PAST SEMINARS

Title: Automatic detection of potential adverse drug events in unstructured texts: state-of-the-art and challenges

Speaker: Luca Toldo

When: 6 Dec 2012

Abstract: The European Union requires drug manufacturers to conduct adequate ongoing monitoring and benefit/risk evaluation during the post-authorization period" with the aim of ensuring that "safety hazards are minimized and the benefits of treatment maximized by appropriate action" (CIOMS reporting group, 1998). Therefore, a crucial activity after a product is launched, is the monitoring of a variety of information sources (scientific, medical, internet, others) to detect adverse potential adverse drug events. If the signals so collected are demonstrated to be due to the specific drug, then changes in the drug label must be implemented, and withdraw of the medicament is also a possible scenario. The potential adverse drug events generated with the highest quality are those resulting from the expensive "clinical trials". After the market launch, however any physician / patient / could file to the authorities a so called "spontaneous adverse events report", in highly structured format, of good quality.The scientific literature is also an important source of information about potential adverse drug events, through the so called "Case Reports" since scientists, generally speaking physicians, would report unusual cases they have encountered in their clinical practice. Most recently, also the social media has become a potential source of information about adverse drug events, although quite some discussions are ongoing about its practical use due to the generally accepted low quality (due to spamming, duplications, etc). In this talk, the state of the art in the application of automatic methods for detection of potential adverse drug events in text will be reported, including with visions for the future and limitations.

Title: FBK's Machine Translation Systems for IWSLT 2012's TED Lectures

Speaker: Nick Ruiz

When: 29 Nov 2012

Abstract: In this presentation, we will report on FBK's Machine Translation (MT) submissions for the TED talk translation tasks in the IWSLT 2012 Evaluation. In total, FBK participated in the English-French and the Arabic-, Dutch-, German-, and Turkish-English translation tasks. We focus primarily on FBK's contributions to the Arabic-, Dutch-, and Turkish-English language pairs. In addition to using fill-up combinations of phrase-tables for domain adaptation, we explore the use of corpora filtering based on cross-entropy to produce concise and accurate translation and language models. We describe challenges encountered in under-resourced languages (Turkish) and language-specific preprocessing needs.

Title: From Wikipedia to multifaceted meaning representations

Speaker: Vivi Nastase

When: 22 Nov 2012

Abstract: Many natural language processing tasks rely on some representation of lexical semantics through a word's relation with others, through dictionary definitions, or through collocates from a large corpus. Because of their different requirements, e.g. relational models of meaning rely on differentiating word senses, while distributional representations do not (cannot) make such distinctions resources that capture word meanings in such different manners are separate from one another. By using Wikipedia and exploiting its various structured/semi-structured sources of information, we built WikiNet, a multi-lingual network of concepts. WikiNet combines the three types of meaning representations mentioned above for an inventory of concepts obtained from Wikipedia. I will show how WikiNet was built, and how its combination of representations makes possible an unsupervised approach to the task of metonymy resolution.

Title: Integrated active noise control and noise reduction in hearing aids

Speaker: Romain Serizel

When: 28 Sep 2012

Abstract: In everyday life conversations and listening scenarios the desired speech signal is rarely delivered alone. The listener most commonly faces a scenario where he has to understand speech in a noisy environment. Hearing impairments, and more particularly sensorineural losses, can cause a reduction of speech understanding in noise. Therefore, in a hearing aid compensating for such kind of losses it is not sufficient to just amplify the incoming sound. Hearing aids also need to integrate algorithms that allow to discriminate between speech and noise in order to extract a desired speech from a noisy environment. A standard noise reduction scheme in general aims at maximising the signal-to-noise ratio of the signal to be fed in the hearing aid loudspeaker. This signal, however, does not reach the eardrum directly. It first has to propagate through an acoustic path and encounter some perturbations which can override the action of the noise reduction in the hearing aid. This presentation introduces an integrated active noise control and noise reduction scheme for hearing aids to tackle secondary path effects and effects of noise leakage through an open fitting.

Title: Context-aware Machine Translation for Computer Aided Translation

Speaker: Katharina Waeschle

When: 20 Sep 2012

Abstract: Most current Statistical Machine Translation (SMT) systems translate each sentence in a document in isolation. While this reduces the complexity of translating large documents, it introduces the problem that important discourse-level information is lost. The MT component of a Computer Aided Translation (CAT) system should be able to benefit from implicit user feedback collected on approved translations. We investigate the impact of an SMT system providing suggestions which are consistent with respect to the already edited segments of a document. The context information is embedded in the statistical models and allows to better disambiguate among lexical alternatives. The context-based models combine information about recurrent terms and expressions extracted during a document analysis with the corresponding chosen and confirmed translations as soon as they become available.

Title: Weighting lexical (mis-)alignments for cross-lingual textual entailment

Speaker: Miquel Esplà-Gomis

When: 19 Sep 2012

Abstract: Cross-lingual textual entailment (CLTE) is the task of deciding, given a text T and a hypothesis H in different languages, if the meaning of H can be inferred from the meaning of T. The task combines challenging problems relevant to computational semantics and machine translation, opening to interesting applications in both fields (e.g. cross-lingual question answering, and document summarization, as well as machine translation evaluation). Most previous works on CLTE addressed the task as a classification problem, building on pivoting or cross-lingual methods that compute word/n-gram matching features over T and H. The work described in this seminar addressed the task by: i) training word-alignment models to map words in T and H, and ii) exploiting different types of linguistic information to properly "weight" aligned and unaligned fragments (under the assumption that matching non-relevant words and not-matching relevant ones provides useful information to support the entailment decision process). Experiments have been carried out over the SemEval 2012 CLTE task dataset, showing significant improvements of the state of the art. This is the result of a three-months summer internship that Miquel spent in our group.

Title: Online adaptation in Computer Aided Translation

Speaker: Patrick Simianer

When: 14 Sep 2012

Abstract: MateCat will provide methods for the automatic self-correction of MT exploiting the implicit feedback of the user. The segments of text already post-edited by the user will be analysed and compared with the corresponding automatic translations by MT in order to spot the errors together with their corrections and the portions accepted by the translator.The MT models will be modified accordingly by penalizing the former, reinforcing the latter, or, more drastically, by removing the source

of errors. Although ad-hoc transformations could be similar to those for the project adaptation, the goal is here to make them very precise and consistent with the actual translator. Through this on-line adaptation, which is performed real-time and sentence-by-sentence, MT should automatically translate the following segments more and more coherently with respect to the previous ones from the point of view of user's lexical and stylistic preferences. We will present a reranking approach to tackle this online adaptation in the CAT setting.

Title: Can Machine Traslation (MT) benefit from Machine Linking (ML) and viceversa?: A Pilot Study

Speaker: Angeliki Lazaridou

When: 14 Sep 2012

Abstract: ML consists in recognizing and linking terms (mainly nouns and name entities) in text to Wikipedia and other resources of the Linked Open Data. This process can be perceived as term disambiguation and provides a framework for enriching individual terms with supplementary information. We think that this aspect of ML can be of valuable source of information for MT, especially in cases where MT fails to semantically identify terms, thus introducing translation errors. The preliminary experiments showed that ML, even in a simplistic way, has the potential to provide semantic information missing from current MT systems. Our second goal was to investigate the use of MT for providing an automatic way of evaluating ML systems, which currently relies mostly on human-annotated data the collection of which is a time-consuming and expensive process. In the experiments that we conducted, we found that the use of parallel corpora can provide a good approximation of the quality of the system.

Title: Incremental Machine Translation Adaptation

Speaker: Ondřej Plátek

When: 13 Sep 2012

Abstract: MateCAT is a EU funded project with the goal of increasing the productivity of human translators by integrating statistical machine translation (SMT) into a Computer Assisted Translation (CAT) tool. The amount of document-related data continuously grows due to the new translations supplied by the users, allowing the incremental update of the models employed by the translation engine. During the internship, several baselines for translating English into Italian were set up for the Information Technology domain, and many experiments were carried out to find out how different amounts of in-domain data from specific projects affect machine translation quality.

Title: Analysis of the Characteristics of Talk-show TV Programs

Speaker: Diego Giuliani

When: 4 Sep 2012

Abstract: We examined the content of 2 talk-show TV programs in order to better understand the challenges posed by this program genre to automatic transcription. Six talk-show episodes were first segmented, transcribed and annotated by experts. Most of the speech content was found in conversational style with a significant portion of overlapped speech, about 18%. Then, automatic speech recognition experiments were conducted showing that recognition performance on talk-show programs is much worse, 28.3% word error rate (WER), in comparison with that achieved on broadcast news programs, 10.9% WER. For talk-shows performance varied tangibly between non-overlapped speech, 21.8% WER, and overlapped speech, 58.5% WER. On clean, non-overlapped speech an 18.7% WER is achieved, this result is significantly worse than the result achieved for the dominant condition in broadcast news programs represented by clean read/planned speech from the anchormen, 7.6% WER.

Title: Separating Transliterations from Non-Transliterations in Transliteration Mining and Translation Context

Speaker: Hassan Sajjad

When: 31 Aug 2012

Abstract: Transliteration is a process of converting a word written in one script to another script in such a way that pronunciation remains almost the same. It is useful in major applications of natural language processing such as machine translation and cross language information retrieval. The transliteration system is generally built using two types of manually created resources -- hand-crafted transliteration rules and a list of transliteration pairs. These are language pairs dependent resources which are not available for all language pairs. Using transliteration mining, one can automatically extract a list of transliteration pairs from a parallel corpus. However, all the state-of-the-art transliteration mining techniques are supervised or semi-supervised and require language dependent information for training. I solved this issue by showing that transliteration mining can be done in an unsupervised fashion. The proposed method does not require any language pairs dependent resources. I incorporated transliterations to machine translation and word alignment and showed that it improves the performance of the systems.

Title: PLIS – a Probabilistic Lexical Inference System for Textual Entailment

Speaker: Eyal Shnarch

When: 31 Aug 2012

Abstract: Can your research gain from having background knowledge such as Xinjiang is in China or Interleukin-1 is a cytokine which is a protein? Is your system too, needs to bridge text understanding gaps cause by lexical variations such as astronomical and astronomer or by acronym such as NBA which means National Basketball Association? Would you like to be able to make semantic inferences such as ignition implies the existence of fire, that with migraine we feel pain and that the wordcrowd makes a reference to people? All these relations are examples of Lexical Entailment rules. Acquiring such knowledge, integrating it and estimate its validity probability is the topic of the research I will present in this talk.

