POS Tagging for German: How Important is the Right Context?

Steliana Ivanova, Sandra Kübler

Umbria Inc., Boulder, CO, Indiana University Bloomington, IN, USA
sivanova@umbrialistens.com, skuebler@indiana.edu

Abstract

Part-of-Speech tagging is generally performed by Markov models, based on bigram or trigram models. While Markov models have a strong concentration on the left context of a word, many languages require the inclusion of right context for correct disambiguation. We show for German that the best results are reached by a combination of left and right context. If only left context is available, then changing the direction of analysis and going from right to left improves the results. In a version of MBT (Daelemans et al., 1996) with default parameter settings, the inclusion of the right context improved POS tagging accuracy from 94.00% to 96.08%, thus corroborating our hypothesis. The version with optimized parameters reaches 96.73%.

1. Introduction

Part-of-Speech (POS) tagging is generally performed by Markov models, based on bigram or trigram models. One of the best performing POS taggers based on Markov Models is TnT (Brants, 2000). The use of Markov models for this task rests on the assumption that a local context of one or two words to the left of the focus word is sufficient in the majority of cases to successfully disambiguate a word. The high accuracy rates of such POS taggers prove that this assumption is justified.

However, many languages, such as German, exhibit systematic ambiguities of words that can be correctly disambiguated only by the right context. In German, the definite determiners also serve as relative and demonstrative pronouns. In such cases, the left context generally does not provide enough information for the disambiguation of these words. Section 2. gives an example for this ambiguity class, as well as for further phenomena that fall into the same category.

The high frequency of these words necessitates investigating how important the right context is for POS tagging German. In this study, we conducted experiments to determine the optimal context for POS tagging. Since POS tagging is performed in a left-to-right fashion, the context to the right of the focus word cannot be treated identically to the left context. For the left context, the previously assigned POS tags are used. For the right context, in contrast, we use the ambiguity classes instead. POS taggers based on Markov models generally use only the left context (but see (Church, 1988) for an exception). And since the inclusion of right context in a Markov model would require a major redesign of the probability model in available Markov model POS taggers, we decided to use a POS tagger with independent features, which is more easily adaptable to our needs. The memory-based POS tagger, MBT (Daelemans et al., 1996), is an ideal candidate for this investigation.

Th structure of the paper is as follows: In the next section, we motivate linguistically why the use of the right context in German is necessary. in Section 3., we will discuss approaches to POS tagging that go beyond standard bigram and trigram tagging. Section 4. describes the experimental setup, Section 5. describes the results, and Section 6. investigates the types of errors that occur.

2. The Need for Right Context

In German, many high frequency words are ambiguous. As described above, definite determiners belong to this group. In the following example, all occurrences of *der* and *den* are ambiguous between definite determiners, relative pronouns, and demonstrative pronouns. In one case, the first occurrence of *den*, a context of two words to the left rules out the relative pronoun reading, but the remaining ambiguities can only be resolved by the right context.

(1) Beide wissen, der Anpassungsdruck an den
Both know, the peer pressure at the
High-Schools ist bereits jetzt enorm hoch -
high schools is already now enormously high -
der, den Mitschüler ausüben.
the one which classmates exert.

Another example for a systematic ambiguity that can only be resolved by right context is the ambiguity between separable verbal prefixes and prepositions: verbal prefixes occur at the end of clauses, while prepositions are followed by noun phrases. In the following example, there are two prepositions, *Im* and *in*, and two verbal particles, *durch* and *aus*. The only clear case is the first preposition because it is a merger of the preposition with a determiner. The other three cases are ambiguous between preposition and verbal particle. They can only be disambiguated with reference to the right context, which is a noun phrase for the preposition, and a comma and the end of the clause for the verbal particles.

(2) Im vorigen Jahr ging der Vorschlag knapp
In the previous year went the proposal barely
durch, in Schottland fiel das Ergebnis hingegen
through, in Scotland turned the result however
deutlich aus.
noticeable out.

'In the previous year, the proposal went through with a slim margin; in Scotland, however, the result was very clear.'

