Automatic Tagger Evaluation

NGSLT: NLP

Syntax Assignment Report

Kaarel Veskis
Institute of Estonian and General Linguistics,
University of Tartu, Estonia
kaarel.veskis@ut.ee

Erkki Liba
Institute of Estonian and General Linguistics,
University of Tartu, Estonia
erkki.liba@ut.ee

Abstract

This paper presents an overview of testing the new version of the disambiguator for Estonian ESTYHMM. We describe the disambiguator as well as the morphologically disambiguated corpus, and tagsets that were used in our tests. We conducted three tests, including a standard 10-fold cross-validation test, and a test that involved using subcorpora of specific language types as test sets. Our tests results prove that disambiguation result is in close correlation with the language use, or some other aspects of the genre that the disambiguator was tested on, rather than with the size of the training corpus.

1 Introduction

The aim of our work was to evaluate the Estonian disambiguator ESTYHMM. ESTYHMM is a statistical part-of-speech (POS) tagger created for the Estonian language (Kaalep & Vaino, 2001).

ESTYHMM is not the only morphological disambiguator of Estonian. Another approach is the Constraint Grammar disambiguator for Estonian (Müürisep, 2001). However, many applications, e.g. automatic summarizing, and detection of nominal phrases, depend on ESTYHMM rather than the Constraint Grammar disambiguator. Therefore, the work on testing ESTYHMM is of considerable importance from the point of view of both the existing language applications as well as for developing new implementations that involve the levels of syntax, semantics, and pragmatics of the Estonian language.

As yet, there has been no thorough evaluation of the new version of ESTYHMM. Also, the tagger has not been trained on the larger version of the morphologically disambiguated corpus of the University of Tartu. Thus, the practical value of our work lies mostly in estimating the correctness of the tagger output on a larger corpus.

2 Method

2.1 Tools and Resources

Our 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.

We chose to use the disambiguator for Estonian called ESTYHMM (Kaalep & Vaino, 2001), a morphological analyzer and trigram HMM-disambiguator for Estonian.

In the case of Estonian, as of any agglutinative-inflectional language, the role of the tagger is to disambiguate among the tags that are given by morphological analysis.

Morphological analysis for the ESTYHMM disambiguator is provided by the morphological analyzer ESTMORF (Kaalep, 1997). ESTMORF is based on dictionary look-up and involves no heuristics or two-level rules. Adequate morphological descriptions are assigned to about 97% of tokens in a running text. The remaining 3% that are not analysed are rare words such as proper names, abbrev-

iations, acronyms, specific terminology, slang etc (Kaalep, 1997).

As for the corpus, we chose the morphologically disambiguated corpus of the University of Tartu[1].

This corpus was previously morphologically tagged with the ESTMORF tagger, and the output was manually disambiguated by human linguists. All texts have been disambiguated by two persons while a third person compared and corrected the result. Therefore, the corpus provides a solid Gold Standard for testing the disambiguator.

The corpus contains texts belonging to different genres (fiction, newspaper texts, legal texts, science, texts from a magazine of popular science, and reference texts; 513 000 words in total) and is divided into 6 subcorpora accordingly (see Table 1).

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 changes involving tagsets between the earlier and the current version of the disambiguator.

ESTYHMM treats a sentence not as a sequence of words but as a sequence of special disambiguating tags (M's) that have been obtained by transforming the morphological tags. The most probable sentence is selected, according to the probabilities that were calculated on the basis of the manually tagged training material. As a final step, the disambiguator special tags are being retransformed into morphological tags. In the case of very complicated situations, the disambiguator preserves ambiguity. Such cases make up 13.5% of input words (Kaalep, 1997).

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.

A new version of ESTYHMM has been recently released by its authors H. J. Kaalep and T. Vaino. In our experiments, this new version of ESTYHMM is under observation.

2.2 Tagsets

Estonian words can be divided into three main inflectional groups: declinable words, conjugable words, and uninflected words. These three groups can be divided into smaller units, depending on syntactic and/or semantic properties, but there is no one and correct classification scheme in this respect (Kaalep, 1997).

Additionally, there is a significant variation in terms of morphological tagsets used in annotating Estonian texts. As a part of our evaluation, we complemented the table that describes the morpho-syntactic categories of the morphologically disambiguated corpus of the University of Tartu[2], with the corresponding ESTMORF morphological tags, and the ESTYHMM M's. Table 3 presents a portion of this updated data.

Currently the tagset of ESTYHMM, when trained on all of the morphologically disambiguated corpus, includes 118 disambiguator tags (M's).

