Bootstrapping a Free Part-of-Speech Lexicon Using a Proprietary Corpus

Jonas Sjöbergh
KTH Nada
SE-100 44 Stockholm, Sweden
Email: *jsh@nada.kth.se*

Abstract

In this paper a method for automatically creating a free lexicon for a part-of-speech tagger from a proprietary lexicon is described. The method uses an ensemble of taggers to tag a new free corpus and retrain the tagger on the new data. The method requires almost no manual labour. The new lexicon gives slightly worse tagging accuracy than the original proprietary lexicon, 95.6% accuracy compared to 96.0%. Different variations on how to create the training data, such as using only those sentences where all taggers agree or adding ungrammatical language to the training data, are evaluated.

1 Introduction

This paper is motivated by having the implementation of a part-of-speech tagger [Carlberger and Kann, 1999], developed for Swedish but easily retrainable for reasonably similar languages, such as English. The intention is now to make this tagger freely available. The tagger is of little use without a lexicon, created automatically from annotated training data. So unless one already has large quantities of annotated text, one would also want to get a freely available lexicon together with the tagger. Currently the tagger uses a lexicon created from the Stockholm-Umeå Corpus (SUC), [Ejerhed *et al.*, 1992], which can be used if you have a license for the SUC. The SUC is freely available for academic use.

This paper describes a method used to create a new totally free lexicon that can be distributed along with the tagger. The method uses an ensemble of taggers, all freely available and automatically trained on the SUC, to tag new texts to be used when creating the new lexicon. The texts were taken mainly from the KTH News Corpus [Hassel, 2001], a collection of news articles from the web editions of several Swedish newspapers. These texts were automatically collected from the web sites of the newspapers. The texts thus annotated were then used to create a new lexicon for the tagger.

How the size of the training data and other factors influence the quality of the lexicon was then evaluated. As expected, the new lexicon gives worse performance of the tagger than the one created from the manually corrected SUC, but still good enough to be useful.

2 Creating a New Lexicon

2.1 How the Tagger Creates a Lexicon

When the Granska tagger [Carlberger and Kann, 1999], which the new lexicon is to be used by, creates a lexicon it needs an annotated corpus. Preferably, each word should be annotated with its part-of-speech tag and its lemma, and sentence boundaries should be marked. The tagger has a lemma guesser and a sentence boundary detection mechanism, so only part-of-speech information is strictly necessary. Also, the lemma information is not used much in actual tagging (but for some unknown words it can be used), but mostly for guessing lemmas of new text (which in turn is used by the grammar checking framework the tagger is normally used in).

From the annotated training data the following information is extracted:

- Word frequencies. Word-tag pair frequencies and word-tag-lemma triple frequencies.
- Tag n-grams. Unigram, bigram and trigram frequencies.
- Suffix information. All suffixes of lengths between two and five letters, from words in any open word class are collected, as suffix-tag pair frequencies.

This data is what is referred to as a “lexicon” in this paper. The tag n-grams and the word-tag frequencies are used when tagging known words, while the tag n-grams and the suffix information is used when tagging unknown words. For unknown words, other factors like capitalization is also used.

2.2 Creating the Training Data

First, several freely available taggers were collected. The taggers used were:

- fnTBL [Ngai and Florian, 2001], a transformation based tagger.
- Granska [Carlberger and Kann, 1999], a trigram HMM-tagger.
- Mxpost [Ratnaparkhi, 1996], a maximum entropy tagger.
- Stomp [Sjöbergh, 2003b], a tagger that matches word sequences between training and test data.
- TnT [Brants, 2000], a trigram HMM-tagger.
- TreeTagger [Schmid, 1994], a tagger using decision trees.

These were then trained (automatically) on the SUC, using default options as far as possible. When trained, the taggers all tagged the new texts from the KTH News Corpus. The results were combined by simple voting, giving each tagger one vote and picking the tag which received the most votes. Ties were broken by a predetermined tagger.

Since all taggers were trained on the same corpus, they all tag new text using the same tagset (the one used in the training data). This makes combining their results by voting easy. If instead several taggers that were already trained would be used, they might be trained to use different tagsets. If there is a one to one mapping between the tagsets voting is still no problem, but if the tagsets are in some way different voting is hard. A simple way around this problem would be to let each of the taggers tag an annotated corpus, in this case the SUC, and then train a new classifier on the output of all the taggers and the annotation in the SUC as the correct annotation. This would be a simple way to combine the results and it normally results in slightly higher accuracy than simple voting [Sjöbergh, 2003a], especially since it is easy to add more information to the new classifier, such as the tags of words in the context.

