Gold sponsor

Loquendo
VOCAL TECHNOLOGY AND SERVICES

Silver sponsor

PerVoice

Bronze sponsor

elct

DICIT

QALL-ME

IBM

Media sponsor

Speech TECHNOLOGY

AVIOS

Thanks to

Rai

Audio Video direction
and recording by RAI

Simultaneous translation by
Intersteno - Accademia Aliprandi

Hi Pro
Competence for Evolution

RFID Technology by Hi Pro

LangTech 2008   Rome, 28-29 february 2008

LangTech 2008

Rome, 28-29 february 2008

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

forum fAD

eLDa

EUROPEAN LANGUAGE RESOURCES ASSOCIATION ELRA

Gold sponsor

Loquendo
VOCAL TECHNOLOGY AND SERVICES

LangTech 2008

Rome, 28-29 February 2008

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

forum AD

ELDA

EUROPEAN LANGUAGE RESOURCES ASSOCIATION
ELRA

Gold sponsor

Loquendo
VOCAL TECHNOLOGY AND SERVICES

Edited by Cristina Delogu, *Fondazione Ugo Bordoni*
Mauro Falcone, *Fondazione Ugo Bordoni*

Graphic design Stefania Vinci, *Fondazione Ugo Bordoni*

Index

CD-rom included

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*

SPEAKER SUMMARIES

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

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.

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.

Hypothesized user goals can be considered as preconditions for system actions. Methods and strategies for computing probabilities of hypothesized goals are discussed and some experimental results are presented.

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.

Understanding the Market Movements in Network Speech: Aligning Business and Technology

Daniel Hong
Datamonitor

Speech recognition was once viewed as a futuristic technology that would never leave the realm of science fiction. But over the past 50 years, key technology and commercial achievements in speech recognition along with increased CPU performance and lower hardware costs have helped make speech commercially viable for enterprises and service providers. Today, speech recognition is becoming increasingly prominent as a cost-cutting and value-enhancing solution for customer care and service enablement.

2008 marks an interesting time for speech recognition solutions in the global enterprise and service provider markets. Open-standards are permeating across the industry at a rapid rate, new deployment paradigms are surfacing, development and reporting tools have become more intuitive, application design has vastly improved and video capabilities are becoming embedded in the IVR stack.

Join Daniel Hong, Datamonitor's Lead Analyst and global program manager for Customer Interaction Technologies as he provides an overview of the network-based speech recognition market from both an empirical and trends perspective. The presentation will contain a deep level dive into historic and present data and highlight future forecasts and upcoming trends in speech. It will hone in on the business aspects of speech and align those to technology levers.

Venture Capital and language technology business

Carlo Paris
Paris & Partners

New European Infrastructural and Networking Initiatives

Nicoletta Calzolari
Istituto di Linguistica Computazionale del CNR

I point at two EC infrastructural initiatives in the area of Language Resources (LR) and Language Technologies (LT), which will influence the future of the field:

- 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 and LTs available and usable to scholars of all disciplines, in particular of the humanities and social sciences, ready for an eScience scenario. The purpose is to turn existing, fragmented LRs & LTs into accessible and stable services, providing easy access to language processing resources, that any user can share or adapt and repurpose.
- FLaReNet (Fostering Language Resources Network) is an eContentplus Thematic Network that will facilitate interaction among LR stakeholders. It considers that LRs & LTs present various dimensions: technical, but also organisational, economic, legal, political. It addresses also multicultural and multilingual aspects, essential for use of digital content. It brings together leading experts and groups (national and European institutions, SMEs, large companies), to ensure coherence of LR-related efforts in Europe. FLaReNet will contribute to structuring the area of LRs & LTs by discussing new strategies to: convert existing technologies into economic and societal benefits; integrate partial solutions into broader infrastructures; consolidate areas mature for recommendation of best practices; anticipate needs of new types of LRs. Its outcomes will be of directive nature, to help the EC, and national funding agencies, identifying 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 supporting linguistic diversity in Europe, while 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 LRs & LTs.

ForumTAL initiative

Andrea Paoloni
Fondazione Ugo Bordoni

As part of its efforts to promote the Italian language in the world, the Ministry of Communications has considered appropriate to promote the establishment of a permanent forum to coordinate research and development in the field of Automatic Processing of Language (TAL- Trattamento Automatico del Linguaggio). 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 which presupposes a close interaction and collaboration between humanistic scholars and researchers with a scientific and technical training.

The forumTAL has the following objectives:

- Monitoring the activities of institutions involved in TAL to promote synergies and to stimulate new interests;
- romoting research and development in the field of linguistic tools;
- Studying the initiatives that can lead to an enlargement of the market and to the development of national competitiveness;
- Promoting public and private investment in the area;
- Studying research and tools in the area with particular attention to European initiatives;
- Promoting the use of Italian technology in other countries.

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: CNIPA, CRUI, Expert System, Bordoni Foundation, ILC-CNR, ISTC-CNR, Loquendo, Dante Alighieri Society, and the Ministries of Culture, Environment, Production, Communications, University and Research, and Justice. The Forum (www.forumtal.it) has promoted numerous meetings in the area of language and language technologies.

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
National Court Reporters Association

Realtime speech-to-text technology can help courts and government to reduce costs and become more efficient, however, human judgment is needed to apply technology effectively and support non-technical considerations, such as protection of personal information.

The realtime capture and translation of the spoken word makes a variety of judicial and governmental applications possible, including:

- Litigation support software, which give judges and attorneys immediate and interactive access to the text of testimony during a trial or discovery;
- Creating the Multimedia Record;
- Evidence Presentation;
- Communication Access, for people with hearing loss;
- Case Management; and
- Virtual justice: Court reporters can provide their realtime feed from any location in the world, offering immediate access to the transcript no matter where the parties and participants are situated.

All of these applications are dependent upon extremely high accuracy rates and near instantaneous translation. Speech-to-text methodologies offer varying degrees of accuracy and speed. The session will demonstrate how relatively small variations in speech-to-text translation accuracy impact the quality and utility of the text output and discuss the need for consistent and strict standards for measuring accuracy.

In addition, the often overlooked issue of how to ensure the security of private and confidential information will be addressed.

Specific examples from judicial operations in the United States will be used to demonstrate how these technologies are being integrated to achieve highly modernized and efficient operations of the justice system.

Stentor, a new stenotypie transcription system in french

Thierry Spriet
SténoMédia

We present in this paper the technology used in the new software STENTOR for french stenotypy automatic transcription.
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. French rate homophony is about 1.8 and in the most use french standard method of stenotypy is currently not adapted to reduce that.
In the other way, most of time, we dispose of the punctuation information and have to use a very large dictionary which is more like texts analysis context.
In STENTOR, to disambiguate homophones, we propose actually 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 topic of the speech which have to be processed are various and often very specialized. Even with a very large dictionary, we have often to introduce new words during the transcription process.
We use a specific training corpus about 4,5 million words issue from more than several hundreds of hours of meeting transcription. The word error rate of STENTOR was tested on a corpus about 5000 words only, but the comparison test shown that we are yet a very competitive challenger.

