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

LREC 2008

Outline

Outline

Parsing with Rule-based Precision Grammars

- Precise description with strong linguistic motivation and high generalization
- Usually lacks of robust processing mechanism due to unpredictable *noise* in real world texts

Question

- How to define and extract partial analysis when not all constraints in the grammar are satisfied?

Previous Work

With bottom-up chart-based parsing, partial parse as a set of non-overlapping adjacent passive passing edges that covers the entire input sequence:

- Longest-edge approach: prefer larger fragment analysis
- Shortest-path approach [Kasper et al., 1999]: (heuristically) weighted fragment analyses
- Statistical partial parse selection model [Zhang et al., 2007a]: more elaborated (approximate) disambiguation models for partial parses

Remaining Issues

- Upper-part of the derivation tree is missing
- Disambiguation models have more than one components, and are hard to train

Outline

A Two-Stage Model

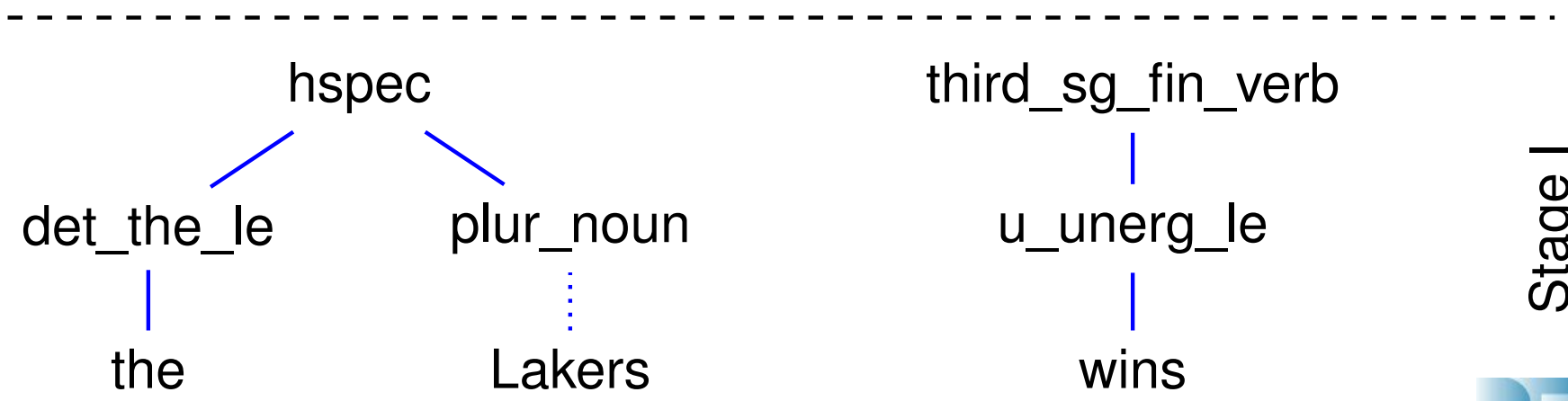
1. HPSG grammar is used to build bottom-up local analyses
2. A CFG backbone grammar extracted from HPSG treebank (LOGON) is used to continue parsing with the passive edges built by HPSG

- Results are complete (pseudo-) derivation trees
- The CFG backbone grammar is generally more relaxed and allows robust construction

