Automatic Acquisition of LFG Resources for German - As Good as it Gets

Ines Rehbein Josef van Genabith
School of Computing, Dublin City University, Ireland
{irehbein,josef}@computing.dcu.ie

Traditionally, deep, wide-coverage linguistic resources are hand-crafted and their creation is time-consuming and costly. Considerable effort has been put into overcoming this problem by automatically inducing linguistic resources such as rich, deep grammars, lexicons or subcategorisation frames from corpora. Most work so far has concentrated on English, like the one of Hockenmaier and Steedman [7], Nakanishi et al. [8] or Cahill et al. [3]. They present approaches for the acquisition of deep linguistic resources from the Penn-II treebank, using different grammar frameworks like CCG, HPSG and LFG. English, however, is a configurational language, where strict word-order constraints help to disambiguate predicate-argument structure. Porting these approaches to a semi-free word order language, the question arises: how good can it get? Can we expect similar results when dealing with (semi-)free word order? Can data-driven methods cope when dealing with ambiguous data structures and sparse data? And, furthermore, what impact has treebank design on the automatic acquisition of linguistic resources such as deep grammars?

This paper describes approaches to treebank-based acquisition of LFG resources for a semi-free word order language, based on and substantially extending the method of Cahill et al., O'Donovan et al. and Burke et al. [2, 9, 1], who presented large-scale acquisition of LFG grammars and lexical resources from the English Penn-II and Penn-III treebanks. They also presented work on data-driven multilingual unification grammar development for Spanish, Chinese and German. While treebank-based deep grammar acquisition can be applied to other languages, results are lower than the ones achieved for English and the Penn treebank. There are different possible reasons: first of all, the size of the English Penn-II treebank, which is much larger than most treebanks for other languages, might be responsible for the good results on English. Another reason might be the configurational English word order, where strict constraints determine the grammatical function of a lexical unit in a certain surface position. Finally, the good results for English might be due to the data structures in the Penn-II treebank, which might be optimised for the task at hand and thus improve performance on the English data.

In this paper we develop different f-structure Annotation Algorithms for German, based on two German treebanks with crucially different annotation schemes, adapted to feature sets of varying granularity as represented in three different gold standards. We discuss problems specific to the annotation schemes of the two treebanks as well as to language-specific properties of German, where the variability in word order and the richer morphology (compared to English) often result in data sparseness, causing severe problems for data-driven methods. We compare the performance of our data-driven grammar acquisition architectures with the hand-crafted German ParGram LFG of Dipper [4] and Rohrer and Forst [10, 5]. The automatically acquired grammars outperform the ParGram LFG with regard to coverage (Table 1). Rohrer and Forst [10] report 81.5% coverage on the NEGRA treebank [11], the automatically induced grammars achieve more than 88% coverage on the same data, but overall f-scores are higher for the hand-crafted LFG. Forst [5] tests parser quality on 1497 sentences from the TiGerDB and reports a lower bound, where a parse tree is chosen randomly from the parse forest, an upper bound, using the parse tree with the highest f-score (evaluated against the gold standard), as well as results for parse selection done by a log-linear disambiguation model. He reports an f-score of 75.7% (preds only) on the TiGer Dependency Bank (TiGer DB) [6], while our best TiGer DB-style grammar currently achieves an f-score of 72.7%. One reason for this are the low PCFG parsing results for German, especially with grammatical function labels assigned to phrasal nodes. These suggest an upper bound to

	ParGram			TiGerDB	DCU250
	up.	log.	low.		
GF	bound	lin.	bound		
da	67	63	55	44	38
gr	88	84	79	71	87
mo	70	63	62	65	73
oa	78	75	65	69	63
quant	70	68	67	67	78
rc	74	62	59	34	30
sb	76	73	68	74	79
preds only	79.4	75.7	72.6	72.7	78.6
	coverage on the NEGRA treebank (>20,000 sentences)				
	81.5	81.5	81.5	88.2	88.7

Table 1: F-scores for selected grammatical functions for the ParGram LFG (upper bounds, log-linear disambiguation model, lower bounds) and for two TiGer grammars

the task of treebank-based grammar acquisition based LFG parsing for German.

Results, however, are not directly comparable. The TiGer DB-based evaluation is biased in favour of the hand-crafted LFG, as the ParGram LFG grammar was used in the creation of the TiGer DB gold standard, ensuring compatibility as regards tokenisation and overall linguistic analysis. Therefore we created the DCU250, a dependency-based goldstandard including 250 sentences from the TiGer treebank with a set of grammatical features more closely adapted to the ones present in the treebank.[1] On the DCU250 the automatically acquired grammar achieves an f-score of 78.6 (preds only).

References

[1] Michael Burke, Olivia Lam, Aoife Cahill, Rowena Chan, Ruth O'Donovan, Adams Bodomo, Josef van Genabith, and Andy Way. Treebank-based acquisition of a chinese lexical-functional grammar. In *Proceedings of PACLIC-18*, Tokyo, Japan, 2004.

[2] A. Cahill, M. McCarthy, J. van Genabith, and A. Way. Automatic annotation of the penn-treebank with lfg f-structure information. In *LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation - Bootstrapping Annotated Language Data, LREC 2002, post-conference workshop*, pages 8–15, Paris, France, 2002.

[3] Aoife Cahill. *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. PhD dissertation, School of Computing, Dublin City University, Dublin 9, Ireland, 2004.

[4] Stefanie Dipper. *Implementing and Documenting Large-Scale Grammars*. PhD dissertation, IMS, University of Stuttgart, 2003.

[5] Martin Forst. Filling statistics with linguistics - property design for the disambiguation of german lfg parses. In *Proceedings of the ACL Workshop on Deep Linguistic Processing*, Prague, Czech Republic, 2007.

[6] Martin Forst, Nria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordoni. Towards a dependency-based gold standard for german parsers - the tiger dependency bank. In *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC '04)*, Geneva, Switzerland, 2004.

[7] Julia Hockenmaier and Mark Steedman. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of ACL 2002*, Philadelphia, PA, 2002.

[8] Hiroko Nakanishi, Yusuke Miyao, and Jun'ichi Tsujii. Using inverse lexical rules to acquire a wide-coverage lexicalized grammar. In *IJCNLP 2004 Workshop on Beyond Shallow Analyses*, Sanya City, Hainan Island, China, 2004.

[9] Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. Large-scale induction and evaluation of lexical resources from the Penn-II and Penn-III treebanks. *Computational Linguistics*, 31(3), 2005.

[10] Christian Rohrer and Martin Forst. Improving coverage and parsing quality of a large-scale lfg for german. In *Proceedings of LREC-2006*, Genoa, Italy, 2006.

[11] Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. An annotation scheme for free word order languages. In *Proceedings of ANLP*, Washington, D.C., 1997.

[1] The TiGer DB has a set of 16 governable and 18 non-governable grammatical functions, while the TiGer250 includes 14 governable and 13 non-governable grammatical functions.