Machine translation in the European Commission

Joseph Bonet
European Commission
Directorate-General for Translation
Unit R.3 – Information Technology

The European Commission (EC) has been involved in machine translation since the mid-seventies. At that time, machine translation was considered a means to overcome language barriers and a Community support was deemed necessary to help advance an enabling technology. In the nineties, Commission investment in machine translation turned to the support of research projects. But development of the EC machine translation system continued in the DG for Translation (DGT) as a support tool for translators. A number of language pairs improved their performance to a level allowing for an acceptable quality after a light post-editing. MT could be used as a productivity tool, increasing the translator's output at the price of reduced quality, and also partially as quality control tool, since it guaranteed that no sentence was forgotten, no figures were changed, etc. But most of the available language combinations had an insufficient quality and therefore MT could not be a global solution. Furthermore, translation memories came into the picture and showed their potential as productivity enhancer. The latter supported all language combinations and yielded high quality results. As a consequence, MT has become in recent year more a gisting tool than a translation tool. A total of 28 language combinations covering 10 languages and with varying levels of quality are presently available in the system."Fully automatic high quality translation" has vanished as a concept, replaced by "fully automated usable translation". Market globalisation, easier, quicker and cheaper development cycles, availability of bigger corpora, like the alignments of EU legislation recently released by DGT, are opportunities for MT. One of the remaining problems, though, is the tendency to concentrate efforts in languages with bigger numbers of speakers.

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.

NEC Machine Translation Service and Technology for Mobile Phones

Akitoshi Okumura
NEC Corporation

In a ubiquitous network society, new forms of communication and new values will emerge. These values will enable us to collaborate, resolve problems, and have better mutual understanding. To create the ubiquitous network society, it is necessary to create technologies capable of breaking through communication barriers, such as barriers of language, culture, values, knowledge, experience, and physical capabilities. Machine translation technology is a technology that will break through these barriers, particularly language barriers.

NEC has been developing machine translation technology since Dr. Koji Kobayashi, NEC's former CEO, presented a concept of C&C, the integration of computers and communications, at INTELCOM 1977. NEC then demonstrated the concept of a speech translation telephone at TELECOM 1983. NEC has also developed text translation and speech technology for several businesses:http://www.nec.co.jp/rd/Eng/innovative/E5/top.html

NEC's technologies have enabled mobile phones to provide new services and functions. My speech will be about the NEC text translation service for mobile phones and speech translation software for PDAs and mobile phones. The service is called J-SERVER Pocket and is provided by NEC and Kodensha. It 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. The service is available via i-mode, EZweb, Yahoo mobile and WILLCOM, which are portals of four major mobile-phone carriers in Japan.

NEC speech translation software helps oral communication between Japanese and English speakers in various situations while traveling. NEC compact large vocabulary continuous speech recognition engine, compact translation engine based on a lexicalized grammar, and compact Japanese speech synthesis engine lead to the development of a Japanese/English bi-directional speech translation system that can be run on the limited computational resources of a mobile phone CPU.

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 language engineering for Business Intelligence

Christian Fluhr
CEA/LIST

To maintain a clear vision of their ever-changing world environment, every company, state and public organization needs to manage data collection from open sources. The subsequent analysis and synthesis of this information helps deciders in their managing activity. This is Business Intelligence.

Business Intelligence concerns observing competitors, finding possible partners, awareness of changes in legislation, following evolution of technologies and patents, evolution of markets in various countries, activity of disinformation from competitors, etc. The world wide economy necessitates 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. Newspapers, television, radio is generally only in local, national languages.

This shows the importance of multilingual language engineering in 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 generates more financial support for the development of tools.

Various tools are used like spoken or written language identification, cross language interrogation, cross language clustering, machine translation.

Technologies of these tools are rapidly changing. A panorama of these technologies will be presented.

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

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).

Putting HLT research and technology into action for European multilingualism

Kimmo Rossi
European Commission
DG Information society and media
Unit E1 – Interaction and interfaces

Research in language technology has been supported for over 20 years in the European technology funding framework. Significant progress has been achieved in areas such as speech technology and data-driven machine translation.

Recent advances in language technology offer great potential to businesses, but a lot of work needs to be done. Further efforts are needed to put language technology into productive use. This will not only justify the investments in the research but will also provide a competitive edge to companies and better services to the citizens.

Business has migrated to the web, and customers require solutions and information in their language. Websites with truly multilingual access and services are likely to sell better in the global market than English-only websites. The efficient use of information and content generated by the public sector requires efficient solutions addressing the language barrier.

One of the main objectives of the i2010 policy framework is to create a single European information space that relies on rich and diverse online content and digital services. An important aspect is the 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 integrated with associated technologies such as content management. Much more is needed than simple machine translation of strings. 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.

Crossing media for improved information access: the REVEAL THIS example

Stelios Piperidis
Language Technology Applications Department
Institute for Language and Speech Processing - Athena R.C.

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.

Challenges of Speech to Speech Translation in the context of Human-Human Communications

Alex Waibel
InterACT

Recognition and Understanding of Meetings: The European AMI and AMIDA Projects

Hervé Bourlard
IDIAP Research Institute

The AMI and AMIDA projects are concerned with the recognition and interpretation of multiparty meetings. Within these projects we have: developed an infrastructure for recording meetings using multiple microphones and cameras; released a 100 hour annotated corpus of meetings; developed techniques for the recognition and interpretation of meetings based primarily on speech recognition and computer vision; and developed an evaluation framework at both component and system levels. In this talk, we will present an overview of these projects, with an emphasis on possible applications and technology transfer activities.

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?

Semantic Search: Content versus Formalism

Christian F. Hempelmann
Hakia

This paper presents a theoretical approach for deep-meaning representation, ontological semantics (OntoSem), for a specific, complex NLP application: a meaning-based Internet search engine. It introduces the resources and technologies of OntoSem and their development. The aim is to provide a general overview of the specific methods in which OntoSem is used in our Internet search approach and give an in-depth account of selected key issues in web search and how we address them. OntoSem parses natural language web content and transposes it into a representation of its meaning, structured around the events described in the text and their participants. Queries can then be matched to this meaning representation in anticipation of any of the permutations in which they can surface in written text. These permutations centrally include overspecification (e.g., not listing all synonyms, which non-semantic search engines require their users to do) and, more importantly, underspecification (as language does in principle). For the latter case, ambiguity can only be reduced by giving the search engine what humans use for disambiguation, namely knowledge of the world as represented in an ontology. The main issue around which the paper will be structured rhetorically is the distinction between semantic content and purportedly semantic formalisms. Meaning for web search requires complex description for automatic generation and can in principle not be extracted from surface text with statistical methods. In contrast to this, formalisms and suggestions for controlled vocabularies like OWL may claim to be semantic, but can, of course, not be, since meaning is content and does not lend itself to automatic extraction from natural language without rich knowledge resources.