Title: Joint Feature Selection in Distributed Stochastic Learning for Large-Scale Discriminative Training in SMT

Speaker: Stefan Riezler

When: 24 Aug 2012

Abstract: With a few exceptions, discriminative training in statistical machine translation (SMT) has been content with tuning weights for large feature sets on small development data. Evidence from machine learning indicates that increasing the training sample size results in better prediction. The goal of this paper is to show that this common wisdom can also be brought to bear upon SMT. We deploy local features for SCFG-based SMT that can be read off from rules at runtime, and present a learning algorithm that applies L1/L2 regularization for joint feature selection over distributed stochastic learning processes. We present experiments on learning on 1.5 million training sentences, and show significant improvements over tuning discriminative models on small development sets.

Title: A Bayesian Approach to Learning the Structure of Human Languages

Speaker: Phil Blunsom

When: 23 Aug 2012

Abstract: Grammar Induction has long been a central challenge of ComputationalLinguistics. Empirically demonstrating the ability of computationalmodels to automatically learn the syntactic structure of humanlanguages will impact upon both our understanding of how childrenlearn language, and our ability to build sophisticated languagetechnologies. In this talk I will describe our recently developedstate-of-the-art approach to syntax induction. Using hierarchicalnon-parametric Bayesian priors we have created probabilistic modelsof syntactic part-of-speech and dependency grammar that are able tointegrate information across a range of granularities. The promisingresults achieved by these models indicate that the great challenge ofGrammar Induction may not be as intractable as long thought.

Title: FrameNet extension for the Semantic Web: creation of the RDF/OWL version of the repository of senses, resource evaluation and lessons learned

Speaker: Irina Sergienya

When: 20 Aug 2012

Abstract: FrameNet is a large-scale lexical resource encoding information about frames (situations) and frame elements (roles, or participants). In the recent work by DKM/HLT researchers, the semantic description of the roles in terms of WordNet synsets was provided. We refer to this enriched representation of FrameNet roles as to the repository of the senses. My internship project concerned the representation of the repository of senses in RDF/OWL format, so as to make the resource available to the Semantic Web. In the seminar, the work conducted during the internship will be presented, with the focus on the translation of the repository to RDF/OWL, the difficulties encountered, and on the outcomes of the (preliminary) evaluation of the obtained resource.

Title: Building cross-lingual distributional similarity models from Wikipedia

Speaker: Roger Leitzke Granada

When: 20 Aug 2012

Abstract: Distributional similarity models (DSMs) assume that words that are close in meaning will occur in similar contexts. Based on this assumption, a variety of approaches (e.g. Latent Semantic Analysis) have been proposed to measure the degree of semantic similarity between two words, with countless applications in NLP (e.g. document classification, information retrieval, word sense disambiguation). Cross-lingual DSMs have the same potential for a number of re search areas where word similarity scores computed between words in different languages (e.g. cross-lingual textual entailment recognition, cross-lingual semantic textual similarity, machine translation). This project aims at developing a DSM for several European languages taking advantage of topically-related Wikipedia articles in the target languages. The outcomes of the project will contribute to several ongoing activities in all the aforementioned areas.

Title: Relation Mining in the Biomedical Domain using Entity-level Semantics

Speaker: Kateryna Tymoshenko

When: 9 Aug 2012

Abstract: This work explores the use of semantic information from background knowledge sources for the task of relation mining between medical entities such as diseases, drugs, and their functional effects/actions. We hypothesize that the semantics of medical entities, and the information about them

in different knowledge sources play an important role in determining their interactions and can thus be exploited to infer relations between these entities. We capture entities' semantics using a number of resources such as Wikipedia, UMLS Semantic Network, MEDCIN, MeSH and SNOMED. Depending on coverage and specificity of the resources, and features of interest, different classifiers are learnt. An ensemble based approach is then used to fuse together individual predictions. Using a human-curated ontology as the gold standard, the proposed approach has been used to recognize ten medical relations of interest.

We show that the proposed approach achieves substantial improvements in both coverage and performance over a distant supervision based baseline that uses sentence-level information. Finally, we also show that even a simple ensemble approach that combines all the semantic information is able to get the best coverage and performance.

Title: Three lectures on SMT

Speaker: Marcello Federico

When: 23 Jul 2012

Abstract: The course is basic and open to everyone. Summer internship students are mostly welcome.

Title: A controlled greedy supervised approach for coreference resolution on clinical text

Speaker: Faisal Chowdhury

When: 11 Jul 2012

Title: INVITED SEMINAR: Shallow discourse parsing

Speaker: Sucheta Ghosh

When: 28 Jun 2012

Abstract: Coherently related set of sentences is defined as discourse. Parsing discourse is a challenging natural language processing task. In this talk, I will describe and discuss our shallow discourse parser. The parsing architecture is based on a cascade of decisions. The parser adopts a two-stage approach where first the local constraints are applied based on conditional random field chunking, and then global constraints are used on a n-best search space. In the second stage, we experiment with different re-rankers trained on the first stage n-best parses, which are generated using lexico-syntactic local features. These re-rankers use inter-sentential or sentence-level (global), data-driven, non-grammatical features to train the system. The two-stage parser yields significant improvements over the best

performing model of discourse parser on the Penn Discourse Tree Bank (PDTB) corpus. Next to this, I will also discuss about the probable applications of this parser at least in two areas: emotion analysis and machine translation.

Title: Modified Distortion Matrices for Phrase-Based Machine Translation

Speaker: Arianna Bisazza

When: 27 Jun 2012

Abstract: We present a novel method to suggest long word reorderings to a phrase-based SMT decoder. We address language pairs where long reordering concentrates on few patterns, and use fuzzy chunk-based rules to predict likely reorderings for these phenomena. Then we use reordered n-gram LMs to rank the resulting permutations and select the n-best for translation. Finally we encode these reorderings by modifying selected entries of the distortion cost matrix, on a per-sentence basis. In this way, we expand the search space by a much finer degree than if we simply raised the distortion limit. The proposed techniques are tested on Arabic-English and German-English using well-known SMT benchmarks.

Title: Evaluating the Learning Curve of Domain Adaptive SMT Systems

Speaker: Nicola Bertoldi

When: 21 Jun 2012

Abstract: The new frontier of computer assisted translation technology is the effective integration of statistical MT within the translation workflow. In this respect, the SMT ability of incrementally learning from the translations produced by users plays a central role. A still open problem is the evaluation of SMT systems that evolve over time. In this paper, we propose a new metric for assessing the quality of an adaptive MT component that is derived from the theory of learning curves: the percentage slope.

Title: Running UIMA 24x7

Speaker: Raffaella Ventaglio - Roberto Franchini

When: 31 May 2012

Abstract: CELI has chosen UIMA as a library for the execution of linguistic analysis pipelines performed every day (24x7). In this talk we will show you how adopting UIMA can simplify the integration of new and legacy components, through the use of convention over configuration, UIMAFit library and some custom code to enable Dependency Injection / Inversion of Control.

Title: A Bayesian Model for Learning SCFGs with Discontiguous Rules

Speaker: Abby Levenberg

When: 30 May 2012

Abstract: We describe a nonparametric model and corresponding inference algorithm for learning Synchronous Context Free Grammars (SCFGs) directly from parallel text. The model employs a Pitman-Yor process prior which uses a novel base distribution over SCFG rules. Through both synthetic grammar induction and statistical machine translation experiments, we demonstrate that our model learns complex translational correspondences--- including discontinuous, many-to-many alignments---and produces competitive translation results. Further, inference is efficient and we present results on significantly larger corpora than prior work.

Title: MT Evaluation and Word Similarity metrics for Semantic Textual Similarity

Speaker: José Guilherme C. de Souza

When: 17 May 2012

Abstract: In this seminar I will present the participation of FBK in the Semantic Textual Similarity (STS) task organized within Semeval 2012. Our approach explores lexical, syntactic and semantic machine translation evaluation metrics combined with distributional and knowledge-based word similarity metrics. Our best model achieves 60.77% correlation with human judgements (Mean score) and ranked 20 out of 88 submitted runs in the Mean ranking, where the average correlation across all the sub-portions of the test set is considered.

Title: Text Generation from Semantic Web Ontologies

Speaker: Nadjet Bouayad-Agha

When: 11 May 2012

Abstract: I will present our work in developing an application for the generation of multilingual personalized environmental bulletins from an OWL-based ontology. We propose a three layer OWL-based ontology framework in which domain, domain communication and linguistic knowledge structures are clearly separated and show how a large scale instantiation of this framework in the environmental domain serves multilingual personalized Natural Language Generation. I will also take the opportunity to quickly review Natural Language Generation research involving semantic web standard representations in order to put our own present and future work into perspective.

Title: Two short talks: Hybrid Language Models for SMT & Turkish Speech Recognition

Speaker: A. Bisazza and M. Federico , A. Bisazza and R. Gretter

When: 10 May 2012

Abstract: In this paper, we address statistical machine translation of public conference talks. Modeling the style of this genre can be very challenging given the shortage of available in-domain training data. We investigate the use a hybrid LMs, where unfrequent words are mapped into classes. Hybrid LMs are used to complement word-based LMs with statistics about the language style of the talk genre. Extensive experiments comparing different settings of the hybrid LM are reported on publicly available benchmarks based on TED talks, from Arabic to English and from English to French. The proposed models show to better exploit in-domain data than conventional word-based LMs for the target language modeling component of a phrase-based statistical machine translation system.

We present an open-vocabulary Turkish news transcription system built with almost no language-specific resources. Our acoustic models are bootstrapped from those of a well trained source language (Italian), without using any Turkish transcribed data. For language modeling, we apply unsupervised word segmentation induced with a state-of-the-art technique (Creutz and Lagus, 2005) and we introduce a novel method to lexicalize suffixes and to recover their surface form in context without need of a morphological analyzer. Encouraging results obtained on a small test set are presented and discussed.

Title: Introduction to PM4

Speaker: Matteo Negri

When: 3 May 2012

Abstract: In this short seminar I will give an overview of the main functionalities of PM4, a "social" tool that the HLT unit will adopt to promote and increase internal communication. In particular, I will provide basic information on how to publish your messages, create and manage distribution lists, and properly use hashtags. This introduction is part of the work that has been carried out within the Focus Group #2 ("Internal communication in HLT").

Title: Ecological Evaluation of Persuasive Messages Using Google AdWords

Speaker: Marco Guerini

When: 12 Apr 2012

Abstract: In recent years there has been a growing interest in crowdsourcing methodologies to be used in experimental research for NLP tasks. In particular, evaluating systems and theories about persuasion is difficult to accommodate within existing frameworks. In this paper we present a new cheap and fast

methodology that allows fast experiment building and evaluation with fully-automated analysis at a low cost. The central idea is exploiting existing commercial tools for advertising on the web, such as Google AdWords, to measure message impact in an ecological setting. The paper includes a description of the approach, tips for how to use Adwords for scientific research, and results of pilot experiments on affective text variations impact, which confirm the effectiveness of the approach.

