Published in: Proceedings of the 11th International Conference on Parsing Technologies (IWPT). 7-9 October 2009 Paris, France. - Stroudsburg, PA: Association for Computational Linguistics, 2009. Pp. 134-137.

Scalable Discriminative Parsing for German

Yannick Versley
SFB 833
Universität Tübingen
versley@sfs.uni-tuebingen.de

Ines Rehbein
Dep. of Computational Linguistics
Universität des Saarlandes
rehbein@coli.uni-sb.de

Abstract

Generative lexicalized parsing models, which are the mainstay for probabilistic parsing of English, do not perform as well when applied to languages with different language-specific properties such as free(r) word order or rich morphology. For German and other non-English languages, linguistically motivated complex treebank transformations have been shown to improve performance within the framework of PCFG parsing, while generative lexicalized models do not seem to be as easily adaptable to these languages.

In this paper, we show a practical way to use grammatical functions as first-class citizens in a discriminative model that allows to extend annotated treebank grammars with rich feature sets without having to suffer from sparse data problems. We demonstrate the flexibility of the approach by integrating unsupervised PP attachment and POS-based word clusters into the parser.

1 Introduction

To capture the semantic relations inherent in a text, parsing has to recover both structural information and grammatical functions, which commonly coincide in English, but not in freer word order languages such as German. Instead one has to make use of morphological features in addition to exploiting ordering preferences such as the (violatable) default ordering of (*subject*<)*dative*<*accusative*.

Because of this fact, many successful approaches for German PCFG parsing (Schiehlen, 2004; Dubey, 2005; Versley, 2005) use *annotated treebank grammars* where the constituent trees from the treebank are enriched with further linguistic information that allows an adequate reconstruction of syntactic relationships, suggesting that probabilistic context-free grammars are an adequate tool for parsing these languages.

In the ACL 2008 workshop on Parsing German (Kübler, 2008), Rafferty and Manning (2008) used a lexicalized PCFG parser using markovization and parent annotation, but no linguistically inspired transformations; Rafferty and Manning did quite well on constituents, but were not successful in reconstructing grammatical functions, with results considerably worse than for other submissions in the shared task.

The framework we present in this paper – annotated treebank grammars with a discriminative model that allows lexicalization based on grammatical function assignment, as well as the addition of features based on unsupervised learning, including PP attachment and word clusters – shows that it is possible to achieve good improvements over generative lexicalized models by using the additional flexibility gained over standard lexicalized PCFG models. Our approach offers more flexibility than generative PCFG models, while computational costs for development and practical use are still acceptable. While we only present results for German, we are confident that the results carry over to other languages where annotated treebank grammars have been used successfully.

2 Parsing German with Morphology and Valence Information

As a base parser, we use BitPar (Schmid, 2004), a fast unlexicalized PCFG parser based on a first pass where non-probabilistic bottom-up parsing and top-down filtering is carried out efficiently by storing the chart in bit vectors, and construct the probabilistic chart only after top-down filtering. We use an annotated treebank PCFG that is de-

Proceedings of the 11th International Conference on Parsing Technologies (IWPT), pages 134–137,
Paris, October 2009.

rived from the Tiger treebank and largely inspired by earlier work on annotated treebank grammars for German (Schiehlen, 2004; Dubey, 2005; Versley, 2005).

Subcategorization With respect to the treebank grammar, we refine the node labels with linguistically important information that is only implicit in the treebank but would be tedious (and pointless) to annotate by hand:

Firstly, we annotate NPs by case; clause nodes (S and VP) are subcategorized by the clause type (*fin,inf,izu,rel*), and NPs and PPs with a relative pronoun are marked. Comparative phrases (e.g., *bigger [than a house]*, marked as NP in Tiger and TüBa-D/Z) are marked by adding a "CC" ending to the node label. Finally, auxiliaries are split according to their verb lemma into sein (*be*), haben (*have*), werden (*become*).

To aid the identification of noun phrase case, we add information related to case/number/gender syncretism to the preterminal labels of determiners, nouns, and adjectives (for details, see Versley, 2005) that allows to accurately determine the set of possible cases while keeping the size of the tagset relatively small .

Verb Valence We use information from the lexicon of the WCDG parser for German (Foth and Menzel, 2006) to mark verbs according to the arguments that they can take. While the WCDG lexicon contains more information, we only encode the possibility of accusative and dative complements, ignoring entries for genitive or clausal complements.

Markovization with Argument Marking It has been noted consistently (Klein and Manning, 2003; Schiehlen, 2004) that using markovization - replacing the original treebank rules by an approximation that only considers a limited context window of one or two siblings - improves results at least for a constituency-based evaluation. However, in some cases this simple markovization scheme leads to undesirable results including sentences with multiple subjects, as predicative arguments also have nominative case. To avoid this, we additionally mark which arguments have already been seen, yielding node labels such as `S_fin<VVFIN_a<RNP_a<sa` in the case of a partial constituent for a finite sentence (`S_fin`) expanding to the right (`<R`) where both subject (`s`) and accusative object (`a`) have already been seen.

