AUTOMATIC DETECTION OF SEMANTIC BOUNDARIES BASED ON ACOUSTIC AND LEXICAL KNOWLEDGE

Mauro Cettolo, Daniele Falavigna

ITC-Irst, Istituto per la Ricerca Scientifica e Tecnologica
via Sommarive 18
38050 Povo, Trento - ITALY
{cettolo,falavi}@itc.it

ABSTRACT

In spoken language systems, the segmentation of utterances into coherent linguistic/semantic units is required when modules following the speech recognizer can only process such units one at a time. In this paper, techniques for semantic boundary prediction, based on both acoustic and lexical knowledge, are presented and tested on a corpus of person-to-person dialogues. Best result gives 62.8% recall and 71.8% precision.

1. INTRODUCTION

In spoken language systems, the minimal unit of analysis does not necessarily correspond to a full sentence. A possible approach for language processing is that of splitting a given sentence in a sequence of units that can be successively processed by linguistic modules one at a time. The goal of the Semantic Boundary (SB) detector is to locate boundaries inside a sentence in order to obtain such "minimal units".

Useful information for SB detection can be extracted either from the waveform of an utterance or from its corresponding word sequence. Some prosodic features, such as the energy contour, the speaking rate, and the fundamental frequency `F0`, can effectively help the boundary individuation [11, 15]. For example, the short-time energy tends to decrease within a segment and to reset at the beginning of the successive segment. Also the speaking rate is generally higher at the beginning of semantic units than at the end, and the `F0` curve shows different trends in presence of questions, exclamations, stressed words, etc. Although the use of prosody is very appealing, the literature shows that the correlation between prosodic features and syntactic-semantic boundaries exists, but it is not reliable enough to assure that prosody can be used alone in boundary detection; however, the presence and the length of filled pauses seem to have a significant correlation with semantic boundaries [12]. An overview on the use of prosody in speech-based systems can be found in [7].

For SB detection, the main source of information lies on the linguistic content of the sentence [10, 2, 14, 9, 4]. However, the detector can use only the recognizer output, in addition to the acoustic parameters, and then it cannot be based on linguistic information requiring sophisticated processing.

Word n-grams can effectively capture statistical relations between words and SBs [10]. In fact, once a word/SB n-gram model is estimated on a training set, simple dynamic programming algorithms can find the best segmentation or the k-best segmentations of a test/input sentence.

In this paper, a detailed description of some techniques aimed at extracting acoustic and lexical information for SB detection is given. Furthermore, a method for combining the two knowledge sources is proposed, together with results obtained on a corpus of hundreds of human-human spontaneous dialogues.

2. SEMANTIC BOUNDARIES

Collected corpus (Section 4) consists of a set of dialogues, each formed by a sequence of turns or utterances. Different communicative intentions, i.e. *speech acts* (SAs), can be enclosed in a turn. In [6], the set of SAs defined for the domain considered in this paper, the "appointment scheduling", is discussed. An example, translated from Italian, of a turn segmented and labeled with SAs follows:

```
[ good morning = greeting ] [ this is ferrari =
introduce-self ] [ i'd like to fix an appointment =
introduce-topic ]
```

Assuming as a working hypothesis that an utterance is a "flat" sequence of SAs, it is of primary interest to investigate the feasibility of automatically detecting boundaries between successive SAs. In fact, segmenting the speech recognizer output in terms of SAs allows the following linguistic modules to process each single SA at a time, reducing greatly the ambiguity.

Given the result of the recognition process, the module based on the techniques discussed in this paper should output a segmented text like:

```
good morning SB this is ferrari SB i'd like to fix
an appointment
```

3. KNOWLEDGE SOURCES

3.1. Lexical Information

Sentence texts of the corpus can be seen as sequences of triples:

$$\cdots w_{i-1}\ b_{i-1}\ n_{i-1}\ w_i\ b_i\ n_i\ w_{i+1}\ b_{i+1}\ n_{i+1} \cdots \qquad (1)$$

where

$$\begin{aligned} & w_j \in W \quad \text{(vocabulary)} \\ & b_j \in \{SB, \lambda\} \\ & n_j \in \{@fp, @hn\}^* \end{aligned}$$

that is, w_js are the non-noise words, b_js are either the symbol `SB` indicating the presence of a semantic boundary, or the empty string λ, and n_js are sequences, possibly empty, of two symbols indicating the presence of filled pauses (eh, ehm, mmm, ah...) or human noises (mainly related to breathing), respectively.

