them, though expectations for much better performance in terms of e.g. speech transcription and translation were evident.

Figure 4: **Screenshots of the REVEAL THIS mobile-application interface**

REVEAL THIS provides a technology suite enabling the development of a fully operational personalized entertainment system. Such a system can be used by content providers, to add value to their content, and directly by end users, for gathering, filtering and categorizing multimedia information.

Acknowledgements

The REVEAL THIS project was funded by the FP6-IST programme of the European Commission, contract number FP6-IST-511689 and was designed and implemented by the REVEAL THIS consortium comprising Institute for Speech and Language Processing / IRIS (Co-ordinator), SAIL LABS Technology AG, Xerox - The Document Company S.A.S, Katholieke Universiteit Leuven R&D, University of Strathclyde, BeTV SA and TVEyes UK Ltd.

Challenges of Speech to Speech Translation in the context of Human-Human Communications

Alex Waibel
InterACT

Recognition and Understanding of Meetings
Overview of the European AMI and AMIDA projects

Herve Bourlard[1] and Steve Renals[2]
[1]*IDIAP Resarch Institute,* [2]*University of Edinburgh*
bourlard@idiap.ch, www.idiap.ch
www.amiproject.org

Abstract

The AMI and AMIDA projects are concerned with the recognition and interpretation of multiparty (face-to-face and remote) meetings. Within these projects we have developed the following: (1) an infrastructure for recording meetings using multiple microphones and cameras; (2) a one hundred hour, manually annotated meeting corpus; (3) a number of techniques for indexing, and summarizing of meeting videos using automatic speech recognition and computer vision, and (4) a extensible framework for browsing, and searching of meeting videos. We give an overview of the various techniques developed in AMI (mainly involving face-to-face meetings), their integration into our meeting browser framework, and future plans for AMIDA (Augmented Multiparty Interaction with Distant Access), the follow-up project to AMI. Technical and business information related to these two projects can be found at www.amiproject.org, respectively on the Scientific and Business portals.

1. Introduction

Over the last few years research interest in recording, archiving, and retrieving of *meeting videos* has increased significantly. This is due to major drops in hardware costs, availability of broadband (for remote meetings), and concerns by corporations about record keeping (auditing decision-making, corporate memory, and complying with regulatory requirements, among others).

Meetings play a crucial role in the generation of ideas, documents, relationships, and actions within an organization. The wealth of information exchanged in meetings, however, is often lost because human-note taking of meeting minutes is subjective and incomplete and captures only a fraction of the information. Audio-visual recording of meetings is therefore attractive, but leads to many practical challenges, from the infrastructure to record the meetings to the archival, indexing, and retrieval of relevant meeting segments. Given the number of meetings in most organizations, efficient and effective recording and access to meeting videos is of extreme importance, making research in content-based indexing and retrieval of meeting videos an important research area, not only because of its potential impact, but also because it requires combining research in several disciplines (e.g., speech recognition, computer vision, etc.).

In this paper, we describe the AMI project. AMI deals with meeting videos throughout the media production chain: from modeling of meetings, to recording infrastructure and recording, to multimodal, automatic indexing, retrieval, and browsing of meeting videos. We give a general overview of each of the components above and discuss use of such AMI technologies within the browsing framework we have developed, which allows browsing, searching, and summarization of meeting videos. The goal of this paper and its main contribution, therefore, is to give an overview of the technologies developed in the project[1] and their integration in the browser framework.

Related Work: Related meeting room projects include [1-3, 5, 12-15, 20, 25, 26, 32, 36], and others. Some works focus use portable recorders (e.g., [36]), others focus only on speech (e.g., 24]), or on particular types of meetings (e.g., [5]). Other projects are particularly focused on modeling (e.g., [4]), labeling (e.g., [6], or video capture (e.g., [27]). The AMI project's components build on and improve the state-of-the art in many areas. Since the goal of this paper is to give a general overview, however, we refer the interested reader to specific AMI publications (available at [9]) for details on how specific techniques developed within AMI differ from related work.

2. INSTRUMENTED MEETING ROOMS & AMI CORPUS

Three standardized meeting rooms were designed and constructed at AMI partners IDIAP, TNO and the

1 A large number of articles that describe technical details in depth have been published by the AMI project and are available in the project website for readers interested in more details.

University of Edinburgh. These rooms, which were designed for the recording of videos of four person meetings, all contained a set of standardized recording equipment (plus additional cameras, microphone arrays, and binaural manikins):

- 7 cameras: 4 providing close-up views of the participants, 3 providing a view of the whole room
- 12 microphones: 4 lapel microphones (one per participant) and an 8-element circular microphone array
- Data projector capture (VGA)
- White board capture digital pen capture

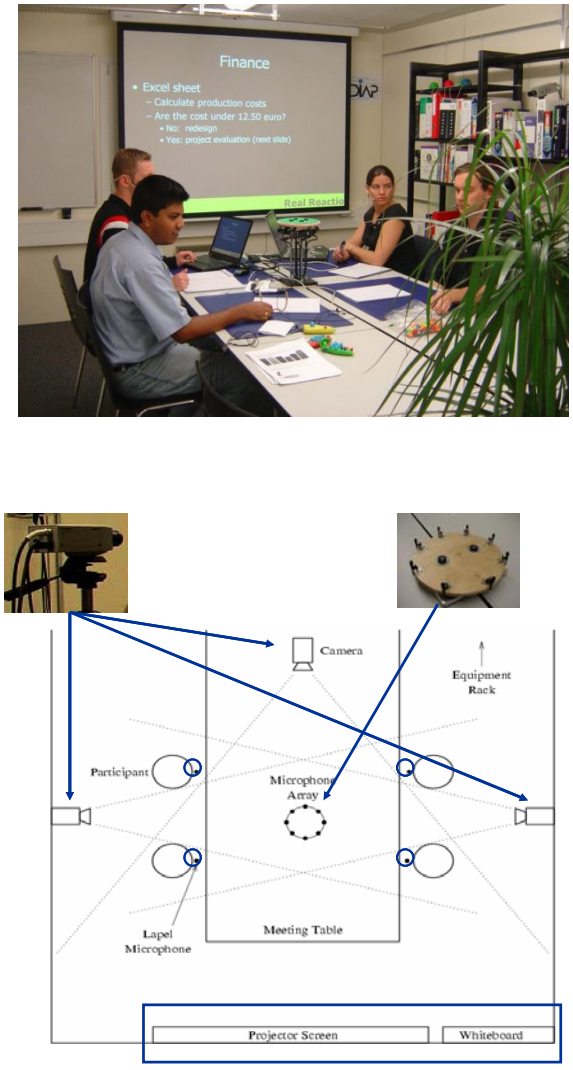


Figure 1: Typical AMI instrumented meeting room setup, integrating and synchronizing 4 close-talking microphones, 4 lapel microphones, 1 circular 8-microphone array, 4 close-view cameras (underneath the circular microphone array), 3 mid-view cameras, one captured projector screen, captured whiteboard activity and captured personal notes.

