Tagging Spoken Language Using Written Language Statistics

Joakim Nivre Leif Grönqvist Malin Gustafsson Torbjörn Lager Sylvana Sofkova
Dept. of Linguistics
Göteborg University
S-41298 Göteborg
Sweden
{joakim,leifg,malin,lager,sylvana}@ling.gu.se

Abstract

This paper reports on two experiments with a probabilistic part-of-speech tagger, trained on a tagged corpus of *written* Swedish, being used to tag a corpus of (transcribed) *spoken* Swedish. The results indicate that with very little adaptations an accuracy rate of 85% can be achieved, with an accuracy rate for known words of 90%. In addition, two different treatments of pauses were explored but with no significant gain in accuracy under either condition.

1 Introduction

What happens when we take a probabilistic part-of-speech tagger trained on written language and try to use it on spoken language transcriptions? The answer to this question is interesting from several points of view, some more practical and some more theoretically oriented. From a practical point of view, it is interesting to know how well a written language tagger can perform on spoken language, because it may save us a lot of work if we can reuse existing taggers instead of developing new ones for spoken language. From a more theoretical point of view, the results of such an experiment may tell us something about the ways in which the structure of spoken language is different (or not so different) from that of written language.

In this paper, we report on experimental work dealing with the part-of-speech tagging of a corpus of (transcribed) spoken Swedish. The tagger used implements a standard probabilistic biclass model (see, e. g., (DeRose 1988)) trained on a tagged subset of the Stockholm-Umeå Corpus of written Swedish (Ejerhed et al 1992). Given that the transcriptions contain many modifications of standard orthography (in order to capture spoken language variants, reductions, etc.) a special lexicon had to be developed to map spoken language variants onto their canonical written language forms. In addition, a special tokenizer had to be developed to handle "meta-symbols" in the transcriptions, such as markers for pauses, overlapping speech, inaudible speech, etc. One of the interesting issues in this context is what use (if any) should be made of information about pauses, interruptions, etc. In the experiment reported here, we compare two different treatments of pauses and evaluate the performance of the tagger under these two different conditions.

2 Background

2.1 Probabilistic Part-of-speech Tagging

The problem of (automatically) assigning parts of speech to words in context has received a lot of attention within computational corpus linguistics. A variety of different methods have been investigated, most of which fall into two broad classes:

- Probabilistic methods, e. g. (DeRose 1988; Cutting et al 1992; Merialdo 1994).
- Rule-based methods, e. g. (Brodda 1982; Karlsson 1990; Koskenniemi 1990; Brill 1992).

Probabilistic taggers have typically been implemented as hidden Markov models, using probabilistic models with two kinds of basic probabilities:

- The *lexical probability* of seeing the word w given the part-of-speech t: $P(w \mid t)$.
- The *contextual probability* of seeing the part-of-speech t_i given the context of $n-1$ parts-of-speech: $P(t_i \mid t_{i-(n-1)}, \ldots, t_{i-1})$.

Models of this kind are usually referred to as n-class models, the most common instances of which are the biclass ($n = 2$) and triclass ($n = 3$) models. The lexical and contextual probabilities of an n-class tagger are usually estimated using one of two methods:[1]

[1] The terms 'RF training' and 'ML training' are taken from Merialdo 1994. It should be pointed out, though, that the use of relative frequencies to estimate occurrence probabilities is also a case of maximum likelihood estimation (MLE).

- Relative Frequency (RF) training: Given a tagged training corpus, the probabilities can be estimated with relative frequencies.
- Maximum Likelihood (ML) training: Given an untagged training corpus, the probabilities can be estimated using the Baum-Welch algorithm (also known as the Forward-Backward algorithm) (Baum 1972).

Of these two methods, RF training seems to give better estimations while being more labor intensive (Merialdo 1994). With proper training, n-class taggers typically reach an accuracy rate of about 95% for English texts (Charniak 1993), and similar results have been reported for other languages such as French and Swedish (Chanod & Tapanainen 1995; Brants & Samuelsson 1995).

