Automatic Tagger Evaluation

NLP course assignment report

Kaarel Veskis, Erkki Liba
March 16, 2008

1 The Task

The task was to choose a tagger and a part-of-speech tagged corpus, train the tagger on a part of the corpus, and evaluate it on a left-out part, experimenting with the size of the training data.

2 The Tools

We chose to use the disambiguator for Estonian called ESTYHHM[1], a morphological analyzer and trigram HMM-disambiguator for Estonian.

As a corpus, we chose the morphologically disambiguated corpus of the University of Tartu[2]. The corpus contains texts belonging to different genres (fiction, newspaper texts, legal texts, etc.) and is divided into subcorpora accordingly.

A previous version of the disambiguator has been trained on a part (130 000 words) of the corpus. The newer version that is applied in our experiments is based similarly to the earlier versions on the Hidden Markov Model (HMM), but in contrast to the earlier versions the current version uses trigrams instead of bigrams. There are also some minor differences involving tagsets between earlier and the current version of the disambiguator.

In the case of Estonian, as of any agglutinative-inflectional languages, the role of the tagger is to disambiguate among the tags that are given by morphological analysis. For the analysis (of the test corpus), we use the morphological analyzer ESTMORF also described in the above mentioned paper.

The tagset used by ESTMORF as well as the disambiguator's tagset compared to other corresponding tagsets are given in the results workbook as a separate worksheet. Currently the tagset of the morphologically disambiguated corpus includes 118 disambiguator tags (M's).

---

[1] Kaalep, H-J., Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9-16, Tartu. http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf

[2] http://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=en

The disambiguator program takes as input a training set file, and produces as output data files that at a later stage are converted to binary-form files used at the testing stage. The program also outputs at the training stage a list of tags, 3-grams, equivalence classes of the tags, and a lexicon file together with corresponding tag probabilities extracted from the training set.

3 The Evalulation Tests

First we used the standard 10-fold crossvalidation test on all of the corpus and then on each subcorpus. For this we randomly partitioned the corpus and all of the subcorpora into 10 sets, each containing sentences from all areas of the corpus (or genre respectively).

Secondly we trained the disambiguator similarly to the crossvalidation tests on all of the corpus except one subcorpus, and then tested on this subcorpus. We did this for all of the subcorpora. The aim of this second test was to determine whether and in what measure is it harder to automatically disambiguate texts of one genre compared to other genres.

Several bash scripts were created to implement different stages of the first two tests in cycles.

To specifically experiment with the size of the corpus, we did a third test that is represented in a table at the bottom of the first worksheet. For this test we trained the disambiguator on only 1/10 of the whole corpus, and then tested it on another 1/10 portion of the corpus. The aim of this test was to see in what extent the reduced training corpus size effects the disambiguation error rate in case of equal test corpora sizes.

In addition to these tests we compared the tagging of the test sets and the original manually disambiguated versions of corresponding sets, extracted the erroneous tag pairs, and sorted them to form frequency lists of erroneous tag pairs. The lists of tag pairs of the second test were also converted to confusion matrices that may be found in the workbook.

4 The Results

The average accuracy in case of the 10-fold crossvalidation test on the whole corpus was 96.23 %. The average accuracy of the second test (corpus/genre) was 94.86 %.

