Internal and External Tagsets in Part-of-Speech Tagging

Thorsten Brants
Universität des Saarlandes
Computational Linguistics
D-66041 Saarbrücken, Germany
thorsten@coli.uni-sb.de

In Proceedings of Eurospeech 97, Rhodes, Greece, September 22 - 25, 1997

Abstract

We present an approach to statistical part-of-speech tagging that uses two different tagsets, one for its internal and one for its external representation. The internal tagset is used in the underlying Markov model, while the external tagset constitutes the output of the tagger. The internal tagset can be modified and optimized to increase tagging accuracy (with respect to the external tagset). We evaluate this approach in an experiment and show that it performs significantly better than approaches using only one tagset.

1 Introduction

The task of part-of-speech tagging is to assign a unique syntactical category (part-of-speech tag) to each word of an input stream. It is used as a component in parsing, for recognition in message extraction systems, for generating intonation in speech production systems, and many others.

Our work focuses on statistical part-of-speech tagging that is based on an underlying n-gram or Markov model. Tags are assigned by maximization of lexical probabilities $p(word_i|tag_i)$ and contextual probabilities $p(tag_i|tagi-1,\ldots)$. These probabilities are learned from a training corpus and are expected to be the same in unseen data (cf. (Church, 1988), (Cutting et al., 1992), (Jelinek, 1990), ...).

It is well known that the tagger misses information when it can only see surrounding tags and work with information contained in these tags. This is especially true for small contexts and small tagset sizes. Generally, more knowledge about a larger context helps in disambiguating the category of the current word (Schütze and Singer, 1994; Brants, 1995), but larger contexts have to be selected with care in order not to increase the sparse data problem. Also, the choice of a different tagset heavily influences accuracy (Elworthy, 1995).

In some cases, finer grained categories of the words in the context deliver the information needed for disambiguation. This fact is exploited in chunk parsing (Abney, 1996). The actual implementation of the parser works on *tag fixes* that change tags for particular words, i.e., the parser does not always use the original tag but a modified one based on the lexical entry of the corresponding word.

We concentrate on this point, and introduce an internal tagset for the representation of the Markov model that contains more information than the external tagset.

2 Internal and External Tagsets

Current work on part-of-speech tagging assumes that there is just one tagset: it is used for annotating the training corpus, and it is also used during actual tagging. But this is not necessarily the case. The manually corrected corpus may be build independently of a particular application. Corpora of this type are described and used, e.g., in (McEnery et al., 1994), (Leech et al., 1994), (Marcus et al., 1994), and (Skut et al., 1997).

Additionally, different applications impose different needs on the granularity of the tagset: the tags used by a parser are preferably different from those used by a speech generation component.

A commonly used procedure in cases where the corpus contains more information than needed is to strip all additional information from the corpus and then use the remaining data to generate a model. The steps of this procedure are shown for part-of-speech tagging on the left side of figure 1.

We propose a second method, shown on the right side of figure 1. It builds the model from all information contained in the corpus, assigns the complete set of information during application, and strips superfluous information as its last step.

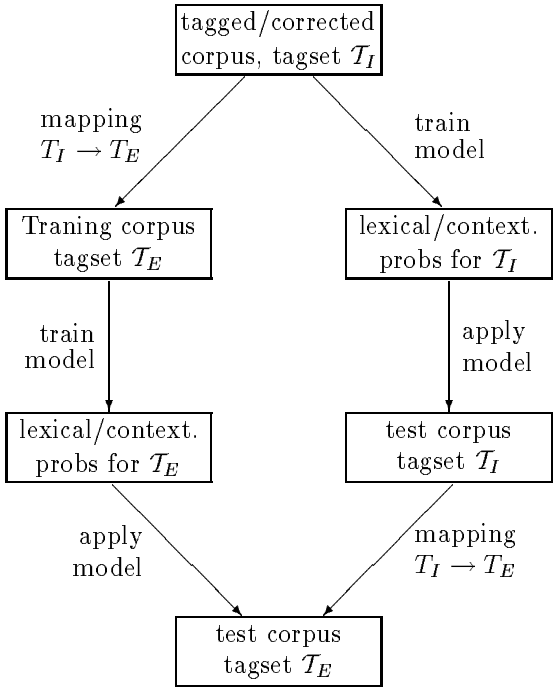

Figure 1: Two ways for handling additional information in the training corpus: stripping → training → application (left) vs. training → application → stripping (right).

Table 1: Sample elements of the four tagsets used in this paper, with different granularities of encoded information. The total counts of tags in the tagsets are shown in brackets.

