[Jump: to the content]

PIRE 2010 Meeting Uppsala, Sweden

[Jump: to the navigation]

Program of the 2010 PIRE meeting in Uppsala

Thursday, July 15 (Main University Building (Venue A) Faculty Room, 1st floor		
Start Time	**Speaker**	**Title**
10:00	Jan Hajic (Charles Univ.)	Welcome
10:15	Micha Elsner (Brown Univ.)	Learning to Fuse Disparate Sentences
10:45		Coffee Break
11:00	Anoop Deoras (JHU)	Empirical Bayes Risk Training for Model Combination
11:30	Carolina Parada (JHU)	A Spoken Term Detection Framework for Recovering Out-of-Vocabulary Words Using the Web
12:00	Scott Novotney (JHU)	Semi-Supervised Speech Recognition
12:30		Lunch (individually in the city)
14:00	Jonny Weese (JHU)	Integrating Syntax and Semantics into Statistical MT
14:30	Saedeeh Momtazi (Saarland Univ.)	Trained Trigger Language Model for Sentence Retrieval in Question Answering Systems
15:00	David Marecek (Charles Univ. in Prague)	Maximum Entropy Translation Model in Dependency-Based MT Framework
15:30		Coffee Break
16:00	Silvie Cinkova (Charles Univ. in Prague)	Speech Reconstruction in Czech and English Dialog Corpora
16:30	Ben Roth (Saarland Univ.)	Wikipedia-Based Cross-Language Information Retrieval
17:00	F. Jelinek, D. Klakow, E. Charniak, J. Hajic, S. Khudanpur, M. Johnson	PIRE PIs' meeting
17:30		End of workshop
19:00	All	Dinner, place TBA

Abstracts of the scheduled talks

Saedeh Momtazi (Saarland Univ.): Trained Trigger Language Model for Sentence Retrieval in Question Answering Systems

We propose a novel language modeling approach for sentence retrieval in question answering systems called trained trigger language model. This model improves the sentence retrieval performance by relaxing the exact matching assumption and reducing data sparsity which are the two major problems of sentence-level retrieval. Our trained trigger language model captures pairs of trigger and target words while training on a large corpus. The word pairs are extracted based on both unsupervised and supervised approaches with different notions of triggering. All trained models are finally used in a language model-based sentence retrieval framework. Our experiments on TREC question answering collection verify this approach significantly improves the sentence retrieval performance compared to the models that do not benefit from this type of word relation. The proposed trained triggering language model achieved 7.23% absolute improvement and 11.48% relative improvement in mean average precision compared to the standard unigram model.

Ben Roth (Saarland Univ.): Wikipedia-Based Cross-Language Information Retrieval

As a growing resource with up-to-date vocabulary, Wikipedia recently gains a lot of attention in different NLP areas. We explore Wikipedia for providing translation mappings for cross-language information retrieval. We will look at methods extracting cross-language information from article level co-occurrence statistics and word-level language models based on the anchor text of links.

Silvie Cinkova (Charles Univ. in Prague): Speech Reconstruction in Czech and English Dialog Corpora

The currently available taggers and parsers perform significantly worse on spontaneous speech than on written text due to disfluencies, such as false starts, syntactic errors and numerous ellipses, typical of spontaneous speech. Speech reconstruction removes the disfluencies and smoothes the resulting text to fit the written-language standards. We present one Czech and one English corpus of dialogs above family photographs. Both corpora are designed in accordance with the other corpora in the family of the Prague Dependency Treebanks; i.e., they consist of three separate but interlinked layers formatted in the Prague Markup Language (PML): the sound recording, the synchronized manual transcript and the manually created speech reconstruction layer. The focus of the talk will be on the annotation guidelines for the speech-reconstruction layer and on its relation to the raw-transcript layer.