Title: Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation

Speaker: Arianna Bisazza and Nick Ruiz

When: 16 Feb 2012

Abstract: In this seminar, we present our work from the IWSLT 2011 conference, in which we compare techniques to combine diverse parallel corpora for domain-specific phrase-based SMT system training. We address a common scenario where little in-domain data is available for the task, but where large background models exist for the same language pair. In particular, we focus on phrase table fill-up: a method that effectively exploits background knowledge to improve model coverage, while preserving the more reliable information coming from the in-domain corpus. We present experiments on an emerging transcribed speech translation task on the TED talks. While performing similarly in terms of BLEU and NIST scores to the popular log-linear and linear interpolation techniques, filled-up translation models are more compact and easy to tune by minimum error training.

Title: Continuous Space Language Models

Speaker: Holger Schwenk

When: 15 Feb 2012

Abstract: Continuous Space Language Models for Speech Recognition and Statistical Machine Translation

Title: Edit hints: using machine translation to suggest the target-side words to change in computer-aided translation proposals

Speaker: Mikel L. Forcada, Miquel Esplà-Gomis and Felipe Sánchez-Martínez

When: 31 Jan 2012

Abstract: I will show how machine translation (MT) may be used to help users of computer-aided translation systems based on translation memory to identify the target words in the translation proposals that need to be changed or kept unedited. The machine translation system is used as a black

box to obtain a set of features for each target word in the translation proposals and then used by a binary classifier to determine the target words to change or keep unedited (no MT output is presented to the translator). Experiments conducted in the translation of Spanish texts into English with different corpora and a machine translation system still in development shows an accuracy above 96% for fuzzy-match scores above 70%. Results show that the parameters of the binary classifier are basically domain-independent. A comparison of this technique with a previously reported technique based on statistical word alignment shows that the accuracy of both approaches is quite similar when translating in-domain texts, whereas for out-of-domain texts the new MT-based approach achieves higher accuracy.

Title: Populating the Livememories Knowledge Store

Speaker: Bishoksan Kafle

When: 12 Sep 2011

Abstract: Bishoksan will present the activities conducted in FBK during his summer internship. Focus of the presentation is the implementation of the modules who populate the Livememories Knowledge Store. Such modules sense new available data from external data providers, possibly call appropriate Web services to (pre)process and annotate them, and finally populate the Knowledge Store with original and processed data (Resources with metadata) and annotations (Mentions and Entities).

Title: Readability Indices and Text Classification

Speaker: Ke Tran Manh

When: 1 Sep 2011

Abstract: Ke will present the implementation of the Coh-metrix, usually employed for assessing text readability, for Italian texts. He will further present some classification experiments aimed at distinguishing easy from difficult-to-read texts and at assigning a text to a readability level corresponding to Elementary, Middle and Secondary school level.

Title: Topic Adaptation for Lecture Translation

Speaker: Nick Ruiz and Marcello Federico

When: 5 Aug 2011

Abstract: This work presents a simplified approach to bilingual topic modeling for language model adaptation by combining text in the source and target language into very short documents and performing Probabilistic Latent Semantic Analysis (PLSA) during model training. During inference,

documents containing only the source language can be used to infer a full topic-word distribution on all words in the target language's vocabulary, from which we perform Minimum Discrimination Information (MDI) adaptation on a background language model (LM). We apply our approach on the English-French IWSLT 2010 TED Talk exercise, and report a 15% reduction in perplexity and relative BLEU and NIST improvements of 3% and 2.4%, respectively over a baseline only using a 5-gram background LM over the entire translation task. Our topic modeling approach is simpler to construct than its counterparts.

Title: Soft computing based feature selection for environmental sound classification

Speaker: Aamir Shakoor

When: 28 Jul 2011

Abstract: In this talk Aamir Shakoor will present his master thesis work: Environmental sound classification systems have a wide range of applications, like hearing aids devices, handhold devices and auditory protection devices. Sound classification systems typically extract features which are learnt by a classifier. Using too many features can result in reduced performance by making the learning algorithm to learn wrong models. The proper selection of features for sound classification is a non-trivial task. Soft computing based feature selection methods are not studied for environmental sound classification, whereas these methods are very promising, because these can handle uncertain information in a more efficient way, using simple set theoretic functions and because these methods are more close to perception based reasoning. Therefore this thesis investigates different feature selection methods, including soft computing based feature selection and classical information, entropy and correlation based approaches. Results of this study show that rough set neighborhood based method performs best in terms of number of features selected, recognition rate and consistency of performance. Also the resulting classification system performs robustly in presence of reverberation.

Title: Network-based Speech-to-Speech Translation Technology

Speaker: Chiori Hori

When: 20 Jul 2011

Abstract: Japan Network-based Speech-to-Speech Translation Technology

Title: Bootstrapping Arabic-Italian SMT through Comparable Texts and Pivot Translation

Speaker: Mauro Cettolo

When: 27 May 2011

Abstract: In this seminar, I describe efforts towards the development of an Arabic to Italian SMT system for the news domain. Since only very little parallel data are available for this language pair, we investigated both the exploitation of comparable corpora and pivot translation. Experimental evaluation was conducted on a new benchmark developed by extending two Arabic-to-English NIST evaluation sets. Preliminary results show potentials of both approaches with respect to performance achieved by a popular state-of-the-art Web-based translation service.

Title: Parallelization tools

Speaker: Fabio Brugnara

When: 24 Mar 2011

Abstract: In this informal seminar, I will describe some tools the ASR group is using for parallelizing jobs on the grid, as I feel that they may be of interest to other people as well.

Title: Il Senso Comune: che cos'è, perché i computer non ce l'hanno, e come possiamo insegnarglielo?

Speaker: Marco Baroni

When: 15 Mar 2011

Abstract: Presentazione nell'ambito di un incontro con le scuole superiori per il progetto "la ricerca come mestiere". The seminar will be given in Italian.

Title: Dependency Tree-let Translation: Syntactically Informed Phrasal SMT (English to Hindi)

Speaker: Prashant Mathur

When: 4 Mar 2011

Abstract: In this talk I will present an overview of a Statistical Machine Translation system which incorporates syntactic information along with phrasal translation. The system uses a source dependency parser and a word aligned parallel corpus. Tree-let (connected sub graph of a dependency tree) translation pairs which are essentially phrase translation pairs, are extracted by projecting dependencies from source sentence onto target sentence. These phrase translations are used for training and obtaining useful models such as Channel model and Order model. These models along with a proper decoder provide us with a machine translation system equipped with both statistical and linguistic knowledge.

Title: Arabic Morphological Parsing Revisited

Speaker: Suhel Jaber

When: 28 Feb 2011

Abstract: We present a new approach to the description of Arabic morphology with 2-tape finite state transducers, based on a particular and systematic use of the operation of composition in a way that allows for incremental substitutions of concatenated lexical morpheme specifications with their surface realization for non-concatenative processes (the case of Arabic templatic interdigitation and non-templatic circumfixation).
We argue that: 1. the method of incremental substitutions through compositions allows for an elegant description of all main morphological processes present in natural languages including non-concatenative ones in strict finite-state terms, without the need to resort to extensions of any sort; 2. our approach allows for the most logical encoding of every kind of dependency, including traditional long-distance ones (mutual exclusiveness), circumfixations and idiosyncratic root and pattern combinations; 3. a smart usage of composition such as ours allows for the creation of a same system that can be easily accomodated to fulfil the duties of both a stemmer (or lexicon development tool) and a full-fledged lexical transducer.

Title: Uncertainty language in scientific texts based on a corpus with 167 years of full text biomedical research publications

Speaker: Ricardo Pietrobon

When: 11 Feb 2011

Abstract: In this presentation Dr. Pietrobon will provide an overview of the work resulting from the collaboration between the group on Psychology of Communication from the University of Macerata and the group focusing on the study of Research Processes from Duke University. The presentation will cover an (1) overview of Duke University and both research groups from Macerata and Duke, (2) an overview of interdisciplinary informatics methods currently available in the Duke group, including Semantic Web technologies and Linked Open Data as well as information retrieval and semi-automated ontology maintenance and instantiation, (3) psychological theories of uncertainty and hedging in scientific language based on the theory of perceptual and cognitive linguistics indicators (PaCLIs), and (4) a brainstorming session regarding the potential of collaboration toward joint funding proposals with the EU and the USA.

Title: An overview to the TAC Knowledge Base Population Track

Speaker: Silvana M. Bernaola Biggio

When: 3 Feb 2011

Abstract: In 2009 Knowledge Base Population (KBP)was proposed, for the first time, as one of the tracks of the Text Analysis Conference (TAC). In 2010 the number of participants almost duplicated the one of the first edition (from 13 to 23), showing an increasing interest of the NLP community. The task focuses on extracting information about entities to expand an existing knowledge base. Participants are provided with a large corpus and a knowledge base, and can take part in at least one of the two KBP tasks: 1) entity linking, 2) slot filling. In the former, the system has to associate an entity mention to the corresponding knowledge base entry. In the latter, the systems uses the corpus to detect attributes of specified entities of the knowledge base. In this talk, I will give an overview of KBP and discuss the current approaches and some open issues.

Title: PhD Thesis Proposals

Speaker: Arianna Bisazza, Faisal Chowdhury, Christian Hardmeier, Gozde Ozbal, Stefan F. Rigo

When: 20 Dec 2010

Abstract: This is an internal presentation session in which PhD students in the HLT group, who are going to apply for the qualifying exam, will present their research topic and work plan.

Title: Mining Parallel Fragments from Comparable Texts

Speaker: Mauro Cettolo

When: 30 Nov 2010

Abstract: We propose a novel method for exploiting comparable documents to generate parallel data for machine translation. First, each source document is paired to each sentence of the corresponding target document; second, partial phrase alignments are computed within the paired texts; finally, fragment pairs across linked phrase-pairs are extracted. The algorithm has been tested on two recent challenging news translation tasks. Results show that mining for parallel fragments is more effective than mining for parallel sentences, and that comparable in-domain texts can be more valuable than parallel out-of-domain texts.

Title: Modelling Pronominal Anaphora in Statistical Machine Translation

Speaker: Christian Hardmeier

When: 29 Nov 2010

Abstract: Current Statistical Machine Translation (SMT) systems translate texts sentence by sentence without considering any cross-sentential context. Assuming independence between sentences makes it difficult to take certain translation decisions when the necessary information cannot be determined locally. We argue for the necessity to include cross-sentence dependencies in SMT. As a case in point, we study the problem of pronominal anaphora translation by manually evaluating German-English SMT output. We then present a word dependency model for SMT, which can represent links between word pairs in the same or in different sentences. We use this model to integrate the output of a coreference resolution system into English-German SMT with a view to improving the translation of anaphoric pronouns.

