Testing Data-Driven Learning Algorithms for PoS Tagging of Icelandic

by

Sigrún Helgadóttir

Abstract

This paper gives the results of an experiment concerned with training three different taggers on tagged Icelandic text. The taggers fnTBL, TnT and MXPOST were trained on the corpus of the Icelandic Frequency Dictionary that contains over 500 thousand running words that have been tagged with morphological tags. The tagset contains over 600 tags. Different methods for tagger combination were also tested. The TnT tagger obtained best results for tagging or 90.36% accuracy. By applying different strategies for tagger combination a tagging accuracy of 93.65% was obtained.

1. Introduction

The Icelandic Language Technology project was launched in the year 2000 in the wake of a survey, which showed that language technology in Iceland is still in its infancy. The project was managed by a ministry-appointed steering committee and financed in part from the national budget. The project is designed to develop the range of tools necessary to communicate with the computer-based technology already in everyday use, so paving the way for the export of products and knowledge arising from it.

In April 2002 the first grants were announced. The Icelandic NLP group, consisting of persons that had been taking Language Technology courses in the Swedish GSLT, together with the Institute of Lexicography in Reykjavík, applied for and was given a grant to develop a tagger that would tag Icelandic text. The present paper gives an overview of the project whose aim was to develop as quickly as possible a tagger that could tag Icelandic text with at least 92% accuracy.

The project started in October 2002. It was decided to test four different data-driven methods for the tagging of Icelandic 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 used in the experiments is the corpus of the Icelandic Frequency Dictionary published in 1991 (Pind *et al*. 1991).

2. The corpus

The corpus used in the experiments was created in the making of the Icelandic Frequency Dictionary (IFD, *Íslensk orðtíðnibók*), published by the Institute of Lexicography in Reykjavík in 1991. The IFD corpus is considered to be a carefully balanced corpus consisting of just over half a million running words. The corpus contains 100 fragments of texts, approximately 5,000 running words each. All texts were published for the first time in 1980–1989. Five categories of texts were considered, i.e. Icelandic fiction, translated fiction, biographies and memoirs, non-fiction (evenly divided between science and humanities) and books for children and youngsters (original Icelandic and translations). No two texts could be attributed to the same person, *i.e.*, as author or translator and all texts start and finish with a complete sentence.

The tagset used in the printed IFD is more or less based on the traditional Icelandic analysis of word classes and grammatical categories, with some exceptions where that classification has been rationalized. The text file used for the present study contained 590,297 running words with 59,358 word forms and 639 different tags. These figures include punctuation.

Of the word forms in the IFD 15.9% are ambiguous as to the tagset within the IFD. This figure is quite high, at least compared to English, which shows that the inflectional morphology of Icelandic is considerably more complex than English. An Icelandic noun can have up to 16 grammatical forms or tags, an adjective up to 120 tags, and a verb over a hundred tags. Some of the ambiguity is due to the fact that inflectional endings in Icelandic have many roles, the same ending often appearing in many places (e.g. *-a* in *penn**a*** for all oblique cases in the singular (acc., dat., gen.), and accusative and genitive in the plural of the masculine noun *penni* 'pen', producing 5 different tags for one form of the same word). The most ambiguous of word forms in the IFD, *minni*, has 24 tags in the corpus, and has not exhausted its possibilities (Kristín Bjarnadóttir 2002).

3. Methods and taggers used in the Icelandic experiment

Four different data-driven methods for POS tagging were tested and five different POS taggers. It was decided to test two statistical methods, hidden Markov models and maximum entropy learning. The TnT tagger (TRIGRAMS'N'TAGS) of Thorsten Brants (Brants 2000) was chosen to represent the hidden Markov models and MXPOST, an implementation of the maximum entropy framework developed for POS tagging by Ratnaparkhi (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, described by Brill in (Brill 1994) and (Brill 1995) were tested. These were fnTBL by Radu Florian and Grace Ngai (Florian and Ngai 2002) and Torbjörn Lager's µTBL (Lager 1999).

3.1 The experiments

The computer files for the IFD corpus each contain one text excerpt. Each file was divided into ten approximately equal parts. From these ten different disjoint pairs of files were created. In each pair there is a training set containing about 90% of running words from the corpus and a test set containing about 10% of running words from the corpus. Each set should therefore contain a representative sample from all genres in the corpus. The test sets are independent of each other whereas the training sets overlap and share about 80% of the examples. All words in the texts except proper nouns start with a lower case letter.

