The Open Source Tagger HunPoS for Swedish

Beáta B. Megyesi

Department of Linguistics and Philology
Uppsala University
beata.megyesi@lingfil.uu.se

Abstract

HunPoS, a freely available open source part-of-speech tagger—a reimplementation of one of the best performing taggers, TnT—is applied to Swedish and evaluated when the tagger is trained on various sizes of training data. The tagger's accuracy is compared to other data-driven taggers for Swedish. The results show that the tagging performance of HunPoS is as accurate as TnT and can be used efficiently to tag running text.

1 Introduction

In the last decade, several data-driven part-of-speech taggers have been successfully developed, such as MXPOST (Ratnaparkhi, 1996) based on the maximum entropy framework, the memory-based tagger (MBT) (Daelemans et al., 1996), Brill's tagger (TBL) based on transformation-based learning (Brill, 1995), and Trigram 'n' Tags (TnT) based on Hidden Markov models (Brants, 2000). These taggers are freely available for research purposes but not for industrial use, and in many cases they are not open in the sense that the user does not have access to the source files, hence she/he cannot make any changes to fit the tagger to his/her needs.

One of the best performing taggers among the data-driven tools is the Trigrams 'n' Tags, shortly TnT (Brants, 2000). Recently, HunPoS (Halácsy et al., 2007), a reimplementation of TnT was released, allowing the user to tune the tagger by using different feature settings.

The goal of our work is to find out how the open source tagger HunPos performs when applied to Swedish compared to other data-driven taggers. We apply HunPos to Swedish by training the tagger on the Stockholm Umeå Corpus. We vary the size of the training data and the features used for tagging unknown and known tokens. We then compare the results to other data-driven taggers when applied to Swedish.

The paper is structured as follows. First, we briefly describe HunPoS and the data sets used for training and testing the tagger. Then, we present the experiments with different feature settings while we vary the size of the training data followed by the comparison to other taggers. Lastly, we conclude the paper.

2 HunPoS Applied to Swedish

HunPoS is based on Hidden Markov Models with trigram language models similar to TnT with the difference that the tagger also estimates lexical/emission probabilities based on the current tag and previous tags. For the treatment of unseen words, TnT's suffix guessing algorithm is also implemented where the length of the last characters can be varied as well as the frequency required for a particular word to appear in in order to be taken into account in the learning for guessing the tag for unknown words.

In our study, the tagger is trained with various feature settings in order to find out which setting is the most appropriate for Swedish. In addition, these feature settings were tested on training data of various sizes from one thousand tokens to one million.

For the tagging experiments we use the Stockholm Umeå Corpus (Ejerhed and Källgren, 1997), henceforth SUC, which is a balanced corpus, consisting of over one million tokens. The tokens in the corpus are lemmatized, and tagged with their syntactically correct part-of-speech and morphological features. The corpus is publicly available and free for research purposes.[1] For annotation scheme, we use the PAROLE tagset (Ejerhed and Ridings, 1997) consisting of 156 tags.

[1] For more information about SUC, see http://www.ling.su.se/DaLi/projects/SUC/.

data size	t2,e2	t2,e1	t1,e2	t1,e1
1000	68.07	68.25	**68.71**	68.70
2000	75.11	75.23	**75.75**	75.65
5000	81.29	81.41	82.15	**82.19**
10000	84.55	84.67	85.12	**85.23**
20000	88.03	88.10	88.24	**88.26**
50000	91.19	**91.22**	91.11	91.10
100000	93.13	**93.15**	92.95	92.95
200000	**94.35**	94.34	93.98	93.93
500000	**95.34**	95.27	94.87	94.80
1000000	**95.90**	95.79	95.38	95.25

Table 1: Tagger performance with various tag- and emission order given various size of training data.

To train the tagger on various sizes of data, we reuse the same split as has been used previously for the comparison of different data-driven taggers when applied to Swedish, as described in (Megyesi, 2002). From a randomly ordered set extracted from SUC, the training sets and the test data were taken. The size of the training data varies from 1 000, 2 000, 5 000, 10 000, 20 000, 50 000, 100 000, 200 000, 500 000 to 1 000 000 tokens, and the separate test set contains 117 685 tokens containing 7 464 sentences.

2.1 Experiments with Feature Settings

We run several experiments to train the tagger with different feature settings. First, we experiment with the order of tag transitions ($-t$) using either bigram tagging ($-t1$) or default trigram tagging (-$t2$). As for the lexical probabilities, we test emission order e by either setting the tag order NUM to 1, where $NUM = 1 \rightarrow P(w_i|t_i)$ or using the default tag order NUM set to 2 where $NUM = 2 \rightarrow P(w_i|t_{i-1}t_i)$. The results of the combination of these features are shown in Table 1. Not surprisingly, bigram models better fit to smaller training data containing less than 50 000 tokens while trigram models are to prefer when we use larger data sets, over 50 000 tokens, for training.

