SMT of Latvian, Lithuanian and Estonian Languages: a Comparative Study

Maxim KHALILOV [a,1], Lauma PRETKALNIŅA [b], Natalja KUVALDINA [c] and Veronika PERESEINA [d]

[a] *Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, The Netherlands*

[b] *Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia*

[c] *Marine Systems Institute, Tallinn University of Technology, Tallinn, Estonia*

[d] *Jönköping International Business School, Jönköping University, Jönköping, Sweden*

Abstract. This paper is an attempt to discover the main challenges in working with Baltic and Estonian languages, and to identify the most significant sources of errors generated by a SMT system trained on large-vocabulary parallel corpora from legislative domain. An immense distinction between Latvian/Lithuanian and Estonian languages causes a set of non-equivalent difficulties which we classify and compare.

In the analysis step, we move beyond automatic scores and contribute presenting a human error analysis of MT systems output that helps to determine the most prominent source of errors typical for SMT systems under consideration.

Keywords. Machine translation, Error analysis, Statistical methods

Introduction

Unlike many small languages, Latvian, Lithuanian (called Baltic languages together) and Estonian (LLE) languages have been quite well-researched linguistically and possess parallel corpora, which is an indispensable resource for statistical machine translation (SMT). The availability of the bilingual corpus opens the way for the estimation of the SMT models and the development of real-world automatic translation systems.

Until recently, automatic translation from/into LLE languages has not received much attention from the scientific community and, to a certain extent, can be considered still an open research line in the field of automatic translation. Scarce attempts at constructing SMT systems for these languages can be found as of 2007 [1,2,3], that is much later than SMT systems for popular language pairs.

In this study we present a set of full multilingual bi-directional experiments on Latvian↔English, Lithuanian↔English and Estonian↔English SMT, mostly concentrating on more difficult translation tasks in which English language is a source. We compare the outputs of state-of-the-art SMT systems that follow a phrase-based approach to

[1]Corresponding Author: Maxim Khalilov, Institute for Logic, Language and Computation, University of Amsterdam, P.O. Box 94242, 1090 GE Amsterdam, The Netherlands, E-mail: m.khalilov@uva.nl.

MT and report results in terms of automatic evaluation metrics. We also experiment with different parameters of SMT systems and show that their accurate tuning can improve the quality of modeling the deviations between LLE languages and English.

In the following step, we move beyond automatic scores of translation quality and present a manual error analysis of English⇒Latvian/Lithuanian and English⇒Estonian MT systems output that the vast majority of research papers avoid. The translation errors typical for each language pair are detected following the framework proposed in [4]. The results of human evaluation done by native or nearly native speakers of the target languages helps to shed light on advantages and disadvantages of the SMT systems under consideration and identify the most prominent source of errors.

The rest of the paper is organized as follows: Section 1 briefly outlines the most important characteristics of the LLE languages and describes the corpus which was used in experiments, Section 2 introduces the phrase-based approach to SMT, Section 3 details the experiments, Section 4 reports the results of automatic translation quality evaluation, along with the results of human error analysis, while Section 5 presents the conclusions drawn from the study.

1. Languages and data

There is a variety of languages spoken in Baltic states, which includes languages like Lithuanian, Latvian and Estonian. Here, we provide the reader with a brief overview of the three official languages of Baltic countries and their most important grammatical characteristics.

Latvian. Latvian is the official language of Latvia and belongs to the Baltic branch of the Indo-European language family. There are about 1.5 million native Latvian speakers around the world: 1.38 million are in Latvia, while others are spread in USA, Russia, Sweden, and some other countries. Also Latvian language is a second language for about 0.5 million inhabitants of Latvia and several tens of thousands from neighbor countries, especially Lithuania[2]. Latvian is characterized by rich morphology, relatively complex pre- and postposition structures and high level of morphosyntactic ambiguity. There are no articles, two grammatical genders and two numbers in Latvian. Nouns decline into seven cases.