The µTBL tagger had previously been tested on a small portion of the IFD corpus and gave promising results (Sigrún Helgadóttir 2002). It did, however, 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.

3.2 Results

Results for ten-fold cross-validation testing for the three taggers are shown in table 1. As can be seen from the table the TnT tagger gave best results, the MXPOST tagger came second and fnTBL gave the worst results of the three taggers.

It is worth noticing that these results show lower performance rates when the taggers are applied to the Icelandic corpus than is achieved for example for Swedish as reported in (Megyesi 2002:56). In that study the taggers were applied to and tested on the SUC corpus with 139 tags compared to the Icelandic tagset of over 600 tags. Performance rates are also considerably lower than have been reported for the systems trained on the English Penn treebank.

Table 1. *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

Table 1 shows results for known words, unknown words and all words. 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. This is similar to what was seen in the experiment on Swedish text and indicates that the major difficulty in annotating Icelandic words stems from the difficulty in finding the correct tag for unknown words. Words belonging to the open word classes (nouns, adjectives and verbs) account for about 96% of unknown words in the test sets whereas words in these word classes account for just over 51% of all words in the test sets. The three taggers have different procedures for annotating unknown words and this is reflected in the difference in performance.

Extensive analysis was performed of the errors made by the three different taggers. The analysis showed that the taggers make to a certain degree different types of errors. This can be used to combine the results of tagging with different taggers to improve tagging accuracy. In (Halteren *et al.* 2001) there is a comprehensive overview of methods for combining POS taggers. After some experimentation it was decided to use a voting strategy where each tagger is weighted by its overall precision. Since there are three taggers in the experiment this is equivalent to choosing the tag that two or three taggers agree upon. If all three disagree on a tag the tag chosen is the tag that is assigned by the tagger with the highest overall precision, TnT in our case. By voting between the taggers in this way a precision of 91.54% was obtained for all words.

The tagset of the IFD is rather large compared to tagsets for some other languages such as English and the Scandinavian languages. It is possible that for some applications it is not necessary to use such a fine-grained tagset. Some applications may need more accuracy in tagging and a less fine-grained tagset. Tagging accuracy was computed when some of the tags had been simplified. The simplification included ignoring further classification of adverbs and conjunctions and ignoring subcategories of pronouns. Table 2 shows mean tagging accuracy for known words, unknown words and all words for the three taggers after the tags have been simplified. Tagging accuracy for TnT is 91.83% for all words after simplification of tags.

Table 2. *Mean tagging accuracy for all words, known words and unknown words for three taggers after tags were simplified*

Accuracy %	MXPOST	fnTBL	TnT
All words	90.46	89.98	91.83
Known words	92.52	92.62	93.32
Unknown words	62.52	54.05	71.61

To increase tagging accuracy it seems important to improve tagging of unknown words. This can be done in two ways, either by improving the methods that the taggers use for tagging unknown words or increasing the size of the lexicon used by the taggers. Two of the taggers tested allow the use of a backup lexicon that is used before the taggers apply the unknown word handler. 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. Table 3 shows mean tagging accuracy for all words, known words and unknown words for fnTBL and TnT when a backup lexicon is used. The table shows that TnT obtains 91.54% accuracy when utilizing the backup lexicon.

Table 3. *Mean tagging accuracy for all words, known words and unknown words for TnT and fnTBL when using a backup lexicon*

Accuracy %	fnTBL	TnT
All words	90.06	91.54
Known words	91.50	91.93
Unknown words	70.44	86.28

Lars Borin (Borin 2000) advocates the use of a “knowledge-rich method to the problem of combining POS taggers, by formulating linguistically motivated rules for how tagger differences should be utilized in the combination of taggers". While analyzing the outcome of the three different taggers in the Icelandic experiment 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. These forms are e.g. the accusative and dative forms of feminine nouns in the singular and nominative and accusative forms of neuter nouns in the singular. By voting between the taggers as has already been described the advantages of TnT over the other two taggers have already been exploited. It was, however, possible to formulate linguistic rules to choose the outcome of MXPOST rather than the outcome of voting if certain conditions were fulfilled.

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. This gave tagging accuracy of 93.53%. Finally the linguistic rules were applied increasing the accuracy to 93.65%.

