Sigrún Helgadóttir and Eiríkur Rögnvaldsson
Institute of Lexicography, University of Iceland

Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic

Abstract

- We present the main results of an experiment concerned with training three different taggers on Icelandic texts
- The taggers fnTBL, TnT and MXPOST were trained on a corpus that contains over 500,000 running words
- The corpus had been morphologically tagged using a tagset containing over 600 tags
- The TnT tagger obtained best results for tagging or 90.36% accuracy
- Different methods for tagger combination (voting and applying linguistic rules) were also tested
- By applying different strategies for tagger combination a tagging accuracy of 93.65% was obtained

Introduction

We describe a project whose aim was to develop a tagger that could tag Icelandic text with at least 92% accuracy using a large tagset.

We decided to test four different data-driven methods rather than develop a new tagger.

Therefore, five different off-the-shelf POS taggers were tested to find out which approach would be most suitable for Icelandic.

The Corpus and the Tagset

The corpus used in the experiments is a carefully balanced manually tagged corpus consisting of just over half a million running words. This is the corpus of the Icelandic Frequency Dictionary (IFD) published in 1991.

The tagset used in the printed IFD contains more than 600 tags. It is based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized.

Methods and Taggers Used

We tested two statistical methods, hidden Markov models and maximum entropy learning.

The TnT tagger (Brants 2000) was chosen to represent the hidden Markov models and MXPOST (Ratnaparkhi 1996) for maximum entropy learning.

To test memory-based learning the MBT software (Daelemans *et al.* 2002) was used.

Finally, two implementations of the transformation-based learning algorithm were tested. These were fnTBL (Florian and Ngai 2002) and µ-TBL (Lager 1999).

First Results

In preliminary testings, the µ-TBL tagger had shown promising results but did not seem to be able to cope with the whole corpus. The MBT tagger gave for some reason disappointing results and was left out of further experiments. A ten-fold cross-validation test was performed for the three remaining taggers.

Mean tagging accuracy for all words, known words and unknown words for three taggers

Accuracy %	MXPOST	fnTBL	TnT
All words	89.08	88.80	90.36
Known words	91.04	91.36	91.74
Unknown words	62.50	54.03	71.60

Mean percentage of unknown words in the ten test sets was 6.84. TnT shows overall best performance in tagging both known and unknown words.

MXPOST seems to do better than fnTBL at tagging unknown words but does worse on known words than fnTBL.

Improving the Results

Voting

After some experimentation it was decided to use a voting strategy where each tagger is weighted by its overall precision. By voting between the taggers in this way a precision of 91.54% was obtained for all words.

Simplification of Tags

Tagging accuracy was computed when some of the tags had been simplified. The simplification included ignoring subclasses of pronouns and adverbs. After this simplification, tagging accuracy for TnT reached 91.83% for all words.

Using a Backup Lexicon

To test the effect of using a backup lexicon with the taggers a lexicon was made containing about half of unknown words in each test set with respect to the appropriate training set. TnT obtained 91.54% accuracy for all words when utilizing the backup lexicon.

Applying Linguistic Rules

It was observed that the MXPOST tagger seemed to do better at distinguishing between identical word forms that should have different tags than the other two taggers. It was possible to formulate linguistic rules to choose the outcome of MXPOST rather than the outcome of voting if certain conditions were fulfilled.

Final Results

The final step was to apply all procedures for improving tagging results obtained by individual taggers. First the individual taggers were applied by utilizing a lexicon for TnT and fnTBL. The tags were then simplified as described above and a majority voting was performed on the simplified tags. Finally the linguistic rules were applied increasing the accuracy to **93.65%**.

Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. Version 2.2. http://www.coli.uni-sb.de/~thorsten/tnt/

Daelemans, Walter, Jakub Zavrel, Peter Berck, and Steven Gillis. 2002. TiMBL Tilburg Memory-Based Learner, version 4.2. Reference Guide. ILK Technical Report – 02-01.

Florian, Radu, and Grace Ngai. 2002. Fast Transformation-Based Learning Toolkit. http://nlp.cs.jhu.edu/~rflorian/fntbl/tbl-toolkit/tbl-toolkit.html

Lager, Torbjörn. 1999. The µ-TBL System. User's manual. Version 0.9. http://wwww.ling.gu.se/~lager/mutbl.html

Ratnaparkhi, Adwait. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Mehods in Natural Lanugage Processing* (EMNLP-96), pp. 133-142. Philadelphia.