3. Previous Work

There is a long tradition of using bigrams and trigrams for POS tagging. However, relatively little work has been done

on using a more flexible context. Church (1988) presents an early trigram POS tagger that used two context words on the right rather than on the left. Toutanova et al. (2003) introduce bidirectional POS tagging. They use bidirectional dependency networks that have access to one word to the left and one word to the right. Tsuruoka et al. (2005) extend the model to a more flexible strategy that makes use of easy-first decisions, which allows decoding in polynomial time. While these models provide a very flexible architecture, which allows the inclusion of left and right context as needed, they are less suited for a more linguistically oriented investigation of how important certain types of context are. Banko and Moore (2004) introduce an unsupervised Hidden Markov Model that uses a context of one word to the left and one word to the right; however, only for the lexical probabilities.

4. Experimental Setup

The data used for the experiment was taken from the Tübingen Treebank of Written German, TüBa-D/Z (version 3) (Telljohann et al., 2006). TüBa-D/Z is a syntactically annotated corpus, tagged with the STTS tag set (Schiller et al., 1995). The corpus consists of newspaper articles from the German newspaper 'die tageszeitung' (taz) and at present, comprises 27 125 sentences, or 473 747 words. We used 90% of the data as training set and 10% for testing.

The experiment was conducted using MBT (Daelemans et al., 1996), a memory-based POS tagger-generator. MBT proceeds in two phases: generating a tagger using the memory-based learner TiMBL (Daelemans et al., 2004), and tagging text with the previously generated tagger. In the first phase, MBT takes as input a tagged text and creates a lexicon and case bases for known and unknown words. In the lexicon, every word is stored with its ambitag, an ambiguous tag representing the word's ambiguity class in the corpus. In a second phase, this knowledge is used to tag new text. This corresponds loosely to the training and tagging phase of a statistical POS tagger. MBT uses two different models, one for known words, and one for unknown words. The model for the known words is learned from a previously POS annotated training text. For unknown words, the learner uses words that occur infrequently in the text, thus using the assumption that the behavior of such infrequent words is similar to the behavior of words not seen in the training data. Memory-based learning is a learning method that assumes that decisions are based on previously seen events. It belongs to the lazy learning paradigm, i.e. the training instances are stored without modification or abstraction. When a new instance is to be classified, the learner selects the k most similar instances form the instance base (the k nearest neighbors) and uses the majority class assigned to these instances as the class of the new instance. Thus, if for a new word, 7 nearest neighbors with similar context are retrieved and 5 are assigned the POS tag preposition (APPR) and 2 the POS tag verbal particle (PTKVZ), the new word would be assigned the tag APPR.

Daelemans et al. (2003) show that only the joint optimization of system parameters and features gives optimal results. However, such an optimization would obscure the influence which a specific set of features has on tagging accuracy. For this reason, we used the default settings of TiMBL for determining the optimal context for POS tagging: the IGTREE memory-based machine learning algorithm for known words, the IB1 algorithm in combination with the overlap metric and gain ratio feature weighting for unknown words, and the number of nearest neighbors set to 1. In a second step, we optimized the parameter settings for the optimal feature set. The results of the feature optimization are shown in Section 5.1., the results for the parameter optimization in Section 5.2.

For known and unknown words	
d	the tag of left context
a	the ambitag of right context
w	left or right word
For known words only	
f	ambitag
W	word
For unknown words only	
F	position of the unknown word
c	the word contains capital letters
h	the word contains a hyphen
n	the word contains numerical character
p	character at the start of the word
s	character at the end of the word

Table 1: Tagging options

MBT provides a number of options for feature selection. Among them are the number of words from the left and right context, taking into account the actual word in addition to the tag, etc. The options are specified by two strings of symbols, one for known and one for unknown words. The symbols are represented in Table 1.

The string dfWaa, for example, denotes the tag of one word on the left, the ambitags of two words on the right, and the ambitag and the focus word form. For the experiments, we varied the options for context (left or right) and its size (from 0 to 2 words from each side) and kept all other options constant. For known words, we used the tag for left context (d), the ambitag for right context (a), and the ambitag (f) and actual word (W) for the focus word; for unknown words, we chose the tag/ambitag of the left/right context, the position of the unknown word (F), and the three characters from the beginning (p) and the end (s) of the word.

