Estonian-English Statistical Machine Translation: the First Results

Mark Fishel
Department of Informatics
University of Tartu
phishel@gmail.com

Heiki-Jaan Kaalep
Dept of General Linguistics
University of Tartu
Heiki-Jaan.Kaalep@ut.ee

Kadri Muischnek
Dept of General Linguistics
University of Tartu
Kadri.Muischnek@ut.ee

Abstract

This paper describes the experiments that apply phrase-based statistical machine translation to Estonian. The work has two main aims: the first one is to define the main problems in the output of Estonian-English statistical machine translation and set a baseline for further experiments with this language pair. The second is to compare the two available corpora of translated legislation texts and test them for compatibility. The experiment results show that statistical machine translation works well with that kind of text. The corpora appear to be compatible, and their combining – beneficial.

1 Introduction

Machine translation and automatic processing of the Estonian language in general is a considerable challenge. The language is highly inflective, which causes a great number of different word-forms. It has a complex system of joining and splitting compound nouns, which is hard to grasp even for a human learner. Finally, the word order is very heterogeneous.

The work described in this paper focuses on statistical machine translation (SMT) from Estonian into English. It has two aims. The first one is to examine, how well SMT works with this language pair, and to determine the main problems in its output. We thus want to set a baseline, which can be used by further experiments in the same area.

The second aim is to compare and evaluate the available resources. There are two sufficiently large parallel Estonian-English corpora, both consisting of translations of legislation texts. It is therefore necessary to compare them from the perspective of suitability for SMT, and to see whether these are similar enough to be combined to enrich the resulting translation and language models.

2 The grammatical system of Estonian

In this section we will briefly discuss some linguistic features of Estonian in order to better understand the challenges that the Estonian-English machine translation has to face.

Estonian has rich inflectional morphology: the nouns inflect for number and 14 cases, the verbs inflect for person, number, mood and tense. This means that we need great amounts of parallel data as, for example, one noun lemma can have 28 different word-forms in text. Compounding is free and productive in Estonian; orthography of a NP depends largely on semantics.

The morphological richness of Estonian is one of the main reasons for using Moses as we hope that in our future experiments we can split the word-forms into lemmas and grammatical categories and ease the data sparseness problem this way.

The syntactic relations (subject, object etc) in Estonian are coded mostly using morphological devices; the word order does not differentiate between the syntactic functions. The word order or, rather, constituent order of Estonian reveals remarkable heterogeneity. For example, a sentence consisting of three words (or constituents) can have nine different word order variants as exemplified in (1) (all the example sentences mean roughly the same, namely 'The child is eating a bun'). The actual word order in text depends on the pragmatics, information structure, clause type etc.

(2)	Laps	sööb	saia
	child-NOM	eat-3SG	bun-PART

Laps saia sööb.
Saia sööb laps.
Saia laps sööb.
Sööb laps saia.
Sööb saia laps.

Contrary to the constituent order of the clause, the order of the components of a noun phrase is fixed. But this fixed word oder can be diametrically opposite to that in English. For example, in a nominalization (3) the head of the NP, namely the word-form 'hospitalization' begins the phrase in English and ends it in Estonian.

(3)vältimatut	psühhiaatrilist
emergency-PART	psychiatric-PART

abi	vajava	isiku
care-PART	needing-PART	person-PART

haiglasse	paigutamine
hospital-ILL	allocation-NOM

'hospitalization of a person in need of emergency psychiatric care'

If the predicate of the clause is an analytic or perifrastic verb, the parts of the predicate can be separated from each other by several intervening constituents in certain clause types. In example (4) the predicate is a particle verb vastu võtma 'to adopt' .

(4)nõukogu	võttis	13. novembril
council	took	november-ADE

vastu	resolutsiooni
PARTICLE	resolution-GEN

'The Council adopted a resolution on 13th of November'

3 Corpora Description

As mentioned in the introduction, there exist two partially overlapping Estonian-English parallel corpora, which are sufficiently large for training SMT models. The source of both are translated legislation texts. Firstly, this means that it should be possible to combine the two and therefore to enrich the trained SMT models. Secondly, the contained language is considerably more constrained than spoken language – it should therefore be easier to model it. Thirdly, the law text domain potentially has a higher demand for translating huge amounts of texts, and would therefore benefit from a semi- or fully automatic translation system.

