Stagger: A modern POS tagger for Swedish

Robert Östling
Department of Linguistics
Stockholm University
SE-106 91 Stockholm
robert@ling.su.se

Abstract
The field of Part of Speech (POS) tagging has made slow but steady progress during the last decade, though many of the new methods developed have not previously been applied to Swedish. I present a new system, based on the Averaged Perceptron algorithm and semi-supervised learning, that is more accurate than previous Swedish POS taggers. Furthermore, a new version of the Stockholm-Umeå Corpus is presented, whose more consistent annotation leads to significantly lower error rates for the POS tagger. Finally, a new, freely available annotated corpus of Swedish blog posts is presented and used to evaluate the tagger’s accuracy on this increasingly important genre. Details of the evaluation are presented throughout, to ensure easy comparison with future results.

1. Introduction
The task of syntactic disambiguation of natural language, frequently referred to as part of speech (POS) tagging, aims to annotate each word token in a text with its part of speech and (often) its morphological features.

I have implemented a new, freely available POS tagging system for Swedish, named Stagger,1 and used it to evaluate recently developed tagging algorithms on Swedish, as well as the effects of improved corpus annotation on POS tagging accuracy.

2. Data
Two corpora were used to evaluate the accuracy of the POS tagger: an updated version of the Stockholm-Umeå Corpus, and a new corpus of Swedish blog texts.

2.1 Stockholm-Umeå Corpus
The Stockholm-Umeå Corpus (SUC) is a balanced and POS-annotated corpus of about one million words of Swedish text, which was originally developed at the universities of Stockholm and Umeå during the 1990s. Its most recent release (Gustafson-Capková and Hartmann, 2008) has become a de-facto standard for Swedish POS tagging research.

Due to the size of the corpus, multiple annotators have been used, and annotation (in)consistency is an issue. Källgren (1996) explored tagging errors in an earlier version of the corpus, and found that 1.2% of the words sampled contained POS annotation errors. Forsbom and Wilhelmsson (2010) corrected over 1500 errors in SUC 2.0, mostly in common, polysemous grammatical words, and found that this results in a small but significant improvement in POS tagger accuracy.

We have included the changes of Forsbom and Wilhelmsson (2010), as well as over 2500 other changes to the annotation, into version 3.0 of SUC.2

2.2 Swedish blog texts
The language in so-called user-generated content, written by non-professionals in for instance blog posts or online forum posts, may differ considerably from traditional written language and poses a challenge to many Natural Language Processing applications, including POS tagging (Giesbrecht and Evert, 2009).

In order to evaluate the current POS tagger on user-generated content in Swedish, a small corpus (8 174 tokens) of blog texts was compiled and manually annotated with SUC-compatible POS tags and named entities. The corpus is freely available for download from the Stockholm University website.3

2.3 Unannotated data
For semi-supervised training, Collobert and Weston (2008) embeddings were induced from a corpus of about two billion tokens of Swedish blog texts.

2.4 Lexicon
In addition to the vocabulary in the training data, the SALDO lexicon of Swedish morphology (Borin and Forsberg, 2009) is used as a POS tag lexicon. For known words, only POS tags occurring with the word in either the training data or the SALDO lexicon are considered. For unknown words, all POS tags that occur with a token of the same type (e.g. number, emoticon or letter sequence) are considered.

3. Method
The tagger uses a feature-rich model, based on the averaged perceptron tagger of Collins (2002).

A basic feature set similar to Collins’ is used. Details are omitted due to space limitations, but are documented in the software package. In addition, 48-dimensional Collobert and Weston (2008) embeddings (C&W) were used as features in one tagger configuration. Each word can then be

1http://www.ling.su.se/stagger
2More information and instructions for obtaining the corpus can be found at: http://www.ling.su.se/suc
3http://www.ling.su.se/sic
Table 1: POS tagging accuracy in percent, with figures in bold significantly better than the others in the same column.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>SUC2</th>
<th>SUC3</th>
<th>Test3</th>
<th>Blogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>–</td>
<td>95.86</td>
<td>96.04</td>
<td>96.58</td>
<td>91.72</td>
</tr>
<tr>
<td>SALDO</td>
<td>96.32</td>
<td>96.52</td>
<td>96.94</td>
<td>92.45</td>
</tr>
<tr>
<td>SALDO+C&amp;W</td>
<td><strong>96.40</strong></td>
<td><strong>96.57</strong></td>
<td><strong>96.94</strong></td>
<td><strong>92.10</strong></td>
</tr>
</tbody>
</table>

The evaluation also demonstrated the importance of using a good lexicon, where the SALDO lexicon of Swedish morphology made a great contribution to tagging accuracy. Finally, Collobert & Weston embeddings were shown to improve tagging accuracy by a small but significant amount in the cross-evaluation, similar to what Turian et al. (2010) showed for other NLP tasks. Surprisingly, given the fact that the embeddings were computed from an unannotated blog corpus, the accuracy on the annotated blog corpus is instead significantly lower with C&W embeddings. However, since there are only three authors represented in the blog corpus, it would be risky to draw too general conclusions on the basis of this result.

5. Related work

Sjöbergh (2003) evaluated seven different POS tagging systems for Swedish through ten-fold cross-validation on SUC 2.0, where accuracies ranged between 93.8% and 96.0% for single systems (Carlberger and Kann, 1999, being the best), and a voting combination of all taggers reached 96.7%.

However, since the details of his evaluation were not published, and he used a larger training data set in each fold (95%) than the present study (90%), our respective accuracy figures are not directly comparable.

6. Acknowledgments

Thanks to those who found and corrected the thousands of annotation errors in SUC 2.0 that have been fixed in SUC 3.0: Britt Hartmann, Kenneth Wilhelmsson, Eva Forsbom and the Swedish Treebank project. Further thanks to the two anonymous reviewers, who provided useful comments.

7. References


4McNemar’s test with $p < 0.05$ is used throughout to test for statistical significance.