Description	A (424)	B (159)	C (61)	D (14)
...				
her as possessive	APPGf	APPG	AP	pron
my as possessive	APPGi1	APPG	AP	pron
our	APPGi2	APPG	AP	pron
...				
sing count noun	NN1c	NN1	NN	noun
sing mass noun	NN1u	NN1	NN	noun
plur noun	NN2	NN2	NN	noun
...				
modal verb past tense	VMd	VM	VM	verb
model verb present t.	VMo	VM	VM	verb
...				
intr verb base form	VV0i	VV0	VV	verb
trans verb base form	VV0t	VV0	VV	verb
intr verb past tense	VVDi	VVD	VV	verb
...				

On the one hand, this method increases the number of categories and thereby increases the sparse data problem. But on the other hand there is more information encoded in the tags, and this information should help to disambiguate words in context. This can increase the overall tagging accuracy even if the additional information is assigned less reliably than the basic information contained in the tagset, since the additional information is stripped before emitting the tag.

We use the Susanne corpus for our investigations, which is part-of-speech tagged with a very fine-grained tagset comprising 424 tags (Sampson, 1995). Can we exploit the information contained in this fine-grained tagset even if we concentrate on the tagging accuracy for a smaller tagset?

To answer this question, we consider four tagsets: the original tagset A, consisting of 424 tags; tagset B, 159 tags; tagset C, 61 tags; and tagset D, 14 tags. Table 1 shows examples of the tags used in these tagsets and gives an impression of the granularity. Each tag in a larger tagset uniquely identifies a tag in a smaller tagset, i.e., the larger tagsets are proper extensions of the smaller ones.

In the following, two of these tagsets are selected at a time and our tagger will employ the larger one as its internal tagset $\mathcal{T}_I$, and the smaller one as its external tagset $\mathcal{T}_E$. The first one, $\mathcal{T}_I$, is used for training, in the internal representation of the Markov model and during tagging, but before emitting the tags are mapped to the external tagset $\mathcal{T}_E$ (according to the right half of figure 1).

If $\mathcal{T}_I$ and $\mathcal{T}_E$ are identical, the tagger performs like a standard tagger and nothing is changed. But if $\mathcal{T}_I$ is a proper extension of $\mathcal{T}_E$ (i.e., each tag in $\mathcal{T}_I$ uniquely identifies a tag in $\mathcal{T}_E$; e.g., both singular and plural noun in $\mathcal{T}_I$ identify a noun in $\mathcal{T}_E$), we use information that is not contained in the external tagset.

So, instead of calculating

$$\operatorname*{argmax}_{t_1 \ldots t_n} \prod_{j=1}^{n} p(t_j | t_{j-1} \ldots) p(w_j | t_j)$$

$w_j \in$ words, $t_j \in \mathcal{T}_E$, we now calculate

$$\operatorname*{argmax}_{t'_1 \ldots t'_n} \prod_{j=1}^{n} p(t'_j | t'_{j-1} \ldots) p(w_j | t'_j)$$

$t'_j \in \mathcal{T}_I$, thereafter map the tags

$$t_j = f(t'_j)$$

Table 2: Results of the standard tagging experiment for each of the four tagsets. The table shows the size of the tagset, the number of tags per word before disambiguation, the number of unknown words, the tagging accuracy, and the standard deviation.

tagset	size	avg tgs/wd	avg. unkn.	average accuracy	std-dev.
A	424	5.5	8.0%	93.8%	0.8
B	159	3.8	8.0%	95.1%	0.7
C	61	2.9	8.0%	95.0%	0.7
D	14	2.3	8.0%	94.5%	0.8

and emit t_j.

The additional information contained in $\mathcal{T}_I$ should be useful for disambiguation, thus it should increase the tagging accuracy. The experiments of the following section are intended to test this hypothesis.

3 Experiments

We have performed our experiments on the Susanne corpus (Sampson, 1995), which consists of approximately 150,000 tokens and is annotated for part of speech using 424 different tags. These tags can be grouped to form three smaller tagsets, consisting of 159, 61, and 14 tags (cf. figure 1 for examples).

Our tagger is a standard statistical, trigram-based part-of-speech tagger. The optimal Markovian path is calculated using the Viterbi algorithm. Sparse data is handled by linear interpolation of unigrams, bigrams, and trigrams. The weights for the interpolation are derived by deleted interpolation (Brown et al., 1992). The distinction of upper and lower case of characters is ignored. Unknown words are handled by using a suffix trie according to (Samuelsson, 1993).

The performance of the tagger in the standard technique of using a single tagset is shown in table 2. In order to yield reliable estimates of tagging accuracy, tagging is repeated 15 times for each tagset. Each time, 140,000 tokens of the corpus are used for training, and the rest (10,000 tokens) is used for testing (training and test part are ensured to be disjoint). The table shows the tagset, its size, the average number of tags per word before disambiguation in running text (this includes the number of tags for unknown words, which is generally much larger than for known words, but heavily depends on the suffix), the percentage of unknown words (i.e., words that were not seen during training but occur during testing), the average tagging accuracy and its standard deviation calculated from the 15 test runs for each tagset.