The meeting rooms were used to record the AMI Meeting Corpus (publicly available at http://corpus.amiproject.org), which consists of 100 hours of meeting recordings. The corpus includes manually produced orthographic transcriptions of the spoken dialogues, aligned at the word level with the common time line, and annotations describing participant behaviour during the meetings[2] (e.g., dialogue acts; topic segmentation; extractive and abstractive summaries; named entities; limited forms of head gesture, hand gesture, and gaze direction; movement around the room; emotional state; location of the heads in the video frames).

The corpus consists of two types of meetings: (1) design scenario (approx. 2/3 of AMI corpus), (2) free topic. In the design scenario, each group of four participants had four meetings and given tasks to complete between meetings. Participant roles were driven in real-time by emails and web information. This control made it easier to understand the content of the meetings, enabled the construction of ontologies, and the building of outcome measures (e.g., preferred design output). The meetings are also replicable, enabling system-level evaluations. Free topic meetings were naturally occurring meetings in a range of domains.

The project also further developed NXT (NITE XML Toolkit), an open source XML-based infrastructure for the annotation and management of multimodal recordings. NXT consists of libraries from which user interfaces for annotating and searching annotations of multi-modal data sets can be easily built. Within AMI, new tools for annotation were created, for instance for dialogue acts, named entities, topic segmentation, summarization, and a generic time-aligned coder and display.

3. AUDIO-VISUAL PROCESSING

AMI work in audio-visual processing was primarily concerned with the development of algorithms that can automatically answer each of the following questions from the raw audio-video streams:

- What has been said during the meeting? (Speech recognition)
- What acoustic events and keywords occur in the meeting? (Keyword spotting)
- Who and where are the persons in the meeting? (Localization and tracking)
- Who in the meeting is acting or speaking? (Speaker tracking)
- How do people act in the meeting? (Gesture and action recognition)

2 Not all 100 hours of meetings have been marked with all kinds of annotations (e.g., linguistic annotations cover 70% of the corpus).

- What are the participants' emotions in the meeting? (Emotion)
- Where or what is the focus of attention in meetings? (Focus of attention).

Speech recognition: Automatic transcription of speech in meetings is of crucial importance for tasks such as meeting analysis, content analysis, analysis of dialogue structure, and summarization.

AMI developed systems for the two types of microphone configurations in the instrumented meeting rooms (close-talking headset microphones and tabletop microphone arrays), focusing on the headset microphone conditions to develop core acoustic modeling approaches, but with an overall orientation to tabletop microphone arrays, which are less intrusive. In particular, the AMI speech recognition effort has addressed several research issues including the following:

- Microphone array beamforming: filtering and combining the individual microphone signals to enhance signals coming from a particular location (and suppressing competing locations);
- Development of novel acoustic parameterizations, including approaches based on posterior probability estimation
- Automatic construction of domain-specific language models using text extracted from the web
- Acoustic segmentation
- Development of a flexible large vocabulary decoder, based on a weighted finite state transducer formalism

AMI has developed an evaluation framework that is generic, flexible, comparable, and that allows us to conduct research and development in a stable environment. Using this framework, our system obtains exceptionally good results on AMI meeting data; in international technology evaluations organized by NIST, no other system was significantly more accurate than the AMI system on close-talking microphones. This system has been used to decode the complete AMI corpus (using an n-fold cross-validation technique), and these transcriptions have been used for tasks such as summarization and topic segmentation.

Keyword spotting: In acoustic keyword spotting (KWS), the goal is to find keywords and their position in speech data. AMI has developed three approaches: acoustic, LVCSR, and a hybrid approach.

In the acoustic approach, a keyword score is obtained by comparing the posterior probability of the keyword phonetic model, with a background model. This is very fast since many of the key parameters may be pre-computed. It is relatively precise (the precision increases with the length of the keyword) and any word can be searched provided its phonetic form is available. It is ideal for on-line applications (such as monitoring remote meetings), but it is not suitable for browsing huge archives, as it needs to process all the acoustic data for each search.

The LVCSR lattice approach locates the keywords in lattices generated by a large vocabulary continuous speech recognition system. Given the output of the speech recognizer, this approach is very fast, but it is accurate only for frequently occurring words. There is degradation in performance for less common words, which is a drawback, since these words (such as technical terms and proper names) carry most of the information and are likely to be searched by users. Therefore, this approach has to be complemented by a method unconstrained by the recognition vocabulary.

The hybrid phoneme lattice approach is based on the construction of graphs of phoneme probabilities, from which the phonetic form of the keyword may be extracted. This is a reasonable compromise in terms of accuracy and speed. Currently, AMI work on indexing phoneme lattices using tri-phoneme sequences is advancing and preliminary results show a good accuracy/speed trade-off for rare words.

Speaker tracking: The objective of speaker tracking is to segment, cluster and recognize the speakers in a meeting, based on their speech. The first approach developed in AMI uses the acoustic contents of the microphone signal to segment and cluster speakers. In the NIST evaluations this system produced very good results for speech activity detection (the lowest error rate reported) and for speaker diarization ("who spoke when"). The second approach developed in AMI, based on cross-correlations between microphone signals operates in real time, and has been integrated with the online keyword spotter.

Localization and tracking: Location coordinates of each person in the meeting are an essential input to various meeting analysis tasks, including focus of attention and action recognition. The steps required are identification, localization, and tracking. For identification, generative approaches have proven to be the most robust so in AMI a variety of models with different trade offs between speed and accuracy have been used (e.g., based on Gaussian mixtures and HMMs). The algorithms have been developed as a machine vision package for the open source machine learning library, TORCH (extended within AMI (http://www.torch.ch)). For localization and tracking AMI developed, applied, and evaluated four different methods including approaches based on dynamic Bayesian networks, active shape trackers using particle filters, and face trackers based on skin colour.

Gesture and action recognition: We have defined a set of actions and gestures that are relevant for meetings (e.g., hand, body, and head gestures such as pointing, writing, standing up, or nodding). Special attention has been paid to negative signals, such as a negative response to a yes-no question, usually characterized by a head shake. This kind of gesture contains important information about the decision making in meetings, but can be very subtle and involve little head movement, making automatic detection very difficult.

For gesture recognition two methods were applied: Bayesian Information Criterion and an Activity Measure approach. We extracted, for each person in the meeting, the 2D location of the head and hands, a set of nine 3D joint locations, and a set of ten joint angles. In addition we performed classification of the segmented data. Due to the temporal character of gestures we focused on different HMM methods. Gestures like standing up and important speech supporting gestures produced satisfactory results (100% and 85% recognition rate, respectively). However the results for the detection of negative signals were not significantly better than guessing. Detecting gestures such as shaking or nodding and negative signals is still a challenging problem that requires methods capable of detecting very subtle head movements.