After annotating the new data with part-of-speech tags, the Granska tagger was run on the tagged text, in lemma guessing mode, to add lemma information (which is not really used by the tagger for tagging, but is expected to be found in the training data, see Section 2.1) and sentence boundaries.

The Granska tagger, which the new lexicon is to be used by, was then trained on the new data. The quality of the lexicon was evaluated on a test set from SUC, consisting of 60 000 words and disjoint from the part of SUC used when training the taggers. The test set was tagged with the Granska tagger using different lexicons and the quality of the lexicons compared by comparing the accuracies on the test set.

This method of creating a new lexicon requires very little manual labour, only acquiring the taggers and the texts, and then starting the training and tagging (which is then automatic).

3 Evaluation of Lexicons

A few variations where tried when creating the new training data, to see what effect different methods have on the resulting lexicon. The following expected behaviors were evaluated: the more training data used, the better the lexicon; using several genres is better than using only one (i.e. newspaper texts); using well written texts is better than using texts with many errors; using many taggers to annotate the new texts is better than using only one.

This was evaluated by using the following strategies for annotating the training data:

- Using only the tagger itself when tagging the training data.
- Using the voting ensemble when tagging the training data.
- Using only sentences where all taggers agree on the tagging as training data. This gives less training data (about 90% of the data is discarded) but the tagging accuracy on the remaining data is very high.

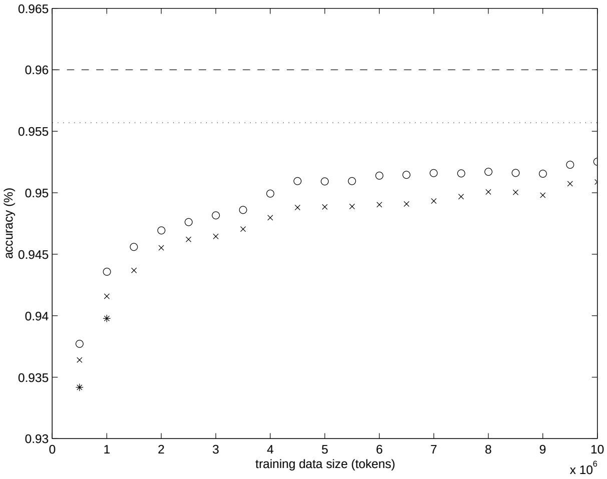

Figure 1: Tagging accuracy when using different lexicons, as a function of the size of the training data used when creating the lexicon. The dashed line is the accuracy when using the manually corrected SUC as training data and the dotted line is when using all available new data.
x using only the tagger itself to tag the training data
o using an ensemble of taggers
***** using only the sentences where all taggers agree on the tagging.

To test some of the assumptions above, other texts were added to the KTH News Corpus texts (but some of these can likely not be included if the lexicon is to be freely available). The following are examples of variations of the training data:

- Adding other genres, in this case text from an encyclopedia.
- Adding malformed language, in the form of essays written by second language learners, to the training data (to see if the quality of the texts is important).
- Adding text of high quality but which is not modern Swedish, in the form of a novel from 1879.

The performance of the tagger using lexicons constructed by these methods can be found in Figures 1 and 2. In Figure 3 a comparison of how much manually corrected data is needed to achieve a certain accuracy is shown, by using different amounts of data from the SUC.

As expected the new lexicon gives worse performance than the one created from the manually corrected SUC. This is because the taggers strengthen their misconceptions from the original training data on the new texts. The best accuracy achieved (when using all available data, 10 million tokens of newspaper texts, 3 million tokens of encyclopedia text, 100 000 tokens of student essays and 100 000 tokens from a novel) was 95.6% and the best accuracy when using only the KTH News Corpus texts was 95.3%. The accuracy when using the manually corrected SUC was 96.0%. To achieve the same accuracy as the best result of the automatic method using manually corrected annotation would require roughly

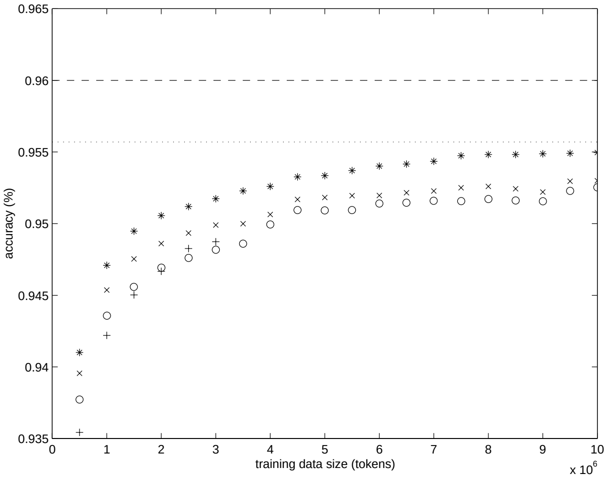

Figure 2: Tagging accuracy when using different lexicons, as a function of the size of the training data used when creating the lexicon. The dashed and dotted lines are the same as in Figure 1.
o using only newspaper texts
+ using only encyclopedia texts
x using newspaper texts and some malformed text (second language learner student essays)
***** using both encyclopedia and newspaper texts
In all cases the ensemble of taggers tagged the text.