3.1 The UT Corpus

The first of the abovementioned corpora[1] was created at the university of Tartu. The corpus contains 7.8 million words in English and 5.0 million in Estonian.

The corpus is sentence-aligned using the Vanilla aligner (Danielsson and Ridings, 1997), based on the algorithm by Gale and Church (1993). The total number of aligned units is 435 700.

3.2 The JRC-Acquis Corpus

The second used corpus consists of the Estonian and English parts of the JRC-Acquis multilingual parallel corpus (Steinberger et al., 2006). The used corpus contains 7.6 million English and 5 million Estonian words. The corpus is initially aligned on the level of paragraphs, but these are usually short and do usually contain one sentence, or even only part of a sentence. Automatic alignment was also performed using the Vanilla aligner. The total number of aligned units is 295 000. Regardless of the amount of words being almost the same as in the UT corpus, the more general alignment level causes the number of the alignment units to be smaller (and the units themselves, longer on the average).

4 Experiments and Results

4.1 Experiment setup

To ensure statistical significance of the results both corpora were randomly split into the training and

[1] http://www.cl.ut.ee/korpused/paralleel

the testing set; the latter initially consisted of 0.1% of the corresponding corpora. We further filtered both testing sets manually, leaving out alignment errors, pairs with one of the sentences empty, sentences witout a single word in the source or target language and paragraph and section numbering sentences. This way the results would show the performance of the SMT system applied to natural language sentences only. Finally, we removed the testing sentences that also appear in the training set. As a result, the size of the test sets was reduced to 749 sentences in the UT, and 649 – in the JRC-Acquis corpus.

Since manual filtering of the training sets wasn't feasible due to the set sizes, only automatic filtering was performed. The excluded sentence pairs were the ones which included sentences longer than 100 words and the ones where the ratio of the word numbers exceeded 9. This left 429 000 and 272 000 parallel units in the UT and JRC-Acquis training sets, respectively. In order for the corpora to suit with the requirements of the used software they were preprocessed in the following way. The UT corpus was converted to UTF-8 encoding and HTML entities in both corpora were replaced with corresponding UTF-characters. All sentences were lower-cased. Finally, the punctuation was separated from the words in order for the translation model training script to recognize them as separate words.

We used n-gram language models, trained with the SRI LM package (Stolcke, 2002). Word alignments were obtained using GIZA++ (Och and Ney 2003). Phrase table composition and decoding was done with Moses[2] and the software included with it.

The automatic evaluation metric, used in the experiments, is BLEU. However, in order not to limit the comparison to that, we performed a limited human evaluation of the output. In addition, the testing results are available online[3].

4.2 Results

We trained three models: on the UT corpus, on the JRC-Acquis corpus and combining both corpora. These models we evaluated against the UT corpus and the JRC-Acquis corpus.

Table 1 presents the quality of the translations measured by BLEU.

Trained on	UT	JRC	Combined
Tested on			
UT	39.26	29.80	41.60
JRC	38.45	42.38	45.22

Table 1. Translation quality of SMT systems trained /tested on different corpora as measured by BLEU.

Intra-corpus translation

In the first set of experiments we trained and tested the SMT model on the same corpora. This would show the relative corpus performance when used in SMT.

The BLEU scores for UT and JRC-Acquis corpora were 39.26 and 42.38 respectively. The scores are noticeably higher than the ones, published for spoken/written language baseline translation – e.g. (Bojar et al, 2006), (Koehn and Knight, 2003) – which is most probably explained by the highly constrained nature of the legislation language.

Inter-Corpus Translation

We continued by taking a SMT model trained on one corpus and testing it on another. This would show how similar the two corpora are from the SMT perspective.

Training the model using the UT corpus and testing it on the JRC-Acquis test set produced a BLEU score of 38.45, this is only slightly lower than the JRC-Acquis-trained model score. On the other hand, the JRC-Acquis-trained model gave a 29.8 BLEU score when trained on the UT test set. This suggests that the SMT model, trained on the UT corpus is more applicable to the extra-corpus language phenomena. We suggest that the reason is in the more detailed alignment in the UT corpus: this most probably causes less corpus-subjective word alignments and phrase table entries.