Focus of Attention: Gaze detection requires higher resolution of facial images than what is available in the AMI corpus. As an approximation, we have developed algorithms for tracking the head and estimating its pose, based on a Bayesian filtering framework, which is then solved through sampling techniques. Results (evaluated on 8 minutes of meeting recordings involving a total of 8 people) were good, with a majority of head pan (resp. tilt) angular errors smaller than 10 (resp. 18) degrees. As expected, we found a variation of results among individuals, depending on their resemblance with people in the appearance training set.

In addition, we formulated focus of attention (FoA) as a classification task by automatically classifying FoA into one of the following categories: meeting participants, objects in the meeting room, and an “unfocused” location. Experiments using the ground truth head-pose pointing vectors resulted in frame-based classification rate of 68% and 47%, depending on the person's position in the smart meeting room. Accuracy is lower than reported in other works, mainly because of the complexity of the scenes and number of categories. Exploiting other features/modalities (e.g speaking status) in addition to the head pose can be used to disambiguate FoA classification. We found that using the estimated head-pose instead of the ground truth did not degrade the results strongly (about 9% decrease, thus much less than the differences w.r.t. position in the meeting room), which was encouraging given the difficulty of the task. We also found that there was a large variation of recognition amongst individuals, which directly calls for adaption approaches such as Maximum A Posteriori techniques for the FoA recognition. These adaptation techniques, along with the use of multimodal observations, are the topic of current research.

4. CONTENT EXTRACTION

Dialogue act recognition: Dialogue acts are labels for utterances which roughly categorize the speaker's intention. They are useful, for example as part of a browser which highlights all points where a suggestion or offer was recognized. Dialogue acts also serve as elementary units, upon which further structuring or discourse processing may be based (e.g., summarization). Each dialog act in a meeting is given one of the following labels:

- Information exchange: giving and eliciting information;
- Possible actions: making or eliciting suggestions or offers;
- Commenting on the discussion: making or eliciting assessments and comments about understanding;
- Social acts: expressing positive or negative feelings towards individuals or the group;
- Other: a remainder class for utterances which convey an intention, but do not fit into the four previous categories;
- Back channel, Stall and Fragment: classes for utterances without content, which allow complete segmentation of the material.

We have used combinations of machine learning based on a multimodal set of features, including a word-based language model, prosodic features (based on duration, energy and intonation), context features (e.g., speaker overlap), and discourse features (history of previously recognized dialogue acts). Using generative models that explicitly take account of the dependence on multiple streams of data (such as dynamic Bayesian networks, factored language models, and hidden event language models) we have obtained state-of-the-art results for dialogue act segmentation. Interestingly, although the best approach to dialogue act segmentation involves jointly segmenting and labeling the dialogue act sequence, we have found that the labeling may be substantially improved by re tagging using discriminative approaches, in particular conditional random fields. Comparing the performance on automatically transcribed speech with human transcribed speech, we find that the performance of dialogue act recognition drops by about 10%.

Topic segmentation: The aim of topic segmentation is to automatically infer the sequential structure of the meeting by topic (and sub-topic); it differs from dialogue act

recognition in that the fundamental units (topics) are typically many minutes in duration.

We have explored two basic approaches to this task. An unsupervised approach, LCSeg automatically infers (without training) topic boundaries as points where the statistics of text change significantly. The supervised approach, on the other hand, learns topic boundaries based on a hand-annotated training set. An advantage of the supervised approach is that it is possible to use additional features relating to prosody (e.g., pauses) and the structure of the conversation (e.g., speaker overlap). These additional features are also relatively independent of errors in the automatic speech transcription. We have also developed approaches to automatically generate labels for topics, based on the statistics of the automatically transcribed words that make up a topic.

Summarization: We have investigated two distinct ways of constructing summaries of a meeting. Extractive techniques construct summaries by locating the most relevant parts of a meeting and concatenating them together to provide a 'cut-and-paste' summary, which may be textual or multimodal. Abstractive summaries, on the other hand, are similar to what a human summarizer might construct, generating new text to succinctly describe the meeting. Abstractive summarization is more challenging than extractive summarization, and requires relatively deep domain knowledge.

Our approach to extractive summarization is based on automatically extracting relevant dialogue acts from a meeting. It thus requires (as a minimum) the automatic speech transcription and the dialogue act segmentation modules described above. Lexical information is clearly extremely important for this task, but we have found it beneficial to augment information derived from the transcription with speaker features (relating to activity, dominance and overlap), structural features (the length and position of dialogue acts), prosody, and discourse cues (phrases which signal likely relevance). All of these features are important to develop accurate methods for extractive summarization. We have also explored reduced dimension representations of text, based on latent semantic analysis, which also add precision to the summarization. Using an evaluation measure referred to as weighted precision, we have discovered that it is possible to reliably extract the most relevant dialogue acts, even in the presence of speech recognition errors. We have explored "dialogue act compression," in which the extracted dialogue acts are condensed by removing irrelevant portions. Again, taking account of speech features such as the overall intonation contour of the dialogue act helps to improve the overall performance.

We have also implemented a prototype abstractive summarization system, based on ontology of the AMI scenario meetings, together with annotations of propositional content, and the topic structure of the meetings. Given these annotations an ontological representation is built, which is then passed to a natural language generation component which produces a one paragraph summary of the meeting.

Influence and dominance detection: Person-to-group influence (i.e., influence of a person over the group) is estimated from audio features with a framework based on a two-level Dynamic Bayesian Network, in which an influence distribution is defined as the prior probability of individual state streams contributing to the group state stream. Such a distribution can be automatically estimated from data and was tested on AMI spoke data. Dominance relations between meeting participants has also been inferred. Using SVMs we were able to predict who is more, less or normally dominant in a meeting with an accuracy of 75%.

Video content extraction: We have developed "automatic camera operator" algorithms based on extracted video and audio features to perform this operation. Subjective evaluation with users indicated that the deployed algorithms were functionally acceptable, but were of significantly lower aesthetic quality compared with human production. We have also developed methods for identifying "hot-spots" such as laughter, directly from video features based on things such as motion and texture.

5. AMI Meeting Browsers.

Many AMI technologies are demonstrated within a Java-based browsing framework, referred to as JFerret. As illustrated in Figure 2, JFerret is a multimedia browser that is extremely flexible, enabling almost any user interface to be composed, using a combination of plug-in modules. An XML configuration specifies which plug-in components to use, how to arrange them visually, and how they communicate with each other. JFerret comes with a library of pre-defined plugins, for presentation of video, audio, slides, annotation time-lines, controls, and so on, and it is straightforward to write new plugins. This has been the main route to demonstration for many of the technologies described in previous sections. Java allows the application to run cross-platform, either as a applet (inside a web-browser) or as a stand-alone application.

An example JFerret configuration, enables browsing via keyword search on the speech-recognized transcript, search within captured slides, and browsing by speaker activity. Time-synchronized recordings that may be browsed include multiple video and audio streams and white board

capture. Other semantically rich browser components that have been constructed include direct keyword-spotting, video hot spots, and argumentation.