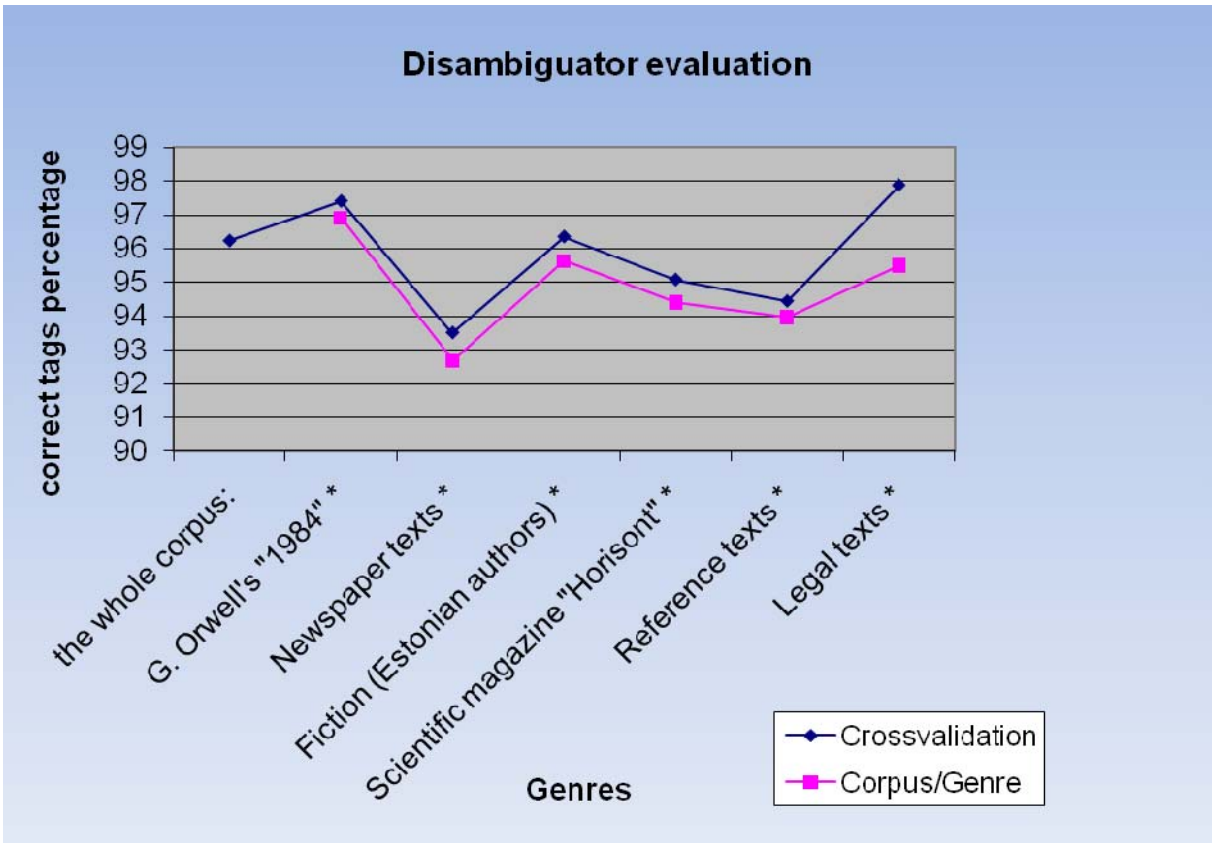

Fig. 1 The crossvalidation results

The diagram (Fig. 1) shows that the average result of each subcorpus is greatly dependent on the test corpus itself and only slightly dependent on the training corpus. The upper line of the diagram represents the test in the case of which the tagger was trained only on the texts of the same genre while the lower line represents the results of training the tagger on all of the texts **except** the test set subcorpus.

The difference between the two tests was greatest in the case of legal texts and lowest for reference texts. Newspaper texts give the worst results as a test corpus irrespectively of the training corpus size or genre.

The result of the third test confirmed the notion that the effect on tagging accuracy of the test set language type is predominant when compared to the effect of the training set size. The accuracy of the tagger, when trained on only 1/10 of the corpus, was only 1.44 % less than in the case of training on 9/10 of the corpus.

The frequency lists of erroneous tag pairs as well as the confusion matrices show that the most common tagging errors include some major problems facing all current taggers, e.g. NCSG (common noun singular genitive) vs. NPSG (proper noun singular genitive) and other noun-related problems, especially in case of legal texts.

However, most frequent tagging errors of newspaper texts are WOQ:X and WCQ:X. WOQ and WCQ represent opening-quotes and closing-quotes respectively, and X is a string that the tagger does not recognize as anything significant. Therefore, we can assume that the low score of tests on newspaper texts derives from a simple bug of the disambiguator program not recognizing one kind of quotes used in some subcorpora.

5 Conclusion

The average accuracy of this version of the disambiguator for Estonian is slightly lower than reported in (Kaalep and Vaino 2001)[3], but some bugfixes may likely result in a more accurate disambiguator.

6 Acknowledgements

We would like to thank the authors of the disambiguator Heiki-Jaan Kaalep and Tarmo Vaino for kindly instructing us on using the program and preparing our tests. We also thank Heiki-Jaan Kaalep for creating a script for converting tag pair lists into confusion matrices.

[3] Kaalep, H-J., Vaino, T. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V*, pp. 9-16, Tartu. http://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf