Porting a Stochastic Part-of-Speech Tagger to Swedish

Douglass R. Cutting
Cupertino

Abstract

The Xerox Part-of-Speech Tagger (XPOST) claims to be *practical*. One aspect of practicality as defined here is *reusability*. Thus it is meant to be easy to port XPOST to a new language. To test this, XPOST was ported to Swedish. This port is described and evaluated.

Practical Part-of-Speech Tagging

In previous work on part-of-speech tagging, a *practical* part-of-speech tagger was defined as one with the following set of properties (Cutting *et al* 1992):[1]

- **accurate**

```
A     tagger should assign the  correct part  of
det/2 n      modal  v      det  adj/2   n/2   prep

speech to     every word  in     the  text.
n      prep/2 det   n     prep/4 det  n
```

  While 100% accuracy is desirable, it may not in fact be achievable. When text is manually tagged by several linguists, the tags assigned differ by a few percent, suggesting an effective upper-bound for tagging accuracy (Church 1989).

- **fast**
  Ideally, the addition of part-of-speech tagging to a system will not significantly alter the speed with which text is processed. This may be difficult to evaluate, as systems which incorporate tagging may not operate at all without tagging. As a surrogate, one may compare the cost of assigning tags with that of simply extracting words from text—*tokenization*. If tagging is not significantly slower than tokenizing then its performance impact on complex text processing systems should certainly be minimal.

[1] The Xerox Part-of-Speech tagger is available for anonymous FTP from parftp.xerox.com.

Proceedings of NODALIDA 1993, pages 65-70

- **robust**
  A tagger should correctly tag text previously unseen by the system. It must accommodate previously unseen words, as unseen texts frequently contain unseen words. New grammatical constructions may also be encountered. Ideally a tagger will accommodate these too. However, before addressing these, one should ask: do previously unseen items occur at such a rate that handling them incorrectly affects overall accuracy? Taggers typically answer this affirmatively for new lexemes and negatively for new constructions.

- **reusable**
  It should be possible to easily configure a tagger to handle a broad range of texts and tasks. Texts vary in things as mundane as typographic conventions and as fundamental as natural languages. Different tasks may require different tagsets, e.g., some may need to distinguish subject and object pronouns, while others may not.

The author previously helped construct the Xerox Part-of-Speech Tagger (XPOST) in an attempt to meet these criteria. The present paper first reviews XPOST in the light of recent Scandinavian work on part-of-speech tagging. It then describes the author's experiences porting XPOST to Swedish while visiting the Swedish Institute for Computer Science (SICS).

Stochastic Part-of-Speech Tagging

Stochastic part-of-speech taggers operate by constructing a probabilistic model of text; then estimating the probabilities of the model, or *training*, and finally, using the trained model to assign parts-of-speech to previously unseen text. The models employed typically contain two sorts of probabilities: *transition probabilities* and *symbol probabilities*. Transition probabilities are recorded for sequences of tags, usually pairs, and indicate the probability of that sequence occurring in text, e.g. the probability that a determiner is followed by a noun. Symbol probabilities record the likelihood that a given input item, typically a word, assumes a given part of speech, e.g. the probability that "bank" is a verb. Given an input sequence (e.g. a sentence) and these two sets of probabilities, one may compute the probability of each possible tag assignment by multiplying all the applicable symbol and transition probabilities. The tag assignment with the highest such product is selected as most likely. This simple methodology has been shown to work quite well (Church 1988).[1]

[1]Comparable results have been achieved with non-stochastic methods (Eineborg *et al* 1993, Voutilainen *et al* 1992).

A weakness with this approach is that symbol probabilities are difficult to estimate for words. A substantial portion of text is composed of low-frequency words. For these words, there are not enough observations make accurate estimates of symbol probabilities. And words which are unknown when training have no observations at all. Compounding this problem, Samuelsson has shown that symbol probabilities are more significant in improving accuracy than transition probabilities (Samuelsson 1993). Together these suggest that, if one is not satisfied with the accuracy of a stochastic part-of-speech tagger, one should attempt to improve symbol probability estimation.

A common approach to such sparse-data problems is to develop an alternate representation which pools data into coarser categories, increasing the number of observations of each of a smaller set of phenomena. In XPOST, each word is represented by its *ambiguity class*—the set of tags it may assume. All words in an ambiguity class are considered identical, and their observations may thus be pooled to provide better estimates.

XPOST guesses ambiguity classes for unknown words based on their suffixes. Frequencies of suffixes of a known words in a text are analyzed to generate a table which, given a suffix, names the ambiguity class which accounts for the vast majority of the words with that suffix. This is similar to the method for handling unknown words proposed by Eklund (1993; 1994).

Another weakness of many stochastic taggers is their reliance upon hand-tagged corpora for training. While hand-tagged corpora do provide accurate estimates, they are very expensive to produce. XPOST avoids reliance on hand-tagged corpora by using a hidden Markov model (HMM).[1] The Baum-Welch (or *forward-backward*) algorithm enables one to estimate symbol and transition probabilities of an HMM without hand-tagged training data (Baum 1972).

The Baum-Welch algorithm operates by incrementally adjusting probabilities to make the training data more likely. One can steer it out of local-maxima by initializing some of the probabilities manually. For example, one might initialize the transition probability between determiner and noun to be higher than the transition probability between determiner and verb. In effect, this permits one to specify simple *a priori* grammatical constraints. Here we see that stochastic taggers are not purely data-driven and self-organizing, as is sometimes claimed by those

[1]HMMs have been used in other taggers, but not in combination with ambiguity classes (Jelinek 1985).

promoting grammar-based taggers, but rather permit integration of linguistic knowledge.

Performance of XPOST on English

On the Brown text collection (Francis *et al* 1982) XPOST achieves the following results:

- **accuracy**: the correct tag is assigned to 96% of the words (88% of the ambiguous words). This accuracy is comparable to that achieved by other stochastic part-of-speech taggers trained on tagged data.

- **robustness**: In experiments on texts with unknown words, 77% of unknown words are tagged correctly.

```
Corandic is    an   emurient grof  with  many  fribs ; it
n        v3sg  det  adj      n     prep  adj   pl      pro

granks  from  corite, an   olg  which  cargs  like   lange .
v3sg    prep  n       det  n    pro    v3sg   prep   n
```

- **reusability**: has been ported to French, German[1] and Swedish (described subsequently).

- **speed**: in Common Lisp on a Sun SPARCStation2 the tagger requires approximately one millisecond per word tagged with the Brown tagset. With 38 tags in 174 ambiguity classes, this tagset is reasonably large. Tagset size is a factor in speed, so one can expect better performance with a smaller tagset. Note that, even with this tagset, tagging (including lexicon lookup) operates at approximately the same speed as tokenization.

Average μseconds per word			
tokenizer	lexicon	tagging	total
604	388	233	1235

XPOST thus appears to meet most of the criteria for practical part-of-speech tagging.

Porting XPOST to Swedish

Teleman's corpus of tagged Swedish was used to evaluate XPOST on Swedish (Teleman 1974). The methodology was similar to that used for

[1]For more information contact Helmut Schmid <schmid@ims.uni-stuttgart.de>.

the Brown corpus. First a lexicon was induced from the entire collection containing, for each word, the list of the tags which it may be assigned. The corpus was then divided into two sections, one containing the even numbered sentences and one containing the odd numbered sentences. The former were used, without tags, to train XPOST. The latter sentences were then automatically tagged by XPOST. The tags thus assigned were then compared with the tags assigned by Teleman.

The Teleman tagged corpus contains around 85 thousand words tagged with 259 unique tags. Many of these tags occur very infrequently in the corpus, making parameter estimation difficult. The tagset was thus initially recoded to the 13 tags specified by Samuelsson (Samuelsson 1993). With this tagset, XPOST tagged 91% of the words correctly.

Examination of the errors suggested that XPOST might do better with a somewhat more refined tagset. This is not usually a good idea, as it creates more parameters to be trained, and hence, less evidence per parameter. The addition of some distinctions may not change the number of ambiguous forms, but may provide more precise grammatical contexts for disambiguating neighboring ambiguous forms. By this logic, genitive names and nouns were broken out as separate tags, and pronouns were broken into four categories: relative, personal, genitive and object. After these changes accuracy rose to 95%.

Issues

Some issues which remain to be examined in stochastic taggers include:

- Should common words be modeled individually? Some authors have proposed (e.g. Kupiec 1992) have proposed that high-frequency words should have their own ambiguity class, even if the set of tags in the class is not distinct from that in other classes. The Swedish word "om" might benefit from this treatment. It is a high-frequency word which may be used as an adverb, a conjunction or a preposition. Other words with the same ambiguity, e.g. "efter" and "sedan", are infrequent enough to benefit from having their statistics pooled, while "om" is frequent enough that it may fare better on its own.

- Voutilainen *et al* (1992) have developed a tagging method which achieves high accuracy, but which moreover, can accurately predict its errors. In other words, rather than generating the wrong tags, it is able to pass the ambiguity along so that it may be resolved by higher-level processing. This is clearly a superior property. It remains to be seen if a stochastic tagger can implement this.

Acknowledgments

Thanks to Jussi Karlgren for suggesting that I visit SICS, to Christer Samuelsson for arranging my stay, and to the whole SICS NLP group for making that stay fun.

Bibliography

Baum, L.E. 1972. *An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process.* In INEQUALITIES. 3: pp. 1–8.

Church, K. 1989. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.* In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing.*

Church, K.W. 1988. *A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.* In *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas.

Cutting, D., J. Kupiec, J. Pedersen and P. Sibun. 1992. *A Practical Part-of-Speech Tagger.* In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Eineborg, Martin and Björn Gambäck. 1993. *Tagging Experiments Using Neural Networks.* In Eklund (ed.) *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistik-dagarna', Stockholm 3-5 June 1993.* Stockholm..

Eklund, Robert. 1994. (ed) *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistik-dagarna', Stockholm 3-5 June 1993.* Stockholm.

Eklund, Robert. 1994. *A Probabilistic Word Class Tagging Module Based On Surface Pattern Matching.* In Eklund (ed), *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3-5 June 1993.* Stockholm.

Eklund, Robert. 1993, *A Probabilistic Word Class Tagging Module Based On Surface Pattern Matching.* Stockholm University, Department of Linguistics, Computational Linguistics.

Francis, W.N. and F. Kucera. 1982. *Frequency Analysis of English Usage.* Houghton Mifflin.

Jelinek, F. 1985. *Markov Source Modeling of Text Generation.* In *Impact of Processing Techniques on Communication*, J.K. Skwirzinski, Editor. Nijhoff: Dordrecht.

Kupiec, J.M. 1992. *Robust Part-of-Speech Tagging Using a Hidden Markov Model.* Xerox Palo Alto Research Center.

Samuelsson, Christer. 1993. *A Morphological Tagger Based Entirely on Bayesian Inference.* In Eklund (ed): *Nodalida '93 – Proceedings of '9:e Nordiska Datalingvistikdagarna', Stockholm 3-5 June 1993*, Stockholm.

Teleman, U. 1974. *Manual för Grammatisk Beskrivning av Talad och Skriven Svenska.* University of Lund.

Voutilainen, Atro, Juja Heikkilä and Arto Anttila. 1992. Constraint Grammar of English. Department of Linguistics, University of Helsinki.