Combined corpora experiments

Finally we tested the compatibility of the two corpora from the perspective of combining them for SMT training. Although the corpora have overlapping sources, only 18 000 and 27 000 unique parallel units coincide completely in the Estonian and

[2] http://www.statmt.org/moses

[3] http://ats.cs.ut.ee/u/fishel/smt

English corpora parts, respectively. Therefore corpora are combined by simple concatenation.

The BLEU score of the SMT models trained on the combined corpus is 41.6 and 45.22 when tested on the UT and JRC-Acquis test set, respectively. When compared to the intra-corpus translation results, the improvement of the UT test set score (2.34 BLEU) is slightly lower than the improvement of the JRC-Acquis one (2.84 BLEU). This supports the hypothesis made in the inter-corpus experiment section: the models built on the UT training set generalize better on the JRC-Acquis test set than vice versa.

4.3 Manual Output Evaluation

It has been pointed out (Callison-Burch et al, 2006) that while BLEU attempts to capture allowable variation in the translation, it allows random permuting of phrases in the hypothesis compared with the reference translation. In our opinion it also explains the relatively high BLEU score in our experiments.

In order to balance these shortcomings, we carried out a limited human evaluation of the results. The human evaluator gave 6 x 250 output sentences one of the following ratings: 1) good translation, i.e. expresses the same meaning as the source sentence and is grammatically correct; 2) an acceptable translation with minor errors, i.e. expresses the same meaning as the source sentence, but has some grammar errors; 3) does not express the same meaning as the source sentence. The third group covers both the cases if the output is an unintelligible mess of words and if the sentence has a meaning, but that is different from that of a source sentence.

While the UT corpus test set was evaluated as it is, the JRC-Acquis one had the paragraphs split manually into sentences before evaluating; however, approximately every 10th paragraph contained more than one sentence. Results of the human evaluation are presented in table 2.

The shortcomings of the human evaluation are that the sub-clauses of a long sentence have not been evaluated separately. If one sub-clause of a long sentence consisting of several sub-clauses is unintelligible, the sentence gets the overall “wrong” rating.

Trained on	UT	Cmb	JRC	UT	Cmb	JRC
Tested on	UT	UT	UT	JRC	JRC	JRC
Good	16%	15%	8%	13%	15%	11%
Acceptable	11%	15%	9%	9%	15%	14%
Wrong	73%	70%	83%	78%	70%	75%

Table 2. Human evaluation of the SMT output. Cmb – combined corpora.

5 Discussion

The results of human evaluation mostly support the conclusion, initially based on BLEU results: combining the corpora results in slight improvement in the SMT output. This conclusion, however, remains so far subjective to the used corpora and requires further testing. In addition, we believe that the sources of the corpora might overlap much more than indicated in the subsection 3.2, which doesn’t show due to differences in version/encoding etc. This has to be regarded in the further experiments.

The main problem that distorts the meaning and grammar of the resulting translations is the failure to place the parts of the translation in the right order. The legislative language that the corpora contains is characterized by heavy use of nominalisations, the resulting noun phrases are long and tend to have a complicated structure. So, if the word order (constituent order) in Estonian source sentence is too different from the correct English one, the system fails to make the needed permutations in long sentences. We had hoped that using a phrase-based statistical machine translation system helps us to overcome the word order differences in the source and target languages, but apparently additional techniques are required to do so.

To exemplify the problems with word/constituent order, let's take an Estonian sentence from the JRC-Acquis test corpus and have a closer look at its reference translation and the output of our system. In order to see what has gone wrong in our translation, the phrases (in the meaning used in the phrase-based SMT) that represent the same meaningful units in both Estonian and English have been numbered according to their order in the Estonian sentence.

```
(5) source:
[1 euroopa majandusühenduse ja
Šveitsi konföderatsiooni]
[2 vaheliste kokkulepete]
[3 kohaldamisel]
[4 rakendatakse ühenduses]
[5 ühiskomitee]
[6 otsust nr 5 / 81]

SMT output:
[1 the european economic
community and the swiss
confederation]
[2 of agreements between]
[3 the application]
[5 of the joint committee]
[4 shall apply in the community]
[6 decision no 5 / 81]

reference:
[3 for the purposes of
application]
[2 of the agreements between]
[1 the european economic
community and the swiss
confederation] ,
[6 decision no 5 / 81]
[5 of the joint committee]
[4 shall apply in the community]
```

We can see that the output of the SMT system contains all the correct phrases except for the translation of the word *kohaldamisel* 'applying', translated in the reference translation as 'for the purposes of application' and as 'the application' in the system output. But the order of the phrases in the system output follows too much the phrase order in the source sentence - the system has failed to make the long-distance permutations. The phrase order of the source sentence is 1-2-3-4-5-6; the order of these constituents in the reference sentence is 3-2-1-6-5-4, but our system produces 1-2-3-5-4-6.

