

Site Update: DFKI & CoLi, Saarbrücken/Berlin, Germany

Yi Zhang

LT-Lab, DFKI, Germany

Computational Linguistics, Saarland University, Germany



Overview of Activities

- ▶ Linguistic Resource Building
 - ▶ WSJ treebanking with ERG (Valia Kordoni)
 - ▶ Metagrammar engineering (Antske Fokkens)
 - ▶ Slavic grammars (Tania Avgustinova)
 - ▶ Mandarin Chinese grammar (Yi Zhang)
- ▶ Development of Processing Technologies
 - ▶ HPSG chart pruning (Bart Cramer)
 - ▶ Corpus-driven CFG approximation of HPSG (Yi Zhang)
- ▶ Applications of Deep NLP
 - ▶ Project TAKE: ACL Anthology Searchbench (Ulrich Schäfer)
 - ▶ Deep DARE (Peter Adolphs, Feiyu Xu)
 - ▶ IE-oriented HPSG Parse Reranking (Feiyu Xu, Yi Zhang)

TAKE: Technologies for Advanced Knowledge Extraction

Ulrich Schäfer, Bernd Kiefer, Christian Spurk, Jörg Steffen, Rui Wang

... since last Summit

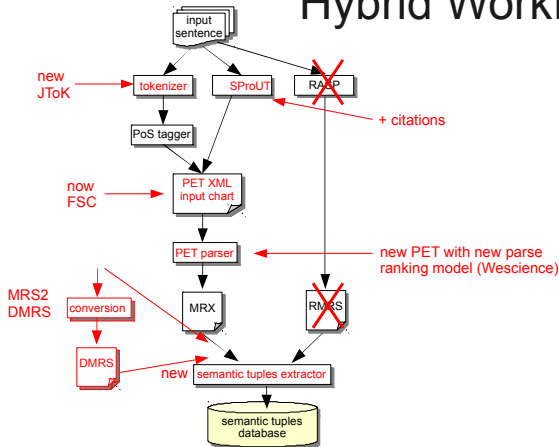
- ▶ Boost in deep parsing coverage of scholarly paper content
- ▶ New public Searchbench application, released at ACL 2011
- ▶ Ongoing related work: coreference resolution, citation classification

Boost in Deep Parsing Coverage

- ▶ by combination of chart mapping + pruning, new ERG, new parse ranking model (WeScience), and domain-specific NER, including citation patterns
- ▶ Drastic improvement of overall parsing coverage to now 83% including fragments - despite fully automatic PDF extraction and preprocessing!

New Architecture

Hybrid Workflow '10



Extraction of Semantic Tuples

- ▶ So far (2009): one “flat” predicate-argument tuple per sentence, extracted directly from (R)MRS
- ▶ New algorithm, based on MRS-to-DMRS conversion result
 - ▶ Recursive, i.e. now also suited for long sentences with subclauses etc.
 - ▶ More robust and complete: covers all successfully parsed sentences
 - ▶ Treatment of special phenomena: adjectives vs. verbs, coordination, adverbs

ACL Anthology Searchbench



Add and remove filters for the documents you are interested in – or remove all currently set filters:

Statements

obtain improvements

Keywords

Topics

dependency parsing

Publication

Authors

Year

2008

Affiliations

Affiliation Sites

0 documents found that match your search filters:

Task-oriented Evaluation of Syntactic Parsers and Their Representations (2008)

Miyao, Yusuke, Sætre, Rune, Sagae, Kenji, Matsuzaki, Takuya, Tsujii, Jun'ichi

... Interestingly, the accuracy improvements are observed even for ensembles of different representations from the same parser. ... We also measure the accuracy improvements obtained by parser retraining with GENIA, to examine the domain portability, and to evaluate the effectiveness of domain adaptation. ...

Dependency Parsing with Short Dependency Relations in Unlabeled Data (2008)

Wenliang Chen, Daisuke Kawahara, Kiyotaka Uchimoto, Yujie Zhang, Hitoshi Isahara

... They obtain consistent improvements by penalizing dependencies between two words that are farther apart in the string. ...

