Evaluating POS Tagging under Sub-optimal Conditions. Or: Does Meticulousness Pay?

Sandra Kübler, Andreas Wagner
University of Tübingen
SFB 441: Linguistische Datenstrukturen
Köstlinstr. 6, D-72074 Tübingen, Germany
{kuebler, wagner}@sfs.nphil.uni-tuebingen.de

Abstract

In this paper, we investigate the role of sub-optimality in training data for part-of-speech tagging. In particular, we examine to what extent the size of the training corpus and certain types of errors in it affect the performance of the tagger. We distinguish four types of errors: If a word is assigned a wrong tag, this tag can belong to the ambiguity class of the word (i.e. to the set of possible tags for that word) or not; furthermore, the major syntactic category (e.g. 'N' or 'V') can be correctly assigned (e.g. if a finite verb is classified as an infinitive) or not (e.g. if a verb is classified as a noun). We empirically explore the decrease of performance that each of these error types causes for different sizes of the training set. Our results show that those types of errors that are easier to eliminate have a particularly negative effect on the performance. Thus, it is worthwhile concentrating on the elimination of these types of errors, especially if the training corpus is large.

1 Introduction

In the last few years, part-of-speech taggers have become widely used, efficient, and reliable tools in corpus linguistics. One of the problems most users encounter is the acquisition of a big enough corpus for training. If there does not already exist a tagged corpus for the language in question, usually an incremental approach to training the tagger is applied. For the incremental approach, one starts with a small amount of manually tagged data for the initial training, then runs the tagger on more data, which afterwards need to be corrected manually. This data set is then used as the next training set. The whole process continues until an acceptable level of performance is reached. (Another possibility, of course, would be unsupervised training methods, cf. [3], [5]; but even these papers suggest supervised training if a tagged training corpus is available, since this yields better performance.)

Although the manual correction of tagged data is less costly than tagging manually from scratch, it is still a very time consuming process. So the question arises if it is worth going through hours and hours of manual work or how this task could be rendered more efficient.

Questions concerning the proper size of the training corpus or the impact of noisy data are mainly ignored, usually out of practical considerations, e.g. the availability of already tagged data, or financial and temporal constraints.

In this paper, we will investigate these questions. In particular, we examine the influence of different sizes of the training corpus and different types of errors on the tagging performance. The error types we consider can be grouped along the following dimensions:

- Is the (wrong) tag part of the ambiguity class of the word (i.e. does it belong to the set of possible tags for that word) or not? (short: AMBI vs. NOT-AMBI)
- Is the major part-of-speech category correct (but not the finer distinctions or the morphological information) or not? (short: MCAT vs. NOT-MCAT)

All these four error types occur in the output of a tagger to such a degree that they cannot be neglected. Furthermore, it is possible to detect them automatically by comparison with a 'gold standard', i.e. the correctly tagged counterpart of the tagger output.

We assume that the 'negative' error types (NOT-AMBI and NOT-MCAT) will cause more damage in the tagging performance than the 'positive' types. If this hypothesis holds, these cases should receive more attention than the others in the process of correcting data for the next training round. Fortunately, both negative types are easier to check semi-automatically.

For NOT-AMBI, the checking can be done by comparing the current tag to the ambiguity class, i.e. the list of possible tags for the word. If it is not part of the ambiguity class, a manual inspection should be initiated. In this case, a lexicon with the ambiguity class for each word is a prerequisite. In case such a lexicon is not available, one can extract a (possibly imperfect) approximation of the ambiguity class for each word from the training corpus and then go through to look for suspicious tags, which presumably do not belong to the 'true' ambiguity class of the word. The occurences of these word-tag pairs must then be inspected manually.

For NOT-MCAT, the checking can also be done via the approximate ambiguity classes, again extracted from the current training corpus. Here, all words whose ambiguity class contains tags of more than one major category must be marked for further inspection.

2 The experimental setup

2.1 Resources

The two major decisions we faced here were the choice of the part-of-speech tagger and the choice of the training data.

As part-of-speech tagger we chose the transformation-based error-driven tagger developed by Eric Brill ([1], henceforth: Brill-tagger). This approach offers the advantage that unknown words are treated systematically by rules learned in the first training phase.

