

## Chapter 3

---

# Unsupervised Part of Speech Tagging with Extended Templates

MARKUS BECKER

**ABSTRACT.** In this paper we describe an extension of the unsupervised learning algorithm for automatically training a rule-based part of speech tagger, originally proposed by [Brill,1995]. We claim that the employment of templates with wider contexts will yield both linguistically more satisfying results and also higher precision rates.

The research underlying this paper was partially supported by research grants from the German Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) to the DFKI project PARADIME, FKZ ITW 9704.

## 1 Introduction

Recently quite a number of methods for the training of part of speech taggers have emerged. Most of these are statistically based, employing the training of Hidden Markov Models. [Brill,1992, Brill,1994] has shown that rule based Part-of-Speech (PoS) Taggers may perform comparably, with the added advantage of the rules being both more human readable and less space consuming than probability matrices of HMM's. However, the described algorithm is supervised, needing as input annotated corpora of a considerable size, hence the term "supervised", conceiving of the annotation as an act of human supervision to the algorithm.

In [Brill,1995] a rule based approach is described which is unsupervised. This means no manually tagged input is required. In this respect it is comparable to stochastic taggers which were trained using the Baum-Welch Algorithm [Baum,1972]. The algorithm is simply fed with an untagged corpus and

a dictionary providing the possible PoS tags for every word in the corpus. However, the rule based learner does not tend to overtrain in contrast to the stochastic algorithms, as Brill reports.

Though unsupervised training yields lower precision rates than supervised training, it has the following advantages. For most languages large annotated corpora are not available, whereas unannotated corpora in electronic form are abundant for most languages. Furthermore, the training of a tagger is domain dependent: Applying a trained tagger to a text of another domain usually results in a degradation of precision. Even for English not for every possible domain annotated corpora are available.

The unsupervised training algorithm originally proposed by Brill was extended in several directions by different authors. [Aone & Hausman,1996] describe an extension that allows for the parametrization of learning and tagging rules. They also handle the problem of unknown words, which was superficially circumvented by Brill, simply assuming a complete lexicon. [Satta & Henderson,1997] address efficiency problems of the learning algorithm through the use of suffix trees.

In our work we explore the impact different templates have on the accuracy of the trained tagger. The templates proposed by Brill examine only the immediate left and right contexts, thereby only setting up bigram statistics. Our templates are user definable and allow both for wider contexts and for stricter firing conditions of rules through conjunctions of contexts. Along the lines of [Aone & Hausman,1996], we also provide for parametrization of learning and tagging rules.

## 2 Unsupervised Training

The algorithm starts with the *Initial State Annotator* assigning all possible syntactic categories to every word of the raw corpus, using some electronic dictionary. This might either be readily available, or we might simply extract it from the tagged version of the raw corpus. In this way, there will be no unknown words in the initially tagged corpus. Taking this move for comparability with Brill's results, at least the *Initial State Annotator* hardly deserves the title "unsupervised".

The algorithm is supplied with a set of templates, which in [Brill,1995] is specified as follows:

Change the tag of a word from  $\chi$  to X if:

- The previous tag is T

- The previous word is W
- The next tag is T
- The next word is W

We call the set  $\chi$  the source categories, and X the target category.

In each iteration, the templates are expanded into rules on all ambiguous tokens in the corpus. From all expanded rules the best one is chosen according to a given scoring formula and applied to the whole corpus. The algorithm goes on until the best score drops below a specified threshold or a maximum number of iterations is exceeded. A typical rule may look like:

- (3.1) Change the tag from {NOUN or VERB} to NOUN if the previous tag is DETERMINER.

Training this algorithm on the *Penn Treebank* and the *Brown Corpus*, Brill claims to reach accuracy rates between 95.1% and 96.0%, depending on the size of the training corpus. Note that he assumes a complete lexicon.

### 3 Corpus and Tagset

For our training, we used a fraction of about 8000 words from the so called “German Collection”, a collection of texts from a variety of German newspapers. The corpus is tagged with the Stuttgart/Tübinger Tagset (STTS), which is described in [Schiller & Thielen, 1994]. This tagset comprises 52 part of speech tags, from which 3 tags are used only for the tagging of punctuation marks.

Let us have a look at an example. After the *Initial State Annotator* has performed tagging and the corpus was already trained for some iterations, the following sentence still has some ambiguities. Here an underscore denotes an ambiguity between some rivaling tags.<sup>1</sup>

- (3.2) ./ . Die/ART\_PRELS Kronen/NN ragen/VVFIN in/APPR  
den/ART\_PRELS unendlichen/ADJA Himmel/NN ./ .

