Tagging a Norwegian Speech Corpus

Anders Nøklestad
The Text Laboratory
University of Oslo
P.O. Box 1102 Blindern
0317 Oslo, Norway
anders.noklestad@iln.uio.no

Åshild Søfteland
Department of Linguistics and Scandinavian Studies
University of Oslo
P.O. Box 1102 Blindern
0317 Oslo, Norway
ashildso@hfstud.uio.no

Abstract

This paper describes work on the grammatical tagging of a newly created Norwegian speech corpus: the first corpus of modern Norwegian speech. We use an iterative procedure to perform computer-aided manual tagging of a part of the corpus. This material is then used to train the final taggers, which are applied to the rest of the corpus. We experiment with taggers that are based on three different data-driven methods: memory-based learning, decision trees, and hidden Markov models, and find that the decision tree tagger performs best. We also test the effects of removing pauses and/or hesitations from the material before training and applying the taggers. We conclude that these attempts at cleaning up hurt the performance of the taggers, indicating that such material, rather than functioning as noise, actually contributes important information about the grammatical function of the words in their nearest context.

1 Introduction

In this paper we describe a number of experiments on tagging a Norwegian speech corpus. The corpus, called NoTa (*Norsk talespråkskorpus* "Norwegian Speech Corpus"), contains 900,000 words of present-day Norwegian speech from informants located in the Oslo area. The corpus has a web-based search interface that enables queries to be restricted by a wide variety of informant properties, and the search results are linked to audio and video recordings of the informants.

The corpus is transcribed in standard orthography, and is tagged using a modified version of the tag set used by the Oslo-Bergen tagger (Hagen et al., 2000), a rule-based tagger for written Norwegian. In addition to part-of-speech, the tagset encodes detailed information about morphosyntactic features (e.g. gender, number, definiteness, tense) and certain lexical features (e.g. whether or not a certain pronoun is used to denote human beings). Because of this high level of detail, the tagset is relatively large, counting a total of 302 different tags. For more information about the NoTa corpus, see Johannessen and Hagen (to appear).

2 The taggers

In our experiments, we have used three different data-driven taggers. The first is the Memory-Based Tagger (MBT)[1] (Daelemans et al., 2003), which uses memory-based learning and is built on top of the Tilburg Memory-Based Learner (TiMBL) (Daelemans et al., 2004). The second one is the TreeTagger (Schmid, 1994)[2], which is based on decision tree technology. We used the TreeTagger in its default setting as a trigram tagger (running it as a bigram tagger yielded slightly inferior results). Finally, QTag (Tufis and Mason, 1998) is a trigram Hidden Markov Model (HMM) tagger. HMM tag-

[1] The Memory-Based Tagger can be downloaded from http://ilk.uvt.nl/mbt/.

[2] The TreeTagger is available from http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 245–248

gers are probably the most widespread type of part-of-speech tagger, and is the technology that was used, for instance, to tag the Swedish Gothenburg Spoken Language Corpus, as described by Nivre and Grönqvist (2001).

3 Creating the training corpus

All of the taggers we have tested make use of supervised learning, meaning that they need to be trained on a manually tagged corpus. We used the Memory-Based Tagger to create such a corpus using an iterative procedure which is commonly employed in cases where no pre-tagged material is available.

We started by tagging a small part of the corpus completely by hand, and trained a first version of the tagger on this material. We then ran the tagger on a different part of the corpus, manually corrected the tagger output, and added the corrected material to the training corpus. We were then in a position to re-train the tagger on this bigger training corpus, so that it could be applied to yet another part of the corpus, yielding slightly better results than it did the first time. By repeating this process, we simultaneously obtained an increasingly better tagger and an increasingly larger manually corrected corpus, which now contains approx. 190,000 tokens.

4 Cross-validation experiments

We have run 10-fold cross-validation experiments (Weiss and Kulikowski, 1991) with the different taggers on the manually tagged corpus of 190K tokens. In 10-fold cross-validation, 90% of the data is used for training and the remaining 10% for testing, and this procedure is repeated ten times, using a diffent 10% for testing each time. The data were shuffled before the distribution into training and test data began.

The MBT and QTag require a training corpus where each token is accompanied by its manually disambiguated tag, and a test corpus consisting only of tokens (however, if disambiguated tags are given in the test corpus as well, the MBT will report on its accuracy rate at the end of testing).