A trigram Language Model (LM) can be trained on such sequences. Given a test/input sentence of the form (1):

$$\vec{\mathbf{v}} = v_1 \cdots v_n \ , \quad v_i \in V = W \cup \{SB\} \cup \{@fp, @hn\}$$

its probability can be computed by using the trigram LM:

$$\begin{aligned} \Pr{}^{LM}(\vec{\mathbf{v}}) \quad & = \textstyle\prod_{i=1}^{n} \Pr(v_i \mid h_i) \\ & \approx \Pr(v_1)\Pr(v_2 \mid v_1) \textstyle\prod_{i=3}^{n} \Pr(v_i \mid v_{i-2} v_{i-1}) \end{aligned}$$

where the history (or context) h_i of v_i is limited to the two words preceding v_i.

Once a n-gram LM is estimated on a training set, the most likely segmentation of a test/input sentence can be found by scoring and sorting all its possible segmentations.

If a sentence consists of m words, all the possible segmentations are 2^{m-1}, an infeasible number for large m. Simple heuristics can be introduced to limit the number of segmentations to be scored, such as that of allowing an SB between words w_{i-1} and w_i only if the difference of the two probabilities $Pr(w_{i-n+1}\ w_{i-n+2} \cdots w_i)$ and $Pr(w_{i-n+1}\ w_{i-n+2} \cdots w_{i-1}\ SB)$ is less than a given threshold, or allowing only the $q < m$ boundaries corresponding to the q lower differences.

Once all the allowed segmentations are scored and ordered, the best one can be hypothesized. Another possibility is to rescore the k-best segmentations on the basis of another knowledge source, for example by using prosodic features.

The number k can be fixed a-priori, or made variable by considering the k segmentations whose scores differ from the best one of less than a certain percentage defined by a factor $\delta \in [0, 1]$.

3.2. Prosodic Information

Given a test/input sentence of the form (1), a vector $\vec{\theta}_i$ of prosodic features can be computed for each b_i. A label can be associated to the vector: `True` if b_i is `SB`, `False` if b_i is the empty string λ; in particular, the label `True` occurs when w_i is the last non-noise word of an SA and w_{i+1} is the first non-noise word of the successive SA; the label `False` occurs when w_i and w_{i+1} belong to the same SA, or when w_i is the last non-noise word of the sentence.

A Binary Classification Tree (BCT) [3] can be trained to recognize the presence of an SB on the basis of the feature vector $\vec{\theta}$. Given a sentence $\vec{\mathbf{v}}$ of the form (1) with its segmentation defined by the sequence of b_is, the BCT is asked to give the probability $\Pr(b_i \mid \vec{\theta}_i)$ for each b_i. The product of these probabilities over all b_is:

$$\mathtt{Pl}^{\mathtt{pros}}(\vec{\mathbf{v}}) = \prod_{i=1}^{m} \Pr(b_i \mid \vec{\theta}_i)$$

gives the "prosodic plausibility" of that particular segmentation.

Computed prosodic features are related to speaking rate, energy and `F0` contours. Their brief description follows.

Speaking Rate For each phone-unit **p** used in lexical transcriptions, its average duration μ_p is computed on the training speech data Ω:

$$\mu_p = \frac{1}{N_p} \sum_{p \in \Omega} D_p$$

where D_p is the actual duration of a particular realization of the phone-unit p and N_p is the number of its occurrences in Ω.

The average duration μ_w of each non-noise word w of the lexicon is computed as the sum of the μ_p of the phone-units in its transcription.

Information about the speaking rate can be obtained by comparing actual duration D of a word with its average duration. Three values are computed and included in the feature vector $\vec{\theta}_i$:

- $\frac{D_{w_i}}{\mu_{w_i}}$
- $\frac{D_{w_{i+1}}}{\mu_{w_{i+1}}} - \frac{D_{w_i}}{\mu_{w_i}}$
- $\frac{1}{2}\left(\frac{D_{w_{i+1}}}{\mu_{w_{i+1}}} + \frac{D_{w_{i+2}}}{\mu_{w_{i+2}}}\right) - \frac{1}{2}\left(\frac{D_{w_{i-1}}}{\mu_{w_{i-1}}} + \frac{D_{w_i}}{\mu_{w_i}}\right)$

Energy contour The short time energy is derived from the speech signal using an Hamming window of 20 ms, at a rate of 10 ms. Its contour is then low pass filtered and normalized with respect to the maximum value along the utterance. For each b_i, 25 features are derived from speech segments corresponding to the sequence of words $w_{i-2} \cdots w_{i+2}$. As an example, the minimum, average and maximum values in the three segments w_i, $w_{i-1}w_iw_{i+1}$ and $w_{i-2}\cdots w_{i+2}$ are calculated.