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 Binnenpoorte and Dorota Iskra
LogicaCMG

The key function of a speech-based 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 DTMF-based routing applications. A DTMF application allows the caller 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 callers 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 correctly recognised utterances, 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 routing application. And thirdly, there is the level of involvement within the service-providing organisation; do call centre agents understand and use the advantages of speech 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 routing application. So, in this paper we will not only discuss the traditionally examined technical performance factors, but also a number of other factors based on our experience.

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.

POSTER SUMMARIES

A description language at the accentual unit level for Romanian intonation

Doina Jitcă, Vasile Apopei, *Institute for Computer Science, Romanian Academy, Iasi Branch, Romania*
Magdalena Jitcă, *University "Alexandru Ioan Cuza", Iasi, Romania*

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 Fo 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 Fo contour characterizations as attributes. The structured text will be used as input for the phonetic module that generates the Fo contour for the synthesizer.

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

Longbiao Wang, Seiichi Nakagawa, *Department of Information and Computer Sciences, Toyohashi University of Technology, Japan*
Norihide Kitaoka, *Department of Media Science, Nagoya University, Japan*

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 (VSSUMCLMS) 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.

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

Matthew P. Aylett, *Centre for Speech Technology Research, University of Edinburgh, U.K., Cereproc Ltd., U.K.*
Junichi Yamagishi, *Centre for Speech Technology Research, University of Edinburgh, U.K.*

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.

Conceptual maps and Computational Linguistics: the Italian ALTI project

Francesco Di Maio, *Dipartimento di Scienze della Comunicazione, Università degli Studi di Salerno - Italy*
Johanna Monti, *Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico, Università degli Studi di Napoli "L'Orientale" - Italy*

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 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.

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.

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.

Coupling Speech Recognition and Rule-Based Machine Translation with Chart Parsing

Selçuk Köprü, *Applications Technology, Inc., Turkey*
Adnan Yazıcı, *Dept. of Computer Engineering, Middle East Technical University, Turkey*
Tolga Çiloğlu, *Dept. of Electrical and Electronics Engineering, Middle East Technical University, Turkey*
Ayşenur Birtürk, *Dept. of Computer Engineering, Middle East Technical University, Turkey*

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.

Human Language and Semantic Web Technologies for Business Intelligence Applications

Thierry Declerck, Hans-Ulrich Krieger, *Language Technology Lab, DFKI GmbH*
Marcus Spies, *Digital Enterprise Research Institute, Universität Innsbruck*
Horacio Saggion, *NLP Group, Department of Computer Science, Sheffield University*

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 ìMUSINGî 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.

Improving Third Generation Translation Memory Systems Through Identification of Rhetorical Predicates

Ruslan Mitkov, *University of Wolverhampton*
Gloria Corpas, *University of Malaga*

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 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.

Language Engineering for Basque in a Visual Communication Technologies Context

Maider Lehr, Kutz Arrieta, *VICOMTech Research Center, Donostia-San Sebastian*
Andoni Arruti, *Signal Processing Group, University of the Basque Country, Donostia-San Sebastian*

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 and a priority 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 this 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 will 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.

We describe past, present, and future projects in the areas of Speech and Natural Language and its integration in existing fields of expertise in VICOMTech. We also take into account the issue of Basque, as a minority language, with entailed disadvantages and advantages. On the one hand, fewer resources are available, and on the other, Basque speakers are native speakers of other languages such as Spanish, French, English, etc., and are particularly well positioned to deal with multilingual issues in advanced technology applications.

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, *Computer Science & Engineering, SASTRA Deemed University, Thanjavur*

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, Gujarati and the different type of dialects which are related to Hindi in one respect or the other. The flexibility is possible due to the flexible word structure proposed for the dictionary of the language. Due to the closeness of the system with the Indian Languages, the project has been named as NATIVE LANGUAGE PROCESSING. The base of the project involves storing base information (basic words) in an XML file which is subsequently used in the later stages of the Language Processor. In addition to all the conventional parts a NLP system viz. Lexical Analyzer, Syntactic & Semantic Parser, POS Tagger, this system has an optional answering part (proposed) for generating a proper response to the given output. Hence, the system learns and uses a sentence generator to generate appropriate answers based on sentence structures. Additionally, the semantic parser has been split into proper and improper sections which analyze syntactically proper and improper statements, thereby understanding the correct but improper statements. The system is also supposed to incorporate a word recognition algorithm which takes as input an unknown word and gives out the probability of the word being a possible word in the language. Thus, the system is supposed to learn at runtime and identify valid/invalid words and accept or reject based on the context. The system has been modeled 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 and the proposed extension of the system make it possible to be compatible with most of the modern Indian languages following the Hindi grammar.

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

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.

On Integration of Terminological Data in Translation Systems

Signe Rirdance, Andrejs Vasiljevs, *Tilde, Latvia*

In today's translation practice, a significant gap exists between traditional desktop translation tools and terminological data aon internet. Translators spend up to 60% of translation time on terminology research, therefore it is vital to ensure use of terminology resources in the right format and environment.. Computer assisted translation tools demonstrate some major drawbacks of regarding handling of terminology for translation.

EuroTermBank project facilitates terminology data accessibility and exchange by collecting, consolidating and disseminating dispersed terminology resources through an online terminology data bank. It applies the concept of federation in linking portals and data repositories, which reaches out to the level of semantic interoperability.

Automated entry compounding is an innovative mechanism proposed by EuroTermBank in unification of potentially matching terminology entries from different resources. It carries important implications for new web-based approaches to efficient handling of terminology entries from multiple sources.

Integration of term banks with translation environments is only possible by rigorous implementation of international standards. EuroTermBank project uses the TBX standard as import/export and data storage format.

Currently, the richness of terminology resources on internet does not translate into the expected increased productivity and quality levels of translation. It is important to establish the basic principles that enable easier integration of term banks with the variety of translation environments. A new layer of tools and technologies is required to significantly enhance the current productivity of human translation.

Opentrad: bringing to the market opensource based Machine Translators

Ibon Aizpurua Ugarte, *Eleka Ingeniaritza Linguistikoa, S.L.*
Gema Ramírez Sánchez, *Prompsit Language Engineering*
Jose Ramom Pichel, *Imaxin | Software*
Josu Waliño, *Elhuyar Fundazioa*

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.

Relation Extraction in an Intelligence Context

Bénédicte Goujon, *Thales Research & Technology, France*

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.

Speech technology for language tutoring

Helmer Strik, Ambra Neri and Catia Cucchiarini
Department of Language and Speech, Radboud University Nijmegen, The Netherlands

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.