Carolina Parada (JHU): A Spoken Term Detection Framework for Recovering Out-of-Vocabulary Words Using the Web

Vocabulary restrictions in large vocabulary continuous speech recognition (LVCSR) systems mean that out-of-vocabulary (OOV) words are lost in the output. However, OOV words tend to be information rich terms (often named entities) and their omission from the transcript negatively affects both usability and downstream NLP technologies, such as machine translation or knowledge distillation. We propose a novel approach to OOV recovery that uses a spoken term detection (STD) framework. Given an identified OOV region in the LVCSR output, we recover the uttered OOVs by utilizing contextual information and the vast and constantly updated vocabulary on the Web. Discovered words are integrated into system output, recovering up to 40% of OOVs and resulting in a reduction in system error.
This is joint work with Fred Jelinek (JHU), Mark Dredze (JHU), and Abhinav Sethy (IBM)

Jonny Weese (JHU): Integrating Syntax and Semantics into Statistical MT

Baseline machine translation systems generally do not include linguistic information in their translation models. In this talk, we discuss how translation models based on synchronous context-free grammars (SCFGs) can be extended to take into account both the syntactic and semantic information that is present in the training data. On an Urdu-English translation task, including syntactic information gives a very significant improvement in performance over a non-syntactic SCFG baseline. Adding semantic information also shows promising improvement.

Micha Elsner (Brown Univ.): Learning to Fuse Disparate Sentences

We present a system for fusing sentences which are drawn from the same source document but have different

content. Unlike previous work, our approach is supervised, training on real-world examples of sentences fused by professional journalists in the process of editing news articles. Like Filippova and Strube (08), our system merges dependency graphs using Integer Linear Programming. However, instead of aligning the inputs as a preprocess, we integrate the tasks of finding an alignment and selecting a merged sentence into a joint optimization problem, and learn parameters for this optimization using a structured online algorithm. Evaluation by human judges shows that our technique produces fused sentences that are both informative and highly readable.

Scott Novotney (JHU): Semi-Supervised Speech Recognition

State of the art performance automatic speech recognition requires thousands of hours of manually transcribed speech and billions of words of language modeling text. Deploying to new domains is prohibitively expensive. We seek to reduce this cost by bootstrapping a small initial model with unlabeled audio. By decoding large amounts of audio, we can improve both the acoustic and language model to within 80% of supervised performance. We highlight trends where "self-training" is most effective and the successes and failures of the acoustic and language model.

Anoop Deoras (JHU): Empirical Bayes Risk Training for Model Combination

In this work, we combine various recognition models in a log linear framework (similar in spirit to Beyerlein, 1998) to minimize empirical Bayes Risk instead of 1-best error rate. Experiments are carried on WSJ speech task and in it we show significant improvements in recognition accuracy, obtained when models are trained to minimize Bayes Risk rather than 1 best error. We will also discuss the extension of the method by applying Deterministic Annealing technique during model search wherein it is possible to constrain the search to a minimum entropy on the test data. Finally, we will demonstrate the use of Iterative Decoding (presented during the last PIRE meeting) to carry out rescoring of lattices under this log linear model.

David Marecek (Charles Univ. in Prague): Maximum Entropy Translation Model in Dependency-Based MT Framework

We propose a forward translation model consisting of a set of maximum entropy classifiers: a separate classifier is trained for each (sufficiently frequent) source-side lemma. In this way the estimates of translation probabilities can be sensitive to a large number of features derived from the source sentence (including non-local features, features making use of sentence syntactic structure, etc.). When integrated into English-to-Czech dependency-based translation scenario implemented in the TectoMT framework, the new translation model significantly outperforms the baseline model (MLE) in terms of BLEU. The performance is further boosted in a configuration inspired by Hidden Tree Markov Models which combines the maximum entropy translation model with the target-language dependency tree model.

Content: Jan Hajic.
This project has been supported by the U.S. NSF OISE PIRE grant and by the project ME838 of the Ministry of Education, Youths and Sports of the Czech Republic.

[Jump: to the navigation | to the content]

[Jump: to the content]

[Jump: to the navigation]

PIRE Meeting 2010 Uppsala: The Venue

The meeting will be held at the Main Building of the Uppsala University, where also the main session of the ACL 2010 conference will take place. It is marked as "Venue A" on the ACL venue map.

Thanks to the local organizers of ACL 2010, we were given a nice faculty room for "Humanities and Social Sciences", normally closed to the public. It has blinds, which will be useful for projection (and perhaps because of the weather, too... quite unusual [high] temperatures are expected in Uppsala).

The room is located on the upper floor of the main building, with entrance from the narrow balcony (in its north corner) close to the northwest area of the poster sessions (above the ground-floor ACL Office and Registration).

The Dinner

The dinner has been reserved fro 7pm in William's pub (British pub / Indian Cuisine) just next to the main Univeristy Building (Venue A), where also the talks will take place (see above). The pub is open since 5pm on Thursdays, so that you can go directly there for a pre-dinner beer or two right after the talks are over (if you do not want to walk back to your hotel).