‘The crowns reach into the infinite sky.’

---

<sup>1</sup>The tags have the following meanings: ART = article, PRELS = relative pronoun, NN = noun, VVFIN = finite verb, APPR = preposition, ADJA = attributive adjective.

After expanding the templates to rules and scoring these, the following rule emerges as the one with the highest score. It states that a tagging ambiguity between article or relative pronoun can be resolved to the tagging as an article, when the preceding token is a full stop.

(3.3) Change {ART\_PRELS} to ART  
when word in position (-1) is ”.”

This claim is in fact meaningful, since in German relative pronouns can hardly be found after a full stop. They rather follow a noun, separated by a comma.

(3.4) ... der Mann , der ...  
... the man who ...

When applied to our sentence, the rule resolves the ambiguity on the first word. Note that the ambiguity on the word “den” cannot be resolved, since the context of the word does not fit the description of the rule.

(3.5) ./ . Die/ART Kronen/NN ragen/VVFIN in/APPR  
den/ART\_PRELS unendlichen/ADJA Himmel/NN ./ .

## 4 Extensions

In this section we are going to describe the extensions we have implemented, in order to support the use of templates with expanded contexts.

### 4.1 User Definable Templates

We believe that an extension of the hardcoded templates described in Brill, which look only into the direct adjacent context, will yield both linguistically more satisfying results and also higher precision rates. For example, the trivial observation that in a noun phrase the noun can reliably be identified as such through the existence of a determiner *somewhere* to its left, possibly divided by a adjectival phrase, calls for wider contexts. Furthermore, phenomena like long distance dependencies can only be treated by corresponding *long distance* templates.

We implemented user definable templates in a Lisp-like syntax. In these templates only the contextual descriptions are coded. In order to characterize template (3.6), we have to specify two things in a list, as in (3.7). The

first item gives the positions that are to be inspected, the second item states whether we look at the word or its category.

(3.6) Change the tag of a word from  $\chi$  to X if the previous tag is T.

(3.7) ((-1) :cat)

This coding scheme allows us to specify disjunctive positions, as in (3.9), with the intended meaning:

(3.8) Change the tag of a word from  $\chi$  to X if one of the three previous words is W.

(3.9) ((-3 -2 -1) :word)

Since we provide for the conjunctive connection of templates, it is possible also to express more restrictive conditions, having the expressive power of  $n$ -gram conditions, with  $n > 2$ .

(3.10) Change the tag of a word from  $\chi$  to X if the previous tag is T and the following word is W.

(3.11) (:and ((-1) :cat)  
          (( 1) :word))

## 4.2 Parametrization for Training and Application

In our implementation we took care to keep the behaviour of the program as parametrizable as possible. We defined several switches, that control the behaviour of the scoring and applicational mechanism.

```
#define USE_AMBIGUOUS_CATS_FOR_FREQUENCY 1
```

While scoring, this switch tells whether frequency measures of categories are done only on base of unambiguously tagged words or also with ambiguously tagged words. When ambiguous categories are considered, they are counted not as a full hit but as the reciprocal of the number of categories on the

same token.

```
#define TESTPRED 0
```

This is a ternary switch determining the relation that must hold between the categories of a rule and the ambiguous categories on a token, so that the rule can fire. Between these two sets, the following relations are possible, ordered by the strength of their constraints:

- Identity
- Subset: The categories on the token form a subset of the rule categories
- Non-empty Intersection: The intersection of the two sets must be non-empty. Since by definition the target category is always included in the set of source categories, this amounts to the only requirement, that the target category of the rule must be a member of the token categories. This means, that the specification of the source categories  $\chi$  is redundant.

### 4.3 Efficiency

Efficiency issues were of special interest during the implementation, since the expanded templates called for more computational effort. Thus, we took care to save as many frequency computations as possible. For example, the frequencies are cached away in each iteration, thus avoiding multiple computations. Furthermore, it is possible to keep frequency results from one iteration to another, and to do only minor adjustments at those places, where the current rule was applied.