2.2 Tagging Spoken Language

Spoken language transcriptions are essentially a kind of text, and can therefore be tagged with the methods used for other kinds of text. However, since the transcription of spoken language is a fairly labor-intensive tasks, the availability of suitable training corpora is much more limited than for ordinary written texts. One way to circumvent this problem is to use taggers trained on written texts to tag spoken language also. This has apparently been done successfully for the spoken language part of the British National Corpus, using the CLAWS tagger (Garside).

However, the application of written language taggers to spoken language is not entirely unproblematic. First of all, spoken language transcriptions are typically produced in a different format and with different conventions than ordinary written texts. For example, a transcription is likely to contain markers for pauses, (aspects of) prosody, overlapping speech, etc. Moreover, they do not usually contain the punctuation marks found in ordinary texts. This means that the application of a written language tagger to spoken language minimally requires a special *tokenizer*, i. e., a preprocessor segmenting the text into appropriate coding units (words).

A second type of difficulty arises from the fact that spoken language is often transcribed using non-standard orthography. Even if no phonetic transcription is used, most transcription conventions support the use of modified orthography to capture typical features of spoken language (such as *goin* instead of *going*, *kinda* instead of *kind of*, etc.). Thus, the application of a written language tagger to spoken language typically requires a special *lexicon*, mapping spoken language variants onto their canonical written language forms, in addition to a special tokenizer.

The problems considered so far may be seen as problems of a practical nature, but there is also a more fundamental problem with the use of written language statistics to analyze spoken language, namely that the probability estimates derived from written language may not be representative for spoken language. In the extreme case, some spoken language phenomena (such as hesitation markers) may be (nearly) non-existent in written language. But even for words and collocations that occur both in written and in spoken language, the occurrence probabilities may vary greatly between the two media. How this affects the performance of taggers and what methods can be used to overcome or circumvent the problems are issues that, surprisingly, do not seem to have been discussed in the literature at all. The present paper can be seen as a first attempt to explore this area.

2.3 Tagging Swedish

As far as we know, the methods for automatic part-of-speech tagging have not before been applied to (transcribed) spoken Swedish. For written Swedish, there are a few tagged corpora available, such as the Teleman corpus (see, e. g., (Brants & Samuelsson 1995)) and the Stockholm-Umeå Corpus (Ejerhed et al 1992). A subpart of the latter has been used as training data in the experiments reported below.

3 Method

3.1 The Tagger

The tagger used for the experiments is a standard HMM tagger using the Viterbi algorithm to calculate the most probable sequence of parts-of-speech for each string of words according to the following probabilistic biclass model:

(1) $$P(w_1, \ldots, w_n, t_1, \ldots, t_n) = P(t_1)P(w_1 \mid t_1) \prod_{i=2}^{n} P(t_i \mid t_{i-1})P(w_i \mid t_i)$$

The tagger is coupled with a tokenizer that segments a transcription into utterances (strings of words), that are fed to the tagger one by one. Besides ordinary words, the utterances may also contain markers for pauses and inaudible stretches of speech.[2]

3.2 Training the Tagger

The lexical and contextual probabilities were estimated with relative frequencies in a tagged corpus of written Swedish, a subpart of the Stockholm-Umeå Corpus (SUC) containing 122,377 word tokens (18,343 word types). The tagset included 27 parts-of-speech.[3]

[2] The original transcriptions also contain information about overlapping speech, marking of certain aspects of prosody, and various comments. This information is currently disregarded by the tokenizer.

[3] For a more detailed description of the linguistic annotation system of the Stockholm-Umeå Corpus, see (Ejerhed et al 1992).

3.3 The Spoken Language Lexicon

As noted earlier, the spoken language transcriptions contain many deviations from standard orthography. Therefore, in order to make optimal use of the written language statistics, a special lexicon is required to map spoken language variants onto their canonical written forms. For the present experiments we have developed a lexicon covering 2113 spoken language variants (which are mapped onto 1764 written language forms). We know, however, that this lexicon has less than total coverage and that many regular spoken language reductions are not currently covered.[4]

3.4 Unknown Words and Collocations

The occurrence of "unknown words", i. e., words not occurring in the training corpus, is a notorious problem in (probabilistic) part-of-speech tagging. In our case, this problem is even more serious, since we know beforehand that some words will be treated as unknown although they do in fact occur in the training corpus (because of deviations from standard orthography). In the experiments reported below, we have allowed unknown words to belong to any part-of-speech (which is possible in the given context), but with different weightings for different parts-of-speech. More precisely, when a word cannot be found in the lexicon, we replace the product in (2) (cf. equation 1 above) with the product in (3), where $TTR(t_i)$ is the type-token ratio of t_i (in the training corpus).

(2) $P(t_i \mid t_{i-1})P(w_i \mid t_i)$

(3) $P(t_i \mid t_{i-1})P(t_i)TTR(t_i)$

In this way, we favor parts-of-speech with high probability and high type-token ratio. In practice, this favors open classes (such as nouns, verbs, adjectives) over closed classes (determiners, conjunctions, etc.), and more frequent ones (e. g., nouns) over less frequent ones (e. g., adjectives).

In addition to "unknown words", we have to deal with "unknown collocations", i. e., biclasses that do not occur in the training data. If these biclasses are simply assigned zero probability, then — in the extreme case — a word which is in the lexicon may fail to get a tag because the contextual probabilities of all its known parts-of-speech are zero in the given context. In order to prevent this, we use the following formula to assign contextual probabilities to unknown collocations:

(4) $P(t_i \mid t_{i-1}) = P(t_i)K$

The constant K is chosen in such a way that the contextual probabilities defined by equation (4) are significantly lower than the "real" contextual probabilities derived from the training corpus, so that they only come into play when no known collocation is possible.

3.5 Pauses and Inaudible Speech

As indicated earlier, the utterances to be tagged included markers for pauses and inaudible speech, since these were thought to contain information relevant for the tagging process. The symbol for inaudible (and therefore untranscribed) speech — (...) — was simply added to the lexicon and assigned the "part-of-speech" `major delimiter` (`mad`), which is the category assigned to full stops, etc. in written texts. The result is that the tagger will not treat the last word before the untranscribed passage as immediate context for the first word after the passage.

For pauses we have experimented with two different treatments, which are compared below. We refer to these different treatments as tagging condition 1 and 2, respectively:

- Condition 1: Pauses are simply ignored in the tagging process, which means that the last word before a pause is treated as immediate context for the first word after the pause.
- Condition 2: Pause symbols are added to the lexicon, where short pauses are categorized as `minor delimiters` (`mid`) (commas, etc.), while long pauses are categorized as `mad` (full stops, etc.), which means that the contextual probabilities of words occurring before and after pauses in spoken language will be modelled on the probabilities of words occurring before and after certain punctuation marks in written language.

It was hypothesized that, in certain cases, the tagger might perform better under condition 2, since pauses in spoken language often — though by no means always — indicate major phrase boundaries or even breaks in the grammatical structure.

3.6 Test Corpus

The test corpus was composed of a set of 47 utterances, chosen randomly from a corpus of transcribed spoken Swedish containing 267,206 words. The utterance length varied from 1 word to 688 words (not counting pauses as words), with a mean length of 29 words. The test corpus contained 1360 word tokens and 498 word types.

4 Results

The number of correctly tagged word tokens under condition 1 was 1153 out of a total of 1360, i. e., 84.8%. The results for condition 2 were slightly better: 1248/1457 = 85.7%. However, the latter figures also include the tagged pauses, for which only one category was possible. If these tokens are subtracted, the results for condition 2 are: 1151/1360 = 84.6%.

[4] A common example is the ending *-igt*, which appears in many adjectives (neuter singular) and adverbs and which is usually reduced to *-it* in ordinary speech.