For unknown words, there is a possibility to vary the length of the suffixes that the tagger uses to build a suffix tree. In this study, we tested suffixes of length 10 (default), 9, and 5 to see if a decrease in suffix length can increase performance. Looking at the results given in the second and third columns in Table 2, we can conclude that there is an increase in error rate by reducing the length of the suffixes independently of the size of the training data. For Swedish, suffix length set to 10 yields best results.

As the next step, we also vary the frequency with which a word can occur to be added to the suffix tree. Column four and five in Table 2 show that for small amounts of training data consisting of less than 100 000 tokens, tagger performance can be improved by reducing the frequency requirement for words to be added to the suffix tree. For larger training corpora, the default setting of the tagger can be used, i.e., setting the frequency to 10.

2.2 HunPoS Compared to other Data-Driven Taggers

Lastly, given the default feature setting of the tagger, we compare the result achieved by HunPoS to other taggers' performance when trained on the same data set and evaluated on the separate but same test set. Table 3 lists the data size, the baseline—calculated by assuming unknown words to be common nouns (NCUSNIS), and when capitalized, proper nouns (NP00N0S) and known words receiving their most frequently occurring tag—followed by the accuracy of the MBT tagger (MBT), the MXPOST tagger (ME), Brill's tagger (TBL), TnT, and lastly HunPoS with default settings (HP-default) and HunPoS optimized (HP-best). HunPoS has highest accuracy when trained on small training data consisting of less than 20 000 tokens, while TnT achieves highest performance for the other data sets with the exception of training on one million tokens where both taggers achieve comparable results. The difference in performance between TnT and HunPoS when trained on the largest data set is not significant using McNemar's test ($p <= 0.827$ with 95% confidence level) and the freely available open source system is therefore a good alternative to use.

3 Concluding Remarks

We applied a freely available open source tagger HunPoS to Swedish and trained with different feature settings for the tagging model and lexical probabilities, as well as for the treatment of unknown words. We can conclude that for larger training data consisting of above 200 000 tokens, the default settings of the tagger can be used while for smaller data sets, features for lexical proba-

data size	s10-s9+f10	s5+f10	s10+f9	s10+f5
1000	68.07	**68.08**	68.17	**68.45**
2000	**75.11**	75.10	75.11	**75.32**
5000	**81.29**	81.27	81.29	**81.47**
10000	**84.55**	84.52	84.56	**84.57**
20000	**88.03**	88.02	**88.04**	**88.04**
50000	**91.19**	91.14	91.18	91.18
100000	**93.13**	93.08	93.13	**93.14**
200000	**94.35**	94.30	**94.35**	94.34
500000	**95.34**	95.29	**95.34**	**95.34**
1000000	**95.90**	95.86	**95.90**	95.89

Table 2: Tagger performance for unknown words with different feature settings and using the default model ($t2, e2$) given various size of training data.

data size	baseline	MB	ME	TBL	TnT	HP-default	HP-best
1000	48.68	62.91	53.41	61.10	67.98	68.07	**68.71**
2000	50.90	69.36	61.86	63.44	74.87	75.11	**75.75**
5000	58.19	75.90	72.73	70.49	81.72	81.29	**82.19**
10000	63.60	79.30	78.08	74.62	85.05	84.55	**85.23**
20000	67.19	82.84	82.96	80.32	88.25	88.03	**88.26**
50000	72.77	86.47	88.06	85.33	**91.34**	91.19	91.22
100000	76.89	88.87	90.69	89.84	**93.23**	93.13	93.15
200000	80.18	90.51	92.53	92.40	**94.41**	94.35	94.35
500000	83.55	92.30	94.18	93.45	**95.39**	95.34	95.34
1000000	85.49	93.94	—	92.74	95.89	95.90	**95.90**

Table 3: Performance of taggers given various size of training data.

bilities and treatment of unknown words shall be adapted. Lastly, we also compared the tagging accuracy of HunPoS to the performance of other data-driven taggers applied to Swedish. We conclude that HunPoS is a good alternative to TnT which is one of the best performing taggers today.

References

Thorsten Brants. 2000. Tnt - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-00)*, Seattle, Washington, USA.

Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21:543–566.

W. Daelemans, J. Zavrel, P. Berck, and S. Gillis. 1996. A memory-based part of speech tagger generator. In E. Ejerhed and I. Dagan, editors, *Proceedings of the Fourth Workshop on Very Large Corpora*, pages 14–27.

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Eva Ejerhed and Daniel Ridings. 1997. PAROLE→SUC and SUC→PAROLE. http://spraakbanken.gu.se/parole.

Péter Halácsy, András Kornai, and Csaba Oravecz. 2007. Hunpos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume Companion Volume, Proceedings of the Demo and Poster Sessions, pages 209–212, Prague, Czech Republic. Association for Computational Linguistics.

Beata Megyesi. 2002. *Data-Driven Syntactic Analysis: Methods and Applications for Swedish*. Ph.D. thesis, KTH: Department of Speech, Music and Hearing.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-sppech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.