At the moment reordering is purely the task of the distortion model of the SMT algorithm, and as indicated by the results, this is not enough. One of the ways to solve the problem is described in (Nießen and Ney, 2001). According to this method the input sentence can be reordered using morpho-syntactic information, so that the word order resembles better that of the target language. Another approach to the same problem would be to re-rank the n-best output list and/or reorder the output sentences.

The incapability of our baseline-model to consider grammatical information creates translations where adverbial NP is translated into subject NP (as the first NP in an English sentence is usually the subject, but in Estonian the order of the syntactic constituents is more varied (cf. also example 2).

```
(6) source:
selles   tunnistuses
this-INE certificate-INE
esitatakse
are-reproduced
kontrollimise    tulemused
verification-GEN result-PL.NOM

output:
this certificate shall be sub-
mitted to the results of verifi-
cation

reference translation:
this certificate shall reproduce
the findings of the examination
```

Another frequently examined disadvantage of the SMT output is the failure to translate several Estonian word-forms into English. The probable cause is the data sparseness, caused by the Estonian morphology and free compounding.

So the systems gives a correct translation of the noun *eeskiri* ‘regulation’, but fails to give any translation of the compound *finantseeskiri* ‘financial regulation’. Needless to say that a case-form of a noun that has appeared several times in the training corpus, but not in this particular form is a new unknown word for the system.

One of the possible solutions is to use the factored translation models of the Moses decoder by translating vectors of base forms and morphological features instead of the words themselves. Also, several preprocessing techniques exist that can reduce the problematic effect, e.g. (Koehn and Knight, 2003), (Perez et al, 2006).

6 Conclusions

This paper described a set of experiments, in which statistical machine tanslation was applied to the Estonian language. The first objective of this work was to test, how well SMT translates from Estonian into English, when trained on the available corpora, and to determine the main output problems. The second one was to compare two existing parallel corpora for this language pair, and to test whether combining the two can bring benefit to the resulting SMT models.

The experiment results show that SMT is applicable to Estonian and the domain, represented in the corpora. The output of the SMT was analyzed, and the main output problems were determined: these being the wrong order of phrases and sparse data. Still, the BLEU scores of the output are higher than the ones reported for spoken language translation, most probably due to the constrained nature of the language of the corpora. Furthermore, combining the two corpora appears to improve the translation output.

Future work includes testing the techniques, used for reducing the data sparsity problem and output quality improvement. In addition, the opposite translation direction has to be inspected.

References

Ondrej Bojar, Evgeny Matusov and Hermann Ney. 2006. Czech-English phrase-based machine translation. In *Proceedings of the 5th International Conference on NLP, FinTal* 2006, pp 214-224. Turku, Finland.

Chris Callison-Burch, Miles Osborne and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of the 11thh Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp 249-256. Trento, Italy

Pernilla Danielsson and Daniel Ridings. 1997. Practical presentation of a “Vanilla” aligner. In *TELRI Workshop in alignment and exploitation of texts*, Ljubljana, Slovenia. Available at http://nl.ijs.si/telri/Vanilla/doc/ljubljana/

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. Computational Linguistics, 19 (1), pp 177-184.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp 187-193, Budapest, Hungary.

Sonja Nießen and Herman Ney. 2001. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of MT Summit VIII*, pp 1081-1085, Santiago de Compostela, Galicia, Spain.

Franz J. Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29 (1), pp 19-51.

Alicia Pérez, Inés Torres and Francisco Casacuberta. 2006. Towards the improvement of statistical translation models using linguistic features. In *Proceedings of the 5th International Conference on Natural Language Processing FinTal*, pp 716-725. Turku, Finland.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, pp 2142-2147, Genoa, Italy.

Andres Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, vol 2, pp 901-904. Denver, Colorado, USA.