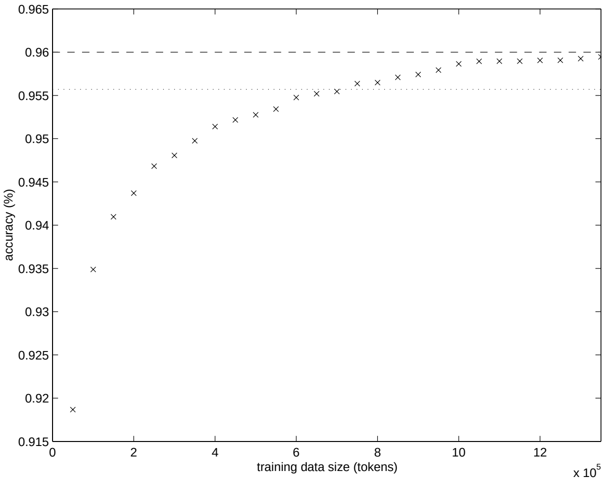

Figure 3: Tagging accuracy when using different lexicons, as a function of the size of the training data used when creating the lexicon. The training data is the SUC, a manually corrected, balanced corpus of Swedish text. The dashed and dotted lines are the same as in Figure 1.

700 000 tokens of annotated text, if the new texts are from the same (balanced) domain as the test data. This can be seen in Figure 3, where different amounts of training data from the SUC were evaluated.

The more data used for training, the less the degradation in tagging accuracy with the new lexicon. While adding more training data improves the lexicon, the rate of improvement decreases when the training set becomes large.

3.1 Annotation Methods

As can be seen in Figure 1, using an ensemble of taggers when annotating the new training data gives better performance than using the tagger alone. This is because the taggers correct each other, so the annotation of the training data contains fewer errors than when using only one tagger. The best result using the ensemble on the newspaper texts was 95.3% and the best result on the same texts using only the best single tagger alone was 95.1%.

For words where all the taggers in the ensemble agree on the tagging, the tagging accuracy is very high, above 99%. To see if using only such data is a good idea a training set was created by using only those sentences where all taggers agreed on the tagging of all words. This results in much less training data, about 90% of the data is discarded, but with less tagging errors in the remaining data. This did not improve the lexicon, even when compared to a lexicon created from an equal amount of training data tagged by voting. In fact, it is even worse than using only one tagger. The highest accuracy when using only data where all taggers agree was 94.0%, and the accuracy when using the same amount of data (though of course not the same sentences) was 94.2% when tagged by one tagger and 94.4% when using the ensemble.

One reason for the perhaps surprisingly bad performance when using only the highly accurately annotated sentences, is that the data consist mostly of simple constructions. The difficult constructions are not represented at all in the training data, since some tagger usually makes a mistake on the difficult constructions. This means that the tagger only learns how to tag simple language constructions, while the test data (and most real texts) contain difficult constructions too, which the tagger cannot handle well without seeing them in the training data.

3.2 Genres

Since different text genres have somewhat different characteristics, using only one genre (in this case newspaper texts) when training was expected to be worse than using several genres, since the test data consists of balanced material, not just newspaper texts. To test this, texts from an encyclopedia, the Nationalencyklopedin [NE, 2000], was tagged in the same way as the newspaper texts. These texts consisted of about 3 million tokens. The performance of lexicons created from only newspaper texts, only encyclopedia texts and from a mix of both was then evaluated. The results are shown in Figure 2.

As expected, using texts from both genres gave much better results than using only one genre. Using only encyclopedia texts was roughly as good as using only newspaper texts, though when using small amounts of text, the encyclopedia texts were worse, likely caused by selecting only the first part of the encyclopedia and thus getting a skewed distribution (i.e. too many words starting with the letter "a"). Using only newspaper texts the best accuracy was 95.3% and for the same amount of training data using both encyclopedia and newspaper texts, the accuracy was 95.5%.