Lithuanian. Lithuanian language is most closely related to Latvian and from linguistic point of view there is no much difference in treating Latvian and Lithuanian. A small difference between them from MT perspective is that the latter has a higher number of declensions, inflectional types of nouns and adjectives and a comparison system of adjectives. Another minor distinction between the two live Baltic languages is that there is no neuter gender in Latvian, while there is a number of neuter obsolete (but still used) word forms in Lithuanian. Linguistic topology of Latvian and Lithuanian is SOV, however, word order is relatively free. Lithuanian is one of the official languages of the European Union. There are about 2.96 million native Lithuanian speakers in Lithuania and about 170,000 abroad[3].

[2]Source: State Language Agency `http://www.valoda.lv/lv/latviesuval`
[3]Source: Wikipedia `http://en.wikipedia.org/wiki/Lithuanian_language`

Estonian. While Lithuanian and Latvian are closely related and descend from the same ancestor language, Estonian differs from them in many aspects and does not even belong to the same language family[4]. Estonian is a highly inflectional agglutinative language characterized by a large number of cases (14 productive cases), and absence of grammatical genders. This language is characterized by rich structure of declensional and conjugational forms. The number of these forms is significantly higher than in Latvian and Lithuanian. Basic word order is SVO. There are about 1.1 million Estonian speakers in Estonia and tens of thousands in other countries[5].

All the languages under consideration are characterized by a relatively free order of sentence constituents (non-configurational languages), however the number of ways how a sentence can be rearranged without becoming ungrammatical is much higher for Estonian than for Lithuanian and Latvian languages.

1.1. Data

We used JRC-Acquis parallel corpus [5] of about one million parallel sentences. Development set contains 500 sentences randomly extracted from the bilingual corpus, test corpus size is 1,000 lines. Development and test are provided with 1 reference translation. Basic statistics of the bilingual corpus can be found in Table 1.

	Latvian	Lithuanian	Estonian	English
Training				
Sentences	1,09M			
Words	23.87M	23.90M	21.15M	28.21M
Vocabulary	338.65K	355.17K	507.80K	237.94K
Development				
Sentences	0.5K			
Words	10.82K	11.56K	9.49K	13.6K
Vocabulary	1.28K	1.90K	2.32K	1.14K
Test				
Sentences	1.0K			
Words	20.09K	21.55K	20.37K	27.74K
Vocabulary	3.86K	4.64K	4.85K	2.43K

Table 1. Basic statistics of the JRC-Acquis corpus.

2. Phrase-based SMT

SMT is based on the principle of translating a source sentence ($f_1^J = f_1, f_2, ..., f_J$) into a sentence in the target language ($e_1^I = e_1, e_2, ..., e_I$). The problem is formulated in

[4]Estonian belongs to the Baltic Finnic branch of the Uralic languages and its most close relative is Finnish. Estonian is one of the few languages in Europe which does not belong to the Indo-European family.

[5]Source: Estonian Institute `http://www.einst.ee/publications/language/`

terms of source and target languages; it is defined according to equation (1) and can be reformulated as selecting a translation with the highest probability from a set of target sentences (2):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ p(e_1^I \mid f_1^J) \right\} = \tag{1}$$

$$= \arg\max_{e_1^I} \left\{ p(f_1^J \mid e_1^I) \cdot p(e_1^I) \right\} \tag{2}$$

where I and J represent the number of words in the target and source languages, respectively.

Modern state-of-the-art SMT systems operate with the bilingual units (phrases) extracted from the parallel corpus based on word-to-word alignment. They are enhanced by the *maximum entropy approach* and the posterior probability is calculated as a *log-linear combination* of a set of feature functions [6]. Using this technique, the additional models are combined to determine the translation hypothesis $\hat{e}_1^I$ that maximizes a log-linear combination of these feature models, as shown in (3):

$$\hat{e}_1^I = \arg\max_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \tag{3}$$

where the feature functions h_m refer to the system models and the set of λ_m refers to the weights corresponding to these models.