Title: RTE-6@TAC2010: the Sixth Recognizing Textual Entailment Challenge

Speaker: Luisa Bentivogli

When: 12 Nov 2010

Abstract: This talk presents an overview of the Sixth Recognizing Textual Entailment Challenge. The Recognizing Textual Entailment (RTE) task consists in developing a system that, given two text fragments, can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other text. Proposed for the first time in 2005, the Challenge is now at its sixth round. After the first three PASCAL RTE Challenges held in Europe, in 2008 RTE became a track at the Text Analysis Conference (TAC), bringing it together with communities working on NLP applications. The interaction has provided the opportunity to apply RTE systems to specific application settings and move them towards more realistic scenarios. In RTE-6 major innovations were introduced. First, the traditional Main Task was replaced by a new task in which Textual Entailment is performed on a real corpus: given a corpus and a hypothesis H, RTE systems are required to identify all the corpus sentences that entail the H. The task was set up in the Summarization setting, with the goal of analyzing the potential impact of Textual Entailment on a real NLP application. A Novelty Detection Subtask was also proposed, based on the Main Task and specifically aimed at detecting novel information in the corpus. To continue the effort of testing RTE in NLP applications, a Knowledge Base Population (KBP) Validation Task was proposed as a Pilot within RTE-6. This task was based on the output of systems participating in the TAC KBP Slot Filling Task and was meant to show the potential utility of RTE systems for Knowledge Base Population. Finally, following the positive experience of the fifth challenge, all the participants in the RTE-6 Main Task were asked to carry out ablation tests on the knowledge resources used by their systems, with the aim of studying the relevance of such resources in recognizing Textual Entailment. This year ablation tests were also extended to tools, such as parsers, coreference resolvers, Named Entity recognizers.

Title: Finding, Linking and Organizing Resources with Linked Data & Natural Language Processing

Speaker: Paul Buitelaar

When: 4 Nov 2010

Abstract: The talk will be concerned with the use of 'Linked Open Data' and natural language processing in an object-driven approach to the discovery and organization of information resources and links between them. The Linked Open Data initiative aims at making available (all) freely available information in a meaningful way, i.e., in a standardized format and with formal links to semantic metadata. To illustrate the potential and current limits of Linked Open Data I will give a brief overview of some recent applications and tools that are currently under development at DERI. An interesting research direction that emerges from this is the combined use of and interaction between Linked Open Data and natural language processing as will be discussed in this talk.

Title: Information Extraction from Medical Texts

Speaker: Pierre Zweigenbaum

When: 22 Oct 2010

Abstract: I will describe recent work at LIMSI on the extraction of medical entities and relations from clinical texts (texts in patient records, e.g. discharge summaries) in English and in French and from the scientific literature (MEDLINE abstracts and full-text articles). This work uses both machine-learning approaches and expert-based development of lexicons and patterns. It was developed in the context of the i2b2 2009 and 2010 challenges, the French-ANR-funded Akenaton project, and the French-German Quaero project. The targeted information includes: - the extraction of prescription information: drug, dosage, mode of administration, etc.; - the detection of medical problems, tests, treatments; - hedging phenomena: asserted fact, negated fact, hypothetical, etc.; - relations between medical problems, tests, treatments.

Title: Information Extraction from Medical Texts

Speaker: Pierre Zweigenbaum

When: 12 Oct 2010

Abstract: I will describe recent work at LIMSI on the extraction of medical entities and relations from clinical texts (texts in patient records, e.g. discharge summaries) in English and in French and from the scientific literature (MEDLINE abstracts and full-text articles). This work uses both machine-learning approaches and expert-based development of lexicons and patterns. It was developed in the context of the i2b2 2009 and 2010 challenges, the French-ANR-funded Akenaton project, and the French-German Quaero project. The targeted information includes: the extraction of prescription information: drug, dosage, mode of administration, etc.; the detection of medical problems, tests, treatments; hedging

phenomena: asserted fact, negated fact, hypothetical, etc.; relations between medical problems, tests, treatme.

Title: Credal classifiers and their empirical evaluation

Speaker: Kristina Gulordava

When: 8 Sep 2010

Abstract: Credal classifiers are recently introduced robust classifiers that can have a set of classes as an output. This new approach for classification arises a problem of empirical evaluation of credal classifiers as traditional metrics such as predictive accuracy cannot be directly applied. In this work we propose a novel approach to evaluate and compare credal classifiers based on discounted accuracy met- ric and utility functions used on top of it. The overall process is separated in two steps where the first step introduces a framework that allows to derive the discounted accuracy measure as a risk-neutral measure and the second step is used to incorporate possible risk-aversion of assessor, i.e., a desire to have more reliable classification. We present two formal frameworks that formalize these steps and yield the uniform approach to evaluate credal classifiers. We then ap- ply this approach in experimental study on a number of data sets to verify our theoretical observations.

Title: Webservice Integration in LiveMemories with ServiceMix

Speaker: Alessio Giori

When: 22 Jul 2010

Abstract: This talk focuses on the integration of the RESTful Web Services developed within the LiveMemories project using Service Mix as Enterprice Service Bus. First, Web Services and their main features will be described. Then, Enterprice Service Bus (ESB) as integration framework will be introduced as well as Service Mix, the Apache implementation of ESB. Finally, the integration activity and the developed applications will be described with some details.

Title: “Bio-entity Identification in Biomedical Literature: Tasks and Trends”

Speaker: Faisal Mahbub Chowdhury

When: 27 May 2010

Abstract: The massive growth of the volume of biomedical literature has made the development of biomedical text mining solutions indispensable. Biomedical texts present characteristics which make them different from texts such as newspaper articles. Ranging from the vocabulary to the valency of

verbs, these texts are inherently complex. One of the essential requirements for any text mining application is the ability to identify relevant entities, i.e. named entity recognition. There have been many efforts in the BioNLP community for developing accurate biomedical named entity recognition (BNER) systems. In this talk, we will discuss the trends in BNER. The talk will also describe the work on BNER carried on in the HLT unit in the context of the eOnco project and our participation in CALBC (Collaborative Annotation of a Large-Scale Biomedical Corpus) Challenge I, recently organised by the European Bioinformatics Institute (EMBL-EBI), U.K., where our system obtained encouraging results.

Title: Multilingual Event Extraction and Semi-automatic acquisition of related resources

Speaker: Hristo Tanev

When: 24 May 2010

Abstract: The never-stopping stream of online content in different languages increases the importance of efficient multilingual approaches to information extraction. In this talk I will present a nearly real time multilingual event extraction system in the domain of crises and security. Event extraction is a higher-level information extraction task whose purpose is to automatically identify events in free texts and to extract information about the type and the participants in these events. In this talk, I will describe the overall structure of an event extraction system from online news. I will also describe several weakly supervised multilingual algorithms for automatic acquisition of domain specific lexica from un-annotated text corpora. These algorithms were used to deploy the event extraction system in several languages.

Title: The new architecture of the FBK cluster (i.e. how to use the cluster from June 2010)

Speaker: Roldano Cattoni

When: 6 May 2010

Abstract: The FBK cluster is growing, with a continuous increase of machines, storage servers, users and groups of users. In order to offer a stable and efficient computational environment, a group of technical people (gsc + 1/2 members from the units using the cluster) has defined a new architecture, taking inspiration from standard clusters in the world. The set-up of new architecture is tentatively scheduled for end of May, begin of June. The features of the new architecture and the policy for its usage will be described with some details.

Title: Exploiting many facets of Wikipedia for building a very large-scale multi-lingual concept network

Speaker: Vivi Nastase

When: 27 Apr 2010

Abstract: Wikipedia is being built by people, and for people. While no structure has been imposed on it, it has emerged in various forms through the continuous editing process from numerous users. In this talk I will show how we exploit some of these obvious and less obvious facets of Wikipedia for building a very large scale multi-lingual concept network.

Title: Entity Mention Detection

Speaker: Silvana Bernaola

When: 16 Apr 2010

Abstract: The Entity Mention Detection Task consists on the segmentation of a text into segments which are then annotated as part of a mention or not. It is basically a classification problem where the possible classes correspond to the type of mentions (e.g. Person, Organization, etc.) and the outside class (the text does not correspond to any type of mention). In this work, It will be presented the Entity Mention Detection System for italian language which has participated on the EVALITA 2009 evaluation campaign. The system detects mentions and classifies them from the syntactic point of view i.e. proper name (NAM), nominal name (NOM) or pronominal name (PRO), as well as from the semantic point of view, in this case, if the mention corresponds to a Person (PER), Organization (ORG), Location (LOC) or Geo-Political entity (GEO). I.e. "Vladimir Putin" (NAM-PER), "paese" (NOM-GPE), "piazza" (NOM-LOC), etc. Also, it will be showed how the EMD system is applied into the liveMemories project and in the Italian wikipedia; finally, a brief explanation about the web service version of the system will be said.

Title: A clustering approach to domain adaptation for SMT

Speaker: Ioannis Klasinas

When: 18 Mar 2010

Abstract: Statistical Machine Translation systems' performance is dependent on the relation between the train and test corpora. If they are not of the same domain, translation quality is quite low. It is not, however, always possible to find in-domain training corpora; on the other hand, creating them is generally a time/money costly process. In this work the above situation is addressed with a clustering technique. Training corpora are clustered, and cluster specific models are trained and used to translate the test set. The motivation is that clustering results in more coherent subsets of the corpus, which can provide more accurate estimates for domain specific models. Afterwards, each part of the test set is translated using the most relevant models. Results are reported on French to English translation, using the Europarl and News corpora for training.

Title: On the Maximalization of the witness sets in Independent Set readings

Speaker: Livio Robaldo

When: 19 Feb 2010

Abstract: Among the readings available for NL sentences, those where two or more sets of entities are independent of one another - termed here as Independent Set readings - are particularly challenging. In the literature, examples of those readings are known as Collective and Cumulative readings. A new logical framework for NL quantification, based on Generalized Quantifiers, Skolem-like functional dependencies, and Maximality of the involved sets of entities is proposed. The framework seems to adequately deal with both Independent Set readings and standard linear readings in a scalable, natural, and uniform fashion.

Title: Morphological Pre-Processing for Turkish-to-English Statistical Machine Translation

Speaker: Arianna Bisazza

When: 21 Jan 2010

Abstract: Morphology plays a fundamental role in any NLP application involving agglutinative languages. This is particularly true for statistical machine translation (SMT) from Turkish into English, because of the severe mismatch between word formation mechanisms of the two languages. We approached this problem through morphological segmentation of Turkish, by taking advantage of linguistic knowledge of both the source and target languages. In particular we focused on the comparison of different segmentation rule sets in order to find an effective preprocessing scheme for the Turkish-English task organized by the IWSLT09 workshop. By minimizing differences between lexical granularities of source and target languages, we could produce more refined alignments and a better modeling of the translation task, which resulted in a considerable improvement of the translation quality. This work shows how a specific linguistic preprocessing can benefit a purely statistics-based, language-independent NLP application like SMT.