The corpus used for our tests was taken from the VERBMOBIL German treebank. VERBMOIL is a long term project for machine translation of spoken language. Within this project, a set of 30,000 German sentences is being syntactically annotated and compiled into a treebank ([7]). For the purpose of our tests, we ignore the syntactic annotations and only use the POS tags, which were automatically assigned and manually corrected. In order to assure high quality annotations, an additional automatic consistency check was run on the sentences.

The corpus as of now comprises approximately 27,000 sentences. As VERBMOBIL deals with spoken language, the decisions about sentence boundaries were left to the annotators who transliterated the data. Furthermore, VERBMOBIL is restricted to the domain of business appointments, travel scheduling, and hotel reservation. Therefore the corpus consists of a fairly homogenuous subset of the German language with a restricted vocabulary. The tagset used in the VERBMOBIL German treebank is the Stuttgart-Tübingen tagset (STTS, [6]). The STTS is based exclusively on the syntactic distribution of word forms. It also includes a certain amount of morphological information although the extended version of the tagset (cf. [8]) was not used. The STTS is widely accepted as a quasi-standard tagset for German and has found its way into the EAGLES guidelines (cf. [2]).

(1) Dienstag/NN würde/VAFIN mir/PPER gut/ADJD
Tuesday would me fine
passen/VVINF ./.
suit .
'Tuesday would suit me fine.'

In example (1), the POS tag 'VAFIN' signifies a finite auxiliary, 'VVINF' is an infinite full verb. 'NN' is a common noun, 'PPER' a personal pronoun, and 'ADJD' an adverbial adjective. The first character of the tag represents the major part-of-speech, e.g. 'N' for noun, 'V' for verb while the following characters make finer distinctions and provide some morphological information. For example, the 'A' in 'VAFIN' encodes the auxiliary, 'FIN' the finite form.

2.2 The different error types in the VERBMOBIL corpus

For each error type we would like to present an example from the faulty training corpus we created for our experiments (cf. below).

AMBI

(2) der/ART vierte/ADJA Januar/NN ist/VAFIN
the fourth January is
aber/KON ein/ART Dienstag/NN ./.
however a Tuesday .
'the fourth of January, however, is a Tuesday.'

'aber' can be a conjunction (KON) or an adverb (ADV) but in this case it should be tagged as an adverb.

NOT-AMBI

(3) wenn/KOUS ich/PPER eine/ART Uhrzeit/NN
if I a time
vorschlagen/VVINF darf/PIS ./.
suggest may .
'if I may suggest a time of day.'

'darf' is a finite modal (VMFIN), it can never be an indefinite pronoun (PIS).

MCAT

(4) das/PDS wären/VAFIN also/ADV für/APPR
that would-be therefore for
mich/PPER die/ART günstigsten/ADJD
me the most-suitable
Termine/NN ./.
dates .
'that would therefore be the most suitable dates for me.'

‘günstigsten’ is an adjective. In this case, however, it is attributive (ADJA), not predicative (ADJD).

NOT-MCAT

(5) und/KON ab/APPR der/ART
and from the
dreiunddreißigsten/VVPP Woche/NN
thirty-third week
fahre/VVFIN ich/PPER in/APPR Urlaub/NN ./.
go I in vacation .
‘and from the thirty third week on I am on vacation.’

‘dreiunddreißigsten’ is an adjective (ADJA), it can never be a participle (VVPP).

2.3 The test settings

We divided the above mentioned VERBMOBIL corpus of manually corrected sentences randomly into a test set of 3,010 sentences and a training set of 24,082 sentences. The training with the full training set resulted in an error rate of 1.95% on the test set.

Our next task was to create a version of the training set that contains a certain amount of errors. Basically, there are two alternatives to achieve this goal: the first is to run a badly trained tagger on the data, the second to randomly assign a wrong tag to one out of every n words. We chose the first alternative to obtain a more ‘realistic’ distribution of errors. As mentioned above, the training of a tagger is normally done incrementally, i.e. the output of the tagger is manually corrected and used as the training set in the next round. As the manual correction is a very tedious process, one can expect that some errors will be overlooked. Therefore the output of a badly trained tagger should be closer to an imperfect data set actually used for training a tagger.