In many applications it seems to be sufficient to know the word class of a word. Table 4 shows tagging accuracy obtained by the three taggers when only considering the word class. There are 11 word classes in the material of the IFD. Classification of running words by word class with 98.14% accuracy can be very useful in many applications.

Table 4. *Mean tagging accuracy for all words, known words and unknown words for three taggers when only considering the word class*

Accuracy %	MXPOST	fnTBL	TnT
All words	97.29	97.25	98.14
Known words	97.96	98.35	98.55
Unknown words	88.19	82.15	92.47

4. Further experiments

To investigate whether it is possible to use the procedure described above for texts that are different from the texts of the IFD some further experiments were performed. The procedures were applied to four different text segments. One consisted of just over 6,000 running words of text from 13 literary works from the last part of the 19th century and first part of the 20th century. The second segment consisted of about 3,600 running words of text from 9 literary works from the period after 1980. The third segment consisted of about 2,900 running words of text about computers and information technology taken from the newspaper *Morgunblaðið*, a newsletter from the University of Iceland Computing Services and web sites of several information technology companies. The fourth segment consisted of about 2,700 running words of text about law and business taken from various sources. All the texts were tagged by the three taggers using the results of training the taggers on the whole corpus of the IFD. Tagging was then corrected so that tagging accuracy could be found.

The TnT tagger obtained the best results for tagging all the texts. The results obtained by the other taggers were not sufficient to improve the results of TnT in the combination procedures described previously. TnT obtained 93.93% accuracy when tagging the older literary texts, 98.01% when only considering the word class. For the more recent literary texts TnT obtained 92.91% accuracy, 98.38% when only considering the word classes. For the computer text TnT obtained 86.77% tagging accuracy, 97.33% when only considering the word class. For the legal and business text TnT obtained 88.54% accuracy, 97.37% when only considering the word class.

This experiment confirms that the texts of the IFD have a strong literary bias. To be able to tag other types of texts it is necessary to train the taggers on different types of texts and to be able to have a more extensive lexicon available.

5. Future developments

One of the projects funded by the Icelandic Language Technology Project was the development of a morphological description of modern Icelandic. On the basis of this description a lexicon will be developed and used as a backup lexicon for tagging. The Icelandic Language Technology Project will also fund the establishment of a tagged Icelandic corpus with 25 million words. The project was started during the summer of 2004 and will be concluded in June 2007. The results of the tagging experiment and the lexicon developed on the basis of the morphological description will be used when tagging the corpus. It is also planned to obtain various lists of proper names, such as names of persons, places, companies and institutions to be able to improve the tagging of proper names and develop a tool for Named Entity Recognition.

Acknowledgements

This project received a grant from the Language Technology steering committee of the Icelandic Ministry of Education, Science and Culture. Participants in the project were the Icelandic NLP group together with the Institute of Lexicography in Reykjavík. The Institute of Lexicography provided office facilities and made the corpus of the IFD available for the project. Many thanks go also to the authors of the taggers already tested for making their systems available.

References

Borin, Lars. 2000. Something borrowed, something blue: Rule-based combination of POS taggers. In *Second International Conference on Language Resources and Evaluation. Proceedings*, pp. 21-26. Athens 31 May - 2 June, 2000.

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

Brill, Eric. 1994. Some Advances in Rule-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, pp. 722-727. Seattle.

Brill, Eric. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. *Computational Linguistics* 21:543-565.

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

van Halteren, Hans, Jakub Zavrel, and Walter Daelemans. 2001. Improving Accuracy in Wordclass Tagging through Combination of Machine Learning Systems. *Computational Linguistics* 27:199–230.

Jörgen Pind (ed.), Friðrik Magnússon and Stefán Briem. 1991. *Íslensk orðtíðnibók*. Orðabók Háskólans, Reykjavík. [Referred to as the Icelandic Frequency Dictionary, IFD.]

Kristín Bjarnadóttir. 2002. The Icelandic µ-TBL Experiment: Preparing the Corpus. Term paper in NLP 1, GSLT.

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

Megyesi, Beáta. 2002. Data-Driven Syntactic Analysis – Methods and Applications for Swedish. Ph.D. Thesis, Department of Speech, Music and Hearing, KTH, Stockholm.

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.

Sigrún Helgadóttir. 2002. The Icelandic μTBL Experiment: Learning rules from four different training corpora by using the μ-TBL System – Further developments. Term paper in NLP 1, GSLT.