Title: Discriminative Spoken keyword Detection

Speaker: Joseph Keshet

When: 11 Dec 2009

Abstract: The current state-of-the-art automatic speech recognizers are mostly based on hidden Markov models (HMMs). Despite their popularity, HMM- based approaches have several known drawbacks such as training objective which is not aimed at optimizing the evaluation objective. We proposes a new approach for spoken keyword spotting, which is based on large margin and kernel methods rather than on HMMs. Unlike previous approaches, the proposed method employs a discriminative learning procedure, in which the learning phase aims at achieving a high area under the ROC curve, as this

quantity is the most common measure to evaluate keyword spotters. The keyword spotter we devise is based on mapping the input acoustic representation of the speech utterance along with the target keyword into a vector space. Building on techniques used for large margin and kernel methods for predicting whole sequences, our keyword spotter distills to a classifier in this vector-space, which separates speech utterances in which the keyword is uttered from speech utterances in which the keyword is not uttered. We describe a simple iterative algorithm for training the keyword spotter and discuss its formal properties, showing theoretically that it attains high area under the ROC curve. Experimental results suggest that on variety standard speech recognition datasets our discriminative system outperforms the conventional context-independent HMM-based system.

Title: People as Content

Speaker: Anton Nijholt

When: 10 Dec 2009

Abstract: Sensor-equipped environments are able to detect, interpret and anticipate our intentions and feelings. On the one hand this allows more natural interaction between humans and intelligent environments that support human activity, on the other hand it allows these environments to collect more information about their human partners than they may find desirable. Environments collect our lives, environments process our lives. In human-human interaction there are situations where it is quite acceptable or even desirable that part of the intentions and feelings of an interacting partner remains hidden for the other. This can happen in everyday life, but also in sports and entertainment. Non-cooperation is often more natural than cooperation. We will discuss the many useful uses of non-cooperative behavior, both from the point of view of the smart environment and from the point of view of its human partners.

Title: Towards combining statistical and symbolic learning: a kernel approach

Speaker: Andrea Passerini

When: 18 Nov 2009

Abstract: Symbolic and statistical approaches to learning have rather opposite characteristics: the former relies on an expressive and structured representation of the domain at hand and aims at producing interpretable models of the underlying concepts; the latter aims at maximizing the predictive perfomance and robustness of the learner, building on a sound generalization theory, and trades interpretability for effectiveness of the learned models. Statistical relational learning is a recent research field trying to combine the advantages of the two approaches. We take a kernel viewpoint and develop a number of algorithms where the kernel acts as an interface between a logical representation of the domain and a statistical learner. Proof tree kernels define the similarity between instances as similarity

between the proofs of logical predicates they satisfy. kFOIL defines features as the truth value of clauses dynamically generated by a greedy search alogorithm. Declarative kernels rely on an axiomatic theory in order to decompose entities into parts and express relationships between them. We discuss mereotopological kernels as well as possible extensions including temporal relationships as those defined in interval temporal logic.

Title: RTE-5@TAC2009: the Fifth Recognizing Textual Entailment Challenge ,

Speaker: Luisa Bentivogli and Milen Kouylekov

When: 12 Nov 2009

Abstract: This talk presents an overview of the Fifth Recognizing Textual Entailment Challenge. The Recognizing Textual Entailment (RTE) task consists in developing a system that, given two text fragments, can determine whether the meaning of one text is entailed, i.e., can be inferred, from the other text. Proposed for the first time in 2005, the Challenge is now at its fifth edition. Following the positive experience of the last campaign, RTE-5 was proposed for the second time as a track at the Text Analysis Conference (TAC) organized by NIST. RTE-5 presented a mixture of innovation and continuity with the previous competitions. Besides the traditional Main Task, a Pilot Search Task was also proposed, consisting of finding all the sentences in a set of documents that entail a given hypothesis. The two tasks were aimed on the one hand at allowing new and old participants to test their systems against the classic RTE task setting, and on the other to keep the interest of the research community high by introducing a more realistic scenario, where textual entailment recognition is performed on a real text corpus. Another important initiative was proposed in the context of this year's campaign. In order to study the relevance of knowledge resources in recognizing textual entailment, participants were required to perform ablation tests on all major knowledge resources employed by their systems. The results of this experiment represent a first step towards the definition of a new pilot task focused on knowledge resource evaluation, which will be proposed in the next campaign.

Title: EDITS: An Open Source Framework for Recognizing Textual Entailment

Speaker: Luisa Bentivogli and Milen Kouylekov

When: 12 Nov 2009

Abstract: This talk is about FBK's participation in the RTE 5 Evaluation Campaign, in the main (two-way classification) and pilot task, using EDITS (Edit Distance Textual Entailment Suite) package, the first freely available open source RTE software, with different configurations. The main sources of knowledge used, the different configurations, and the achieved results will be described, together with ablation tests representing a preliminary analysis of the actual contribution of different resources to the RTE task.

Title: Peculiarita' riscontrate in corpora vocali multicanale: Euronews e Jumas

Speaker: Elena Bresolin, Stefania Tabarelli de Fatis, Roberto Gretter

When: 30 Oct 2009

Abstract: Da sempre, nell'ambito del riconoscimento del parlato si utilizzano corpora vocali. Normalmente si tratta di corpora costituiti da segnali mono, nei quali preferibilmente un unico parlatore pronuncia una o piu' frasi. Recentemente abbiamo avuto a che fare con corpora costituiti da diversi canali audio paralleli, che possono essere interessanti per molti aspetti. In particolare:

- Euronews e' un canale satellitare che trasmette notizie simultaneamente in 8 lingue - un unico video, 8 canali audio in cui le notizie sono allineate temporalmente.

- Jumas e' un progetto europeo che affronta la trascrizione di procedimenti giudiziari, nel quale 4 canali audio sono associati a 4 attori differenti: giudice, avvocato, pubblico ministero, teste. Durante il seminario verranno illustrati alcuni problemi che si incontrano nel gestire dati di questo tipo.

Title: Statistical Parsing of Italian

Speaker: Alberto Lavelli

When: 29 Oct 2009

Abstract: In the seminar I will describe the past and current activities related to the application of statistical parsing techniques to Italian. The work started in 2003 with the first preliminary experiments on the Italian Syntactic-Semantic Treebank (ISST), the only Italian treebank available at the time, and proceeded later with further experiments on the Turin University Treebank (TUT). More recently, we participated in the EVALITA evaluation campaigns, in 2007 for constituency parsing only and in 2009 both for constituency and for dependency parsing. In 2009, we obtained the best result for constituency parsing (with a substantial improvement with respect to the 2007 results) and we were among the best systems for dependency parsing.

Title: Natural Reasoning for Natural Language Processing

Speaker: Roberto Garigliano

When: 22 Oct 2009

Abstract: The talk is in the logic/KB tradition of NLP, so issues of statistical learning will be touched upon only marginally. The author discusses the problems which arise from traditional forms of knowledge representation and inferencing in NLP, such as rigid ontologies, reduction to primitives, normal forms for reasoning, material logic, reasoning by contradiction and plausible reasoning. These issues create

problems of naturalness, which in turn become practical difficulties in rule development, maintenance, portability, learning and ability to produce explanations. In the final part of the seminar, the author will show some aspects of his team's most recent work, SenseGraph, which embeds their approach to tackling these issues.

Title: Two internship projects

Speaker: Elena Cabrio and Sara Tonelli

When: 9 Oct 2009

Abstract: This talk will present two internship projects: one was carried out by Elena Cabrio at the Xerox Research Center Europe in Grenoble for a period of six months, and the other was a two-month summer internship done by Sara Tonelli at the University of Pennsylvania. Goal of the internship at XRCE was to apply the Textual Entailment approach to support document creation, in particular to support Xerox internal pre-sales services in writing responses to Requests For Proposal. The presentation will describe the work done in this direction exploiting the EDITS system, and the attempts of improving the current solution basing on Xerox Incremental Parser. The internship at UPenn was aimed at the study of the Penn Discourse Treebank paradigm and its adaptation to a corpus of spontaneous dialogs in Italian developed in the framework of the LUNA project. The presentation will briefly introduce the Penn Discourse Treebank and some preliminary results obtained on the Italian corpus. The talks will mainly focus on the practical organization of internships. This meeting is intended to be an occasion to present positive experiences and propose a "best practice" for future internships in the HLT group.

Title: Wikipedia as Frame Information Repository

Speaker: Claudio Giuliano

When: 30 Jul 2009

Abstract: In this paper, we address the issue of automatic extending lexical resources by exploiting existing knowledge repositories. In particular, we deal with the new task of linking FrameNet and Wikipedia using a word sense disambiguation system that, for a given pair frame - lexical unit (F, l), finds the Wikipage that best expresses the the meaning of l. The mapping can be exploited to straightforwardly acquire new example sentences and new lexical units, both for English and for all languages available in Wikipedia. In this way, it is possible to easily acquire good-quality data as a starting point for the creation of FrameNet in new languages. The evaluation reported both for the monolingual and the multilingual expansion of FrameNet shows that the approach is promising.

Title: Automatic Cost Estimation for Tree Edit Distance: with Focus on Recognizing Textual Entailment

Speaker: Yashar Mehdad

When: 23 Jul 2009

Abstract: This seminar introduces a new method to improve tree edit distance approach to textual entailment recognition, using particle swarm optimization. Currently, one of the main constraints of recognizing textual entailment using tree edit distance is to tune the cost of edit operations, which is a difficult and challenging task in dealing with the entailment problem and datasets. We tried to estimate the cost of edit operations in tree edit distance algorithm automatically, in order to improve the results for textual entailment. Automatically estimating the optimal values of the cost operations over all Recognizing Textual Entailment (RTE) development datasets, we will show a significant enhancement in accuracy obtained on the test sets. Moreover, we demonstrate that exploring and studying the estimated cost of operations, could be interesting from a linguistics point of view. In addition, this approach could be used to any tree edit distance based task.

Title: Emotion Recognition Methods for the Judicial Domain

Speaker: Elisabetta Fersini

When: 16 Jul 2009

Abstract: Affective Computing, and in particular emotion recognition methods, have gained increasing attention in recent years due to their wide range of applications such as talking toys, call centers, e-learning platforms, intelligent interfaces and others. A new challenging area, in which emotion recognition could find an interesting application, is related to the judicial domain. In this talk we will tackle the problem of finding the predictive model that, with respect to courtroom debates characteristics, is able to produce the optimal recognition performance by considering audio features extracted by speaker utterance. Both static approaches - such as Bayesian Networks, K-NN and Support Vectors Machine - and dynamic classification models - such as HMMs and moving window-based approaches - will be considered.