System ZENON – Semantic Analysis of Intelligence Reports

Matthias Hecking, *FGAN/FKIE*

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.

Technologies for Simultaneous Acquisition of Speech Articulatory Data: Ultrasound, 3D Articulografh and Electroglottografh

Mirko Grimaldi, Barbara Gili Fivela, Francesco Sigona, *Centro di Ricerca Interdisciplinare sul Linguaggio University of Salento, Lecce, Italy*
Michele Tavella, Giorgio Metta, Giulio Sandini, *Laboratory for Integrated Advanced Robotics University of Genoa, Italy*
Paul Fitzpatrick, Laila Craighero, Luciano Fadiga, *NeuroLab, University of Ferrara, Italy*

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) there is the availability of the main equipment used for articulatory studies in the most advanced international research centres: 3D articulograph, ultrasound system, electroglottograph, electropalatograph [Stone, 2005; Wrench, 2007; Zierdt et al. 1999]. Within the European CONTACT project, some of these instruments have been synchronized and are usable to register various material simultaneously [Grimaldi et al. 2007]. The aim of the CONTACT project is to verify if the development of language can be linked to the motoric control development, especially the control that is necessary for precision movements. In this perspective, 3D articulograph, ultrasound system and electroglottograph have been synchronized in order to acquire speech material, which - supplied as input to an automatic speech recognition system - could allow to verify eventual improvement in identification abilities thanks to the presence of articulatory information. We will present some considerations about a corpus of pseudowords read by 9 speakers of the area of Lecce, 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 phonetic-phonologic field.

XGate and XRG: tools for visually editing, querying and benchmarking XML linguistic annotations.

Francesco Cutugno, *Department of Physics - NLP Group, University "Federico II" of Napoli, Italia*
Leandro D'Anna, *Department of Linguistics and Literature, University of Salerno, Italia*

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" (http://www.parlaritaliano.it).

A Calendar Interface in French: XipAGENDA

Claude Roux, *Project Leader, Xerox Research Centre Europe*

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 and querying your agenda.

Acquiring Legal Ontologies from Domain-specific Texts

Felice Dell'Orletta, Simonetta Montemagni, Simone Marchi, Vito Pirrelli,
Giulia Venturi, *Istituto di Linguistica Computazionale, CNR, Pisa, Italy*
Alessandro Lenci, *Department of Linguistics, University of Pisa, Italy*

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-po wered incremental systems like T2K for accurate large- scale semi- - automatic extraction of legal ontologies.

Advances in NLP applied to Word Prediction

Carlo Aliprandi, *Synthema*
Nicola Carmignani, Nedjma Deha, Paolo Mancarella, Michele Rubino, *University of Pisa*

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.

An Online Linguistic Journalism Agency – Starting Up Project

Annibale Elia, *University of Salerno, Salerno, Italy*
Ernesto D'Avanzo, *University of Salerno, Salerno, Italy*
Tsvi Kuflik, *University of Haifa, Haifa, Israel*
Giovanni Catapano, *University of Salerno, Salerno, Italy*
Mara Gruber, *Istituto Italiano Scienze Umane, Napoli, Italy*

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.

Boosting the Recall of Descriptive Phrases in Web Snippets

Alejandro Figueroa, *Deutsches Forschungszentrum für Künstliche Intelligenz - DFKI, Saarbrücken, Germany*

WebQA is a Web Question Answering System, 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, including “Who is Althea Gibson?” and “What are fractals?” This module searches for answers by means of a query rewriting strategy, which considerably boosts the recall of descriptive utterances. This study compares this rewriting strategy with two new search strategies based on Google n-grams and additional search engines. Results show that Google n-grams are promising for improving the recall of descriptive utterances. Additionally, this work deals at greater length with the challenges posed by the assessment of web-based definition Question Answering Systems.

COLDIC a generic tool for the creation, maintenance and management of Lexical Resources

Núria Bel, Sergio Espeja, Montserrat Marimon, Marta Villegas, *Institut Universitari de Linguística Aplicada, Universitat Pompeu Fabra, Barcelona, Spain*

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.

Fast and easy development of pronunciation lexicons for names

Henk van den Heuvel, *CLST, Radboud University Nijmegen, The Netherlands*
Jean-Pierre Martens, *ELIS, Ghent University, Belgium*
Nanneke Konings, *CLST, Radboud University Nijmegen, The Netherlands*

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.

Grammar Systems as Interfaces

Gemma Bel-Enguix, M. Dolores Jiménez-López, *Research Group on Mathematical Linguistics, Rovira i Virgili University*

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. We introduce a Grammar System Interface (GSI) that may contribute to the building of more efficient human-computer interaction tools through the simulation of human-human conversations.

Even though 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. 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.

Many researchers believe that natural language interfaces can provide the most useful and efficient way for people to interact with computers. A challenging solution for accessibility of information anywhere anytime is to provide machines with human-like capabilities.

The GSI we introduce 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. The model uses simple grammars in order to generate a dialogue structure. This is not a psychologically realistic cognitive model, but a model that might successfully emulate human linguistic behaviour in some specific situations such as natural language interfaces.

HLT and communicative disabilities: The need for co-operation between government, industry and academia

Catia Cucchiarini, *Nederlandse Taalunie (Dutch Language Union), The Netherlands*
Dirk Lembrechts, *MODEM, Consultancy Centre on Communicative Disabilities, Wilrijk, Belgium*
Helmer Strik, *Department of Language and Speech, Radboud University Nijmegen, The Netherlands*

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.

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

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 to treat the specific impairment of the user. "Vocaliza" is the result of the joint work of the Aragon Institute for Engineering Research (I3A) and the Public School for Special Education "Alborada" in their goal of introducing different technologies to improve the standard of living of disabled people.

The HLT systems used in the application are Automatic Speech Recognition (ASR), speech synthesis, speaker adaptation and utterance verification. Speech is the main interface for the user to interact with the games and HLT is also used to give the users feedback about the quality of their speech. 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 labelling of a recently acquired database containing 3,192 utterances of impaired speech. The results show that accuracy rises from 60% when using a speaker independent model to 87.66% when using speaker dependent models, due to the ability of these methods to include the speaker variability of every speaker in the acoustic modeling but not the pronunciation errors made by the speaker.

Legal Taxonomy Syllabus: Handling Multilevel Legal Ontologies

Gianmaria Ajani, *Dipartimento di Scienze Giuridiche, Università di Torino - Italy*
Guido Boella, Leonardo Lesmo, Alessandro Mazzei, Daniele P. Radicioni,
Dipartimento di Informatica, Università di Torino - Italy
Piercarlo Rossi, *Dipartimento di Studi per l'Impresa e il Territorio, Università del Piemonte Orientale - Italy*

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.