5. Results

5.1. Results for Different Context Sizes

The results of the experiments with regard to the ideal context are shown in row 1 of Table 2.

Traditionally, POS tagging is performed with bigram or trigram models. The MBT models closest to Markov models are the dfW (bigram) and the ddfW models (trigram). Our results for these models are somewhat lower than the results from TnT (Brants, 2000), which reached 97.04% on the same data set. This is due to the global optimization in Markov models as well as to TnT's elaborate unknown words module. Note, however, that the MBT exper-

	ddfWaa	dfWaa	ddfWa	fWaa	ddfW	dfW	fWa
forward	96.08	96.05	96.06	95.27	94.00	93.81	95.29
backward	96.06	96.05	96.02	95.97	95.26	95.28	95.39

Table 2: The results of tagging with different contexts.

iments were conducted with default parameter settings, as explained in Section 4..

As described in Section 2., there are a number of phenomena that give rise to the hypothesis that the use of right context can improve POS tagging results. This can be shown by the experiment in which two right context words are used instead of the left context words (cf. model fWaa vs. ddfW in row 1 of Table 2). This setting results in an increase of the tagger's accuracy from 94.00% to 95.27%. The results improve further with a context of two words on the left and one word on the right. In this configuration, the tagger reaches an accuracy of 96.06%. Marginally different results are reached with a context of one word to the left and two words to the right as well as with two words on both sides. Now, one might argue that the improvement from model ddfW to ddfWa might result from adding more context, not necessarily to the right of the focus word. However, a comparison of the results from a context of one word to the left (dfW) to a context of two words to the left (ddfW) provides a good counter-argument: Adding the second context word to the left improves results only marginally, from 93.81% to 94.00%. It is therefore very unlikely that adding more left context would result in a noticeable improvement.

Since adding right context gives the largest improvement, it is worth considering the change of directions in POS tagging, i.e. instead of performing the analysis from left to right, going from right to left. This approach has the advantage that the right context of a word is then already disambiguated, i.e. in POS tagging form right to left, we can use the previously assigned POS tags of the right context instead of the ambiguity classes for these words. This may provide more important information than having the ambiguity classes for the words on the right and the POS tags for the words on the left. For this reason, we conducted the same experiments again with the sentences in reverse order. The results of these experiments are shown in the second row in Table 2. Here, the sentences were presented in reverse word order. As a consequence, the model ddfWa means that the context consists of two words to the right, already disambiguated, and one word to the left, for which only the ambiguity class is known. It can be seen from the results that the change in tagging direction does not result in any improvement over the best results in forward tagging. Note that we have to compare the symmetrical cases: ddfWaa forward and backward; ddfWa forward as compared to dfWaa backward, and dfWaa forward as compared to ddfWa backward. These cases use the same context words, but with different portions already disambiguated. These comparisons show that there are only two feature setting for which there is an improvement, the trigram case (ddfW vs. fWaa) and the bigram case (dfW vs. fWa). As soon as context from both sides is available, there are no improvements gained form a right to left processing order. However, if only left context is available, then it is important that the right context is disambiguated first.

POS tag	number of improvements
PTKVZ	160
VVFIN	152
PDS	141
ART	141
NN	124
APPR	98
VVINF	89
ADJA	83
ADV	64
PIS	60

Table 3: The POS tags with the highest improvement from the ddfW model to ddfWa.

5.2. Results for Optimized Parameter Settings

In order to find the optimal machine learning parameter settings, we experimented with other options available in TiMBL. Among them are a variety of algorithms (IB1, IB2, IGTREE, TRIBL and TRIBL2), metrics (Overlap, Levenshtein, Modified Value Difference Metric (MVDM), Jeffrey divergence, dot product, cosine, in addition to metrics for numeric values and the option to ignore select features), as well as a number of TIMBL parameters. Not all of them are suitable for our tagging experiments. For all tests, we used the feature set consisting of 2 words of left and right context, which we found to be the optimal features setting in the previous experiments. The best results were obtained with the following settings: IB1 with Jeffrey divergence and k=5. With these settings, we reached an accuracy of 96.73%.