Unknown Words For the base PCFG parse, we use a decision tree with 43 regular expressions as features, five of which are tailored towards recognizing the past and *zu*-infinitive form of separable prefix verbs (*ab*arbeiten ⇒ *ab*gearbeitet, *ab*zuarbeiten), which cannot be recognized by considering suffixes only. The extended part of speech tags for verbs (which contain valency information) are interpolated between the distribution at the concrete leaf of the decision tree and the global valency distribution for the (coarse) part-of-speech tag.

Additionally, we use SMOR (Schmid et al., 2004) in conjunction with the verb lexicon and a gazetteer list containing person and location names to determine possible fine-grained part-of-speech tags for unknown words.

Restoring Grammatical Functions Adding edge labels to the nodes in PCFG parsing easily creates sparse data problems, as reported by Rafferty and Manning (2008), who witness a drop in constituent F-measure (excluding grammatical function labels) when they include function labels in the symbols of their PCFG. On the other hand, the informativity of grammatical function labels for the contents of the node does not always justify their cost in terms of data sparseness. Thus, we chose an approach where we include linguistically relevant information in the node labels (see above), and use the finer categorization to restore the grammatical function labels automatically: Using the most frequent function label sequence associated with a rule yields good results even in the presence of markovization, where some of the surrounding context is lost. Furthermore, this approach allows us to use the grammatical function label assignments in the subsequent discriminative model, thus yielding typed dependencies rather than the unlabeled dependencies that are used in the lexicalization model of the Stanford parser.

3 Discriminative Parsing

Generative parsing models are based on few distributions that use different feature combinations based on smoothing; incorporating additional features into these is very difficult at best.

As a result, the use of external preferences in such parsers is usually limited to approaches that reattach dependents in the output of the parser rather than integrating them in the parsing process.

Settings	no GFs	with GFs
Rafferty and Manning (2008)	77.40	NA
—, training with GFs	72.09	60.48
markov[unlex]	74.66	62.47
markov+parent[unlex]	73.94	61.63
markovGF[unlex]	75.00	63.58
markov[lex]	77.68	66.05
markovGF[lex]	77.55	66.69
markovGF[+pp]	**78.43**	**67.90**

Table 1: Evaluation results: PARSEVAL F_1 on PaGe development set

fW-*w-pos*, CW-*w-pos*	word form, cluster
f-s_p, fS-s_p-*size*	node label, node size[(1)]
f-s_p-*RHS*	rule expansion
LD*dir*-s_p-s_d-hs_d	daughter attachment
LH-s_p-s_d-hs_d-hl_d	head projection
Ld*dir*-hs_p-hs_d	dependency (pos-pos)
Ld*dir*-hs_p-hs_d-*dist*	attachment length[(1)]
Le*dir*-hs_p-hs_d-hl_d	dependency (pos-lemma)
Lf*dir*-hs_p-hl_p-hs_d	dependency (pos-lemma)
Lf*dir*-hs_p-hl_p-hs_d-*GF*	typed dep. (lemma-pos)
Lh*GF*-hc_p-hl_p-hc_d-hl_d	typed dep. (lemma-lemma)
MIpp-*prep*, MIpp0-*prep*	PP attach (noun)
MIppV-*prep*, MIppV0-*prep*	PP attach (verb)

[1]) node sizes and attachment distances are discretized.
dir: one of H(head), L/R(head dep), B/I/E(nonheaded dep)
$s_{p/d}$ constituent symbol (parent/dep), $hs_{p/d}$ head cat, hc head cat (coarse), hl head lemma

Table 2: List of Features

Discriminative parsing for unification-based grammar commonly uses the conditional random field formulation introduced by Miyao and Tsujii (2002) and Geman and Johnson (2002), which uses local features to select a parse from a packed forest. The much larger cost in terms of memory and time compared to generative models has until recently made this approach largely unattractive (but see Finkel et al., 2008, who distributes the learning process over several powerful machines).

An alternative use of discriminative models has been to incorporate global features, either by reranking (e.g. Charniak and Johnson, 2005, or Kübler et al., 2009 for German) or by beam search over a pruned parse forest (Huang, 2008). However, Huang shows that a discriminative model using only local features reaps most of the benefits of the global model and performs at a similar level than earlier reranking-based approaches, pointing to the fact that local ambiguities often result in the n-best list not containing the correct parse.

The model we propose here extracts a pruned parse forest from a simple unlexicalized parser and then uses a factored discriminative model to apply a rich set of features using the lexicalized parse tree and its typed dependencies.

CRF parsing on pruned forests We extract a pruned forest that contains exactly those nodes and edges that can occur in trees that have a probability $\geq p_{best} \cdot t$, where in practice a threshold of $t = 10^{-3}$ ensures that no good parse is pruned away while at the same time, the resulting forest has only few nodes and edges.

