Robust Parsing with a Large HPSG Grammar

Yi Zhang  Valia Kordoni

Language Technology Lab
German Research Center for Artificial Intelligence, Germany

Department of Computational Linguistics
Saarland University, Germany

FEAST 11.02.2009
Lexical acquisition model (together with the handling of MWEs) fills in the missing lexical knowledge with certain confidence level.

Studies have shown strong correlation between the lexicon quality and parser performance (especially parsing coverage/robustness).

Unpredictable irregularities in real world texts adds to the difficulties in parsing with hand-written grammars.

**Question**

- How to define and extract partial analysis when not all constraints in the grammar are satisfied?
A Two-Stage Robust Parser [Zhang and Kordoni, 2008]

1. A HPSG grammar is used to build bottom-up local analyses.

2. A CFG backbone grammar extracted from HPSG treebank (LOGON) is used to continue parsing with the passive edges built by HPSG.

- Results are complete (pseudo-) derivation trees.
- The CFG backbone grammar is generally more relaxed and allows robust construction.
| the | Lakers | wins |
An Example

Stage I

Y. Zhang, V. Kordoni
Robust Parsing with a Large HPSG Grammar
An Example

Y. Zhang, V. Kordoni
Robust Parsing with a Large HPSG Grammar
Implementation Issues

The two-stage parsing model is implemented as extension to the PET parser, and experimented with ERG

- Disambiguation model
- Efficiency Concerns
- Semantic Composition
Disambiguation model

- Most of the features used in [Toutanova et al., 2002]'s discriminative model can be obtained from derivation tree (feature structures are not necessary), although the model is approximate, for the differences in tree language ($T$) of CFG and HPSG.

- A separately estimated generative PCFG model can be applied in the second parsing phase.
Efficiency Concerns

- Packing is used to reduce local structural ambiguity
  - Subsumption-based packing for Stage I (HPSG parsing)
  - Equivalence-based packing for Stage II (CFG parsing)
- Selective unpacking [Zhang et al., 2007] is invoked to extract best partial readings from pseudo-parse forest
- Must handle cyclic unary productions in the PCFG backbone grammar
  - \( X \rightarrow X \) (e.g. noptcomp \( \rightarrow \) noptcomp)
  - \( X \rightarrow Y, Y \rightarrow X \)
Robust Semantic Composition

- **CFG** rules can be paired with semantic composition rules from the original grammar to compose the MRS analyses.
- Robust unification will be applied to guarantee the unifiability between any pair feature structures, by dynamic extension to the existing signature (type hierarchy).
For Further Reading I

Toutanova, K., Manning, C. D., Shieber, S. M., Flickinger, D., and Oepen, S. (2002).
Parse ranking for a rich HPSG grammar.
In *Proceedings of the 1st Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 253–263, Sozopol, Bulgaria.

Zhang, Y. and Kordoni, V. (2008).
Robust Parsing with a Large HPSG Grammar.
In *Proceedings of the Sixth International Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco.

Zhang, Y., Oepen, S., and Carroll, J. (2007).
Efficiency in unification-based N-best parsing.
In *Proceedings of the 10th International Conference on Parsing Technologies (IWPT 2007)*, pages 48–59, Prague, Czech.