Adapting Stochastic LFG Input for Semantics

Annette Hautli
Universität Konstanz
annette.hautli@uni-konstanz.de

Tracy Holloway King
Powerset Inc./Microsoft
tracy.king@microsoft.com

LFG c-structure and f-structure analyses provide the detailed syntactic structures necessary for subsequent semantic analysis (Dalrymple (1999, 2001)). The f-structure encodes grammatical functions as well as semantically relevant features like tense and number. The c-structure, in conjunction with the ϕ-mapping, provides the information on linear precedence necessary for semantic scope and anaphora resolution.

LFG has also proven an excellent theory for use in computational linguistics due to its computational and mathematical tractability. Cross-linguistic theoretical and implementational work has resulted in large-scale LFG grammars for typologically varied languages (Butt et al. (2002)). Such systems face two challenges. First, some constructions may be outside the scope of the grammar. This can arise when the input is ungrammatical, e.g. contains typos, or when a construction is not covered by the grammar, e.g. the construction is too computationally costly or too rare in the corpus to warrant inclusion in the grammar. Second, even highly efficient LFG implementations can be significantly slower than state-of-the-art stochastic parsers.

In this paper, we present an experiment in which a stochastic LFG-like grammar is the input to the semantics used in two meaning-sensitive applications, question answering (Bobrow et al. (2007)) and search. The semantics (Crouch and King (2006)) expects as input well-formed LFG c- and f-structures as created by the English ParGram grammar in conjunction the XLE LFG parser (Crouch et al. (2008)). The LFG-like grammar, created at Dublin City University and referred to here as the DCU English grammar, uses stochastic methods to create a c-structure and a pseudo f-structure (Cahill et al. (2002)). These pseudo f-structures do not necessarily obey LFG's completeness and coherence conditions, especially when long-distance dependencies are involved, and do not have all of the f-structure features that the XLE grammar provides.

A set of ordered rewrite rules (Crouch et al. (2008)) augments and reconfigures the output of the DCU English grammar in order to add more information to the stochastic output, thereby creating true LFG f-structures with all of the features that the semantics requires. For example, the following rule transfers the f-structure facts for any English common noun, adding NTYPE-related features:

```
pred(%1,%pred), num(%1,%NumVal), pers(%1,3), -proper(%1, %%)
==>
PRED(%1,%pred), NUM(%1,%NumVal), PERS(%1,3),
NTYPE(%1, %Ntype), NSYN(%Ntype, common), NSEM(%Ntype, %Nsem), COMMON(%Nsem, count).
```

For most open class items, general rules, such as the one above for common nouns, can be used; other lexical items, such as pronouns and determiners, require more specific, lexicalized rules to create the appropriate f-structure facts. In total, the system consists of ∼150 rules which use the same mechanism as the semantic rules. These rules primarily add additional features: the existing features in the DCU output are highly parallel to those of the XLE output since both systems follow ParGram project analyses (Butt et al. (2002)). Major differences between DCU and XLE grammar output exist for interrogative and imperative sentences, due to the lack of appropriate training data on the DCU side for these constructions; some of these differences could be corrected in the rules.

To test the system we determined how accurately the transferred DCU f-structures match the information in the XLE LFG f-structures. The results are shown below:

F-structure Match Results

	precision	recall	f-score
indicatives	87.13	82.67	84.84
imperatives	61.83	49.10	54.74
interrogatives	45.17	43.15	44.13

The ultimate goal of the experiment was to provide f-structures which allow a proper semantic analysis of the sentence. To match the semantic representation of transferred DCU f-structures and original XLE f-structures we created a corpus of queries and answer passages (e.g. *Although Mary likes vegetables, she eats them raw. Does Mary like vegetables?*). High scores for the passages indicate that the DCU English grammar can be used by the semantics in place of the XLE English grammar. The query semantics is much less accurate. To deal with this issue either the XLE grammar can be used on the query side or the DCU English grammar can be retrained on an augmented corpus, resulting in a grammar targeted specifically at queries (Judge et al. (2005)).

The success of this initial experiment suggest that pseudo LFG structures produced by stochastic grammars such as the DCU English grammar can be used for meaning-sensitive applications. They provide the advantages of LFG structures, e.g. the explicit encoding of grammatical functions in f-structure, in conjunction with the advantages of stochastic systems, e.g. providing connected parses in the face of less-than-ideal input. This suggests that the pseudo LFG structures can be used when no LFG grammar is available but a treebank of the language is: it can be faster to create a stochastic grammar instead of a rule-based one (Cahill et al. (2005)). When an LFG grammar does exist, the DCU grammars can be used in conjunction with the LFG grammar to replace it in out-of-coverage sentences: the XLE LFG grammars produce distinctive fragment parses when sentences are out of coverage; these can be replaced by the DCU pseudo f-structures to provide spanning c- and f-structures.

References

D. Bobrow, B. Cheslow, C. Condoravdi, L. Karttunen, T. H. King, R. Nairn, V. de Paiva, C. Price, and A. Zaenen. PARC's Bridge and Question Answering System. In *Grammar Engineering Across Frameworks*, pages 46–66, 2007.

M. Butt, H. Dyvik, T. H. King, H. Masuichi, and C. Rohrer. The Parallel Grammar Project. In *COLING2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7, 2002.

A. Cahill, M. McCarthy, J. van Genabith, and A. Way. Parsing with PCFGs and Automatic F-Structure Annotation. In *LFG02 Proceedings*, pages 76–95, 2002.

A. Cahill, M. Forst, M. Burke, M. McCarthy, R. O'Donovan, C. Rohrer, J. van Genabith, and A. Way. Treebank-Based Acquisition of Multilingual Unification Grammar Resources. *Journal of Research on Language and Computation*, pages 247–279, 2005.

D. Crouch and T. H. King. Semantics via F-structure Rewriting. In *LFG06 Proceedings*, 2006.

D. Crouch, M. Dalrymple, R. Kaplan, T. King, J. Maxwell, and P. Newman. XLE Documentation. 2008.

M. Dalrymple. *Lexical Functional Grammar*. Academic Press, 2001.

M. Dalrymple, editor. *Semantics and Syntax in Lexical Functional Grammar*. MIT Press, 1999.

J. Judge, M. Burke, A. Cahill, R. O'Donovan, J. van Genabith, and A. Way. Strong Domain Variation and Treebank-Induced LFG Resources. In *LFG05 Proceedings*, pages 186–204, 2005.