A phrase-based translation [6] is considered a three step algorithm: (1) the source sequence of words is segmented in phrases, (2) each phrase is translated into target language using translation table, (3) the target phrases are reordered to be inherent in the target language.

A phrase-based system which we experiment with within the framework of this study employs feature functions for a phrase pair translation model, a language model (LM), a reordering model, and a model to score translation hypothesis according to length. The weights λ_m are usually set to optimize system performance [7] as measures by BLEU [8].

Two word reordering methods are considered: a distance-based distortion model [9] and lexicalized MSD block-oriented model [10].

An alternative decoding technique is Minimum Bayes Risk (MBR), the approach that seeks for hypothesis which is similar to the most likely translations using optimization functions that measure translation performance [11].

3. Experiments

Experimental setup The system built for the English⇔LLE translation experiments is implemented within the open-source Moses toolkit [12]. Standard training and weights tuning procedures which were used to build our system are explained in details on the Moses web page: `http://www.statmt.org/moses/`. Word alignments have been estimated using GIZA++ [13] tool assuming 4 iterations of the IBM2 model, 5 HMM model iterations, 4 iterations of the IBM4 model, and 50 statistical word classes (found with mkcls tool [14]). Target LMs with unmodified Kneser-Ney backoff discounting

were generated using the SRI language modeling toolkit [15]. Automatic evaluation was case insensitive and punctuation marks were not considered.

Systems Apart from unfactorized phrase-translation, the set of systems considered in this paper includes alternative configurations. We investigate the impact that different ingredients of a phrase-based translation system have on the final system performance. We experiment with (1) different orders of target-side LM, (2) the way to reduce the search space during decoding (beam size) and (3) MBR decoding.

4. Results

Evaluation of the system performance is twofold. In the first step, we report the standard automatic translation scores, namely BLEU, NIST and METEOR (MTR) scores for the tasks in which English is a target language, and BLEU and NIST scores for the English⇒Latvian/Lithuanian/Estonian tasks. In the next step, we look at the human analysis of translation output, that, in the general case, provides a comprehensive comparison of multiple translation systems and reveals the most prominent source of errors generated by phrase-based systems.

4.1. Automatic evaluation

The evaluation results for the test datasets are reported in Tables 2 and 3.

System	EnLv		EnLt		EnEst	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
Baseline	19.07	4.81	13.29	4.06	11.84	3.76
LM: 3-gram	18.36	4.74	13.21	3.94	10.99	3.66
LM: 4-gram	18.37	4.74	14.14	4.15	11.39	3.78
S1000	18.95	4.77	13.05	3.98	11.58	3.77
MBR	19.15	4.83	13.50	4.21	11.56	3.71

Table 2. Automatic translation scores for English⇒LLE translations.

The systems considered include: (1) a *baseline* configuration (5-gram target-side LM); (2-3) *LM: 3(4)-gram* systems, considering lower order target-side LMs; (4) *S1000* system with increased stack size (beam) for histogram pruning (100 is the default value) and (5) *MBR* configuration where MBR algorithm is used during decoding.

The major conclusion that can be drawn from the results of automatic evaluation is that modification of default MOSES parameters does not significantly change translation systems' performance. However, when using MBR algorithm instead of standard optimization procedure leads to a slight improvement in terms of translation scores for English⇔Latvian and English⇔Lithuanian tasks. For all translations into English there is a consistent improvement of systems' performance with an increase of the target-side LM order, that is not the case for English⇒LLE translations.

As expected, translation from and into Latvian is a less complex task comparing to other directions, while Estonian⇔English tasks are the most complicated from the SMT perspective. Increased beam size has a positive impact on translation scores for

the majority of the systems under consideration but at the cost of translation speed that increases significantly (in 3-4 times).