META-MultilanguagE Text Analyzer

Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile, Leo Iaquinta,
Pasquale Lops, Giovanni Semeraro,
Department of Computer Science, University of Bari, Italy

Natural Language Processing (NLP) has significant a impact on many relevantWeb-based and SemanticWeb applications, such as information filtering and retrieval. Tools supporting the development of NLP applications are playing a key role in textbased 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.

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

Semantic Press is 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.

Text Processing Tools and Services from iLexIR Ltd

Ted Briscoe, Paula Buttery, John Carroll Ben Medlock, Rebecca Watson
iLexIR Ltd, Cambridge, UK

The RASP (robust accurate statistical parsing) toolkit is developed by research groups based at the universities of Cambridge and Sussex (www.informatics.sussex.ac.uk/research/nlp/rasp). iLexIR is the sole commercial agent for 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), named entity recognisers (e.g. Lingpipe, www.alias-i.com/lingpipe) and XML-based document metadata handling systems (e.g. UIMA, uima-framework.sourceforge.net, www.digitalpebble.com/resources.html), to solve a diverse range of real-world text processing tasks.

iLexIR also licenses the timed aggregate perceptron classifier, an innovative model with accuracy comparable to support vector machines but training time closer to a naive bayes classifier. This is a significant advantage for real world text classification problems where reductions in training time allow vital experimentation intoenhancing feature generation and selection from the range of feature types made available by RASP, as well as frequent retraining as data is accumulated. TAP or other classifiers can be deployed for document passage, and (sub)sentential classification tasks utilising features derived from RASP in a (semi-)supervised fashion dependent on the training data available.

The Impact of Standards on Today Speech Applications

Paolo Baggia, *Loquendo, Torino, Italy*

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 was another important driver in promoting the adoption of a novel architecture, then 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.

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 all the fields. 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 W3C was begun and it forced an abrupt change in the industry.

The paper describes the Voice Browsing approach to speech applications, with a short description of the major standards, then the Voice Browsing Platforms in use today and finally other standardization areas and future evolutions are discussed.

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

This paper describes the design, implementation and population of the BioLexicon, a comprehensive lexical resource especially designed in the framework of BOOTStrep to support text mining in the bio-domain. The BioLexicon presents some major novelties. It has been conceived to meet both the domain requirements and the recent international standards for lexical representation: the overall data model is compliant to the ISO Lexical Markup Framework. The semantic layer adheres to innovative theoretical approaches in computational lexicography. It is an extendable resource: initially populated with data collected from available biomedical sources, bio-terms (and variants) are enriched with morphological, syntactic and lexical semantic features automatically extracted from texts, also including sub-categorization patterns and predicate-argument structure of bio-events. Hence, the BioLexicon is an integrated resource which combines features of both terminologies and lexicons. The other novelty is the alignment of term-related syntactic and semantic information with concepts of the BioOntology.

These two tightly-bound resources will constitute a terminological backbone linked to a BioFact repository and, all together, they will form the Bio Knowledge Store for supporting text mining and information extraction applications. The BioLexicon implementation consists of a flexible, extensible relational database which comes equipped with automatic population procedures. Population relies on a dedicated input data structure, an XML Interchange Format, allowing to structure terms gathered from existing sources and “pull-and-push” them in the database. As a matter of fact, the BioLexicon teaches that the state-of-the-art both in the bio-domain and in NLP lexicons is mature enough for us to aim at creating a lexicon which aspires to become ìtheî standard in this domain. Finally, although conceived for a special domain, thanks to conformity to standards, the BioLexicon accounts for interoperability and extendibility to other areas.

Using Model Trees for Online Monitoring of Human-Computer Call Dialogues

Woosung Kim and Jayant M. Naik, *Convergys Corporation*

The recent growth of the call center market has made automatic call monitoring essential. In particular, automatically monitoring human-computer call dialogues is promising as it can be used practically, i.e., to first detect a bad call when a caller has trouble with dialogue systems and then bring in a live human agent to salvage the caller. The challenge is how to decide, if and when, to bring in a human agent, but more importantly, how to do it in real-time (online) before the caller gives up the call.

This paper particularly concerns practical issues in online call monitoring: how to detect bad calls during the call without computational overhead. We approach this task as a regression problem. I.e., given a test call, we estimate the likelihood of the call being good or bad. If the likelihood drops below a certain threshold, we decide the call as bad and bring in a human agent. More specifically, we propose using model trees as they result in simple decision rules, which can be easily used for practical applications.

Our experiments show over 86% classification accuracy and yet the decision rules are simple. Furthermore, our analysis reveals that the problematic calls may be detected as early as at the caller's 3rd or 4th turn. We also demonstrate that heterogeneous features may be incorporated, yielding a gain in classification performance. This is encouraging as further gains are expected if better features are used. Finally, we conclude this paper by emphasizing the importance of automatic call monitoring of human-computer calls, as this could push dialogue systems toward a more advanced direction.

Sponsors and Exhibitors

Gold Sponsor and Exhibitor

Loquendo
VOCAL TECHNOLOGY AND SERVICES

With over 30 years cutting-edge R&D experience in speech technology, Loquendo is the leading and multi-award winning innovator at the forefront of the global speech market.
For the resolution of security issues Loquendo has developed Loquendo outstanding Speaker Recognition technology.
Everyone has a unique voiceprint: Loquendo technology makes use of this voiceprint in biometric verification to ascertain an individual's identity.
Loquendo is a Telecom Italia company headquartered in Turin; for more information about Loquendo visit www.loquendo.com.

Silver Sponsor and Exhibitor

PerVoice

Pervoice S.p.A., start-up of the “Fondazione Bruno Kessler” Research Labs and other companies, is the first Italian service company for the processing, analysing and evaluating of spoken language.
Pervoice is operating in three market segments:

- In the Transcription Market, Pervoice provides semi-automated transcription services using spontaneous speech-to-text technology, in any situation where it may be necessary to produce notes or reports in a short time.
- In the Media Sector, Pervoice offers a system of “Media Intelligence” for the monitoring of TV broadcasting in relation to advertising campaigns, brands, names, and personalities.

Bronze Sponsor and Exhibitor

CELCT, the Center for the Evaluation of Language and Communication Technologies, is a joint enterprise established by FBK (Bruno Kessler Foundation) ex ITC-Irst, and DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz), and funded by the Autonomous Province of Trento.

The mission of CELCT are to set up infrastructures and develop skills in order to operate successfully in the field of the evaluation. In particular its goal is studying and devising evaluation procedures for human language and multimodal communication technologies such as, for instance, question answering, speech-to-speech translation, spoken language technologies, cross-language information retrieval, word sense disambiguation, web pages accessability and usability, etc.

Bronze Sponsor and Exhibitor

Fondazione Bruno Kessler (FBK), born on March 1st 2007 and located in Trento, is a non-profit body with a public interest mission having private law status and inheriting the history of Istituto Trentino di Cultura (ITC – founded in 1962 by the Autonomous Province of Trento). Scientific excellence and innovation as well as technology transfer to companies and public services are FBK's main objectives. Research in the three main areas of Information Technology (Engineering, Content, and Interaction) is organized in Research Units.