We have also begun to explore techniques for time-based media compression, since this can clearly contribute to efficient browsing of recorded meetings. Time-based compression can be done in three major ways: 1) speech speedup, 2) excision of less important parts, and 3) simultaneous presentation of speech from two locations. Two interactive prototypes for accelerated listening of recorded speech have been implemented. One prototype provides support for speed controls as well as skipping ahead and back based on speaker segmentations. The other prototype presents two parts of the meeting simultaneously using binaural in two different locations so that the user can listen to one part of the meeting while monitoring another part. We also devised a PDA-based wireless presentation system, including recording of slide presentations, which was integrated with the meeting browser using VNC.

6. EVALUATION

In AMI, we have adopted different levels of evaluation: component evaluation and system evaluation. All of our evaluation work is well supported by the AMI corpus and its annotations. AMI scientists have been closely involved in several international evaluation efforts such as the NIST Meeting Recognition evaluation of speech recognition and speaker diarization in meetings, for which the AMI corpus has been one of the main data sources. AMI has also participated in the CLEAR evaluations of focus of attention and face detection. Additionally, the AMI corpus, together with speech recognition output, has been provided to the Cross Language Evaluation Forum (CLEF) for their 2007 evaluation on cross-lingual question answering.

Content extraction tasks, such as summarization or topic segmentation, are somewhat artificial as a stand-alone task, and are often carried out within some other context (such as browsing). In such cases, extrinsic evaluation approaches may be preferred, in which a task is evaluated in the context of a larger scenario. In AMI we have developed a framework for extrinsic evaluation of browser components, that we call the Browser Evaluation Test (BET). The BET provides a framework for the comparison of arbitrary meeting browser setups, where setups differ in terms of which content extraction or abstraction components are employed. The BET consists of a set of experiments in which test subjects have to answer true/false questions about observations of interest for a meeting recording. The test subject uses the browser under test to answer these questions, given a time limit (typically half the meeting length). This framework has proven to be a successful way to evaluate browser components.

We have also developed a task-based evaluation that is supported by the design of the AMI corpus (about 70% of corpus meetings are based on a replicable design team scenario). In the task-based evaluation, a new team takes over for the fourth meeting, with access to the previous three meetings. The evaluation compares team performance in the existing case with basic meeting records (including powerpoint files, emails and minutes), with a basic AMI meeting browser, and with a task-based browser. The task-based evaluation is in terms of both objective measures such as design quality, meeting duration, assessment of outcome, and behaviourial measures of leadership, and subjective measures including browser usability, workload (mental effort), and group process.

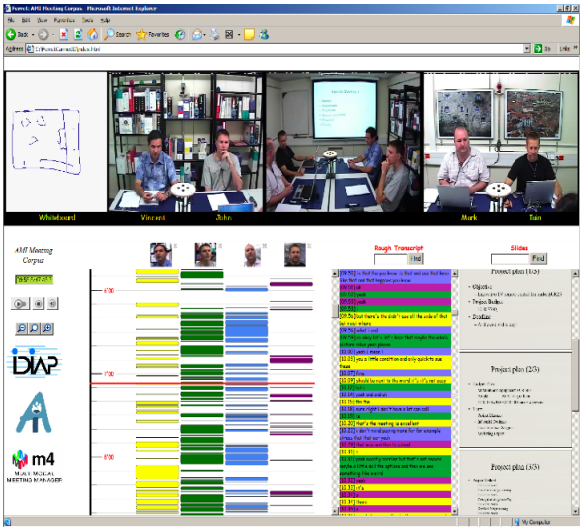


Figure 2: A typical AMI browser, integrating many of the AMI technologies, including speaker diarization ("who spoke when?), automatic speech recognition (speaker-dependent color-coded transcript), and several audio-visual processing techniques. The Jferret framework allows customizable construction of browsers.

7. CONCLUSIONS AND FUTURE WORK

We have provided an overview of the AMI project. The major achievements of AMI are in six areas:

Instrumented meeting rooms (development of a recording infrastructure, based on instrumentation of meeting rooms, in which we can capture all aspects of interaction in a meeting, in a time synchronized manner), **the AMI Corpus** (a 100 hour corpus of recorded meetings,

with multiple time synchronized signals across several modalities, annotated at many different levels), **audio-video processing** (significant advances in several areas including speech recognition, audio-video localization and tracking, and detection of focus of attention), **content extraction** (new state-of-the-art techniques in several areas such as summarization and dialogue act recognition), **integrated demonstrations** (AMI has developed an integrated browsing framework in which the outputs of multimodal recognition and content extraction modules may be incorporated as plugins or data streams), and **evaluation** (novel frameworks for system evaluation).

For each of the areas described there are many ongoing improvements and plans for future work. In general, improving robustness, speed, and accuracy are important issues, as well as scaling the techniques to deal with larger amounts of data. Within the new AMIDA project (see also [9]) we are working on improving many of the techniques, paying particular attention to their integration into a framework of "meeting assistants" that can perform in close-to real-time (i.e., in some cases delays of several seconds or even minutes may be acceptable). In AMIDA we are interested in building applications that integrate these techniques for use during, and between meetings, in remote and co-located settings.

REFERENCES

[1] M. Chen, "Achieving Effective Floor Control with a Low-Bandwidth Gesture-Sensitive Videoconferencing System," Proc. *ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[2] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, "Room with a Rear View: Meeting Capture in a Multimedia Conference Room," *IEEE Multimedia*, Vol. 7, No. 4, pp. 48-54., Oct-Dec 2000.

[3] J. Connell, A.W. Senior, A. Hampapur, Y-L Tian, L. Brown, and S. Pankanti, "Detection and Tracking in the IBM PeopleVision System," *IEEE ICME 2004*, June 2004.

[4] C. Costa, P. Antunes, and J. Dias, "A Model for Organizational Integration of Meeting Outcomes," in *Contemporary Trends in Systems Development*, M.K. Sein, B.-E. Munk-vold, T.U. Ørvik, W. Wojtkowski, W.G. Wojtkowski, J. Zupancic, and S. Wrycza, Eds. Kluwer Plenum, 2001.

[5] R. Cutler, et. al., "Distributed Meetings: A Meeting Capture and Broadcasting System," Proc. *ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[6] R. Dhillon, S. Bhagat, H. Carvey, and E. Shriberg, "Meeting Recorder Project: Dialog Act Labeling Guide," *ICSI Technical Report TR-04-002*, 2004.

[7] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting Group Interest-Level in Meetings," *IDIAP Research Report 04-51*, September 2004.

[8] W. Geyer, H. Ritcher, and G. Abowd, "Making Multimedia Meeting Records More Meaningful," *IEEE ICME 2003*, Baltimore, MD, July 2003.

[9] http://www.amiproject.org/

[10] http://corpus.amiproject.org.

[11] http://sourceforge.net/projects/nite/

[12] http://www.is.cs.cmu.edu/meeting room/.

[13] http://www.m4project.org/.

[14] http://www.nist.gov/speech/test beds/mr proj/.