The TreeTagger accepts the same kinds of files, but also requires a lexicon that lists the set of possible tags for each known word, as well as a set of open class tags that can be used for unknown words. Assuming that the lexicon is only supposed to contain known words, i.e., words that occur in the training data, we create a different lexicon for each fold by extracting all tags that occur for each word in the training corpus used in the fold. The set of open class tags includes all noun, verb, and adjective tags[3] that occur in the manually tagged corpus, reaching a total of 112. This presents the tagger with a fairly high number of choices for unknown words—compare this number, for instance, to the mere 17 tags that are suggested for the Penn Treebank tagset by the documentation for the TreeTagger.

Table 1 shows the averages and standard deviations for the 10-fold cross-validation experiments with the Memory-Based Tagger and the TreeTagger. The TreeTagger shows the best performance, while the performance of the HMM-based QTag lags far behind the other two taggers. Using McNemar's test, we have found all differences between the taggers to be statistically significant at the 0.01 level.

TreeTagger's superior performance over the HMM tagger agrees with Schmid's (1994) findings for written English, where TreeTagger outperformed the HMM tagger presented by Kempe (1993). In Megyesi's (2002) experiments on written Swedish, on the other hand, an HMM tagger (Brants, 2000) outperformed all other taggers, including one that used memory-based learning, thus showing results that differ considerably from those obtained here (she did not test any decision tree tagger). Our TreeTagger results are better than the best results obtained in either of these studies (which were 96.36% and 93.55%, respectively), and better than the 95.29% accuracy obtained by the HMM tagger in Nivre and Grönqvist's (2001) experiments on spoken Swedish with their largest tagset of 23 tags.

5 Removing pauses and hesitations

One of the properties that characterize spoken as opposed to written language is the presence of pauses and hesitations, as illustrated in (1), where *e* represents a hesitation sound and # represents a pause:

(1) men i hvert fall # det er *e*
"but anyway # it is *e*"

[3] In NoTa, all traditional adverbs that may be inflected are treated as adjectives (in accordance with Faarlund et al. 1995); hence, adverb is not counted among the open classes.

	Avg. accuracy	Standard deviation
TreeTagger	96.89	0.56
MBT	95.19	0.15
QTag	89.96	0.30

Table 1: Average accuracy and standard deviation for the 10-fold cross-validation experiments using the TreeTagger, the Memory-Based Tagger, and QTag.

We wanted to examine the effect of such phenomena on the performance of a statistical tagger for Norwegian spoken language. If pauses and hesitations tend to occur more or less randomly throughout an utterance, we would expect them to have a negative impact on the tagger. This is so because a pause or a hesitation occurring between two words will reduce the confidence of the tagger with respect to the propensity of these words to occur together. If, on the other hand, they tend to occur at certain structural positions in the sentence, they may actually contribute important information about the grammatical properties of the surrounding words.

In order to investigate this question, we have created versions of the cross-validation data in which we remove either pauses or hesitations or both, and we have re-run the 10-fold cross-validation experiments on these data. The results are shown in Table 2. For each tagger, the first row repeats the performance given in Table 1 on the original corpus; the second row shows the performance when hesitations are filtered out; the third row lists the performance with pauses removed, and the fourth row shows the results when both hesitations and pauses are filtered out.

Interestingly, removal of hesitations and pauses deteriorates the performance of the MBT and the TreeTagger, indicating that this material does not in fact function as noise for these taggers, but rather provides useful information about the grammatical status of surrounding words. This is particularly true for pauses, where removal leads to the largest drop in performance. Removing pauses is detrimental for QTag as well, but deleting hesitations actually improves the performance of this tagger. The indication that pauses in particular may provide important grammatical information supports the findings by Strangert (1993) that pauses tend to occur at positions that are relevant to the underlying message, including syntactic boundaries.

	Accuracy (std.dev.)
TreeTagger	96.89 (± 0.56)
TreeTagger **-hesitations**	96.86 (± 0.58)
TreeTagger **-pauses**	96.67 (± 0.60)
TreeTagger **-both**	96.61 (± 0.61)
MBT	95.19 (± 0.15)
MBT **-hesitations**	95.10 (± 0.18)
MBT **-pauses**	94.78 (± 0.19)
MBT **-both**	94.68 (± 0.17)
QTag	89.96 (± 0.30)
QTag **-hesitations**	90.11 (± 0.28)
QTag **-pauses**	89.10 (± 0.36)
QTag **-both**	89.19 (± 0.31)

Table 2: Average accuracy and standard deviation for the experiments involving removal of hesitations and pauses. See the text for explanation.