3.3 Text Quality

To see if the quality of the texts is important, 100 000 words from second language learner essays were added to the training data. These texts contain many erroneous constructions and misspelled words. Somewhat unexpectedly, adding malformed language to the training data improved the lexicon when compared to a lexicon created from an equal amount of data consisting only of newspaper texts. This is probably caused by the addition of new genres to the training data, which as seen earlier is good, even when some of the new material is ungrammatical. Replacing the essays with the novel "Röda rummet" ("The Red Room") by August Strindberg [Strindberg, 1879], which consists of well written nineteenth century Swedish (i.e. not modern Swedish), gave the same accuracy, while replacing it with an equal amount of encyclopedia text (very high quality text), gave even higher accuracy. This indicates that while high quality text is better than low quality, the benefits of using several genres is greater than the drawbacks of errors in the text, and that using the same genres in the training data as in the test/application domain is better than other genres (i.e. in this case modern Swedish of the encyclopedia vs. Strindberg).

Some reasons for text quality not being very important can be seen from the properties of the tagger. The errors in the essays consist of spelling errors, which do not effect the tagger, and grammatical errors, which do effect the tagger. The reason spelling errors do not effect the tagger is that they will contribute a lot of misspelled words to the lexicon, but these will have no effect on the tagging of other words, unless the new words are the same misspelled words. Theoretically, spelling errors could possibly help the tagger, by having the ensemble guessing the tag and thereby having more input on how to tag unknown words (most misspelled words will be uncommon and thus used in training the taggers behavior on "unknown" words, so there will be more words for "unknown" word training). Also, the training data only needs to be locally (grammatically) correct to be useful, and many local constructions are correct in the essays.

4 Conclusions

The method gives a slightly less useful lexicon, when considering tagging accuracy, compared to the lexicon created from the proprietary corpus, but the new lexicon is freely distributable. The accuracy is still high enough to be useful. Since the method requires almost no manual work, it is a cheap way to create a new free lexicon.

Using several taggers when annotating the new training data gives a noticeable improvement in tagging accuracy when compared to using only one tagger. The extra manual workload is very low, since all the taggers are automatically trained and freely available.

The more new training data used, the less the degradation in tagging accuracy. Using more data requires no extra manual work (though the automatic tagging and training take more time) other than gathering the data, which was in this case also done automatically. Using as much data as possible seems reasonable, though the rate of improvement when adding more data decreases when the training set becomes large. When the training set already is large, it is more important to add new genres of text than to add large amounts of text, since using several genres is much better than using only one genre.

Adding texts with a lot of errors to the training data actually improved the lexicon. This was probably because these texts were from a different domain, and since the test data did not consist solely of newspaper texts, the positive effects from a broader spectrum of genres outweighed the negative effects of ungrammatical language constructions in these texts. Adding the same amount of high quality text from a new domain is even better than adding the poor quality texts.

Acknowledgments

This work has been funded by The Swedish Agency for Innovation Systems (VINNOVA). I would like to thank my supervisor professor Viggo Kann. I would also like to thank Stockholm University for letting me use the Stockholm-Umeå Corpus and NE Nationalencyklopedin AB for letting me use the Nationalencyklopedin.

References

[Brants, 2000] Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference, ANLP-2000*, pages 224–231, Seattle, USA, 2000.

[Carlberger and Kann, 1999] Johan Carlberger and Viggo Kann. Implementing an efficient part-of-speech tagger. *Software – Practice and Experience*, 29(9):815–832, 1999.

[Ejerhed *et al.*, 1992] Eva Ejerhed, Gunnel Källgren, Ola Wennstedt, and Magnus Åström. The linguistic annotation system of the Stockholm-Umeå Corpus project. Technical report, Department of General Linguistics, University of Umeå (DGL-UUM-R-33), Umeå, Sweden, 1992.

[Hassel, 2001] Martin Hassel. Internet as corpus - automatic construction of a Swedish news corpus. In *Proceedings of Nodalida 2001*, Uppsala, Sweden, 2001.

[NE, 2000] NE. *Nationalencyklopedin*. NE Nationalencyklopedin AB, 2000.

[Ngai and Florian, 2001] Grace Ngai and Radu Florian. Transformation-based learning in the fast lane. In *Proceedings of NAACL-2001*, pages 40–47, Carnegie Mellon University, Pittsburgh, USA, 2001.

[Ratnaparkhi, 1996] Adwait Ratnaparkhi. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania, Philadelphia, USA, 1996.

[Schmid, 1994] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994.

[Sjöbergh, 2003a] Jonas Sjöbergh. Combining pos-taggers for improved accuracy on Swedish text. In *Proceedings of NoDaLiDa 2003*, Reykjavik, Iceland, 2003.

[Sjöbergh, 2003b] Jonas Sjöbergh. Stomp, a POS-tagger with a different view. In *Proceedings of RANLP-2003*, pages 440–444, Borovets, Bulgaria, 2003.

[Strindberg, 1879] August Strindberg. Röda rummet (The Red Room), 1879.