[15] http://www.icsi.berkeley.edu/Speech/mr/.

[16] N. Kern, B. Schiele, H. Junker, P. Lukowicz, G. Tröster, "Wearable Sensing to Annotate Meeting Recordings," In *Personal and Ubiquitous Computing: Selected papers from the ISWC2002 Conference*, 2003.

[17] A. Hakeem, M. Shah, "Ontology and Taxonomy Collaborated Framework for Meeting Classification", in proc. *ICPR 2004*, Cambridge, UK, August 2004.

[18] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection Of Agreement vs. Disagreement In Meetings: Training With Unlabeled Data," Proc. *HLT-NAACL Conference*, Edmonton, Canada, May 2003

[19] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata, "Memory Cues for Meeting Video Retrieval," *1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, New York, NY, USA, October 2004.

[20] R. Jain, P. Kim, and Z. Li, "Experiential Meeting System," in *ACM Multimedia Workshop in Experiential Telepresence (ETP 2003)*, Berkeley, CA, Nov. 2003.

[21] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, B. Wrede. "The ICSI Meeting Project: Resources and Research," *NIST ICASSP 2004 Meeting Recognition Workshop*, Montreal, May 2004.

[22] L. Kennedy and D. Ellis, "Laughter Detection in Meetings," *NIST Meeting Recognition Workshop at ICASSP 2004*, Montreal, May 2004.

[23] L. Kennedy and D. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Automatic Speech Recognition and Understanding Workhop IEEE ASRU 2003*, St. Thomas, December 2003.

[24] H. Koike, S. Nagashima, Y. Nakanishi and Y. Sato, "EnhancedTable: Supporting a Small Meeting in Ubiquitous and Augmented Environment," in *proc. PCM 2004*, Tokyo, Japan.

[25] D.-S. Lee, B. Erol, J. Graham, H.J. Hull, and N. Murata, "Portable Meeting Recorder," in *proc. ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[26] Q. Liu, D. Kimber, J. Foote, L. Wylcox, and J. Boreczky, "FLYSPEC: A Multi-User Video Camera System with Hybrid Human and Automatic Control", in *proc. ACM Multimedia 2002*, Juan Les Pines, France, 2002.

[27] S. Marchand-Maillet, "Meeting Record Modelling for Enhanced Browsing," *Technical Report Computer Vision and Multimedia Laboratory*, Computing Centre, University of Geneva, March, 2003.

[28] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard. Modeling human interaction in meetings. In *IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP*, 2003.

[29] D. McNeil, *Hand and Mind: What Gestures Reveal About Thought.* U. of Chicago Press, 1995.

[30] I. Mikic, K. Huang, and M. Trivedi, "Activity Monitoring and Summarization for an Intelligent Meeting Room," in *proc. IEEE Workshop on Human Motion*, Austin, Texas, Dec. 2000.

[31] S. Mukhopadhyay, and B. Smith, "Passive Capture and Structuring of Lectures," in *proc. ACM Multimedia '99*, Orlando, FL, 1999.

[32] R.Ogata, Y.Nakamura, Y.Ohta, "Computational Video Editing Model based on Optimization with Constraint-Satisfaction," *proc. Fourth Pacific-Rim Conference on Multimedia*, 2003.

[33] M.Ozeki, Y.Nakamura, Y.Ohta, "Automated Camerawork For Capturing Desktop Presentations – Camerawork Design And Evaluation In Virtual And Real Scenes," *Proc. 1st European Conference on Visual Media Production*, 2004.

[34] F. Quek, D. McNeill, R. Bryll, C. Kirbas H. Arslan, K.E. McCullough, N. Furuyama, R. Ansari, "Gesture, Speech, and Gaze Cues for Discourse Segmentation," *in proc. of CVPR 2000*, Hilton Head Island, South Carolina, USA, 2000.

[35] Y. Rui, A. Gupta and J.J. Cadiz, "Viewing Meetings Captured by an Omni-Directional Carema," in *Proc. of ACM CHI 2001*, Seattle, WA, March, 2001.

[36] Y. Rui, A. Gupta, J. Grudin and L. He, "Automating lecture capture and broadcast: technology and videography," *ACM Multimedia Systems Journal*, 3-15, Springer V., 2004.

[37] X. Sun, and B.S. Manjunath, "Panoramic Capturing and Recognition of Human Activity," in *proc. of IEEE ICIP 2002*, Rochester, NY, USA, September 2002.

[38] http://www.quindi.com/

[39] M. Trivedi, I. Mikic, S. Bhonsle: "Active Camera Networks and Semantic Event Databases for Intelligent Environents", *IEEE Wsp on Human Modeling, Analysis and Synthesis (in conj. with CVPR)*, Hilton Head, South Carolina, June 2000

[40] Uchihashi S, "Improvising Camera Control for Capturing Meeting Activities using a Floor Plan," in *proceedings of ACM Multimedia 2001*, pp. 12-18, 2001.

[41] A. Waibel, et. al., "SmaRT: The Smart Meeting Room Task at ISL," in *IEEE ICASSP 2003*.

[42] A. Waibel et. Al., "Advances in Automatic Meeting Recording and Access," in *ICASSP 2001*, SLC, UT, 2001.

[43] B. Wrede and E. Shriberg, "Spotting Hot Spots in Meetings: Human Judgements and Prosodic Cues," in *EUROSPEECH 2003*, Geneva, September 2003.

[44] Zobl, M, Wallhoff, F and Rigoll, G, "Action Recognition in Meeting Scenarios Using Global Motion Features." Proc. *IEEE Intl. Wkshp on Perf. Eval. of Tracking and Surveillance (PETS-CCVS) Graz, Austria, March* 2003.

Language Technology in Tomorrow's Search Applications

Hans Uszkoreit
DFKI, German Research Center for Artificial Intelligence
Saarland University

As texts constitute the fabric of the web, advanced web search is one of the most exciting and challenging application areas for language technology.
But experts enthusiastically disagree on the exact role of language processing in advanced search and they differ even more in their predictions on the maturity of the proposed solutions. In our panel discussion, several experts working with relevant players in the development of tomorrow's search applications will explain their positions.
Among the topics to be addressed in the discussion are the following questions: How much can search technology be improved by language technologies in the near future? When will web-wide "natural language search" become reality?
How much language technology is really needed for tomorrow's search applications?
Which technologies are needed most? How are they to be combined with machine learning, multimedia processing and semantic technologies?

Advanced speech and language technology for complex customer care automation and self-service

Roberto Pieraccini
SpeechCycle

The past decade has witnessed the evolution of spoken human-machine communication from the first research prototypes to commercial high volume applications. Since the early exploitation of this technology in the mid 1990s, we can identify three distinct generations of systems of different complexity, scope, and architecture. Today, the third generation of spoken dialog systems includes the most challenging applications in the area of problem solving: self-service troubleshooting is one of them. In order to interact with users using voice and provide them effective help for the resolution of technical problems, a troubleshooting automated agent needs to have a deep knowledge of the type of systems and devices for which it is designed, and of the linguistic variety of expressions that describe the user perception of problems and observations. Moreover, any advanced troubleshooting system needs to be integrated with the available diagnostic tools and user knowledge databases in order to automate the problem solving problem with minimal intervention from the user's side. In this talk I will give a short overview of the history of spoken dialog systems, after which I will concentrate on today's most sophisticated troubleshooting automated agents. I will show how a good combination of human-computer interaction and voice user interface, knowledge representation, and integration with external devices and back-ends, can lead to successful automation of technical support which, eventually, drives substantial customer care cost reduction and improves customer experience.