4.2. Manual error evaluation

We performed error analysis on the 1,000 lines test dataset for English⇒LLE baseline systems. The analysis of typical errors generated by each system was done following the error classification scheme proposed in [4] by contrasting the systems output with the reference translation. The comparative statistics of errors is reported in Table 4.

System	LvEn			LtEn			EstEn		
	BLEU	NIST	MTR	BLEU	NIST	MTR	BLEU	NIST	MTR
Baseline	29.69	6.38	55.07	26.27	6.04	49.59	18.52	4.42	45.74
LM: 3-gram	27.78	6.18	54.65	25.55	5.82	49.23	17.32	4.39	45.25
LM: 4-gram	29.47	6.25	54.88	26.01	5.93	49.57	18.21	4.41	45.54
S1000	29.75	6.43	55.09	26.12	5.93	49.57	18.55	4.44	45.79
MBR	29.73	6.39	55.05	26.33	6.01	49.55	18.40	4.34	45.71

Table 3. Automatic translation scores for LLE⇒English translations.

Type	Sub-type	EnLv	EnLt	EnEst
Missing words		**631 (10.16 %)**	**622 (10.35 %)**	**884 (12.21 %)**
	Content words	272 (4.38 %)	244 (4.06 %)	422 (5.83 %)
	Filler words	359 (5.78 %)	378 (6.29 %)	462 (6.38 %)
Word order		**885 (14.27 %)**	**868 (14.44 %)**	**1,216 (16.80 %)**
	Local word order	181 (2.92 %)	194 (3.23 %)	300 (4.14 %)
	Local phrase order	317 (5.11 %)	270 (4.49 %)	459 (6.34 %)
	Global word order	241 (3.89 %)	216 (3.59 %)	340 (4.70 %)
	Global phrase order	146 (2.35 %)	188 (3.13 %)	117 (2.44 %)
Incorrect words		**4,294 (69.18 %)**	**4,164 (69.30 %)**	**4,653 (64.27 %)**
	Wrong lex. choice	348 (5.60 %)	292 (4.86 %)	758 (10.47 %)
	Incorrect disambig.	865 (13.94 %)	920 (15.31 %)	525 (7.25 %)
	Incorrect form	2,237 (36.05 %)	2,472 (41.14 %)	2,927 (40.43 %)
	Extra words	750 (12.08 %)	430 (7.16 %)	351 (4.85 %)
	Style	94 (1.51 %)	50 (0.83 %)	85 (1.18 %)
	Idioms	0 (0.00 %)	0 (0.00 %)	7 (0.09 %)
Unk. words		**85 (1.37 %)**	**107 (1.78 %)**	**341 (4.71 %)**
Punctuation		**198 (3.19 %)**	**248 (4.13 %)**	**86 (1.19 %)**
Total		**6,206**	**6,009**	**7,240**

Table 4. Human made error statistics for a representative test set.

Distribution of errors for the languages under consideration is quite similar, however the total number of errors generated by Estionian system is about 20% higher than for Lithuanian and Latvian systems.

The most prominent class of errors is related to incorrect words/word forms that is typical for morphologically rich languages, while the prevailing type of errors within this class is "incorrect word forms" that can be re-phrased as if the system is able to generate the correct word lemma but can not find the correct lexical form.

Rich morphology, high level of morpho-syntactic ambiguity and relatively complex pre- and postposition structures typical for Latvian and Lithuanian cause a significant number of errors typical for morphologically-rich languages, namely, incorrect word forms and wrong lexical choice.

The minor linguistical distinction between Lithuanian and Latvian is reflected in similar total number and distribution of errors when translating into Latvian and Lithuanian. In case of Estonian language that differs from Latvain and Lithuanian in many aspects, many errors come from erroneous grammatical choice, i.e. the translation system is not able to generate the correct word on the target side. The major difficulty that either an Estonian-English or an English-Estonian SMT system faces is a rich structure of word forms, whose number is much higher than in Latvian and Lithuanian languages.