Title: Which place is this place? An overview of toponym disambiguation

Speaker: Davide Buscaldi

When: 2 Jul 2009

Abstract: Lexical ambiguity and its relationship to Information Retrieval (IR) has been object of many studies in the past decade. One of the most debated issues is whether Word Sense Disambiguation (WSD) is useful to IR or not, and to which extent. Some studies concluded that in certain retrieval tasks, and in some restricted domain, effective WSD can improve the results. My hypothesis is that Geographic Information Retrieval (GIR) could represent such a task. The ambiguity of place names, or toponym

ambiguity, is an important issue in GIR, representing the direct or indirect reason of many retrieval errors. In this talk I will give an overview of the Toponym Disambiguation (TD) task, its state of the art and the problems encountered in the evaluation of TD methods. I will also describe the work I'm currently carrying out at FBK on the "L'Adige" news corpus. This corpus presents some features that makes the task more interesting, such as its localization and the ambiguities of street names, which raise the overall degree of ambiguity of the toponyms in the collection.

Title: Towards Interactive Question Answering: An Ontology Based Approach

Speaker: Manuela Speranza

When: 25 Jun 2009

Abstract: The ability to provide both rich and natural answers with respect to a given question and clear explanations for failures is a crucial aspect for a future generation of Question Answering systems able to interact with a user. In this talk I will describe an ontology-based approach developed within the QALL-ME project to represent the structure of question-answer pairs, as the above mentioned abilities require a deep analysis of the content of both the question and the answer. The approach is domain- and language-independent and can be assumed as a general framework for both Open Domain QA and Natural Language Interfaces to Databases. It is based on the definition of an Interactive Question Answer (IQA) ontology that captures significant aspects of interaction and uses dialogue templates (based on the IQA ontology), which can model natural dialogues with a user.

Title: Multilinguality and Digital Libraries

Speaker: Nicola Ferro

When: 18 Jun 2009

Abstract: Digital libraries are becoming increasingly complex and they need to satisfy user needs and carry out tasks that are getting more and more complicated. The amount of information managed by such systems, its heterogeneity and variety, and the demand for an insightful access to it are key challenges in the present research agenda. In this context, where multilinguality and multicultural aspects strongly interact with the offered functionalities, multilingual information access systems cannot simply be inserted as ?black boxes? but, on the contrary, they need to effectively interact with the different components of such more complex systems in order to exploit the peculiar features of the managed information resources and the offered functionalities, as well as to effectively impact on the different document workflows supported. The talk will discuss some key issues in the current research scenario about the relationships among multilinguality, quality, and interoperability in digital libraries and will propose some points for further discussion.

Title: Syntax-Based Reordering Model for Statistical Machine Translation

Speaker: Maxim Khalilov

When: 11 Jun 2009

Abstract: Significant improvements have been achieved in machine translation over the past few years. It is mostly motivated by appearance of statistical machine translation (SMT) technology that is currently considered as the best way to do automatic translation of natural languages. This talk focuses on a syntax-based approach to handle the fundamental problem of word ordering for SMT exploiting syntactic representations of source and target texts. The talk begins with the existing idea of taking reordering rules automatically derived through a syntactically augmented alignment of source and target texts. A new approach to hierarchically extract reordering patterns is then proposed. A set of extracted reordering rules is applied in a preprocessing step before translation to make the source sentence structurally more like the target. We evaluate the proposed approach, combined with a POS lattice-based decoding, on the Arabic-to-English and Chinese-to-English translation tasks. Furthermore, a brief introduction to an N-gram-based approach to SMT will be given, along with analysis of major differences between N-gram-based and phrase-based approaches to SMT.

Title: L'importanza dell'analisi della struttura frasale per un sistema di traduzione automatica Italiano-LIS

Speaker: Cristiano Chesi

When: 4 Jun 2009

Abstract: In questo talk verrà descritta una strategia di traduzione tranfer-based che richiede un'analisi articolata della struttura frasale (sia in termini di costituenti che di dipendenze). Per spiegare come ottenere una tale analisi verrà presentato un parser ibrido che impiega sia "regole" (basate su grammatiche di tipo (Phase-based) Minimalist Grammars, Stabler 1997, Chesi 2007) che indizi statistici (basati su corpora allineati del tipo descritto in Chesi et al. 2008) per risolvere l'ambiguità ai vari livelli (PoS tagging, PP attachment e, potenzialmente, individuazione del corretto synset associato al costituente da tradurre).

Title: Syntax-Based Reordering Model for Statistical Machine Translation

Speaker: Maxim Khalilov

When: 28 May 2009

Abstract: Significant improvements have been achieved in machine translation over the past few years. It is mostly motivated by appearance of statistical machine translation (SMT) technology that is currently considered as the best way to do automatic translation of natural languages. This talk focuses on a

syntax-based approach to handle the fundamental problem of word ordering for SMT exploiting syntactic representations of source and target texts. The talk begins with the existing idea of taking reordering rules automatically derived through a syntactically augmented alignment of source and target texts. A new approach to hierarchically extract reordering patterns is then proposed. A set of extracted reordering rules is applied in a preprocessing step before translation to make the source sentence structurally more like the target. We evaluate the proposed approach, combined with a POS lattice-based decoding, on the Arabic-to-English and Chinese-to-English translation tasks. Furthermore, a brief introduction to an N-gram-based approach to SMT will be given, along with analysis of major differences between N-gram-based and phrase-based approaches to SMT.

Title: Automatic Evaluation of Machine Translation: the BLEU and the METEOR metrics

Speaker: Mauro Cettolo

When: 21 May 2009

Abstract: Machine translation, as well as any HLT model/component/system, needs to be evaluated. The quality of a translation can be evaluated either by humans or by means of automatic metrics. Since the quality of a translation is inherently subjective, there is no objective or quantifiable "good". Therefore, the task for any metric is to assign scores of quality in such a way that they well correlate with human judgments. The BLEU score, introduced by IBM people in 2002, is one of the first metrics to report high correlation with human judgments of quality and remains one of the most popular. Since then, many other metrics have been proposed to overcome the BLEU limitations. Among them, the METEOR includes the synonymy matching, based on WordNet, a feature not found in other metrics where the matching is only on the exact word form. In this talk, I will describe in some details the BLEU and the METEOR scores. Since METEOR currently supports evaluation of MT outputs in English, French, German, Spanish and Czech but not in Italian, I also hope to find a volunteer interested in working with us to the extension of METEOR to the Italian language.

Title: Hidden Variable Transforms of Dependency Grammars for Parsing

Speaker: Gabriele Antonio Musillo

When: 14 May 2009

Abstract: Recent work in parsing probabilistic context-free grammars for natural languages has investigated partially supervised techniques to induce hidden grammatical representations that are finer-grained than those that can be read off the parsed sentences in treebanks. This talk presents extensions of such grammar induction techniques to dependency grammars. Our extensions rely on transformations of dependency grammars into efficiently parsable context-free grammars annotated with hidden symbols. Because dependency grammars are reduced to context-free grammars, any

decoding or learning algorithm developed for probabilistic context-free grammars can in principle be applied to them. Specifically, we use the Inside-Outside algorithm to learn transformed dependency grammars annotated with hidden symbols. What distinguishes our work from most previous work on dependency parsing is that our models are not lexicalised. Our models are instead decorated with hidden symbols that are designed to capture both lexical and structural information relevant to accurate dependency parsing without having to rely on any explicit supervision. Our best unlexicalised grammar achieves an accuracy of 88% on the Penn Treebank data set, that represents the state-of-the-art performance with respect to previously reported results on unlexicalised dependency parsing. Such performance shows that our unlexicalised models are able to capture both lexical and structural information that is relevant to parsing accuracy and suggests that we should reassess the relevance of massive lexicalisation to dependency parsing.

Title: NLP Resources for Computational Humor

Speaker: Alessandro Valitutti

When: 7 May 2009

Abstract: In the last years there is an increasing interest in the use of technology for creative, entertaining, persuasive, or emotional applications. In this context, computational humor prototypes were developed in order to exploit the generation and the recognition of funny texts. Exploitation of NLP resources can be crucial for the improving of humorous systems, because they allow us to play with user expectations and linguistic ambiguity. In this talk a specific strategy of humor generation, consisting in the lexical variation of familiar expressions, will be introduced. Then the result of a pilot study about the correlation of text funniness and lexical properties will be presented. Finally, some observations on the notion of ambiguity in NLP systems will be proposed as a way for designing a more general typology of humorous tools.

Title: Practical Frame Semantics for Natural Language Systems

Speaker: Bonaventura Coppola

When: 30 Apr 2009

Abstract: The dissertation reports on extensive work about implementation, evaluation, and application of Fillmore's Frame Semantics to automatic natural language understanding. In very recent years, Frame Semantics theory has been gaining increasing popularity in natural language processing research for its desirable balance between two critical requirements. In fact, while providing sufficient theoretical expressiveness for capturing language meaning beyond many surface variability phenomena, at the same time it allows for robust, effective implementation through statistical machine learning techniques. The emphasis of this doctoral research is on two practical aspects of Frame Semantics. First,

special attention is given to empirical applicability in difficult experimental settings as those with small and/or noisy learning data sets. Second, careful architectural design and system engineering are applied in order to effectively manage the complexity of the underlying machine learning models. Accordingly, the most relevant dissertation's achievements are: 1) realization of automatic, robust, multi-language frame-based Shallow Semantic Parsing of free text, 2) successful evaluation in the standard setting of the Berkeley FrameNet project, 3) exploitation of the above results in real world applied scenarios.These include spoken dialog understanding, cross-language and cross-domain portability of frame-based language analysis, semi-automatic development of FrameNet-like resources in languages different than English, and knowledge acquisition through frame-based domain ontology learning. The parser implementation is oriented to flexibility and scalability, though keeping computation time controlled. The resulting features include multi-language and multi-domain portability, transparent caching and parallel processing, and a mechanism for effective information sharing across different learning models. The presentation also includes a live demonstration of the first Italian frame-based shallow semantic parser, currently operating on a specific user domain.

Title: Wikipedia mining

Speaker: Kateryna Tymoshenko

When: 23 Apr 2009

Abstract: Wikipedia is the largest encyclopedia ever existing. It is updated daily by millions of contributors adding and refining articles in more than 260 languages. Moreover, due to its special structure Wikipedia is a valuable resource for a number of human language technology tasks. The seminar will be organized as follows. Firstly, the degree of Wikipedia's accuracy and its most important features will be introduced. Secondly, an overview of papers in which Wikipedia is used for solving problems, related to the ontology population and ontology learning, will be provided. Then we will consider the most notable and large-scale resources using Wikipedia as a source of knowledge. Finally, the WikiMachine, a semantic annotation tool currently being developed in the framework of the Intelligent Technologies and Cultural Visits (ITCH) project, will be briefly presented.

