Revision of Part-of-Speech Tagging in Stockholm Umeå Corpus 2.0

Eva Forsbom*, Kenneth Wilhelmsson†

*Department of Linguistics and Philology
Uppsala University
evafo@stp.lingfil.uu.se

†Swedish School of Library and Information Science
University of Borås
kenneth.wilhelmsson@hb.se

Abstract
Many parsers use a part-of-speech tagger as a first step in parsing. The accuracy of the tagger naturally affects the performance of the parser. In this experiment, we revise 1500+ proposed errors in SUC 2.0 that were mainly found during work with schema parsing, and evaluate tagger instances trained on the revised corpus. The revisions turned out to be beneficial also for the taggers.

1. Introduction
Many parsers of today rely on a statistical part-of-speech tagger as a preprocessing step, in order to rank or limit the amount of possible analyses for each word. However, the tagger is only as good as the data it is trained on, and could potentially be a bottleneck for the correctness of parser systems. If the data contain errors and inconsistencies, the tag distribution for the affected words and n-grams would be skewed. Some of the errors are likely to harm both tagging and parsing (e.g. sentence initial errors), while others may only harm one of the two.

In this paper, we present an initial attempt to investigate if, and how much, tagging accuracy can be enhanced through revising a set of 1500+ potential errors mainly collected in the work concerning schema parsing (Wilhelmsson, 2010) with the Swedish Stockholm-Umeå corpus (Ejerhed et al., 2006). The corpus, henceforth SUC, has become the de facto standard for training and evaluating part-of-speech taggers, as its annotation has been manually revised, and also improved for version 2.0. It still includes errors and inconsistencies, however.

2. Set of changes
The proposed set of changes particularly includes types with severe consequences for parsing, such as tagging of/to verbs, and tagging of/to the markers of sub-clauses or relative clauses. In SUC 2.0, there are five such markers: subjunction, interrogative/relative pronoun, interrogative/relative adverb, interrogative/relative determiner and interrogative/relative possessive.

The following is a typical example of how som should be changed from conjunction (KN) to interrogative/relative pronoun (HP) to signal the start of a relative clause:

Vad är det som/KN har hänt (kk27-057)
What is it that has happened

In the graphical user interface of the schema parser, these types of errors yield analyses that are often visually recognizable directly. On the other hand, possible errors concerning more subtle aspects, e.g. gender agreement in NP-chunks, have not been detected to the same extent, as the parser is robust enough to ignore these.

The set of suggested changes affects 2% of the sentences in SUC. It is not claimed to reflect the proportions of all the actual errors in SUC 2.0. It is unknown how many these are, what their exact distribution is, and what would be the accuracy for a tagger trained on a corrected, or perfect, corpus.

As it seemed likely that some changes, although linguistically well-motivated, actually would decrease the accuracy, we divided the errors into nine groups (see Table 1). If any of the groups should decrease the accuracy, these groups could be skipped, or the sentences affected could be removed from the training data to increase overall accuracy.

The division was based on error type, with the extra constraints that the number of changes in each group should be large enough to be able to yield significant changes in accuracy score and that the groups should not overlap. Members that could belong to more than one group were therefore placed in the group with the lowest group number. Each group contains 4–15% of the suggested changes.

3. Evaluation
The error groups were evaluated using the statistical TnT tagger (Brants, 2000) and 10-fold cross validation on SUC for three tagsets, as the granularity of the tagset affects tagging performance.

The SUC tagset consists of 150+ tags, but a better tagging accuracy can be achieved with the Granska tagset (Carlberger and Kann, 1999), which is a variation of the SUC tagset, or “Granskaish”, which, in turn, is a subset of the Granska tagset that can be mapped back to the SUC tagset losslessly (Forsbom, 2008). The Granska tagset was altered to fit the needs of the Granska grammar checker, adding some features to the tags, and conflating tags with infrequent features. Granskaish only added features for copulas, auxiliaries, singulars (cardinal), and dates.

For each error group, we performed the changes, divided the corpus into 10 partitions, trained a tagging model
Table 1: Error groups with overall tagging accuracy and standard deviation.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>No. of changes</th>
<th>SUC</th>
<th>Granska</th>
<th>Granskaish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No changes</td>
<td>0</td>
<td>95.52±0.15</td>
<td>95.69±0.15</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td></td>
<td>All changes</td>
<td>1569</td>
<td>95.58±0.15</td>
<td>95.74±0.15</td>
<td>95.67±0.15</td>
</tr>
<tr>
<td>1</td>
<td>Sentence initial changes</td>
<td>192</td>
<td>95.53±0.15</td>
<td>95.70±0.15</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td>2</td>
<td>Changes to interrogative/rel. adverb</td>
<td>258</td>
<td>95.52±0.15</td>
<td>95.69±0.15</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td>3</td>
<td><em>Simp</em> to conjunction</td>
<td>92</td>
<td>95.53±0.15</td>
<td>95.70±0.15</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td>4</td>
<td><em>Som</em> to interrogative/rel. pronoun</td>
<td>111</td>
<td>95.53±0.15</td>
<td>95.69±0.14</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td>5</td>
<td>Changes to conjunction</td>
<td>71</td>
<td>95.52±0.15</td>
<td>95.69±0.15</td>
<td>95.62±0.14</td>
</tr>
<tr>
<td>6</td>
<td>Changes to subjunction</td>
<td>130</td>
<td>95.52±0.15</td>
<td>95.68±0.15</td>
<td>95.61±0.14</td>
</tr>
<tr>
<td>7</td>
<td>Changes to adverb</td>
<td>285</td>
<td>95.55±0.15</td>
<td>95.71±0.15</td>
<td>95.64±0.15</td>
</tr>
<tr>
<td>8</td>
<td>Changes to preposition</td>
<td>193</td>
<td>95.53±0.15</td>
<td>95.69±0.15</td>
<td>95.62±0.15</td>
</tr>
<tr>
<td>9</td>
<td>Other changes</td>
<td>237</td>
<td>95.53±0.15</td>
<td>95.70±0.15</td>
<td>95.63±0.14</td>
</tr>
</tbody>
</table>

for each partition and tagset, and ran the 10-fold cross-validation test (see Table 1).

Altogether, the changes improved tagging accuracy, albeit with a small increase in standard deviation for the Granskaish tagset. Group 7 improved the accuracy most, while group 2 and 5 had no effect at all. Group 6 actually decreased the accuracy, at least for the Granska-based tagsets. All other groups had a minor positive effect.

However small, the improvements in accuracy were all statistically significant (α = 0.001) using the McNemar test (McNemar, 1947).

4. Discussion

This initial experiment showed that part-of-speech errors that cause problems for a parser are troublesome also for statistical part-of-speech taggers. By revising such errors in the training data, it is possible to improve the accuracy of the tagger, and, most likely, consequently the accuracy of the parser.

Contrary to our initial hypothesis, no group of changes was obviously harmful for all tagsets, although some groups did not improve accuracy. It may still be the case, however, that individual errors in a group actually decrease accuracy.

A natural second step would be to study in more detail how the taggers tag the changed occurrence and its nearest context, and to try to find more errors in a systematic way, e.g., using the variation n-gram method proposed by Dickinson (2005).

5. References


Eva Ejerhed, Gunnel Källgren, and Benny Brodda. 2006. Stockholm-Umeå corpus version 2.0. Stockholm University, Department of Linguistics and Umeå University, Department of Linguistics.