There is a substantial number of errors related to generation of the correct word/constituent order within the sentence for all English⇒LLE tasks (≈15%), which is explained by a free word order nature of the target languages. For non-configurational languages, the rich overall inflectional system renders word order less important than in isolating languages like English. Nevertheless, there is only a limited number of acceptable word permutations. Evaluation of the word order correctness for free word order languages is not a trivial task. We considered equally all admissible word order combinations for the translations, hence the clumps are marked erroneous only if the word order is not acceptable or changes the meaning of the sentence.

The total number of errors generated by the English⇒Latvian system is slightly higher than by the one for the English⇒Lithuanian translation that contradicts theory. We explain this phenomenon by sparseness of translation model.

5. Conclusions and discussion

In this paper, we report results of multilingual translation experiments that involve Baltic and Estonian languages on the one side and English on the other. Unsurprisingly, translation scores for Latvian and Lithunian translataions are higher than for transltions into and from Estonian, that is equivalent to the fact that the latter is a more difficult translation task. MBR decoding is slightly more efficient than standard maximum a posteriori decoding for Latvian⇔English and Lithuanian⇔English tasks.

Human-made error analysis, performed on the next step, gives a more complete and fair view of translation quality than automatic scores which just compare a translation output with a reference translation. Surprisingly, all three LLE languages are found to be quite similar in terms of error distribution that can be partly explained by the specificity of the legal domain that the data belongs to. English⇒Estionan system generates more errors than English⇒Latvian and English⇒Lithuanian systems mostly due to richer morphology, different word order and linguistic typology. Latvian and Lithuanian systems

mostly suffer from incorrect word forms, incorrect disambiguation of lexical instances and word order errors. In case of Estonian system, the most frequent errors, in addition to aforementioned errors, include wrong words translation.

The high number of translation errors of all types (6-7 per sentences) leaves room for a lot of interesting research which can potentially lead to a significant improvement of English⇔LLE translations.

References

[1] M. Fishel, H. Kaalep, and K. Muischnek. Estonian-english statistical machine translation: the first results. In *Proceedings of NODALIDA-2007*, Tartu, Estonia, May 2007.

[2] M. Khalilov, J.A.R. Fonollosa, I. Skadina, E. Bralitis, and L Pretkalnina. Towards improving english-latvian translation: a system comparison and a new rescoring feature. In *Proceedings of LREC'10*, pages 1719–1725, Valetta, Malta, May 2010.

[3] Ph. Koehn, A. Birch, and R. Steinberger. 462 machine translation systems for europe. In *Proceedings of the twelfth Machine Translation Summit*, pages 65–72, Ottawa, Ontario, Canada, August 2009.

[4] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney. Error Analysis of Machine Translation Output. In *Proceedings of LREC'06*, pages 697–702, 2006.

[5] S. Ralf, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC'2006*, Genoa, Italy, May 2006.

[6] F. Och and H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of ACL 2002*, pages 295–302, 2002.

[7] F. Och. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*, pages 160–167, Sapporo, July 2003.

[8] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318, 2002.

[9] Ph. Koehn, F. Och, and D. Marcu. Statistical phrase-based machine translation. In *Proceedings of the HLT-NAACL 2003*, pages 48–54, 2003.

[10] C. Tillman. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL'04*, 2004.

[11] S. Kumar and W. Byrne. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of HLT/NAACL 2004*, 2004.

[12] Ph. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: open-source toolkit for statistical machine translation. In *Proceedings of ACL 2007*, pages 177–180, 2007.

[13] F. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[14] F. Och. An efficient method for determining bilingual word classes. In *Proceedings of ACL 1999*, pages 71–76, June 1999.

[15] A. Stolcke. SRILM: an extensible language modeling toolkit. In *Proceedings of the Int. Conf. on Spoken Language Processing*, pages 901–904, 2002.