To obtain such a badly trained tagger, we trained the Brill-tagger with a training set of 2,000 sentences. Then we ran this tagger on the full training set. That way, we introduced an error rate of 4.29% in the data. For our experiments, we recursively divided the set of sentences at random into half so that we got data sets with 24,082, 12,041, 6,020, 3,010, and 1,505 sentences. From these we created four different versions of every set of sentences, one version for each error type introduced above. Thus we obtained for each data set a version that only contained errors that were part of the ambiguity class (AMBI), that were not part of the ambiguity class (NOT-AMBI), that had the correct major category (MCAT), or that did not have the correct major category (NOT-MCAT), respectively. All the errors which were not of the intended type for the version in question were corrected by comparison with the original data. Additionally, we created the corresponding data sets from the original data, in order to have a gold standard for the different sizes of the training sets. This way, we obtained 25 different training sets, varying in size and error type, as well as in the rate of errors. Table 1 gives an overview of these error rates.

It is worth noting that in almost all cases the error rates for the type AMBI are higher than for the type NOT-AMBI. In contrast, the error rates for MCAT are noticeably lower than the ones for NOT-MCAT. The latter contradicts findings by Feldweg (cf. [4]) for a German HMM tagger. Feldweg reports that out of the 20 most common kinds of tagging errors, 12 are within the same major POS category. This discrepancy may be due to the fact that the HMM tagger that is used by Feldweg relies on the ambiguity classes from a lexicon and chooses one of the tags out of the ambiguity class for the word in question, based on the context. If therefore the ambiguity classes often contain tags from the same major category, erroneous choices within a major category are much more probable. [4] confirms the precondition for this assumption: “The elements of the most frequent ambiguity types for German, however, belong to the same major word classes, with only a few exceptions ...”. This fact can be explained to a certain extent by the setup of the tagset (Feldweg uses a predecessor of the STTS). Many German verbs, for example, are ambiguous with respect to finiteness and mood so they can be either ‘VVFIN’, ‘VVINF’, or ‘VVIMP’. The Brill-tagger, on the other hand, does not rely on ambiguity classes as a basis for the choice of tags. The transformations which are induced by the Brill-tagger do not systematically reflect the ambiguity classes.

For each combination of error type and size of the training corpus we trained the Brill-tagger. The trained tagger was then used to tag the test set of 3,010 sentences.

3 Results

Table 2 gives the error rates of the 25 test runs. One surprising result of the test with the 24,000 sentence gold standard is that we reached an accuracy of more than 98%. To our knowledge, this is the best result for any automatically trained tagger without external knowledge sources. This is certainly due to the rather high number of training sentences and to the homogeneity of the language data we used for the tests. Simple syntactic patterns like in greetings or in sentences like “wie sieht es denn bei Ihnen aus?” (“how does it look on your end?”) tend to be repeated several times.

A not very surprising result is that as the size of the training corpus decreases, the error rate increases monotonically. This holds true for all four error types, as well as for the gold standard.

The comparison of the gold standard with the different error types shows that for all sizes and error types, training with the gold standard resulted in the best performance on

	# sentences				
	24,082	12,041	6,020	3,010	1,505
gold standard	0.00%	0.00%	0.00%	0.00%	0.00%
AMBI	2.39%	2.41%	1.64%	2.52%	2.49%
NOT-AMBI	1.90%	1.93%	1.96%	2.05%	2.18%
MCAT	1.61%	1.61%	1.64%	1.71%	1.85%
NOT-MCAT	2.68%	2.73%	2.79%	2.85%	2.81%

Table 1. The error rates in the training sets

	# sentences				
	24,082	12,041	6,020	3,010	1,505
gold standard	1.95%	2.74%	3.15%	3.98%	5.35%
AMBI	2.86%	3.16%	4.06%	4.27%	5.43%
NOT-AMBI	3.57%	4.15%	4.59%	5.11%	6.35%
MCAT	2.87%	3.71%	3.93%	4.47%	5.67%
NOT-MCAT	3.99%	4.09%	4.47%	5.18%	6.15%

Table 2. The error rates in the test set

the test data. (The difference in performance between the gold standard and the different faulty training sets is statistically significant on the 1% level in all cases except for AMBI with 1,505 sentences.)[1] This result implies that a meticulous approach to correcting the training data will improve the accuracy of the tagger.

However, the results also show that eliminating different error types results in different degrees of improvement of the tagging performance.