Bronze Sponsor

IBM's mission is to be a partner in innovation, helping companies and institutions develop infrastructure, processes and new models of business aimed towards growth and competitiveness.

IBM turns to clients with an offer in which the hardware, software and service components are harmonized in the broadest concept of solution that creates and transfers value. To do this, it has developed in depth skills in the various market sectors and integrates them with specific know-how in the different technological and application areas.

Exhibitor

DFKI German Research Center for Artificial Intelligence

Founded in 1988, DFKI today is one of the largest nonprofit contract research institutes in the field of innovative software technology based on Artificial Intelligence (AI) methods. DFKI is focusing on the complete cycle of innovation - from world-class basic research and technology development through leading-edge demonstrators and prototypes to product functions and commercialization. Organized as a public-private-partnership, the center maintains sites in Saarbrücken, Kaiserslautern, Bremen and Berlin. As language technology is one of the key research areas of DFKI, the research company is also home of the German Competence Center for Language Technology. Among the fifty spin-off companies of DFKI are several start-ups commercializing language technology applications.

Exhibitor

Fondazione Ugo Bordoni

The Ugo Bordoni Foundation is involved in research, study and consultancy activity in the information and communication technology sector.
It has been recognised in law 3 of January 16 2003 as a private institution of cultural importance, under the supervision of the Minister for Communications.
In the field of TAL the Foundation operates in different areas: the traditional area of the assessment of systems and technologies and the correlated theme of collecting corpora for use in either evaluating or developing products (with particular reference to the FOCUS, SIVA, CLIPS corpora); the basic themes of speech recognition for which it has developed a specific software called IDEM.

Exhibitor

Istituto di Linguistica Computazionale del CNR (ILC) - Pisa

ILC has a clear international and national leadership, promoting the cooperation with the major initiatives related to Language Resources and Language Technologies. Specific attention is given to inserting the Italian language in an international multilingual network. Its activities span the thematic areas of:

- Design of standards and building of computational language resources
- Models and methods for natural language processing and mono- and multilingual prototypes
- Methods and tools for humanities research

ILC presents some language resources, tools, and application prototypes.

Exhibitor

Born in 1996, Interactive Media delivers to network carriers, service providers and enterprises state-of-the-art solutions that connect media, people and information. Meltemi Communications System, IM software product, is a powerful, scalable, IMS compliant platform integrating advanced speech, video, network and ICT technologies for rapid creation, deployment and operation of network services and applications.
IM solutions include videogateways, media and application servers, video service centres, video/voice interactive services systems supporting traditional as well IP networks.

Exhibitor

INTERSTENO unites the world wide community of those using a full range of speed writing methods to quickly produce high quality texts. Intersteno is now represented in all continents by national groups, and embraces professional reporters, parliamentary reporters, as well as teachers and secretaries, using their skills as personal or professional productivity tools, which play an important part in the multi-media communication processes of our times. They are eager to appreciate new technical issues helping them to increase efficiency in their work.
Accademia Giuseppe Aliprandi, founded in 1927, is involved in researches of technologies which are strictly connected with language and speech, like shorthand, speech recognition, reporting as well as informatics and multimedia. It also fosters the relevant teachings. An important historical library preserving more than 3000 books is available for consultation in Florence and is connected to the Tuscan Library net.

Exhibitor

At the National Institute of Information and Communications Technology (NICT) in Japan, the Computational Linguistics Group is engaged in a wide variety of research projects related to natural language processing, ranging from basic technologies such as extraction of informative linguistic information to applied technologies including information retrieval and machine translation. The group is also compiling and publishing large-scale linguistic resources such as the EDR lexicon.

Exhibitor

At the National Institute of Information and Communications Technology (NICT) in Japan, the Computational Linguistics Group is engaged in a wide variety of research projects related to natural language processing, ranging from basic technologies such as extraction of informative linguistic information to applied technologies including information retrieval and machine translation. The group is also compiling and publishing large-scale linguistic resources such as the EDR lexicon.

Exhibitor

Synthema is an Italian SME established in 1993 as a spin-off of the IBM Scientific Center at Pisa. It operates in the field of Information Technology, providing Software Solutions, Research and Development.
Its main area of activity refers to Human Language Technologies, including Machine Translation, Natural Language Processing, Knowledge Extraction and Management, Text Mining and Speech Recognition.
Synthema main asset is a large portfolio of resources, tools and products, based on its own MT engine and Lexical engine, available for Italian, English and many other languages.

Endorsed by

Created in 1995, ELDA, originally European Language Resources Distribution agency, was set up to identify, classify, collect, validate and produce the language resources which may be needed by the Human Language Technology (HLT) community. Throughout the years, and following the evolution in HLT field, ELDA has broadened its activities and launched evaluation activities, distributing the language resources needed for evaluation purposes in language engineering applications.

Media sponsor

Speech TECHNOLOGY

AVIDS

Thanks to

Rai

Audio Video direction and recording by RAI

INTERSTENO

ITALICA ACADEMIA

Simultaneous translation by Intersteno - “Accademia Aliprandi”

Hi Pro
Competence for Evolution

RFID Technology by Hi Pro

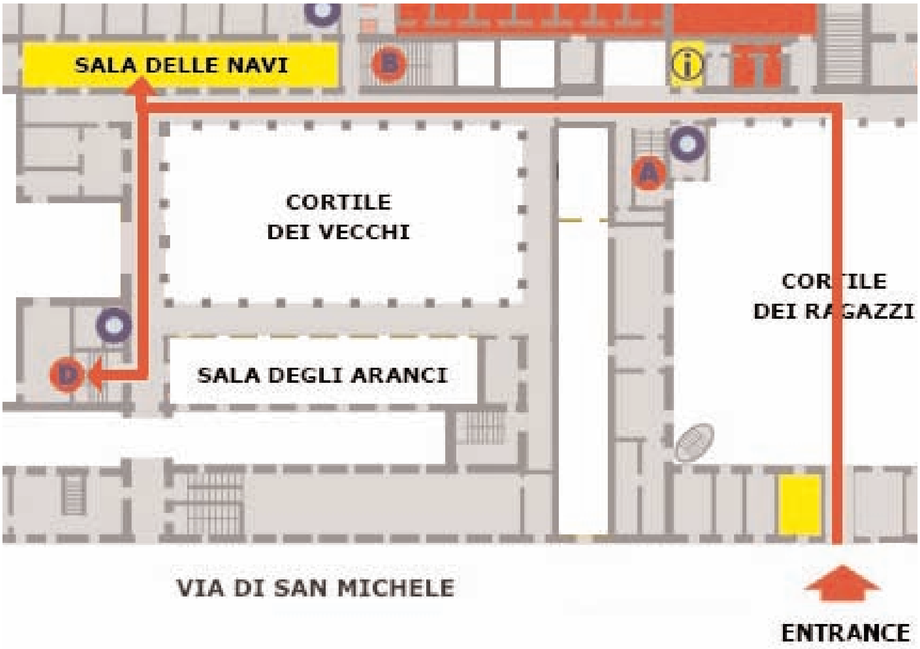
Complesso Monumentale di San Michele a Ripa Grande