3 Evaluation of ESTYHMM

As a first step, we used the standard 10-fold crossvalidation test on all of the corpus, and then the same 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 the 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.

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

2 http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=en

Genre	Number of words
Fiction (Estonian authors)	104 000
G. Orwell's "1984"	75 500
Newspaper texts	111 000
Legal texts	121 000
Texts from the scientific magazine "Horisont"	98 000
Reference texts	4 000
Altogether	513 000

Table 1. *The structure of the morphologically disambiguated corpus of the University of Tartu.*

187	WOQ	X
182	WCQ	X
84	NCSG	NCSN
79	NCSN	NCSG
77	NCSG	NCS1
77	NCS1	NCSG
72	NPSG	NPSN
62	NCS1	NCSA
52	VMAZ	NCSN

Table 2. *A fragment of a sorted frequency list of the result of the first part of the 10-fold cross-validation test.*

To specifically experiment with the size of the corpus, we did a third test for which 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 third test was to see in what extent the reduced training corpus size effects the disambiguation error rate in case of equal training and test corpora sizes.

In addition to these tests we compared the tagging of the test sets of the first two experiments and the original manually disambiguated versions of corresponding sets, extracted the erroneous tag pairs (see Table 2), and sorted them to form frequency lists of erroneous tag pairs. The lists of tag pairs were also converted to confusion matrices (see table 4).

Tag	Explanation	Disambiguator tag (M)	Filosoft morph. analyzer notation	Corpus frequency	Example
_S_ com ?	Noun common	NCSX	_S_ ?	106	kesk-
_S_ com sg nom	Noun common singular nominative	NCSN	_S_ sg n	33515	jalg
_S_ com sg gen	Noun common singular genitive	NCSG	_S_ sg g	39871	jala
_S_ com sg part	Noun common singular partitive	NCS1	_S_ sg p	17027	jalga
_S_ com sg ill	Noun common singular illative	NCSA	_S_ sg ill	954	jalasse
_S_ com sg in	Noun common singular inessive	NCSA	_S_ sg in	8140	jalas
_S_ com sg el	Noun common singular elative	NCSA	_S_ sg el	5007	jalast
_S_ com sg all	Noun common singular allative	NCSA	_S_ sg all	4872	jalale
_S_ com sg ad	Noun common singular adessive	NCSA	_S_ sg ad	8360	jalal

_S_ com sg abl	Noun common singular ablative	NCSA	_S_ sg abl	764	jalalt
_S_ com sg tr	Noun common singular translative	NCS	_S_ sg tr	4591	jalaks
_S_ com sg term	Noun common singular terminative	NCS	_S_ sg ter	546	jalani
_S_ com sg es	Noun common singular essive	NCS	_S_ sg es	714	jalana
_S_ com sg abes	Noun common singular abessive	NCS	_S_ sg ab	330	jalata
_S_ com sg kom	Noun common singular komitative	NCS	_S_ sg kom	4589	jalaga
_S_ com sg adit	Noun common singular aditive	NCSA	_S_ adt	2407	jalga

Table 3. *Comparison of the tagsets used by ESTYHMM and ESTMORF with the mark-up of the morphologically disambiguated corpus of the University of Tartu.*

4 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 %.

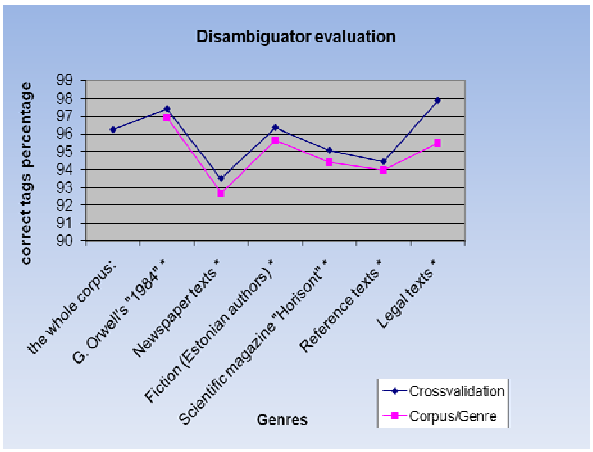

Figure 1. *The crossvalidation results*

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. In other words, the disambiguation result is in close correlation with the language use, or some other aspects of the genre that the disambiguator was tested on.

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 (Jurafsky & Martin, 2007: 157), e.g. NCSG (common noun singular genitive) vs. NPSG (proper noun singular genitive) and other noun-related problems, especially in case of legal texts (see table 4).