5 Discussion

The overall accuracy rate for the tagger is around 85%, which is not too impressive when compared to the results reported for written language. However, if we take a closer look at the results, it seems that an important source of error is the lack of coverage of the lexicon and the training corpus. Of the two hundred or so errors made by the tagger, more than eighty concern tokens that could not be matched with any word form occurring in the training corpus. The most common type of error in this class is that a word is erroneously tagged as a noun. It is likely that this is an artifact of the way we assign lexical probabilities to unknown words and that a more sophisticated method may improve the results for this class of words. More importantly, though, if we only consider the results for words that were known to the tagger, the accuracy rate goes up to about 90%, and most of the errors remaining concern classes that are notoriously difficult even under normal circumstances, such as adverbs vs verb particles and prepositions vs subordinating conjunctions. Taken together, these results seem to indicate that with a more extensive lexicon, a larger training corpus of written language, and perhaps a more sophisticated treatment of unknown words, it should be possible to obtain results approaching those obtained for written language.

As regards the two treatments of pauses, the results are virtually identical in terms of overall accuracy rate. If we look at individual words, however, we find that the part-of-speech assignment differs in 25 cases. In 10 of these cases, the correct part-of-speech is assigned under condition 1; in 9 cases, the correct tag is found under condition 2; and in 6 cases, both conditions yield an incorrect assignment. The conclusion to draw from these results is probably that the treatment of pauses as delimiters yields a better analysis in cases where the pause marks an interruption or major phrase boundary, while it is better to ignore pauses when they do not mark any break in grammatical structure. Unfortunately, these two types of pauses seem to be equally common, which means that neither treatment results in any gain in overall accuracy. However, preliminary observations seem to indicate that it may be possible to get better results if a more fine-grained analysis of pause length is taken into account. This presupposes, of course, that this kind of information is available in the transcriptions.

6 Conclusion

In this paper we have reported on an experiment using a probabilistic part-of-speech tagger trained on written language to analyze (transcribed) spoken language. The results indicate that, with little or no adaptations, an overall accuracy rate of 85% can be achieved, with an accuracy rate of 90% for known words. On the negative side, we found that the treatment of pauses as delimiters (as opposed to simply ignoring them) did not result in a better performance of the tagger.

References

Merialdo, B. (1995) Tagging English Text with a Probabilistic Model. *Computational Linguistics* 20, 155–171.

Baum, L. E. (1972) An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process. *Inequalities* 3, 1–8.

Brants, T. & Samuelsson, C. (1995) Tagging the Teleman Corpus. In *Proceedings of the 10th Nordic Conference of Computational Linguistics, NODALIDA-95*, Helsinki, 7–20.

Brill, E. (1992) A Simple Rule-based Part of Speech Tagger. In *Third Conference of Applied Natural Language Processing*, ACL.

Brodda, B. (1982) Problems with Tagging and a Solution. *Nordic Journal of Linguistics* 5, 93–116.

Chanod, J.-P. & Tapanainen, P. (1995) Tagging French — Comparing a Statistical and a Constraint-based Method. In *Seventh Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 149–156.

Charniak, E. (1993) *Statistical Language Learning*. Cambridge, MA: MIT Press.

Cutting, D., Kupiec, J., Pedersen, J. & Sibun, P. (1992) A Practical Part-of-speech Tagger. In *Third Conference on Applied Natural Language Processing*, ACL, 133–140.

DeRose, S. J. (1988) Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics* 14, 31–39.

Ejerhed, E., Källgren, G., Wennstedt, O. & Åström, M. (1992) The Linguistic Annotation System of the Stockholm-Umeå Corpus Project. Report 33. University of Umeå: Department of Linguistics.

Garside, R., Using CLAWS to Annotate the British National Corpus. [http://info.ox.ac.uk:80/bnc/garside_allc.html].

Karlsson, F. (1990) Constraint Grammar as a System for Parsing Running Text. In *Proceedings of COLING-90*, Helsinki, 168–173.

Koskenniemi, K. (1990) Finite-state Parsing and Disambiguation. In *Proceedings of COLING-90*, Helsinki, 229–232.