The venue of the conference is inside the monumental building of San Michele a Ripa, one of the most interesting and great architectural structure in Rome. It starts in 1686 as Istituto Apostolico San Michele, with the goal of receiving and rehabilitate young orphans and needful children but it was completed as late as 1834. The Complex is composed by: la Chiesa Grande (The Big Church), ideated by Carlo Fontana in 1706; the plant of San Michele a Ripa, hosting flourishing craftsmanlike activities since its creation until 1870; the wool mill, built in 1703, mostly utilizing the work of the prisoners of the Casa di Correzione (Punishment House); the Arazzeria (Tapestry factory); the Scuole di arti e mestieri (School of Arts and Crafts).

Sala dello Stenditoio

Ground floor

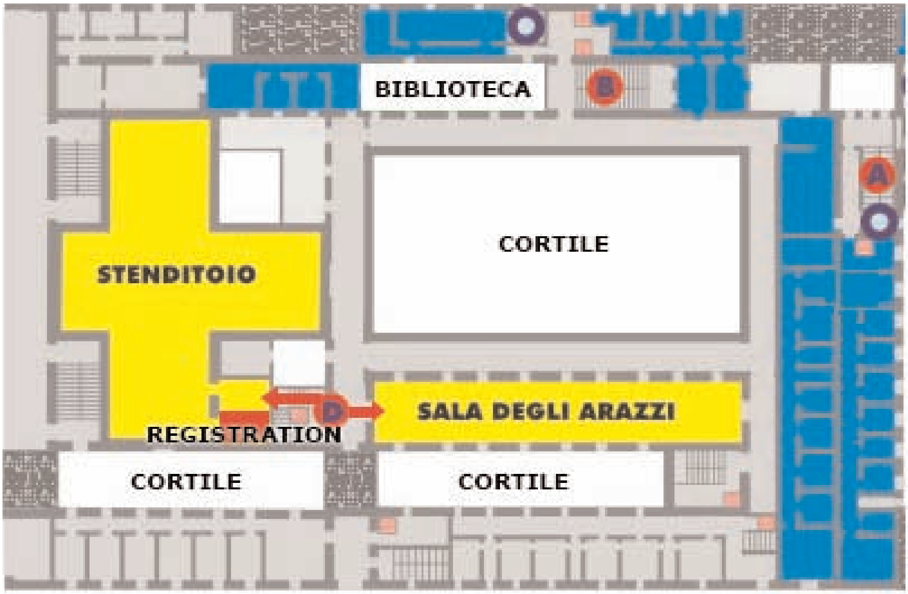
Conference room - Parallel session (Sala delle Navi)



Third floor

Conference room and posters (Sala dello Stenditoio)

Exhibition hall (Sala degli Arazzi)



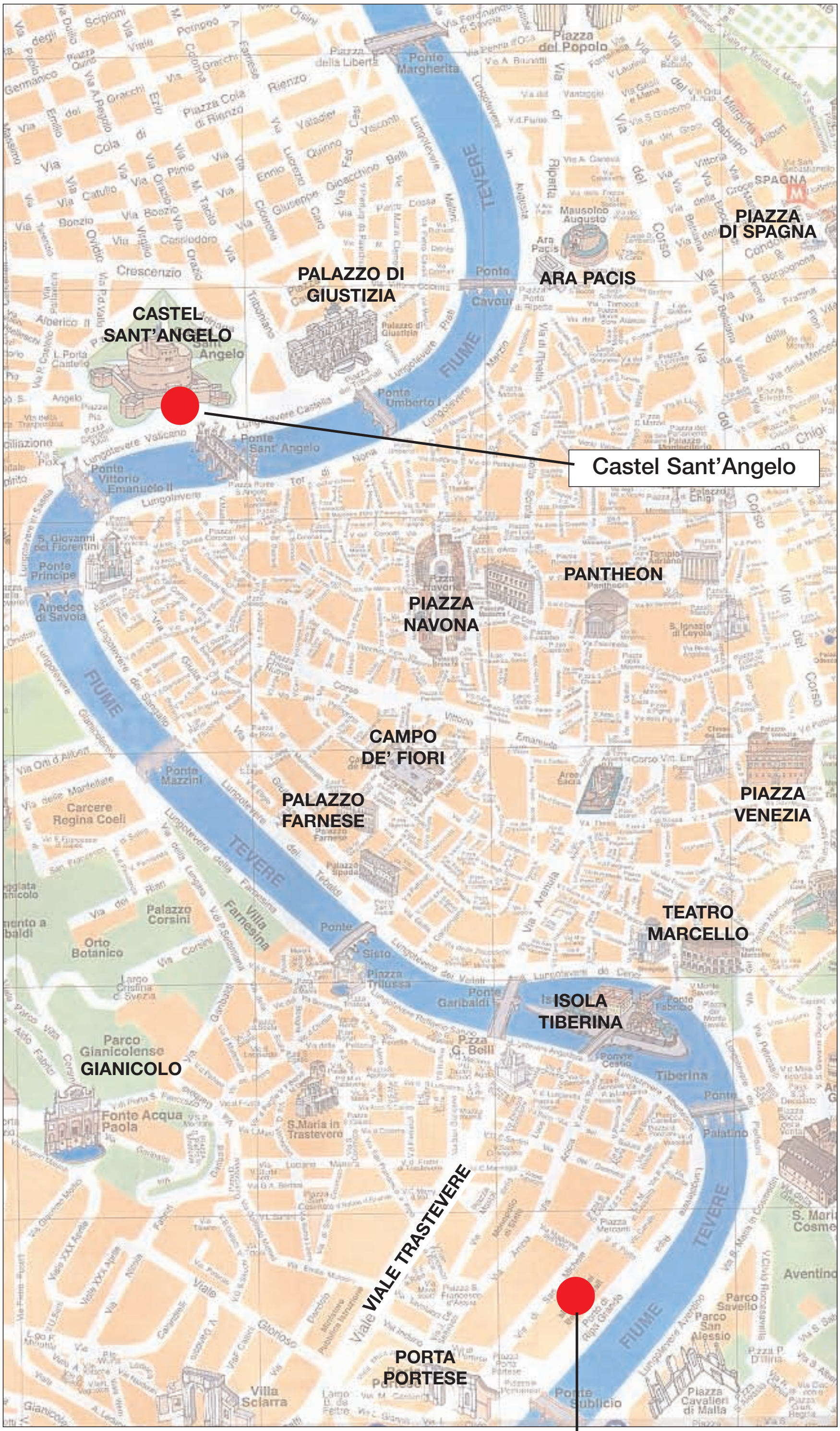
Social event

Castel Sant'Angelo

LangTech2008 has the pleasure to invite you in the magnificent Castel Sant'Angelo, one of the most important architectural building in Rome.
The history from its origin in 136 till today, and the reference to this building in opera Tosca by Puccini, makes it an icon of the eternal city.