What makes a successful speech enabled call routing application?

Diana M. Binnenpoorte, Dorota J. Iskra

Customer Contact Solutions, LogicaCMG, the Netherlands
{diana.binnenpoorte,dorota.iskra}@logicacmg.com

Abstract

The key function of a speech-enabled call routing application is to connect customers who call a contact centre with the service that they want. In principle, a speech-based call routing application has the same functionality as existing and well-known touch-tone based routing applications. A touch-tone application allows the calling customer to select a service from a predefined set by pressing keys, i.e. by making choices in a multi-layered structured menu. A speech-based call routing application allows customers to formulate their question by using natural speech.
The success of a call routing application can be expressed in various ways. First, there is the technical performance; for instance, the percentage of calls that are correctly routed, but also measures that express the impact of incorrectly routed calls on a call centre organisation. Second, there is the level of customer experience; to what extent do the calling customers appreciate the speech enabled routing application? And third, there is the level of involvement within the service-providing organisation; do call centre agents understand and use the advantages of speech enabled routing in their contact with customers?
For each of these measures there are important factors to pay attention to during the design, build and implementation of a speech enabled call routing application.

1. Introduction

1.1. Why speech enabled routing?

"How may I help you?" – this question that customers hear when calling their bank, insurance company, or another service provider is invariably preceded by a number of choices they have to make in an automatic structured dialogue menu (Interactive Voice Reponse, IVR). The aim of this, always one-sided dialogue, where the computer speaks and the customers press the keys on the keypad of their telephone, is to route the customers to an appropriate agent group in the call center that will be capable of answering their question.
The ubiquitous touch-tone IVR menus are typically structured in a way remote from the customer's intuition, as a result of which a large percentage of calls end in time-outs, wrong menu choices or transfer to the operator. Complicated multi-layered touch-tone IVR menus, which are a nuisance from the customer's point of view [1], are indispensable from the point of view of an organisation. For the sake of transparency most service organisations limit the number of contact numbers. As a result various issues can be resolved under a single telephone number by various agent groups. Agent groups in a call center specialise in a restricted area of topics and are frequently physically situated in different call centre locations. Therefore, call routing is necessary.
Speech technology offers a solution satisfying both the service provider and the customer. It facilitates call routing, but in a far more customer-friendly manner. Instead of having to choose from a limited set of menu options the customers are free to speak their question and formulate it in a natural way.

1.2. How does a speech enabled call routing application work?

A speech enabled routing application actually consists of several sub-systems: a. a dialogue manager, b. an automatic speech recognizer (ASR) and c. a classification algorithm.
The dialogue manager regulates the interaction with the customer, and, in an open speech enabled routing application, typically starts by stating the question "How may I help you?" [2]. This open question evokes the customer to utter the intention of his call (see Figure 1).

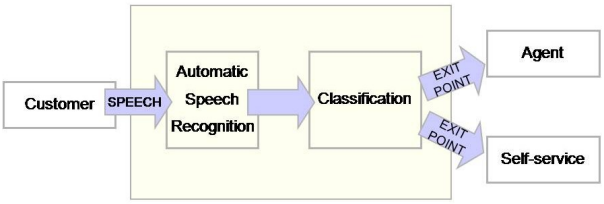


Figure 1 *Schematic diagram of speech enabled call routing*

The ASR system then recognises the spoken words using a lexicon and a grammar. The string of words, including recognition confidence scores, is subsequently given to the classification algorithm. The purpose of this algorithm is to determine the meaning of the string of words. This is accomplished by comparing the incoming word sequence with a database that is filled with word sequences that are already tagged, the so-called training material. A tag in this context is a category, or exit point (or routing goal), indicating a group of agents with specific skills (i.e. who are able to answer certain types of questions) in a call center. The classification algorithm yields one or more categories, and a classification confidence score indicating the degree of certainty for those categories. Dependent on the outcome of the classification algorithm, the dialogue manager determines the next step in the interaction with the customer: either setting up a connection to a human agent that is associated with the solely found exit point, or asking the customer for more (or different) information to find an exit point with high enough certainty. It may be clear that the more exit points, or categories, are defined, the more complex the calculations of the classification algorithm will be. Next to this, the lexicon, the grammar, the inventory of tagged sentences, and the dialogue strategies influence the performance of the application.
In short, a speech enabled routing application allows customers to just say the reason for their call without any restrictions in their choice of words. This can be accomplished by deploying speech technology which allows an open dialogue structure, as opposed to the very restricted form-filling type of dialogue. Since the customer is allowed

to phrase his intention in a natural way, it is no longer the customer who has to make a choice in a structured multi-layered menu by pressing keys, but it is the speech application that has to interpret the customer's intention, and subsequently, has to route him to the right agent or service.

2. Success factors

The performance of dialogue systems is commonly measured as the percentage of calls where the task has successfully been completed (in our case: where the call has correctly been routed), also known as Dialogue Success Rate (DSR) [3,4,5]. Besides the DSR, recognition accuracy expressed as Word Error Rate (WER) is also often referred to as a performance measure [6]. Such measures indicate the success of the application from a technical point of view. A technically successful application is, however, not necessarily one that is widely used and appreciated by a large group of customers [7]. Neither does it imply that the service providing organisation makes the most use of what the application has to offer. In this paper we also want to address other, non-technical, factors that determine the success of a speech enabled routing application, i.e. customer experience (in the broadest sense) and level of involvement within a service providing organisation. These factors, which often represent conflicting interests, play an important role in the different stages of a project that is aimed at the deployment of a speech application. In the remaining of this paper we will list the key success factors and show the way they are intertwined according to the different phases, i.e. design, build, and deployment.

3. Design

During the design of a speech enabled routing application, various issues play an important role. First, one has to establish the strict functionalities of the application; what it should do, given a list of preconditions, from both technical and organisational sources. We will not further outline the functionalities here. Second, the dialogue of the application is designed; how the application should interact with the end user, given preconditions from the supposed end user's point of view as well as the organisation's point of view. The third issue concerns determining the inventory of exit points, or categories. This issue will be presented separately from the dialogue design issue, since it plays an important role in all success factors.
It can be stated that in the design phase the foundation is laid for the following success factors: technical performance, customer experience and involvement of the organisation.