Title: Efficient Linearization of Tree Kernel Functions

Speaker: Daniele Pighin

When: 12 Mar 2009

Abstract: The combination of Support Vector Machines with very high dimensional kernels, such as string or tree kernels, suffers from two major drawbacks: first, the implicit representation of feature spaces does not allow us to understand which features actually triggered the generalization; second, the resulting computational burden may in some cases render unfeasible to use large data sets for training.

We propose an approach based on feature space reverse engineering to tackle both problems. Our experiments with Tree Kernels on a Semantic Role Labeling data set show that the proposed solution can drastically reduce the computational footprint while yielding almost unbiased accuracy.

Title: Hadoop: a framework for Data Intensive Distributed Applications

Speaker: Christian Girardi and Roldano Cattoni

When: 19 Feb 2009

Abstract: Hadoop is a free java software framework for running applications on large clusters of commodity hardware. Inspired by Google's papers on distributed computing, it enables applications to work with thousands of nodes and petabytes of data. In particular Hadoop implements:

(1) a particular storage named HDFS (Hadoop Distributed File System) for storing large files across multiple machines.
(2) (2) a computational paradigm named Map/Reduce, for computing certain kinds of distributable problems using a large number of computers.

The talk is splitted in two parts: first, the basic concepts of the Hadoop framework and architectureare introduced. The second part focused on Hadoop approach to support reliability and manage failure recovering.

Title: Outlook for Machine Translation Research at FBK

Speaker: Marcello Federico and Matteo Negri

When: 12 Feb 2009

Abstract: This short seminar aims at sharing views on future goals of Machine Translation that are currently under discussion in some project proposals of the FP7 under preparation in which the HLT is involved. A not less important goal of the seminar is to check the internal interest for some of the proposed activities.

Title: Learning to Translate: a statistical and computational analysis

Speaker: Marco Turchi

When: 10 Feb 2009

Abstract: In this talk, an extensive experimental study of a Statistical Machine Translation system, Moses, from the point of view of its learning capabilities is presented. Very accurate Learning Curves are

obtained, by using high-performance computing, and extrapolations of the projected performance of thesystem under different conditions are provided. Our experiments suggest:

1. The representation power of the system is not currently a limitation to its performance,
2. The inference of its models from finite sets of i.i.d. data is responsible for current performance limitations,
3. It is unlikely that increasing dataset sizes will result in significant improvements (at least in traditional i.i.d. setting),
4. It is unlikely that novel statistical estimation methods will result in significant improvements.

The current performance wall is mostly a consequence of Zipf's law, and this should be taken into account when designing a statistical machine translation system. A few possible research directions are discussed as a result of this investigation, most notably the integration of linguistic rules into the model inference phase, and the development of active learning procedures.

Title: Populating Senso Comune with TMEO (Tutoring Methodology for the Enrichment of Ontologies)

Speaker: Alessandro Oltramari

When: 5 Feb 2009

Abstract: Senso Comune is a collaborative platform to build and maintain an open knowledge base of Italian language. The knowledge base will be initially populated with a suitable formalization of basic Italian lexicon (2K lemmas, about 10K senses) then it will be integrated with other existing linguistic resources, as well as user supplied information. The project is backed by an association of Italian scientists chaired by Prof. Tullio De Mauro, and is being supported by Fondazione IBM Italia.
Senso Comune depends on two core aspects: 1) a top-down direction, where top-level ontological categories and relations are introduced and maintained by ontologists to constrain lexicalised concepts; 2) a bottom-up direction, where non-expert users are asked to enrich the semantic resource with linguistic information through a wiki-like platform. In this building-up process, users are allowed only to access to the lexical level of the resource (therefore, explicit ontological choices are kept 'opaque' to ease users' task). These access-restrictions produce an epistemological spread between dimensions 1) and 2), a necessary requirement if we want to keep the deep technical aspects of the ontological layer aside from wiki-users. Conversely, to make dimension 2) plainly effective, those lexical concepts and relations which are introduced by users must fit the intended ontological choices underlying the system. For this reason, we designed a tutoring methodology to support linguistic enrichment of ontologies, towards the creation of comprehensive hybrid semantic resources. TMEO is an interactive Q/A system based on general distinctions embedded in DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering).

Title: HLT Discussion Panel: "How to write an academic paper and get published"

Speaker: Daniele Pighin, Claudio Giuliano, Marcello Federico, Bernardo Magnini

When: 22 Jan 2009

Abstract: Apart from conducting the research, writing publications in order to convey the results is a significant method of communication to the scientific community. Composing a good academic paper and publishing in a journal or conference is not a very trivial task. Beside the important factors such as ideas reflected in the article, importance of research and methods, argumentation and results, there are other aspects which involve in the process of paper being accepted and published. In fact, scientific content alone does not guarantee the acceptance nor understanding of the context being described by the author. Pointing out the importance of this issue, Thursday's seminar is dedicated to discuss about the efficient approaches of writing a paper in order to increase the chance of being accepted for publication. Towards this panel discussion, panelists would answer to the main questions about the reasons of paper rejections following by the approaches to overcome this matter aiming to increase the number of publications and enhance the quality of accepted papers.

Title: Tree Edit Distance: Approaches, Applications and Challenges with Focus on Recognizing Textual Entailment

Speaker: Yashar Mehdad

When: 15 Jan 2009

Abstract: Trees are amongst the most common and well studied data structures in computer science. The importance of trees is principally derived from the problem of comparing trees, which occurs in many contexts and applications such as computational biology, structured text databases, image analysis, natural language processing, automatic theorem proving, optimization and approximation. To solve this problem, tree edit distance is a common and significant measurement defining the difference between two tree structures quantitatively. One of the recent applications of tree edit distance algorithms is recognizing textual entailment between a pair of texts, a task which is receiving great attention in computational linguistics. In brief, textual entailment is to determine whether the meaning of a given text passage entails that of another or whether they have the same meaning or not. Given a pair of texts (called the text and the hypothesis), the core of the approach is based on applying a tree edit distance algorithm on the dependency trees of both the text and the hypothesis in order to estimate the distance between them. In this seminar, we initially talk about different methods and algorithms for solving the tree edit distance problem. Then, we introduce the node-by-node approach currently used by the EDITS system in applying tree edit distance in recognizing textual entailment, which allows the use of lexical entailment rules. Finally we discuss future extensions of the node-by-node approach where we consider edit operations on subtrees of the text and the hypothesis, in order to allow the applications of phrase-based entailment rules.

Title: Computing Implicit Entities in Text

Speaker: Rodolfo Delmonte

When: 4 Dec 2008

Abstract: In my talk I will focus on the notion of "implicit" or lexically unexpressed linguistic elements that are nonetheless necessary for a complete semantic interpretation of any text. I will refer to "entities" and "events" because the recovery of the implicit material may affect all the modules of a system for semantic processing, from the grammatically guided components to the inferential and reasoning ones. Reference to the system GETARUNS offers one possible implementation of the algorithms and procedures needed to cope with the problem and allows to deal with all the range of phenomena. I will address in detail the following three types of "implicit" entities and events:

- the grammatical ones, as suggested by a linguistic theories like LFG or similar generative theories;
- the semantic ones suggested in the FrameNet project, i.e. CNI, DNI, INI;
- the pragmatic ones: eventually we will present a theory and an implementation for the recovery of implicit entities and events of (non-) standard implicatures. In particular we will show how the use of commonsense knowledge may fruitfully contribute in finding relevant implied meanings.

Last Implicit Entity to be addressed is the Subject of Point of View which is computed by Semantic Informational Structure and contributes the intended entity from whose point of view is expressed a given subjective statement.

Title: The Center for the Evaluation of Language and Communication Technologies (CELCT) and its activities within the Human Language Technology field

Speaker: Danilo Giampiccolo

When: 27 Nov 2008

Abstract: CELCT, founded by ITC-irst (now FBK) and DFKI in December 2003 by grant of PAT, is a company dedicated to language and communication technology evaluation. Its mission is on the one hand to support the research community, by setting up and developing all the necessary infrastructures and skills to support valuation in Human Language Multimodal Communication Technologies (HL-MCTs) –e.g. organizing evaluation campaigns, producing benchmarks and annotated corpora, and performing web evaluation; on the other hand, to co-operate with public and private partners, providing incentives for the development of HL-MCT applications with commercial purposes. This seminar aims at presenting CELCT's expertise, giving an overview of past and ongoing activities and highlighting the areas where CELCT can interact with FBK, in order to enhance knowledge exchange and further promote collaboration in reaching common goals.

Title: HLT Discussion Panel

Speaker: Marcello Federico, Carlo Strapparava and Diego Giuliani

When: 20 Nov 2008

Abstract: In the last 20 years, researchers in HLT witnesses many changes of their "research habitat", which includes the research community, computing infrastructure, software tools, programming languages, linguistic resources, etc. Changes in the habitat have impacted and will continue to reshape the way we work. For instance, the spread of open source software and benchmarks has for sure cut down the time needed to test hypotheses, thus shortening the "insight-to-publication" time. The rapid growth of the HLT community, the spread of new communication means, the availability of "shared tasks" and open source have also significantly widened the spectrum of potential contributions to the field, by researchers and developers. The subject of the discussion is whether such progress has modified or not the notion of "good research" in our field. Is "good research" independent from technological progress? The question is not a philosophical one, but aims at spotting some practical advices for young and less young researchers.

Title: A history of machine translation

Speaker: Mauro Cettolo

When: 13 Nov 2008

Abstract: Even in the Bible we find traces of the desire of humans to overcome the Babylon of languages: the multitude came together, and were confused, because everyone heard them speak in his own language. [...] "Look, are not all these who speak Galileans? And how is it that we hear, each in our own language in which we were born?" (Acts 2, 1-11). But we have to wait until the XX century for the official birth of Machine Translation, precisely on March 4th 1947, even if first ideas of universals and philosophical languages may be traced back to seventeenth century and Artsrouni and Troyanskii's patents were issued in 1933. In this talk I will present a concise history of machine translation: starting from the pioneering ideas of Warren Weaver, I will review the most significant steps of this fascinating technical and scientific adventure, like the first demonstration of a translation system (1954), the (in)famous ALPAC report (1966), the first operational and commercial systems in 1970s and the corpus-based paradigm with stochastic and example-based methodologies in the most recent years.

Title: Towards Italian FrameNet

Speaker: Sara Tonelli

When: 16 Oct 2008

Abstract: This talk presents work in progress to develop Italian FrameNet. We describe two algorithms for the automatic projection of frame-semantic information from English to Italian texts and discuss

some criteria for the choice of the best parallel corpus to maximize projection performance. We compare our approach and our results to similar experiments carried out on other European languages and point out typical features of the Italian language as regards frame-semantic annotation. Besides, we present the LUNA corpus of Italian dialogs annotated with frame information. Finally, we present future developments, focusing on the use of MultiWordNet to automatically extend the resource created so far.