After a visit, a buffet will be served in the "Giretto Coperto", build by Pope Pio IV around 1555, inside the castel. An hystorical place with unbelievable view on the Tiber river.

Lungotevere Castello, 50
28th February at 7,30 p.m.



Castel Sant'Angelo

Complesso monumentale di San Michele a Ripa Grande

"Langtech 2008"

Programme

THURSDAY 28 February 2008

8.30 Registration

9.30 - 10.30 **Opening Session**
Antonio Sassano, *Director General Fondazione Ugo Bordoni*
Giordano Bruno Guerri, *Conference Chair*
K-J. Lönnroth, *Director General for Translation (CEE)*

10.30 - 11.00 **Exhibition visit and coffee break** (Room: ARAZZI)

11.00 - 12.45 **Speech Technologies**
Chair: **Roberto Pieraccini**, *SpeechCycle*
- **Renato De Mori**, *University of Avignon*
  Using frames in Spoken Language Understanding
- **Geoffrey Zweig**, *Microsoft Research*
  Voice search on mobile devices

12.45 - 13.00 **Sponsor technical talk**
Chair: **Andrea Paoloni,** *Fondazione Ugo Bordoni*
- Loquendo (10')
- Pervoice (5')

13.00 - 14.00 **Exhibition visit** (Room: ARAZZI) **and lunch** (Room: NAVI)

14.00 - 15.00 **Posters** (Room: STENDITOIO)

15.00 - 16.00 **Business, market and strategies** (Room: STENDITOIO)
Chair: **Christian Fluhr**, *CEA/LIST*
- **Daniel Hong**, *Datamonitor*
  The network-based speech market from an empirical and trends perspective
- **Carlo Paris**, *Paris & Partners*
  Venture Capital and language technology business
- **Nicoletta Calzolari**, *ILC/CNR*
  New European Infrastructural and Networking Initiatives
- **Andrea Paoloni**, *Fondazione Ugo Bordoni*
  ForumTAL initiative

15.00 - 16.00 **Reporting** (Room: NAVI)
Chair: **Fausto Ramondelli**, *Senato della Repubblica*
Captioning – accessibility to education for hearing impaired
- **Mark Golden**, *National Court Reporters'Association*
  Realtime Speech-to-Text: A means to an End
- **Thierry Spriet**, *SténoMédia*
  Stentor, a new computer-aided transcription (CAT) software for French language

16.00 - 16.30 **Exhibition visit and coffee break** (Room: ARAZZI)

16.30 - 17.30 **Machine Translation** (Room: STENDITOIO)
Chair: **Joseph Bonet**, *CEE*
Machine translation in the European commission
- **Philipp Koehn**, *University of Edinburgh*
  Open Source Tools for Statistical Machine Translation
- **Akitoshi Okumura,** *NEC*
  NEC machine translation technology

16.30 - 17.30 **Language Technology and Intelligence** (Room: NAVI)
Chair: **Giuseppe Fabbrocino**, *UGCT*
- **Pasquale Angelosanto**, *ROS Carabinieri*
  Speaker recognition for surveillance scenario against terrorism and organised crime
- **Christian Fluhr**, *CEA/LIST*
  Multilingual language engineering for Business Intelligence
- **Mario Caligiuri**, *University of Calabria*
  University and Intelligence: an Italian point of view
- **Mario Coggio**, *Stato Maggiore Difesa, RIS - CII*
  How semantic technology can support intelligence: the applications in OSINT

17.30 - 18.30 **Exhibition visit** (Room: ARAZZI)

Social event
On Thursday 28th February at 7,30 p.m.. A one-hour tour guide will be arranged at Castel Sant'Angelo (Lungotevere Castello, 50). Afterwards a buffet supper will be served in the so-called "Giretto di Pio IV" of the castle.

FRIDAY 29 February 2008

8.30 Registration

9.00 - 10.30 **Major achievements triggered by European Funded projects**
Chair: **Khalid Choukri**, *ELDA/ELDA*
Language technology evaluation in Europe. Key achievements and the need for an infrastructure
- **Kimmo Rossi**, *European Commission*
  Putting HLT research and technology into action for European multilingualism
- **Stelios Piperidis**, *Institute for Language and Speech Processing - Athena R.C.*
  Retrieval of Video and Language for The Home user in an Information Society
- **Alex Waibel**, *InterACT*
  Challenges of Speech to Speech Translation in the context of Human-Human Communications
- **Hervé Bourlard**, *IDIAP*
  Recognition and Understanding of Meetings: The European AMI and AMIDA Projects

10.30 - 11.00 **Exhibition visit and coffee break** (Room: ARAZZI)

11.00 - 12.00 **Language Technology in Tomorrow's Search Applications**

Chair: **Hans Uszkoreit**, *Germany-DFKI*

- **Christian F. Hempelmann**, *Hakia*
- **Thomas Hofmann**, *Google*
- **Mario Lenz**, *Empolis*
- **Geoffrey Zweig**, *Microsoft Research*
  Voice search on mobile devices

12.00 - 13.00 **CRM (Customer Relationship Management)**

Chair: **Dorota Iskra,** *Logica CMG*

- **Roberto Pieraccini**, *Speechcycle*
  Advanced speech and language technology for complex customer care automation and self-service
- **Diana Binnenpoorte**, *Logica CMG*
  What makes a successful speech enabled call routing application?

13.00 - 14.00 **Exhibition visit** (Room: ARAZZI) **and lunch** (Room: NAVI)

14.00 - 15.00 **Posters** (Room: STENDITOIO)

15.00 - 16.00 **SME Elevator**

Chair: **Bente Maegaard**, *University of Copenhagen*

SME's Presentations

16.00 - 16.30 **Exhibition visit and coffee break** (Room: ARAZZI)

16.30 - 17.30 **Knowledge Management**

Chair: **Nicoletta Calzolari,** *ILC/CNR*

- **Enrico Motta**, *Knowledge Media Institute*
  Language Technologies and the Semantic Web: An Essential Relationship
- **Christopher Welty**, *IBM Research*
  Answering Questions from the Semantic Web

17.30 - 18.00 **Closing Session**

Chair: **Giordano Bruno Guerri,** *Fondazione Ugo Bordoni*

- **Carlo Paris,** *Paris & Partners*
  LangTech Prize
- **Dorota Iskra**, *LogicaCMG*
  Next LangTech