3.1. Dialogue design

When it comes to dialogue design, various, sometimes contradicting interests have to be taken into account:

Technical performance: Optimal technical performance is reached when the dialogue is continued until the application is entirely certain about the chosen exit point. In case of ambiguities, a new dialogue step is introduced according to the following conditions:

- When a single exit point is chosen, but the confidence value is low, or when the confidence is high, but the exit point is a critical self-service, a confirmation question is posed. When answered positively, the customer is transferred to an agent or a self-service. In case of a negative answer, the dialogue has to go back to a general question.
- When two or three exit points are chosen with similar confidence levels, a menu consisting of these exit points is asked and the customer is asked to choose one of the given options.
- When the confidence level of any of the exit points does not exceed the minimum threshold, the customer is asked to reformulate his request. As an extra help, an example sentence may be given at this stage.

Customer experience: A long dialogue seldom contributes to positive customer experience. In all usability tests we have conducted it has been confirmed that the customer wants to speak to a live agent and wants to reduce the interaction with a computer system to a minimum. The customer experience may remain positive with multiple dialogue steps when the customer has a feeling that he is progressing in the right direction. Especially annoying are, however, recognition and classification errors when a confirmation or a menu question containing only wrong exit points is asked. From the point of view of customer satisfaction the dialogue length may not exceed a certain maximum defined in terms of dialogue steps or wrong guesses on the side of the application.

Organisational issues: The service providing organisation usually also has its particular interests that need to be attended to at the design stage. An example might be the pricing model for the service: if the telephone costs are paid by the organisation, it is in the interest of the organisation to keep the dialogue short and force a transfer even in case of ambiguities. Alternatively, some exit points may be more costly than others. Exit point A receives 50% of all the calls whereas exit point B only 2%. If X calls are wrongly transferred to exit point A, the impact on this particular agent group may still be negligible. If the same number of calls are wrongly transferred to exit point B, the impact on this agent group may be severe and cause substantial delays and congestion in the call centre. In order to resolve this, the organisation may, for instance, require an additional confirmation question when exit point B is chosen.
The dialogue design is a trade-off between all these often conflicting aspects of the service.

3.2. Selection of exit points

When a customer calls the speech enabled routing application, his question is first recognised and then categorized, i.e. the meaning or intention of the call is determined. Once it is clear to which category the call belongs, the customer is connected to the right agent or service that can actually answer the question. Establishing the categories, or exit points is not a straightforward task.
Often, the way call centers of large organisations are structured in terms of departments, services, agent skills, and locations, is the result of an organic process of growth, change of needs, economical aspects, and so on. Still, from an organisation's point of view the structure within a call center seems as well-founded as it is. And often this structure can be traced back from menu options in a touch-tone IVR application. And this is exactly what customers complain about: a menu that is hard to navigate through because of complex terms or categories [1]. It can be stated that there is a mismatch between the customer's mental model of the structure in an organisation and the actual operational structure. For instance, a customer is convinced that a question about his income insurance can be answered by the department of life insurances, while for the organisation it is absolutely clear the question should be answered by the department of indemnity insurances.

In a speech enabled routing application the problem for the customer to 'file' his question under one of the given menu options such as in a touch-tone IVR is solved: the application determines to which category the question belongs. However, feedback to the customer can sometimes be necessary by means of a confirmation question. In this type of questions, the name of a found category is often mentioned ('Is it true that you are calling for X?'). Therefore, the names of the categories must be chosen with great care as to fit in as well as possible with the customer's idea of an organisation.
Determining the names of the categories is just one side of the medal, determining the number and type of categories is another. When making the inventory of exit points, the organisation of the call center, how agents are grouped based on certain skills, is leading. Before determining the type and number of exit points the organisation needs to carefully examine what type of questions can be answered by which skilled agent. Even within one subject range it is sometimes necessary to define more than one category since agent skills can be differentiated in, for instance, answering more complex questions versus more common questions.
So on the one side, the organisation requires a rather detailed division, while from the technical point of view it holds that the more exit points are defined, the harder it is to find the right one. There is no golden standard regarding the optimal number of exit points, [8] reports on 40 exit points, while [6] reports on only three.
The number and the inventory of exit points influence both customer experience as well as the technical performance.

4. Build

The quality and size of sentence material which is used to train the grammar is of major importance for the technical performance of the system. In terms of size, the rule "the more the better" applies only to a certain extent. At a given point, which must be established experimentally for each application, a saturation point is reached [9]. Above that point, which indicates that an adequate representation of customer questions has been reached, adding more sentences does not contribute to better performance, but has no or even a slightly negative effect. By adding more sentences there is also a risk of overlapping sentences between exit points which makes them harder to be uniquely identified when similar questions are asked.
In this context *quality* refers to the degree in which training material reflects customer questions with regard to the following aspects:

- *Terminology*: do sentences contain terminology that is used by customers rather than the one internally used by the service providing organisation?
- *Formulation*: are sentences formulated in terms of style and syntax the way customers would formulate them? In practice many spoken sentences contain hesitations, repetitions, etc.
- *Variance*: do training sentences cover the wide range of different subjects (and thus exit points) and speaking styles that are used by customers?

Training material can be collected in several ways. The below list of these approaches is ordered according to the quality with the best quality at the top:

Live sentence material: the best technical performance can be achieved if the training material is collected through a dummy application. The dialogue of such a dummy application ideally reflects the design of the future speech enabled routing application so that customers behave similarly to a large degree. The difference is in the handling of the question. The customer sentences are only recorded for the purpose of training and the customers are transferred to broadly skilled operators who can answer their questions or transfer them to another agent.

Simulated live material can be collected in two ways. Firstly, the customers who have just spoken to an agent can be asked to say their question one more time to an especially for this purpose designed application. This approach results in a wide range of sentences, but they are less spontaneous in terms of formulation since the customer has been warned beforehand. Also, not all questions are transferred to agents and a large proportion is handled by self-services. These sentences will be missing from the collection.
Alternatively, customers or employees of the service providing organisation can be requested to make calls to a special application whose sole purpose is recording training sentences. However, the response from customers in this kind of actions is usually very low. The disadvantage of using employees, even agents, is that their way of formulating questions can only be an approximation of what the customers say. The success of this approach is very dependent on the acting skills of the participants involved. In terms of planning it can coincide with the trial period preceding the development of the actual application which is used to test the dialogue design, collect sentences and give an indication of the expected performance.

Written sentence material: if it is impossible to collect spoken sentences, the second best is text. Written sentences can, for instance, be derived from a website application if the latter allows for natural language input. This kind of sentence material, although containing the vocabulary used by customers, costs a lot of time and effort to be made ready for training purposes. Internet sentences contain numerous spelling mistakes, spelling variations, textual signs such as, e.g. "+", and abbreviations. All of these have to be repaired before the sentences can be used for training. Moreover, it has to be kept in mind that the way of formulating questions in a website application is very different from speech. In writing, customers tend to use short expressions, often one or two words, and restrict their enquiries to keywords.

Simulated written sentence material: if it is impossible to collect spoken sentences and the service providing organisation does not have a website allowing for natural language input, a sentence collection can be organized where the employees of a service providing organisation are asked to make up sentences on a number of given subjects. This approach is similar to *simulated live material*, except that the sentences are provided as text.