The evaluation of the previous version of ESTYHMM (Kaalep, 1997) indicated that about 3% of the morphologically analyzed words got a wrong analysis because of disambiguation. One third of these errors was due to the wrong selection among homonymous case forms of the noun (nominative, genitive, partitive, or short illative).

M's	NCS	NCS1	NCSA	NCSG	NCSN	NCSX
NCS	**3631**			1		
NCS1		**4332**	137	159	58	
NCSA		52	**10484**	38	1	
NCSG		492	131	**20002**	451	
NCSN		220	18	211	**10688**	
NCSX				13	7	**36**

Table 4. *A portion of the confusion matrix of the ESTYHMM evaluation results when tested on the legal texts subcorpus.The numbers not in bold type represent the times of occurrence of the erroneous tag pairs.The column labels indicate correct tags, row labels indicate the tagger's hypothesized tags. (The M's that are explained in Table 3.)*

Table 4 shows that this is an actual issue also in the case of the new version of ESTYHMM. Moreover, the same problem is reported as the most difficult one of problems facing the Constraint Grammar disambiguator for Estonian (Muischnek et al, 2001)

Furthermore, according to our evaluation results not one third but 61% of all erroneous tag pairs were noun-related in case of the crossvalidation test on the complete corpus. The percentage of the erroneous tag pairs where both of the tags were noun case forms was in our crossvalidation test 38%, i.e. slightly more than one third.

M's	NCS	NCS1	NCSA	NCSG	NCSN	NCSX
NCS	x					
NCS1		x	0,0224	0,0209	0,0133	
NCSA		0,0115	x	0,0036	0,0015	
NCSG		0,0247	0,0111	x	0,024	
NCSN		0,0080	0,0015	0,0301	x	
NCSX			0,0001		0,0003	x

Table 5. *A portion of the confusion matrix of the ESTYHMM evaluation results when tested on the newspaper texts subcorpus.Each cell indicates percentage of the overall tagging error.*

Table 5 presents the portion of the newspaper texts subcorpus evaluation result that corresponds to Table 4 (legal texts subcorpus). Instead of erroneous tag pair occurrence times that are comparable to those of table 4, table 5 presents percentages of the overall tagging error. We see that, although noun-related problems are common, each erroneous tag pair type actually constitutes a rather small part of the overall amount of errors. One possible reason to this contrast is the relatively large disambiguator tagset that causes the scattering of error types across the confusion matrix. Yet, the ESTMORF tagset is even much larger, as can be seen from Table 3. Thus, errors are likely to multiply at the re-transforming stage of the ESTYHMM tags to ESTMORF tags.

However, most frequent tagging errors in the case of some genres, and also of the crossvalidation test are WOQ:X and WCQ:X. For example, the most frequently occurring erroneous tag pair in newspaper texts subcorpus is WOQ:X (occurred 984 times), and the same tag pair is also the most frequent one in the case of the "Horisont" subcorpus (379 times).

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.

Besides these two error types, the most frequent error in case of the crossvalidation test was NCSN:NCSG that occurs when the tagger chooses the incorrect nominative case of the noun instead of the correct genitive. Also, NCSG:NCSN is in the top of the erroneous tag pairs frequency list.

5 Conclusion

This paper presents the results of the evaluation of the new version of ESTYHMM, the POS tagger for Estonian.

We show that the further development of the tagger should focus on enhancing the performance of the tagger from the point of view of better distinction among noun cases.

The cross-validation tests results prove that disambiguation result is in close correlation with the language register, or some other aspects of the genre that the disambiguator was tested on, rather than with the size of the training corpus.

The average accuracy of this version of the disambiguator for Estonian is slightly lower than reported in (Kaalep & Vaino, 2001), but some bugfixes may likely result in a more accurate disambiguator. After these bugfixes, the tools devised in the course of our evaluation tests can be used to perform a more detailed error analysis of the tagger.

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.

References

Jurafsky & Martin, 2008. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Preprint

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

Kaalep, H-J. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. Computers and the Humanities 31 pp. 115-133. http://www.cl.ut.ee/yllitised/chum1997.pdf

Muischnek, K. Müürisep, K., Puolakainen, T. 2001. Parsing of Estonian: Morphological Disambiguation and Determination of Syntactic Functions. Congressus Nonus Internationalis Fenno-Ugristarum Pars V pp 411-417. Tartu.

Müürisep, K. 2001. Parsing Estonian with Constraint Grammar. Online proceedings of NODALIDA'01. Uppsala. http://stp.ling.uu.se/nodalida01/pdf/myyrisep.pdf