Grammar Engineering for Deep Linguistic Processing SS2012

Lecture 8: Test Suites & Treebanks

Yi Zhang

Department of Computational Linguistics & Phonetics
Saarland University

Language Technology Lab
German Research Center for Artificial Intelligence

July 2012

Comparison: Grammar Engineering vs. Software Engineering

- Design
- Module & Parallel/Distributed Development
- Testing & QA
- Maintenance

How to Properly Test Your Grammar?

- Test Suite: specially designed inputs with purposeful testing targets
- Corpus/Treebank: typically drawn from naturally occurring texts
- With both we can profile the evolution of grammar behavior between revisions

Test Suite

- For precision grammars, test suite should also contain negative test items, which are not easy to acquire from naturally occurring texts
- It is a good practice to harvest examples from available grammar books

Relation Between Grammar and Treebank

- The availability of precise grammar facilitates the quick development of large-scale treebanks
- The availability of treebank enables the development of (statistical) parse selection models for automatic disambiguation

Disambiguation Models

- Discriminative ranking models are trained with manually disambiguated treebanks
- Generative PCFGs are trained with large-scale auto-disambiguated treebank (tree-cache)

Generation

- Reverse process of parsing
- Semantic input in logic form
- Produce realization of natural language sentences with equivalent semantics
- Expensive computation due to the lack of linear order in the LF and semantically empty words

Grammar Checking & CALL

- ‘Mal-rule’ approach (Bender et al. 2004)
- Accept mild ungrammaticality
- Look up robustness symbols in the error code table
- Present appropriate message to student
- Generate grammatical sentences

Machine Translation

- MRS of different languages are aligned with rules
- MRS transfer rules as a graph rewriting system

$$[\mathcal{C} :]\mathcal{I}[!\mathcal{F}] \rightarrow \mathcal{O}$$

- Deep grammars serve as bridges between surface sentences and their LF representations