## 5 Results

Due to time restrictions, we have only run experiments on a small corpus. We used a fraction of the *German Collection* with a size of about 8.000 words. However, from fig. 3.1 the following tendencies can already be seen. The bottom line C1W1.p shows the progression of precision in the training phase with the original Brill templates, having a context width of 1 (hence C1W1: Categorical span = 1, Word span = 1). The initial precision rate is 87.3 %, climbing up to about 90 % after 600 iterations. Especially in the beginning, the graph shows some irregularities, falling quite drastically. The corresponding graph C1W1.c displays what we call the *chance rate* for

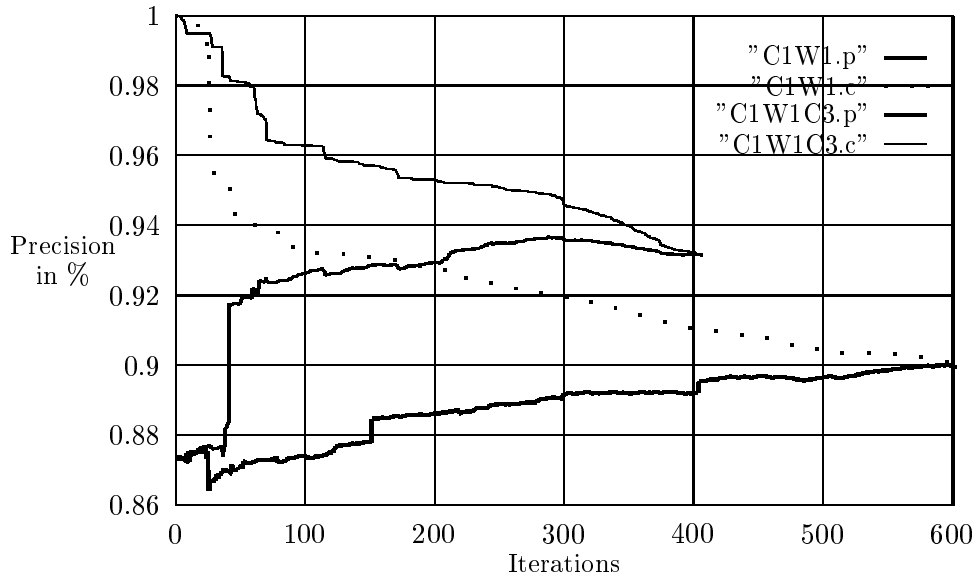


Figure 3.1: This figure displays the development of precision and chance rates for different template sets.

this training run, ie. the inverse of the error rate. We may also view it as the upper limit of the precision rate in each iteration. For example, right after the initial annotation, the tagger has not made any wrong decisions, so the chance to guess right all tags, still is 100 %, whereas in every following iteration we make some mistakes by removing good tags from words, increasing the error rate. Of course, since the trainer is only allowed to remove tags, errors cannot be undone anymore. Thus, the chance rate is a monotonically decreasing function, whereas the precision rate can go up *and* down. As we can see, C1W1.c goes down quite fast in the first 100 iterations.

On the other hand, adding the following two templates to the template set gives rise to the training graphs C1W1C3.p and C1W1C3.c.

Change the tag of a word from  $\chi$  to X if:

- one of the three previous tags is T
- one of the three following tags is T

After only 400 iterations, the training terminates with a precision rate of more than 93 %. Unfortunately, there is some overtraining, since the maximum rate of 93.6 % is reached already after about 300 iterations. The sudden rises from one iteration to another arise, when rule instantiations of wide context templates fire. This accelerates both training and application. Surprisingly, the chance rate (C1W1C3.c) falls much slower than with the restricted template set, even though the disjunctions might have led us to think that we could have a considerable error rate.

## 6 Future Work

We are going to perform systematic training runs on the *German Collection* and on the *Brown Corpus* with more training material, such as 100 K of words, varying the template sizes and the parameters explained above. It would be of interest to find out the relation between template size, different applications, and language. If it holds that the successful application of rules does not depend on the specification of the source categories, we might propose a new template format, where  $\chi$  is omitted, allowing for more general rules.

(3.12) Change the tag of a word to X in *Context*.

## Bibliography

- [Aone & Hausman,1996] Chinatsu Aone and Kevin Hausman. 1996. Unsupervised learning of a rule-based spanish part of speech tagger. In *Proceedings of Coling*.
- [Baum,1972] L. Baum. 1972. An inequality and association maximization technique in statistical estimation for probabilistic functions of a markov process. In *Inequalities*.
- [Brill,1992] Eric Brill. 1992. A simple rule based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*.
- [Brill,1994] Eric Brill. 1994. Some advances in rule based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*.
- [Brill,1995] Eric Brill. 1995. Unsupervised learning of disambiguation rules for part of speech tagging. In *3rd Workshop on Very Large Corpora*.
- [Satta & Henderson,1997] Giorgio Satta and John C. Henderson. 1997. String transformation learning. In *Proceedings of ACL*.
- [Schiller & Thielen, 1994] Christine Thielen and Anne Schiller. 1994. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text. Lexicographica Series Maior. Niemeyer. Tübingen*.