For training, we extract an *oracle* tree, which is selected according to a combination of correct (annotated grammar) constituents, the absence of incorrect constituents, and the likelihood of the tree, to account for the fact that the forest does not always contain the exact gold tree. We then use the AMIS maximum entropy learner of Miyao and Tsujii (2002) to learn the discriminative model by creating a forest from a grammar learned on the remaining 4/5 of the training data.

Efficiency Parsing using the discriminative model is quite efficient, with a memory consumption for the whole system at about 270MB, including the data used to determine the corpus derived features (word clusters, mutual information statistics, semantic role clusters). Parsing speed is at 1.65sec./sentence on a 1.5GHz Pentium M, against 1.84sec./sent for BitPar alone when not using the tag filter for unknown words.

The time needed for learning can be reduced by keeping the pruned parse charts and only rerunning the part of lexicalization and discriminative feature extraction; when reusing the old parameters as a starting point for AMIS' model estimation, the turn-around time including feature extraction is below two hours.

3.1 Clustering for unknown words

To improve the behaviour on unknown words where morphological analyzer and regular expressions do not yield informative preferences, we exploit a large, part-of-speech-tagged corpus to induce clusters which provide robust information that is useful even in our case where preterminals in the PCFG are finer than standard POS tags.

The following features were gathered and used by weighting by the pointwise mutual information between the word and feature occurrences:

The **context** feature retrieves windows of high-frequent words surrounding the word in question

(e.g. *der_mit* for 'der *Mann* mit den Blumen').

The **context2** feature retrieves windows of one high-frequent word and one part-of-speech tag surrounding the word in question (e.g. *der_NN* for 'der *schöne* Mann').

The **postag** feature simply retrieves the part-of-speech tag that is assigned to the word.

The result of using the repeated bisecting k-means implementation of CLUTO (Steinbach et al., 2000) on the resulting features yields syntactically sensible clusters containing years, money sums, last names, or place names.

3.2 Unsupervised PP Attachment and Subject-Object preferences

We used simple part-of-speech tag patterns to gather statistics on the association between nouns and immediately following prepositions, as well as between prepositions and closely following verbs on the DE-WaC corpus (Baroni and Kilgariff, 2006), an 1.7G words sample of the German-language WWW. The mutual information values for PP attachment are made available to the parser as features that are weighted by the mutual information value.

4 Evaluation and Discussion

To evaluate our approach, we use the dataset used for the ACL-2008 Parsing German Workshop (Kübler, 2008) that contains 26,116 sentences of the TIGER treebank (Brants et al., 2002), in a 8:1:1 split of training, testing, and evaluation data, and validate our approach on the development data, where the results published by Rafferty and Manning (2008) provide a useful comparison. All our experiments are done using tags automatically assigned by the parser, which reaches a tagging accuracy of about 97.5% according to the EVALB output.

We find that our final model, combining augmenting the treebank labels with lingustic information in addition to lexicalization and unsupervised PP attachment works better than the best-performing models of Rafferty and Manning, with a very large improvement in grammatical functions that is only surpassed by the Berkeley Parser (Petrov and Klein, 2008), showing that our combination of annotated treebank grammars with a factored discriminative model not only allows great control and flexibility for experimenting with the inclusion of novel features, but also yields very good results compared with the state of the art for German (see table 1 for results on the Tiger treebank). Preliminary results on TüBa-D/Z with a subset of the transformations of Versley (2005) show the same tendency as the results for Tiger, with 91.3% for constituents only, and 80.1% including function labels (compared to 88.9% and 77.2% for the Stanford parser).

Future work will investigate the impact of including additional features into the discriminative parsing model.

References

Baroni, M. and Kilgariff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *EACL 2006*.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. TLT 2002*.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. ACL 2005*.

Dubey, A. (2005). What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *ACL-2005*.

Finkel, J. R., Kleeman, A., and Manning, C. D. (2008). Efficient, feature-based, conditional random field parsing. In *ACL/HLT-2008*.

Foth, K. and Menzel, W. (2006). Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *ACL 2006*.

Geman, S. and Johnson, M. (2002). Dynamic programming for parsing and estimation of stochastic unification-based grammars. In *ACL 2002*.

Huang, L. (2008). Forest reranking: Discriminative parsing with non-local features. In *HLT/ACL 2008*.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *ACL 2003*.

Kübler, S. (2008). The PaGe 2008 shared task on parsing German. In *Proceedings of the ACL-2008 Workshop on Parsing German*.

Kübler, S., Hinrichs, E., Maier, W., and Klett, E. (2009). Parsing coordinations. In *EACL 2009*.

Miyao, Y. and Tsujii, J. (2002). Maximum entropy estimation for feature forests. In *HLT 2002*.

Petrov, S. and Klein, D. (2008). Parsing German with latent variable grammars. In *Parsing German Workshop at ACL-HLT 2008*.

Rafferty, A. and Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL'08 workshop on Parsing German*.

Schiehlen, M. (2004). Annotation strategies for probabilistic parsing in German. In *Proc. Coling 2004*.

Schmid, H. (2004). Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proc. Coling 2004*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Versley, Y. (2005). Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.