6 Problematic words and tags

Table 3 lists the words that are most often mistagged by the TreeTagger, along with the proportion of the total number of errors that these errors constitute. The most problematic word is *så* (eng.: "so"), which may be either verb, conjunction, subjunction, or adverb, and which often occurs at the end of an utterance, making it hard to determine its correct category, as can be seen in (2):

(2) nå er det jo in å bo på østkanten da # altså # *så* ...
"now it is in to live on the east side, you know # then # so ...

Interestingly, the words in Table 3 are also among the words that are most difficult for the human annotators to disambiguate.

Table 4 lists the most common tag confusions made by the TreeTagger. The table shows the erroneous tag produced by the tagger along with the correct tag. The clearest tendency to be extracted from this table seems to be that adverbs, prepositions, and subjunctions are easily confused by the tagger. Also,

word	error (%)	word	error (%)
så	13.1	som	3.3
det	9.4	de	3.0
noe	5.8	da	2.7
den	3.6	noen	2.0
for	3.4	jo	1.9

Table 3: The ten words that are most commonly mistagged by the TreeTagger in the cross-validation experiments.

it is worth noting that all of the words in Table 3 except *jo* exhibit one or more of the ambiguities listed in Table 4.

Output tag	Correct tag
pron_nøyt_ent_pers_3	pron/det
adv	konj/sbu/adv
adj_nøyt_ub_ent_pos	adj_ub_m/f_ent_pos
adv	konj
sbu	prep
konj	adv
adj_ub_m/f_ent_pos	adj_nøyt_ub_ent_pos
prep	sbu
adv	sbu
pron/det	pron_nøyt_ent_pers_3

Table 4: The most common tag confusions made by the TreeTagger.

7 Conclusions and future work

We have described experiments on Norwegian speech data with three data-driven taggers and found that the best performing one is the decision tree-based TreeTagger. This tagger is now being used to tag the rest of the 900,000 word corpus. We have also found that hesitations, and in particular pauses, seem to provide useful information for the taggers. In the future, we would like to modify the Oslo-Bergen rule-based written language tagger (Hagen et al., 2000) to become better suited for spoken language and compare its performance to that of the data-driven taggers.

References

T. Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP-00)*, Seattle, Washington, USA.

W. Daelemans, J. Zavrel, A. Van den Bosch, and K. Van der Sloot. 2003. *MBT: Memory-Based Tagger, version 2.0, Reference Guide*. ILK Technical Report Series 03-13.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2004. *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. ILK Technical Report Series 04-02.

J.T. Faarlund, S. Lie, and K.I. Vannebo. 1995. *Norsk referansegrammatikk*. Universitetsforlaget, Oslo.

K. Hagen, J.B. Johannessen, and A. Nøklestad. 2000. A Constraint-Based Tagger for Norwegian. *17th Scandinavian Conference of Linguistics*, Volume I, no. 19. Odense Working Papers in Language and Communication.

J.B. Johannessen and K. Hagen. To appear. *Språk i Oslo*. Novus forlag, Oslo.

A. Kempe. 1993. *A probabilistic tagger and an analysis of tagging errors*. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

B. Megyesi. 2002. *Data-Driven Syntactic Analysis. Methods and Applications for Swedish*. Doctoral dissertation, Departement of Speech, Music and Hearing, KTH, Stockholm.

J. Nivre and L. Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics* 6(1), 47-78.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*.

E. Strangert. 1993. Clause Structure and Prosodic Segmentation. *FONETIK -93 Papers from the 7th Swedish Phonetics Conference*, Uppsala May 12-14 1993.

D. Tufis and O. Mason 1998. Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger. *Proceedings of the First International Conference on Language Resources & Evaluation (LREC)*, Granada (Spain), 28-30 May 1998, p.589-596.

S. Weiss and C. Kulikowski 1991. *Computer systems that learn*. Morgan Kaufmann, San Mateo, CA.