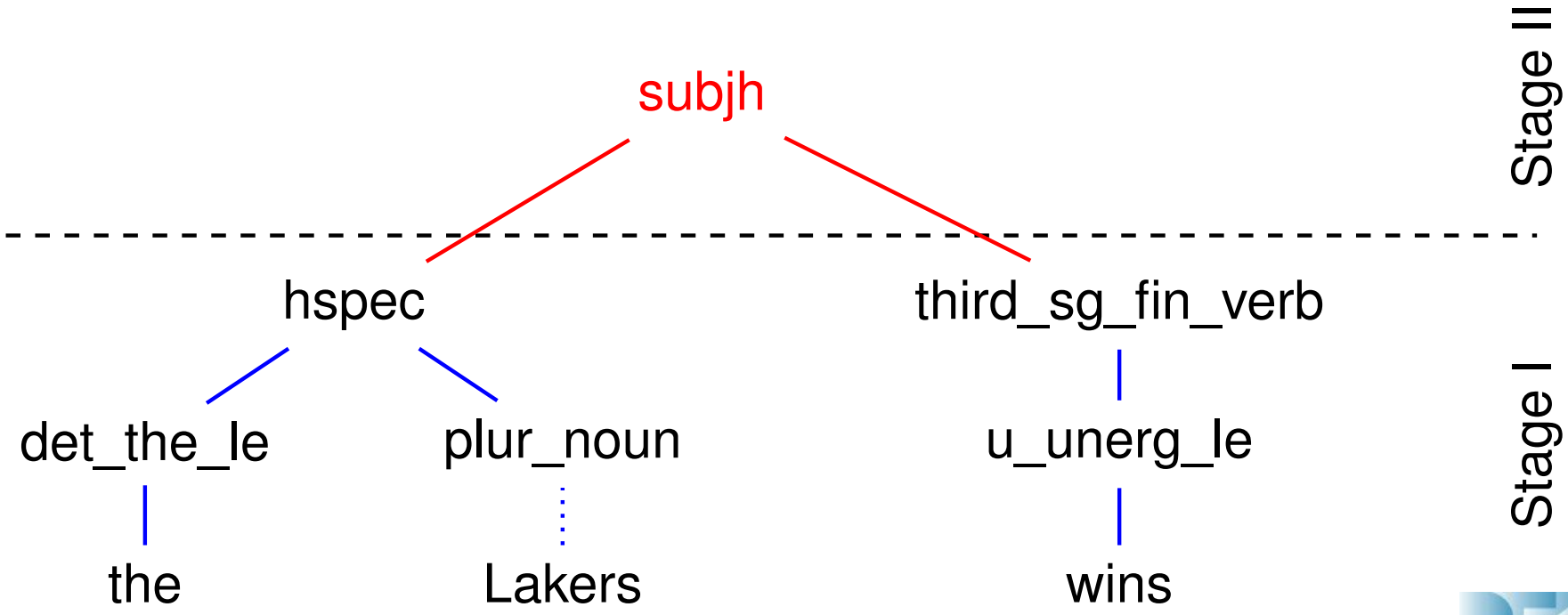
An Example

the Lakers wins

An Example


hspec
third_sg_fin_verb
det_the_le
plur_noun
u_unerg_le
the
Lakers
wins
Stage I

An Example


subjh
hspec
third_sg_fin_verb
det_the_le
plur_noun
u_unerg_le
the
Lakers
wins
Stage II
Stage I

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 model can be obtained from derivation tree (feature structures are not necessary)
- Strictly speaking, the model is approximate, for the difference in tree language (T) of CFG and HPSG.
- Practically, the approximation largely simplified the training process, and the same disambiguation model is used for both full and partial parse disambiguation.

$$P(t|w) = \frac{\exp \sum_{j=1}^{n} \lambda_j f_j(t, w)}{\sum_{t' \in T} \exp \sum_{j=1}^{n} \lambda_j f_j(t', w)}$$

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., 2007b] is invoked to extract best partial readings from pseudo-parse forest

Robust Semantic Composition

- CFG rules can be paired with semantic composition rules
- Can provide informative partial description of semantics in the framework of RMRS

Evaluation

- Manually evaluated a subset of sentences from PARC 700 Dependency Bank with full lexical span in ERG
- 213 parsed by ERG out-of-the-box
- Pseudo-derivation trees are built for 41 out of 54 sentences without full parse
- 13 with no cross-bracketing; 18 with $\leq$ 2 cross-bracketings
- Many errors are related to missing lexical entries

Outline

Summary

- A two-stage parsing algorithm is proposed to achieve robust parsing
- Partial parse as pseudo-derivation tree is more informative than a set of passive parsing edges
- The model can be generalized to use other less restrictive grammar in the second stage
- Robust semantic composition is possible

For Further Reading I

Kasper, W., Kiefer, B., Krieger, H.-U., Rupp, C., and Worm, K. (1999).

Charting the depths of robust speech processing.

In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 1999)*, pages 405–412, Maryland, USA.

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., Kordoni, V., and Fitzgerald, E. (2007a).

Partial parse selection for robust deep processing.

In *Proceedings of ACL 2007 Workshop on Deep Linguistic Processing*, pages 128–135, Prague, Czech.

For Further Reading II

Zhang, Y., Oepen, S., and Carroll, J. (2007b).

Efficiency in unification-based N-best parsing.

In *Proceedings of the 10th International Conference on Parsing Technologies (IWPT 2007)*, pages 48–59, Prague, Czech.