F0 contour The `F0` curve is extracted by means of a high resolution pitch determination algorithm [8] based on the evaluation of the cross-correlation coefficient between two adjacent speech segments[1]. In our implementation the algorithm produces a pitch value every 10 ms. Similarly to energy, 18 parameters are derived from the F0 curve.

3.3. LM and Prosody Integration

The integration of lexical and prosodic information was done by rescoring the k-best segmentations, hypothesized by the LM, with their prosodic plausibilities. In particular, the one giving the best score obtained with the weighted product of its LM probability ($\Pr^{LM}(\vec{\mathbf{v}}_j)$) and its prosodic plausibility ($\mathtt{Pl}^{\mathtt{pros}}(\vec{\mathbf{v}}_j)$) is chosen as follows:

$$\hat{\vec{\mathbf{v}}} = \mathop{\mathtt{argmax}}_{j=1\ldots k} \left(\Pr^{LM}(\vec{\mathbf{v}}_j)\right)^{\alpha} \times \left(\mathtt{Pl}^{\mathtt{pros}}(\vec{\mathbf{v}}_j)\right)^{\beta} \qquad (2)$$

4. CORPUS DESCRIPTION

Experiments were carried out on a dialogue corpus collected at ITC-Irst [1], composed of monolingual person-to-person Italian conversations for which acoustic signals, word transcriptions and linguistic annotations are available. The two speakers were asked to fix an appointment, observing the restrictions shown on two calendar pages they were given; they did not see each other and could hear the partner only through headphones. The conversations took place in an acoustically isolated room and were naturally uttered by the speakers, without any machine mediation.

The dialogues were transcribed by annotating all extra-linguistic phenomena such as mispronunciations, restarts and human noises, with the exception of pauses.

	Training	Test	Whole Corpus
# dialogue	169+12/2	20+12/2	201
# turn	2680	406	3086
# SA	5421	877	6298
# SB	2741	471	3212
size (non-noise)	27786	4683	32469
|W| (non-noise)	1291	627	1433

Table 1: Training and test set statistics.

The whole corpus was then divided into training and test sets (see Table 1), paying attention to avoid speaker overlap between the two sets. The test set consists of all the sentences uttered by 11 speakers, resulting in 20 complete dialogues and 12 half dialogues, for a total of 406 turns.

5. EXPERIMENTS

5.1. Results using LM

In order to make the number of semantic segmentation hypotheses manageable, only a maximum of $q = 14$ SBs (see Subsection 3.1) were allowed inside each sentence. This means that at the most $2^{14} = 16384$ different segmentations had to be scored for each test/input sentence.

In Figure 1, the probability of finding the correct segmentation of a sentence, within the k-best hypothesized segmentations, is given as a function of k. In the experiment, the average number k was determined by varying $\delta \in [0, 1]$ (see Subsection 3.1), as shown along the curve.

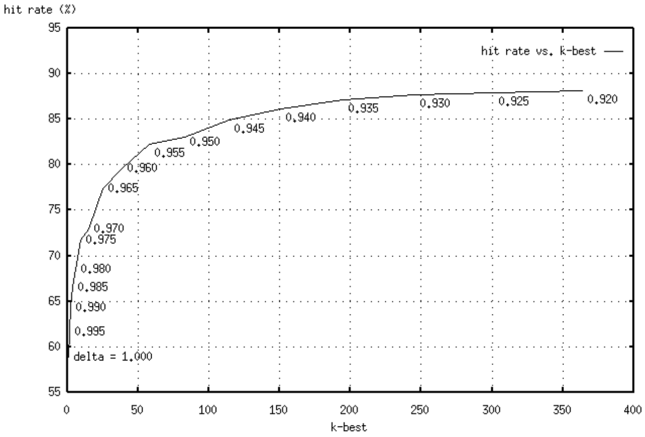

Figure 1: Probability that the actual sentence segmentation falls inside the k-best.

In Table 2 results are reported by aligning the *1*-best output against the hand labelled test data. Performance is given in terms of correct detection (C), insertions (I) and deletions (D) of SBs, and recall and precision measures.

type	C	I	D	Recall	Precision
LM	285	115	186	60.5%	71.3%

Table 2: SB detection results using the LM *1*-best output.

5.2. Results using Prosody

To check the relevance of the three types of prosodic features, three different BCTs were built: one for the 3 speaking rate features (`ros`), one for the 25 features related to the energy contour(`ene`), and one for the 18 features derived from the `F0` curve (`F0`). Finally, a general BCT was trained to handle all the 46 prosodic features considered (`all`).

In Table 3 results obtained on the test set, by aligning the outputs of the BCTs against the hand labelled test data, are reported.

