Comparing Test-Suite Based Evaluation and Corpus-Based Evaluation of a Wide-Coverage Grammar for English

Rashmi Prasad and Anoop Sarkar
Institute for Research in Cognitive Science
University of Pennsylvania
rjprasad,anoop@linc.cis.upenn.edu

Objective

- Which kind of evaluation metric is best suited for evaluating wide-coverage grammars like XTAG?
  - Test Suite-Based methods
  - Corpus-Based methods

Outline

- The grammar system
- Previous evaluations
- Current evaluation
- Comparison of the corpus-based and test suite-based approaches
- Proposal for a *combined* evaluation metric

The XTAG English Grammar

- Wide-coverage grammar based on the LTAG formalism
- Syntactic structures encoded as lexicalized *elementary trees*
- Parsing combines these elementary trees
- approx. 1.8 million lexicalized trees; average of 44.5 trees/word
- Grammar: tree templates and tree template families related by predicate-argument structure.

Previous Evaluations

Corpus-Based

Doran et al., (1994): WSJ and Brown

Error Class	No.	%
Paren./appos.	11	18.3%
Time NP	8	13.3%
Gapless Rel. cl.	3	5%
Comparative	1	1.6%
Bare infin.	1	1.6%
Multiword constr.	7	11.6%
Ellipsis	6	10%
Funny coordination	2	3.3%
Not sentences	6	10%
VP coordination	2	3.3%
Inverse predication	2	3.3%
Missing entry & subcat	9	14.9%
Unclassified	2	3.3%
Total	60	100%

Test Suite-Based

Doran et al., (1997): TSNLP test suite

Error Class	%
POS Tag	19.7%
Missing item in lexicon	43.3%
Missing tree	21.2%
Feature clashes	3%
Tokenization etc.	12.8%
Total	**100%**

Current Evaluations

Corpus-Based
Weather Reports

- Doran et al., (1997): only 20% (10/48) got a correct parse.
- Now, with error analysis and further grammar development, 89.6% (43/48) got a parse
- Problems due to relative clauses: (39.5%)

(a) *A frontal system approaching from the west*

(b) *The disturbance south of Nova Scotia early this morning*

∗ Small size of the test set for analysis convenience

Test Suite-Based
CSLI LKB Test Suite

- (2.7%) (26/966) did not get a correct parse

Error Class	No.	%
Missing Entry	4	0.4%
Lexicalized Tree	4	0.4%
Inverse Predications	2	0.2%
Ellipsis	18	1.8%
Default error	1	0.1%
Total	**26**	**2.7%**

Comparison

Corpora	Test Suites
Novel constructions previously not considered by grammar/test-suite developer Reduced relatives, parentheticals, appositives, time NPs, funny coordination, multiword constructions, etc.	Constructions recognized as linguistically interesting VP coordination, ellipsis, inverse predications, comparatives, etc.
interactions between several phenomena in a sentence reflect real world complexity in parsing 175 minutes for 48 sentences	Usually a single grammatical phenomenon in a sentence 41.5 minutes for 966 gr. sentences
Ample lexical variation	Same lexical items used often

- From the point of view of extending a wide-coverage grammar, a corpus-based evaluation is necessary.

Usefulness of Test-Suites

- Maintaining consistency of the grammar, but,
  - A test-suite tailored to the particular grammar is more desirable
- Coarse metric for comparison with other wide-coverage grammars.
- Accounting for certain rare phenomena that do not occur commonly in corpora: *If managers are not, a consultant is interviewing programmers*.

Disadvantage of Corpus-Based Approach

- Does not provide a method for locating the correct derivation
- Manual search is expensive
- For current evaluation: manual search made somewhat simpler
  - output: a shared forest of parses

Conclusions

- For the evaluation, maintenance and development of a wide-coverage grammar:
  - The test-suite evaluation approach is necessary but not sufficient
  - The corpus-based evaluation makes up for the disadvantages of the test-suite