Title: Fast Speech Decoding through Phone Confusion Networks

Speaker: Daniele Falavigna

When: 9 Oct 2008

Abstract: I will present a two stage automatic speech recognition architecture suited for applications, such as spoken document retrieval, where large scale language models can be used and very low out-of-vocabulary rates need to be reached. The proposed system couples a weakly constrained phone-recognizer with a phone-to-word decoder that was originally developed for phrase-based statistical machine translation. The decoder permits to efficiently decode confusion networks in input, and to exploit large scale unpruned language models. Preliminary experiments are reported on the transcription of speeches of the Italian parliament. The use of phone confusion networks as interface between the two decoding steps permits to reduce the WER by 28%, thus making the system perform relatively close to a state-of-the-art baseline using a comparable language model.

Title: Dealing with Spoken Requests in a Multimodal Question-Answering System

Speaker: Roberto Gretter

When: 18 Sep 2008

Abstract: This talk reports on experiments performed in the development of the QALL-ME system, a multilingual QA infrastructure capable of handling input requests both in written and spoken form. Our objective is to estimate the impact of dealing with automatically transcribed (i.e. noisy) requests on the QALL-ME question interpretation task, namely the extraction of relations from natural language questions. After a brief introduction of the QALL-ME benchmark, the proposed approach based on text entailment will be discussed and some experiments comparing clean and noisy transcriptions will be reported. This work was presented at the beginning of September at the AIMSA08 conference in Varna, Bulgaria.

Title: Discovery of Protein Interactions from Scientific Literature: OntoGene and the BioCreative experience

Speaker: Fabio Rinaldi

When: 24 Jul 2008

Abstract: In this talk I will present activities performed within the scope of the OntoGene project (http://www.ontogene.org/), which aims at supporting the semi-automatic discovery of semantic relations among specific biological entities (such as proteins, genes, diseases) from scientific literature. I will present an environment that supports the interactive development of discovery rules making use of a rich document annotation. I will characterize in detail our participation in the 2nd BioCreative text mining competitive evaluation, describing the nature of the challenge, our own contribution, and the results obtained. Finally, I will conclude with a short overview of current activities.

Title: Question classification systems based on machine learning and shallow linguistic information

Speaker: David Tomas

When: 17 Jul 2008

Abstract: Question Classification is an important task in Question Answering systems, and can be used in a wide range of other domains, such as Helpdesk Services, Online Digital Reference Services and Natural Language Interfaces to Databases. The goal of this task is to assign labels from a taxonomy to questions based on the expected answer type. In the past, most approaches have relied on heuristic rules and hand-made patterns. These systems present two main problems: the human effort needed to define the patterns, and the lack of flexibility and domain dependency. Machine learning techniques can overcome such limitations. The talk will describe several approaches to Question Classification based on machine learning techniques. These approaches are based on shallow linguistic features in other to obtain systems that can be easily adapted to different languages and domains.

Title: Instance-Based Ontology Population Exploiting Named-Entity Substitution

Speaker: Claudio Giuliano

When: 10 Jul 2008

Abstract: We present an approach to ontology population based on a lexical substitution technique. It consists in estimating the plausibility of sentences where the named entity to be classified is substituted with the ones contained in the training data, in our case, a partially populated ontology. Plausibility is estimated by using Web data, while the classification algorithm is instance-based. We evaluated our

method on two different ontology population tasks. Experiments show that our solution is effective, outperforming existing methods, and it can be applied to practical ontology population problems.

Title: Speech Processing for Conversational Systems - Recent Research Activities at Siemens

Speaker: Georg Stemmer

When: 20 Jun 2008

Abstract: After an introduction to the Siemens Professional Speech Processing Group the talk will give an outline of recent research work in the area of speech processing for spoken dialogue systems.Firstly the speaker characterization feature will be described which allows to adapt the dialogue to the current user.Secondly, some experiments on spoken digit recognition will be presented, followed by a short introduction into the project "SemProM - Digital Product Memory"; in the scope of this project we just started to investigate spoken interaction with industrial robots. The talk concludes with two application examples from the medical domain.

Title: The HLT Language Infrastructure

Speaker: Roberto Gretter

When: 5 Jun 2008

Abstract: Linguistic resources are crucial for most of the research activities in HLT. In this talk we present the activities done in the last months, devoted to the collection of information on the linguistic resources that are available in HLT. The internal pages containing the resources will be illustrated in some detail. Particular attention will be given to the corpora under construction, i.e. the written and spoken data that we are collecting day by day and that represent in some sense the evolution of the languages.

Title: Kernels on Linguistic Structures for Question Answering Systems

Speaker: Alessandro Moschitti

When: 22 May 2008

Abstract: Natural Language Processing (NLP) for Information Retrieval has always been an interesting and challenging research area. Despite the high expectations, most of the results indicate that effectively using NLP is very complex. This talk shows how Support Vector Machines along with kernel functions can effectively represent syntax and semantics in learning algorithms. Experiments on

question/answer classification demonstrate that the above models highly improve on bag-of-words over a TREC dataset.

Title: Creating a Gold Standard for Person Cross-Document Coreference Resolution in Italian News

Speaker: Luisa Bentivogli

When: 15 May 2008

Abstract: This talk presents work aimed at the realization of a gold standard for cross-document coreference resolution of person entities in a corpus of Italian news. The gold standard has been created selecting a number of person names occurring in Adige-500K, a corpus composed of all the news stories published by the local newspaper "L'Adige" from 1999 to 2006. The corpus consists of 535,000 news stories, for a total of around 200 million tokens. To sample the person names in the corpus, we identified two dimensions, corresponding to two phenomena we intended to study, namely (i) the fame of the person entities and (ii) the ambiguity of person names. The first version of the gold standard is composed of 209 person names corresponding to 709 entities, for a total of 43,704 annotated documents.

Title: Visual Modeling and Feature Adaptation in Sign Language Recognition

Speaker: Philippe Dreuw

When: 8 May 2008

Abstract: We propose a tracking adaptation to recover from early tracking errors in sign language recognition by optimizing the obtained tracking paths w.r.t. the hypothesized word sequences of an automatic sign language recognition system. Hand or head tracking is usually only optimized according to a tracking criterion. As a consequence, methods which depend on accurate detection and tracking of body parts lead to recognition errors in gesture and sign language processing. Similar to speaker dependent feature adaptation methods in automatic speech recognition, we propose an automatic visual alignment of signers for vision-based sign language recognition. Furthermore, the generation of additional virtual training samples is proposed to reduce the lack of data problem in sign language processing, which often leads to "one-shot" trained models. Most state-of-the-art systems are speaker dependent, and consider tracking as a preprocessing feature extraction part. Experiments on a publicly available benchmark database show that the proposed methods strongly improve the recognition accuracy of the system.

Title: HTMLCleaner: Extracting relevant text from Web pages

Speaker: Christian Girardi

When: 24 Apr 2008

Abstract: The objective of the talk is to show Htmlcleaner, a tool aimed at automatically cleaning HyperText Mark-up Language (HTML) files. Htmlcleaner removes HTML tags and irrelevant text (like some words used as navigation menu, the common header and footer across all pages in a site, etc). It also reformats the discovered relevant text with a basic encoding of the structure of the page using a minimal set of symbols to mark the beginning of headers, paragraphs and list elements. This process can be necessary in order to build a usable written corpus starting from Web pages. This tool has been evaluated in the first CLEANEVAL competition (in september 2007). CLEANEVAL is a shared task and competitive evaluation on the topic of cleaning arbitrary web pages, with the goal of preparing web data for use as a corpus, for linguistic and language technology research and development.

Title: Prosody in Automatic Speech Recognition

Speaker: Dino Seppi

When: 17 Apr 2008

Abstract: Prosody plays a fundamental role in human speech because: 1) It compensates many violations of speech and language rules, such as the absence of strict grammar constrains, and it improves the understanding of fragmented utterances and words, overlapping conversations, etc. 2) Prosody conveys additional latent information such as accent, gender, mood, and other events at functional levels that are otherwise simply ignored. Therefore, prosodic information might be exploited for automatic speech processing purposes. The first aim of this presentation is to show how time-based prosodic features can improve ASR performance. More specifically, two major weaknesses of ASR systems will be addressed: acoustically ambiguous utterances and the well-known unfavorable property of HMMs of inappropriately modeling phoneme and word duration. The former problem is presented for very limited vocabulary tasks by applying and comparing a certain number of approaches, the latter one is extended to a very large vocabulary domain. The second aim of this presentation is to illustrate how prosodic time-based attributes can be extracted and manipulated to improve important speech analytics tasks. Additional latent information obtained from prosodic cues, in conjunction with other segmental and linguistic sources of information, can be used to cope with speech peculiarities. Prosodic-dependent applications will be considered such as punctuation, speech segmentation, confidence measures, and emotion recognition.

Title: Dynamic Clustering for Person Cross-document Coreference

Speaker: Octavian Popescu

When: 10 Apr 2008

Abstract: The aim of this seminar is to discuss the recent advancements we have obtained working on the Cross Document Coreference task (CDC). Particularly we focus on one of the modules of our system, namely the Global Coreference, which is built on the basis of a new clustering technique. We argue that the best way to cluster person name instances (PNM) is to apply a cascade clustering technique and to dynamically adjust the clustering parameters such that the likelihood of previously realized coreferences is maximized at each step. The general approach to CDC is context-similarity driven; to each PNM a set of tokens is associated and the coreference is decided on the basis of the weight of each token of their intersection. We use automatically discovered ontological restrictions to build seed clusters (cascade clustering) that are used for further coreferences. We briefly discuss two properties of context-similarity model, superposition and masking, which are directly responsible for losses in accuracy. Our clustering technique has been developed in order to alleviate the negative effects of superposition and masking, by applying a dynamic weighting procedure. The initial weights are recomputed such that the probability of spurious coreferences is decreased, given the structure of the seed clusters. The algorithm scores better than a very high baseline in terms of purity and wins with 10% with respect to the same baseline using BCubed metrics. The technique behind our approach allows us to propose an evaluation methodology which is a variant of stratified sampling strategy. By default, the classical precision and recall can be used, but we argue that more informative measures are related to confusion and dispersion which we propose. Our main interest is to understand the weakness of the proposed approach, therefore comments and suggestions are highly welcomed.

Title: Report on the ELERFED Johns Hopkins Summer Workshop - exploiting lexical and encyclopedic knowledge

Speaker: Massimo Poesio

When: 3 Apr 2008

Abstract: During the Summer of 2007 a group of us (including Alessandro Moschitti from DISI and Claudio Giuliano from FBK) participated in a workshop on entity disambiguation at Johns Hopkins. The talk will summarize goals and results of the workshop, including (i) resources created - the ACE 05 CDC corpus, the ARRAU corpus, and the BART toolkit for coreference (ii) results concerning the use of tree kernels and wikipedia info for coreference (iii) results concerning new clustering techniques for ED over the Spock dataset.