A Dynamic Annotation of the Wall Street Journal Sections of the Penn Treebank

Valia Kordoni & Yi Zhang
LT-Lab, German Research Center for Artificial Intelligence (DFKI GmbH) & Dept. of Computational Linguistics, Saarland University, Germany
{kordoni,yzhang}@coli.uni-sb.de

Introduction
- An on-going project to annotate texts from WSJ sections of PTB (~1M tokens, ~43K utterances) with HPSG analyses
- Independent annotation from PTB by parsing with ERG and manual disambiguation of the outputs
- Annotation only for sentences covered by the grammar
- Different to conversion-based development of treebanks in rich formalisms (Cahill et al. 2002, Miyao et al. 2004, Hockenmaier 2006, etc)

HPSG Parsing
- English Resource Grammar (Flickinger 2002)
- PET Parser (Callmeier 2001)
- {incr tsdb[]} (Oepen & Callmeier 2000)
- Preprocessing (Adolphs et al. 2008)
- Robust Parsing * (Zhang et al. 2008)

Grammar & Treebank Update
- Update of the disambiguation model: newly annotated sections are used to train better parse selection models that help annotators to find the best tree more quickly
- Update of the Grammar: grammar writer receives feedback from annotators and fixes systematic errors or improves coverage
- Update of the annotation: existing annotations will be semi-automatically synchronized with the new versions of the grammar, with some extra manual annotation required

Treebanking
- For each tree, up to 500 candidates (ranked by a statistical disambiguation model) are recorded
- Discriminant-based treebanking decisions
- Yes/No choice for candidate discriminants
- For n trees, on average need $\log_2 n$ steps to fully disambiguate

Quality Assurance
- Multiple annotators with partially overlapping assignments. Inter-annotator agreement is checked regularly
- Regular treebankers’ meetings to discuss issues in treebanking and annotation strategies to achieve higher inter-annotator agreement

Some Facts in Numbers
- Number of annotators: 3
- Grammar parsing coverage: ~80%
- Average annotation speed: 35–40 sentences/hour
- Expected duration of the project: 15–18 months
- Inter-annotator agreement: exact match agreement of ~50%

References

References