We see that the tagger performs best on the medium-sized tagsets, for which the accuracy is state-of-the-art, and that performance decreases both for the very large and the very small tagset. There is probably insufficient training data for tagset A (424 tags), and insufficient information for disambiguation contained in tagset D (14 tags).

Table 3: Accuracy and standard deviation when using different tagsets for the internal and external representations. As the baseline, we used identical internal and external tagsets.

	tagset (size)	external tagset A (424)	B (159)	C (61)	D (14)
base-line		93.8% 0.8	95.1% 0.8	95.0% 0.7	94.5% 0.7
in-ter-nal	A (424)	base-line	95.6% 0.7	95.9% 0.7	96.1% 0.7
	B (159)	– –	base-line	95.4% 0.7	95.8% 0.7
tag-set	C (61)	– –	– –	base-line	95.4% 0.7

Note that accuracy for tagsets B and C are almost identical despite the different tasks: there are much fewer tags to choose from when using tagset C, and the average number of tags per word in running text before disambiguation is lower (2.9 as opposed to 3.8 for tagset B), thus a priori the chance of making an error is lower for tagset C, but this is compensated for by the different amounts of information encoded in the tags.

In the second set of experiments we use the same tagger and the same tagsets, but before emission the tags are mapped to a smaller tagset. This means that training and the internal representation use the larger tagset, but our output employs the smaller tagset. We thereby simulate a training corpus that is annotated with a larger tagset, and measure the accuracy for a tagset with lower granularity.

The tagging accuracies of the different combinations and the standard deviations are shown in table 3. Each combination of tagsets is tested 15 times, using 140,000 tokens for training and the rest (10,000) for testing. Again, training and test parts are ensured to be disjoint.

As an example, the table entry in row A and column C (95.9%) shows the accuracy of a model that is trained using tagset A (424 tags), and that emits tags of tagset C (61 tags). The accuracy is always measured w.r.t. the external tagset. The result is significantly better than using tagset C for both the internal and external representation, shown as the baseline, which only yields 95.0% accuracy.

Comparing the results in table 3 with the baseline, we see that the tagging accuracy when using a larger

internal representation is always higher than when using the same internal and external representation. We also see that the larger the internal tagset the larger the accuracy for a smaller external tagset (within the bounds of the available tagsets). The increase in accuracy depends on the sizes of the involved tagsets, and ranges from 0.4 to 1.6 percent.

4 Conclusion

We have investigated the benefits of using an additional tagset for the internal representation of a part-of-speech tagger that is invisible to the external application. All information contained in the fine-grained categories of the internal tagset can be used for tagging, even if it is not needed as the output of the tagger. Using the information and stripping it before emitting the tags significantly increases the tagging accuracy (between 0.4 and 1.6% in our experiments).

As a consequence, efforts in manual annotation yield better language models the more information they assign to a text corpus.

Our experiments with internal and external tagsets of different sizes show that the benefit depends on the involved tagsets. Thus, given an external tagset, we assume that there is an optimal internal representation for the Markov model, and further work concerns the automatic modification of the internal tagset in order to find an optimal internal representation.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI'96 Robust Parsing Workshop*, Prague, Czech Republic.

Thorsten Brants. 1995. Estimating HMM topologies. In *Tbilisi Symposium on Language, Logic, and Computation*, Human Communication Research Centre, Edinburgh, HCRC/RP-72.

P. F. Brown, V. J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Kenneth Ward Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proc. Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, Texas, USA.

Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing (ACL)*, pages 133–140.

David Elworthy. 1995. Tagset design and inflected languages. In *Proc. of the Workshop on Very Large Corpora*, pages 1–9, Dublin, Ireland.

F. Jelinek. 1990. Self-organized language modeling for speech recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Kaufmann, San Mateo, CA.

G. Leech, R. Garside, and M Bryant. 1994. Claws4: The tagging of the british national corpus. In *Proceedings of the 15th International Conference on Computational Linguistics COLING-94*, pages 622–628, Kyoto, Japan.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*, San Francisco, Morgan Kaufmann.

A. M. McEnery, M. P. Oakes, R. Garside, J. Hutchinson, and G. N. Leech. 1994. The exploitation of parallel corpora in projects et10/63 and crater. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 108–115, Manchester, UK.

Geoffrey Sampson. 1995. *English for the Computer.* Oxford University Press, Oxford.

Christer Samuelsson. 1993. Morphological tagging based entirely on bayesian inference. In *9th Nordic Conference on Computational Linguistics*, Stockholm University, Stockholm, Sweden.

Hinrich Schütze and Yoram Singer. 1994. Part-of-speech tagging using a variable memory markov model. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pages 181–187, Las Cruces, NM.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of ANLP-97*, Washington, DC.