6. Error Analysis

In Section 5.1., we showed that adding right context improves results for POS tagging German texts. In order to determine which word classes profited the most from the extended context, we performed an error analysis. More specifically, we compared the results of the model that corresponds to a trigram model (ddfW) to the model with one word of right context added (ddfWa). A closer look at the POS tags that improved between the two models shows that the major improvements occur for verbal particles (PTKVZ), finite verbs (VVFIN), substituting demonstrative pronouns (PDS), and determiners (ART). Table 3 shows the ten POS tags that have the highest improvements with their improvement counts.

As explained in Section 2., there is a consistent ambiguity between verbal particles and prepositions (APPR). A look at the confusion sets shows that in 121 cases, the POS tag

was corrected from preposition to verbal particle. There were also 18 cases in which the label was corrected from determiner to verbal particle. These are the cases in which the particle was the word *ein*, which in a majority of the cases is a determiner.

The case of the finite verb is more complex. Here the corrected POS tags result from a set of tags: infinite verbs (VVFIN), past participles (VVPP), and attributive adjectives (ADJA). Many of these cases are corrected as a consequence of the correction of a preceding pronoun. The following sentence shows an example:

(3) Denn auch die gehen davon aus, daß
After all, even those assume *verb part.* that
sie ohne das BLG-Monopol preiswerter
they without the BLG monopoly more economically
arbeiten könnten.
work could.

'After all, even those assume that they could work more economically without the BLG monopoly.'

In the trigram model, *die* was tagged as a determiner, and as a consequence, *gehen* was tagged as an infinitival verb. In the model with a right context word, however, *die* was correctly tagged as a substituting demonstrative pronoun, which led in turn to the correct tagging of the following verb as being finite.

In other cases, finite verbs can be distinguished from infinite verbs by their right context. Infinite verbs are generally located on the right border of a clause while finite verbs are in second position in the main clause. In the following example, the verb *komplettierten* was unknown and tagged as an attributive adjective by the trigram model. The model with the right context had acess to the information that this word is followed by a determiner / relative pronoun / substituting demonstrative pronoun and thus tagged it correctly as finite.

(4) Toilettenhäuschen und Duschcontainer
Portable toilets and showers
komplettierten die Einrichtung dieses für 5 000
completed the facilities of this for 5 000
Menschen geplanten Lagers.
people planned camp.

'Portable toilets and showers completed the facilities of this camp that was planned for 5 000 people.'

7. Conclusions and Future Work

In this study, we investigated the importance of the right context of a word in POS tagging German. Since in German, many high frequency words are ambiguous and have an identical left context, the availability of right context is of great importance for disambiguation. We showed that the optimal model for German requires access to two words on both sides of the focus word. Such a setting improves performance by more than 2 percent points. When also optimizing the parameter settings, we obtain the best result of 96.73%.

For the future, we are planning to conduct equivalent experiments for a range of languages with different typological characteristics, such as English, Bulgarian, and Turkish. We expect to find improvements for all languages. However, we assume that each language will require a specialized context setting in order for MBT to reach optimal performance.

8. References

Michele Banko and Robert C. Moore. 2004. Part-of-speech tagging in context. In *Proceedings of COLING 2004*, Geneva, Switzerland.

Thorsten Brants. 2000. TnT–a statistical part-of-speech tagger. In *Proceedings of ANLP/NAACL'2000*, Seattle, WA.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second ACL Conference on Applied Natural Language Processing*, pages 136–143, Austin, TX.

Walter Daelemans, Jakub Zavrel, Peter Berck, and Steven Gillis. 1996. MBT: A memory-based part of speech tagger-generator. In Eva Ejerhed and Ido Dagan, editors, *Proceedings of the 4th Workshop on Very Large Corpora*, Copenhagen, Denmark.

Walter Daelemans, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. In *Proceedings of ECML-2003*, Cavtat-Dubrovnik, Croatia.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. TiMBL: Tilburg memory based learner – version 5.1 – reference guide. Technical Report ILK 04-02, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen, September.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister, 2006. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Procedings of HLT-NAACL 2003*, pages 252–259, Edmonton, Canada.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 467–474.