[1] The software of the pitch determination algorithm was kindly provided by the Cambridge University Engineering Department.

type	#features	C	I	D	Recall	Precision
ros	3	141	639	330	29.9%	18.1%
ene	25	171	510	300	36.3%	25.1%
F0	18	133	622	338	28.2%	17.6%
all	46	211	520	260	44.8%	28.9%

Table 3: SB detection results using prosody.

5.3. Effects of Integration

The integration of LM and prosody was then applied as explained in Subsection 3.3. The average number k of segmentation hypotheses to be rescored was 5.4, derived setting δ to 0.980. Weights α and β were empirically chosen, and set to 0.8 and 1.0 respectively. Results are reported in Table 4.

type	C	I	D	Recall	Precision
LM$\oplus$prosody	296	116	175	62.8%	71.8%

Table 4: SB detection results using LM and prosody.

6. CONCLUSIONS

A method for automatic semantic boundary detection was presented. It integrates linguistic knowledge, in the form of a trigram language model, and prosodic knowledge, in the form of a BCT model.

Separate performance evaluations were done for each kind of model and for each type of prosodic feature. As expected, the best result is reached with the trigram LM that codes the linguistic knowledge in the form of word sequence probabilities.

Prosodic features give promising results too. Single performance shows that the short term energy curve provides the strongest contribution; also the speaking rate appears to be very helpful, even if in the reported experiments it was modeled only by three features. However, the best result is definitely obtained by training a BCT for all the prosodic features, showing that each one provides an own effective contribution to the classification.

The integration of LM and prosody provides a slight improvement of recall and precision in SB detection. More significant performance increases are expected by investigating different integration methods, such as that proposed in [13] where a tight combination of lexical and acoustic knowledges is implemented through an A^* algorithm.

More efficient representations of both the energy and F0 trajectories could be obtained by expanding them as a linear combination of some interpolating functions. The coefficients of the expansions can be used alternatively or in conjunction with the parameters adopted for this work.

Further efforts will be devoted to a feature selection step, that could suggest what parameters are more useful for SB detection. The selection will be done by inspecting the hierarchical structure of BCTs, and/or by applying a linear transformation to the features in order to maximize their class separability [5].

7. REFERENCES

1. B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, and G. Lazzari. Multilingual Person to Person Communication at IRST. In *ICASSP*, Munich, Germany, 1997.
2. A. Batliner, R. Kompe, A. Kiessling, H. Niemann, and E. Noeth. Syntactic-Prosodic Labeling of Large Spontaneous Speech Data-Bases. In *ICSLP*, Philadelphia, USA, 1996.
3. L. Breiman, J.H. Friedman, R.O. Olshen, and C.J. Stone. *Classification and Regression Trees.* Wadsworth, Pacific Grove, Cal., 1984.
4. M. Cettolo and A. Corazza. Automatic Detection of Semantic Boundaries. In *Eurospeech*, pages 919–922, Rhodes, Greece, 1997.
5. K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, London, 1988.
6. S. Jekat, A. Klein, E. Maier, I. Maleck, M. Mast, and J.J. Qantz. Dialog Acts in Verbmobil. Verbmobil Technical Report 65, Hamburg University, DFKI Saarbrucken, Erlangen University and TU Berlin, Germany, April 1995.
7. R. Kompe. *Prosody in Speech Understanding Systems.* Springer, 1997.
8. Y. Medan, E. Yair, and D. Chazan. Super Resolution Pitch Determination of Speech Signals. *IEEE Transactions on Signal Processing*, 39(1):40–48, 1991.
9. A. Stolcke. Modeling Linguistic Segment and Turn Boundaries for N-Best Rescoring of Spontaneous Speech. In *Eurospeech*, Rhodes, Greece, 1997.
10. A. Stolcke and E. Shriberg. Automatic Linguistic Segmentation of Conversational Speech. In *ICSLP*, Philadelphia, USA, 1996.
11. M. Swerts. Prosodic Features at Discourse Boundaries of Different Strength. *JASA*, 101(1):514–521, 1997.
12. M. Swerts, A. Wichmann, and R.-J. Beun. Filled Pauses as Markers of Discourse Structure. In *ICSLP*, Philadelphia, USA, 1996.
13. V. Warnke, R. Kompe, H. Niemann, and E. Noeth. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Eurospeech*, pages I:207–210, Rhodes, Greece, 1997.
14. S. Wermter and M. Loechel. Learning Dialog Act Processing. Verbmobil Technical Report 139, Hamburg University, Germany, July 1996.
15. C. W. Wightman and M. Ostendorf. Automatic Labeling of Prosodic Patterns. *IEEE Transactions on Speech and Audio Processing*, 2(4):469–481, 1994.