For all different sizes of the training corpus, errors of type AMBI yield a lower error rate than errors of type NOT-AMBI. This also holds true for MCAT and NOT-MCAT, respectively. (The differences in performance are statistically significant on the 1% level in all cases.) The former result is especially striking with regard to the fact that the error rates in the training data are higher for AMBI than for NOT-AMBI, with one exception (6,020 sentences). In other words, less errors of type NOT-AMBI cause more problems than more errors of type AMBI. Thus, our findings provide evidence for our hypothesis explained in the introduction. Fortunately, as sketched above, NOT-AMBI and NOT-MCAT are easier to find and hence to eliminate than AMBI and MCAT.

Another remarkable result is that NOT-AMBI and NOT-MCAT yield comparable error rates on the test set (the differences in performance are not statistically significant on the 1% level in all cases except for the training set with 24,082 sentences) although the error rates in the training data are higher for NOT-MCAT than for NOT-AMBI. In other words, less errors of type NOT-AMBI cause the same amount of problems as more errors of type NOT-MCAT. This would suggest that looking for type NOT-AMBI is more efficient since one needs to correct fewer errors in order to obtain the same rise in accuracy.

Table 3 gives the error rate differences between the four error types and the gold standard for the different sizes of the training corpus. These figures show a tendency for decreasing improvement with decreasing size. These results indicate that for smaller sizes of the training corpus (cf. 1,505 and 3,010 sentences) it may not be worth investing the effort to look for particular error types. The table also shows that even if errors of type AMBI or MCAT are present in the training data, the performance is quite close to the gold standard, which again corroborates that these error types are less crucial.

4 Conclusion and future work

The process of training taggers is usually determined by practical constraints like the availability of resources or financial and/or temporal limitations. The impact of suboptimal training data on tagging performance is usually not considered systematically. We therefore wanted to explore the influence of the size of the training corpus combined with different error types in the training data. We suggested four error types which might influence the performance of the tagger to different degrees. In our tests we found that those error types which are easier to detect are the more harmful ones. Thus eliminating these specific errors in the training data will improve the tagging accuracy noticeably

[1] We used the McNemar Test to test statistical significance.

	# sentences				
	24,082	12,041	6,020	3,010	1,505
AMBI - gold standard	0.91%	0.42%	0.91%	0.29%	0.08%
NOT-AMBI - gold standard	1.62%	1.41%	1.44%	1.13%	1.00%
MCAT - gold standard	0.92%	0.97%	0.78%	0.49%	0.32%
NOT-MCAT - gold standard	2.04%	1.35%	1.32%	1.20%	0.80%

Table 3. The differences of error rates between the error types and the gold standard

at comparably low costs.

For the future, we are planning to extend this work to an HMM tagger and to corpora with different characteristics (e.g. written language or different domains) in order to see whether the regularities we found hold for other tagging methods and corpus types as well. Additionally, it is necessary to have a closer look at the dependencies among the four error types we suggested here. For example, there might be a considerable amount of overlap between the types NOT-AMBI and NOT-MCAT.

5 Acknowledgements

S. Kübler's contribution to this paper was funded both by the German Federal Ministry of Education, Science, Research, and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01 IV 101 N 0 and by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the Sonderforschungsbereich 441. A. Wagner was also funded by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the Sonderforschungsbereich 441.

References

[1] E. Brill. *A Corpus-Based Approach to Transformation-Based Learning*. PhD thesis, University of Pennsylvania, 1993.

[2] EAGLES. Morphosyntactic annotation, 1996. EAGLES document EAG-CSG/IR-T3.1.

[3] D. Elworthy. Does Baum-Welch re-estimation help tagging? In *Proceedings of the 4th ACL Conference on Applied Natural Language Processing*, pages 53 – 58. ACL, 1994.

[4] H. Feldweg. Implementation and evaluation of a German HMM for POS disambiguation. In *EACL Sigdat Workshop*. Dublin, 1995.

[5] B. Merialdo. Tagging english with a probabilistic model. *Computational Linguistics*, 20(2):155 – 171, 1994.

[6] A. Schiller, S. Teufel, and C. Thielen. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen, September 1995. (URL: http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html).

[7] R. Stegmann, H. Schulz, and E. W. Hinrichs. Stylebook for the German Treebank in VERBMOBIL. Universität Tübingen, 1998.

[8] C. Thielen and A. Schiller. Ein kleines und erweitertes Tagset fürs Deutsche. In H. Feldweg and E. Hinrichs, editors, *Lexikon & Text*, pages 215 – 226. Niemeyer, Tübingen, 1994.