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

Yi Zhang

LT-Lab, DFKI, Germany

Computational Linguistics, Saarland University, Germany

DELPH-IN

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
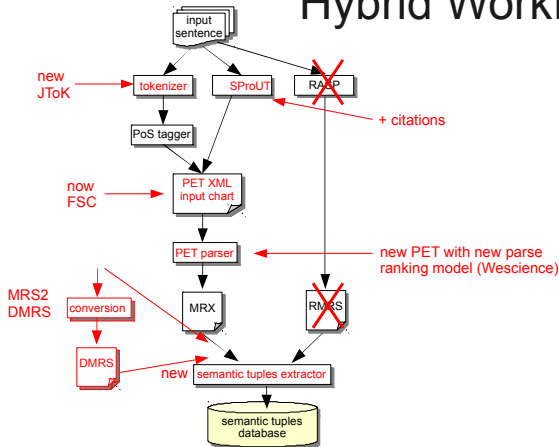
. . . since last Summit

- ▶ 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

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)

Treebank 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

DELPH-IN

Grammar Engineering

- Metagrammar Engineering in a Multi-Lingual Context (Antske Fokkens)
- Slavic Matrix & Russian Resource Grammar (RRG) (Tania Avgustinova)
  - First release avaiable 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`

DELPH-IN

Corpus-Driven CFG Approximation of HPSG

- `http://www.coli.uni-saarland.de/~yzhang/files/jigsaw.tgz`
- Java implementation of BitPar, efficent 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)

DELPH-IN

Bart Cramer: PhD Thesis

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

DELPH-IN