5. Deployment

5.1. Training agents

Once the application is built, tested and put in production, customers will be routed to agents or services. Assume that the customer has perfectly been helped by the application and is directly routed to the right agent. What if that human agent starts the conversation with the question "How may I help you?". The customer will at least have the feeling of having to start all over again. It is important that the agent already knows what the call is about before the actual conversation starts, at least in more detail than on department or skill level. When possible, it is desirable to display the recognised sentence on the agent's telephony desktop. The agent then first reads the sentence before

starting the conversation with the customer. In this way, the conversation is already one step ahead, giving the customer a higher appreciation of his interaction with the organisation.
However, as open speech recognition, as ASR in general, does not perform 100%, the recognised sentence on the agent's desktop deviates from what has actually been spoken by the customer. The degree of deviation determines the usefulness of the displayed sentence. When the sentence contains many recognition errors, the agent can best start the conversation by apologising and trying to retrieve the customer's question himself. The difficulty is, that it is not clear to the agents whether the spoken customer input is recognised very poorly or not. The agent has always to try to interpret the recognised word sequence and decide whether it gives enough clues to start the conversation directly.
The new conversation technique requires some training and change in the way of work of the agents. Once both interest and support from the agents towards the application is fostered, agents will not only appreciate the advantages of a speech enabled routing application, but are also inclined to act as ambassadors towards the customers, making the application successful.

5.2. Management information

The speech enabled routing application can be a valuable source of management information for the service providing organisation. With an appropriate logging and reporting mechanism the application can provide various statistics about, for instance, the number of calls, their distribution with regard to the time of day, or classified exit points.
In a more sophisticated set-up it can also measure, for instance, the effectiveness of an advertising campaign. If a new product is advertised, the service provider can calculate the number of questions posed about this product to the application. If a separate exit point is designed for this product, the calculation is based on the number of calls classified to this exit point. However, if the product is accommodated within an existing exit point, the calculation can be based on the extraction of the words related to that product from the recognition results.
The extent to which the speech enabled routing application is used to provide management information is dependent on the level of involvement of the service providing organisation. The higher this level, the more justified the deployment of a call routing application not only as a means of routing calls, but also as a source of valuable customer data.

6. Conclusions

In this paper we have shown that besides the commonly evaluated technical performance of the speech enabled routing application, a number of other aspects play an important role in determining the final success of this application. These aspects relate to customer experience and the level of involvement of the service providing organisation. We have described a number of issues during the different phases of the project such as dialogue design, selection of the exit points, collection of sentence material, and so forth, where our success factors intertwine and occasionally even represent conflicting interests.

7. References

[1] Suhm, B., and Peterson, P., "A Data-Driven Methodology for Evaluating and Optimizing Call Center IVRs", International Journal of Speech Technology 5, 2002, p 23-37.

[2] Gorin, A.L., Riccardi, G. and Wright, J.H., "How may I help you?", Speech Communication 23 (1997), p 113-127.

[3] Rahim, M., Pieraccini, R., Eckert, W., Levin, E., Di Fabbrizio, G., Riccardi, G., Kamm, C., Narayanan, S., "A Spoken Dialogue System for Conference/Workshop Services", Proceedings of ICSLP 2000.

[4] Sturm, J., den Os, E. and Boves, L. "Dialogue Management in the Dutch A Train Timetable Information System", Proceedings of Eurospeech 1999.

[5] Natarajan, P., Prasad, R., Suhm, B. and McCarthy, D., "Speech-Enabled Natural Language Call Routing: BBN Call Director", Proceedings of ICSLP 2002.

[6] Hessen van, A. and Hinke, J. "IR-based classification of customer-agent phone calls", Proceedings of Interspeech 2005, Lisbon, p 597-600.

[7] Boves, L. and den Os, E., "Applications of Speech Technology: Designing for Usability", Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding 1999.

[8] Kuo, H-K. J. and Goel, V. "A Data Visualisation and Analysis Method for Natural Language Call Routing Systems Design", Proceedings of Interspeech 2007, Antwerp, p 2729-2732.

[9] Schohn, G. and Cohn, D. "Less is more: Active learning with support vector machines", Proceedings of the ICML 2000.

SME Elevator session

Bente Maegaard
CST, University of Copenhagen

Small and Medium-Sized Enterprises (SME) are one of the most important types of entities for creating new ideas and products, not least in the area of information and communication technologies, including language technology. Over the years many small companies have developed fast due to venture capital or due to strategic partnerships. At Langtech you will get the chance to listen to some of the upcoming SMEs and to hear about their newest ideas on products and services. This year, the presentations will in particular feature solutions which involve speech and multimedia, for learning and for services.

Language Technologies and the Semantic Web: An Essential Relationship

Enrico Motta
Knowledge Media Institute
The Open University

A large scale Semantic Web, CST, University of Copenhagen by thousands of ontologies and millions of OWL/RDF documents, is taking shape and is opening the way to a new generation of knowledge-based applications. In this talk I will present some work we are carrying out, which aims to develop concrete end-user applications, which exploit this unprecedented resource. In particular, I will show how large scale semantics can be used to try and enhance typical web-centric activities, such as web browsing, searching for information, and interacting with typical web 2.0 sites. I will also emphasize a number of differences between these new applications enabled by the Semantic Web and "traditional" knowledge-based systems. In particular, while the latter usually rely on closed, high-quality and homogeneous domain models, the Semantic Web is a very large and heterogeneous collection of logical statements exhibiting diverse provenance, diverse conceptualizations, and varying degrees of quality. In contrast with earlier-generation knowledge-based applications, making sense of such heterogeneity often requires not just the application of logical inference methods, but the combination of language, machine learning, and knowledge representation technologies in novel application scenarios. Finally, I will also situate these developments in the wider context of Artificial Intelligence (AI) research and will argue that the Semantic Web provides an exciting opportunity and a new context in which new solutions to classic AI problems can be devised.

Answering Questions from the Semantic Web

Christopher Welty
IBM Research

Researchers and practitioners in NLP often misunderstand the Semantic Web vision, because (understandably) they project themselves in the center of that universe. The semantic web vision is to make explicit the semantics of the back-end databases, which are already structured, from which roughly 80% of the web's HTML pages are generated. It is a simple and reasonable vision that does not require NLP at all. However, language technologies can benefit from the semantic web in a number of ways, and judging by the last International Semantic Web Conference (ISWC), NLP is one of the most successful research and application areas. Most obviously Semantic Web technologies like RDF and OWL provide a standard interlingua for representing and exchanging the results of natural language processing systems. Among the less obvious advantages are that, as these structured sources become increasingly available, they can be used to help address certain problems in NLP that require large amounts of specific knowledge.
We are exploring the use of large, public domain repositories of explicit ground facts to improve the quality of natural language question answering. In this talk, I will discuss the goals and challenges of such an exploration.