Deriving Syntactic and Semantic Dependencies with Two Single-Stage Maximum

Task-oriented Evaluation of Syntactic Parsers and Their Representations

ACL-08: HLT – 2008

Miyao, Yusuke (University of Tokyo, Tokyo Japan)
Sætre, Rune (University of Tokyo, Tokyo Japan)
Sagae, Kenji (University of Tokyo, Tokyo Japan)

Content PDF Citations

conversions between representations when necessary.

We also measure the accuracy improvements obtained by parser retraining with GENIA, to examine the domain portability, and to evaluate the effectiveness of domain adaptation.

Use Sentence as a Filter task to identify protein pairs that are mentioned as interacting in biomedical papers. Recent studies on PPI e-omedical papers is growing rapidly, it is impossible for biomedical researchers to read all papers thus, there is an emerging need for reliable IE technologies, such as PPI identification.

Show Parsing Information: Input Chart (FSC) in names: the former sentence mentions a protein interaction, while the factor does not.

Given a protein pair, PPI Input Chart (XML) binary classification; for example, (IL-8, CXCR1) is a positive example, and (RBP, TTR) is a negative

Deep Analysis (Derivation) that dependency relations between target proteins are effective features for

Deep Analysis (MRS) ans, 2006; Erkan et al., 2007; Sætre et al., 2007).

Deep Analysis (XML) a dependency parser outputs a dependency tree shown in Figure 1.



Extensions released at ACL-11

- ▶ Full content of 19,200 ACL Anthology papers parsed (all including ACL-HLT-2011, except ROCLING) with surprisingly high coverage despite OCR errors
- ▶ Optional removal of negated statements in search results (neg_rel)
- ▶ Title search (search form and in parsed data)
- ▶ Highlighting of matched sentence in original PDF (doesn't work well on older, scanned papers)

Deep processing in relation extraction

- ▶ Deep DARE (Peter Adolphs, Feiyu Xu)
Using DMRS in unsupervised relation extraction
- ▶ IE-oriented HPSG parse selection (Feiyu Xu, Yi Zhang)

Trebank development

- ▶ WSJ annotation with ERG
 - ▶ Switched to ERG 1007
 - ▶ New annotators (Lea & Iliana)
 - ▶ Applying discriminant reranking
 - ▶ Section 00~18 completed (*)
 - ▶ To start an UPDATE round soon (with help from Dan)
- ▶ SynTagRus → Tiger++ (Russian)
 - ▶ Converted from dependency to Tiger-style treebank
 - ▶ Annotated with HPSG derivations (manual + automatic)
 - ▶ Subcategorization frames automatically extracted

Grammar Engineering

- ▶ Metagrammar Engineering in a Multi-Lingual Context (Antske Fokkens)
- ▶ Slavic Matrix & Russian Resource Grammar (RRG) (Tania Avgustinova)
 - ▶ First release available 20110615
 - ▶ Three-layer design with emphasis on the cross-Slavic perspective
 - ▶ <http://www.coli.uni-saarland.de/~tania/rrg/>
- ▶ Mandarin Chinese Grammar (MCG) (Yi Zhang)
 - ▶ First preview release available 20110616
 - ▶ In collaboration with Tokyo group(s)
 - ▶ <http://www.coli.uni-saarland.de/~yzhang/mcg.php>

Corpus-Driven CFG Approximation of HPSG

- ▶ <http://www.coli.uni-saarland.de/~yzhang/files/jigsaw.tgz>
- ▶ Java implementation of BitPar, efficient parsing with millions of CFG production rules trained with over 50 million Wikiwoods trees
- ▶ Naive MLE without construction smoothing
- ▶ Minimal smoothing on the lexicon
- ▶ Context annotation vs. TFS annotation
- ▶ 80% ParsEval labeled bracketing F1
- ▶ Over 92% supertagging accuracy (without an actual sequence labeling model)

Bart Cramer: PhD Thesis



- ▶ Thesis: *“Improving the feasibility of precision-oriented HPSG parsing”*
- ▶ Successfully defended on 08.06.2011