Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-141

Building the World's Best General Domain MT for Baltic Languages

Raivis SKADIŅŠ[a,1], Valters ŠICS[a] and Roberts ROZIS[a]
[a]*Tilde, Latvia*

Abstract. In this paper we present our experience in building machine translation (MT) systems for the languages of the Baltic States: Estonian, Latvian, and Lithuanian. The paper reports on the implementation, research, data, data collection methods, and evaluation of the MT. Results of the evaluation show that it is possible to collect a sufficient amount of data and train MT systems that can compete with Google in quality and even overtake it in general domain MT.

Keywords. Machine translation, Baltic languages, corpora

Introduction

The languages of the Baltic States belong to the class of inflected languages with complex morphology and rather free word order, which makes them complicated subjects for statistical MT [1]. The lack of necessary language technologies and the need for large amounts of parallel corpora makes MT even more difficult. According to a recent report from META-NET, the languages of the Baltic States are at risk of digital extinction and MT technologies are weakly developed for them [2]. At the same time there have been numerous academic and industrial activities to research and build MT systems. The quality of Google and Microsoft MT systems affirms that the quality of statistical MT mainly depends on the amount of training data [3], and the quality level set by Google is difficult to achieve by others. This sets a high challenge for local researchers and industry.

1. MT Systems

To train our SMT systems we used a MT [4] platform which is based on the Moses toolkit [5]. When training general domain SMT systems, we see that a standard phrase-based approach only (even without any language specifics) can result in a good quality MT. To achieve even higher MT quality, we can integrate language pair specific methods which slightly improve SMT quality [6][7], but the improvement from more training data is more convincing. The most promising method to incorporate linguistic knowledge in SMT is to use morphology in factored SMT models. We have improved

[1] Corresponding Author: Raivis Skadiņš, Tilde, Latvia; E-mail: raivis.skadins@tilde.lv

word alignment calculated over lemmas instead of surface forms. An additional language model over morphosyntactic tags can be built in order to improve inter-phrase consistency [6].

We have introduced data filters in the SMT training process that remove suspicious data where the target sentence is equal to the source sentence, too long segments, spaces between each letter, too different word count, too much non-alphabetic characters and characters that are not from the alphabet of the particular language.

There are tokens in the text that cannot be properly translated by SMT because there may not be enough parallel data available to calculate reliable statistics. These tokens are dates, identifiers, currency, and different kinds of numbers, URLs, and e-mail addresses that should not be translated at all. We have introduced a non-translatable token (NTT) detection procedure where we detect different kinds of tokens, and they are not translated but left as in the original text.

Direct speech or citation enclosed in quotes, or explanations enclosed in parentheses are quite independent parts of a wider sentence. We introduce borders around these kinds of phrases to limit word reordering.

Table 1. Amount of training data and results of the automatic evaluation

MT systems	Corpora size, sentences		BLEU
	Parallel	Monolingual	
English – Latvian	8.9 M	60.9 M	37.38
Latvian – English	12.7 M	66.6 M	44.15
English – Lithuanian	5.3 M	24.1 M	28.80
Lithuanian – English	5.3 M	81.0 M	38.42
English – Estonian	12.5 M	33.1 M	24.22
Estonian – English	11.5 M	107.9 M	37.97

2. Training Data

We use both publicly available corpora collected by other institutions and corpora collected by ourselves. The most important sources of data used for MT training are:

- Publicly available parallel and monolingual corpora (see Table 2).
- Parallel and monolingual corpora collected by Tilde (see Table 3).

The collection of publicly available corpora includes: Europarl corpus [8], DGT-TM [9], JRC-Acquis [10], ECDC-TM [11], EAC-TM [12] and other smaller corpora available from the Joint Research Center, the OPUS corpus [13][14], which includes data from the European Medicines Agency (EMEA), European Central Bank (ECB), Open Subtitles, EU Constitution and other smaller corpora. Along with the parallel corpora we also used News Commentary and News Crawl English monolingual corpora (part of WMT 2013 shared task [15] training data) to train English language models.

Parallel and monolingual corpora collected by Tilde includes national legislation, standards, technical documents and product descriptions widely available on the web (some, examples: www.ceresit.net, www.europe-nikon.com), EU brochures from EU BookShop [16], news portals (like www.bnn.lv, www.makroekonomika.lv) and many more. The size of the collected data sets varies significantly, the most important data sets among these are:

- EU BookShop corpus [16]: books, brochures, posters, maps, leaflets, technical documents, periodicals, CD-ROMs, DVDs, etc. on the European Union's activities and policies. The EU Bookshop is an online service and archive of publications from various European institutions. The service contains a large body of publications in the 24 official languages of the EU;
- BookMT corpus: parallel data automatically extracted from comparable corpora containing scanned book pairs, over 3 M parallel segments in English, Latvian, Lithuanian and Estonian;
- WebScrape corpus: English-Latvian parallel data extracted from c.a. 159,000 comparable html and pdf documents crawled from the web (3.48 M sentences);
- Monolingual WebNews corpus: mainly data crawled from the web (state institutions, portals, newspapers etc.);
- ACCURAT Wikipedia corpus: parallel data automatically extracted from Wikipedia data using the ACCURAT Toolkit [17];
- The Bible corpus: a corpus consisting of verse aligned bilingual Bible texts in English, Latvian and Estonian;
- Parallel website corpus: a corpus consisting of parallel data that have been crawled from bilingual and multilingual web sites. The crawled content was aligned using the ACCURAT Toolkit [17] and Microsoft's Bilingual Sentence Aligner [18];
- RAPID corpus: Directorate General Communication press releases (http://europa.eu/rapid/);
- National legislation corpora: Latvian-English legislation corpus of Republic of Latvia[2] and Estonian Acts of Law[3];
- Estonian Open Parallel Corpus (EOPC)[4].

See Table 1 for the total amount of data used in the training of our SMT systems, and Tables 2 and 3 for information about which corpora have been used for which language pairs.

[2] http://metashare.elda.org/repository/browse/latvian-english-ngram-corpus-legislation-of-republic-of-latvia/77492e76a37611e3960f001dd8b71c192245316d09514123af25dcc6acd86c00/

[3] https://www.riigiteataja.ee/tutvustus.html?m=3

[4] http://metashare.dfki.de/repository/browse/estonian-open-parallel-corpus/7e9c6a12a37611e3960f001dd8b71c19d2e99b6816a247a683fa58158006985c/

Table 2. Publicly available corpora used to train the MT systems

Corpora	English-Latvian	English-Lithuanian	English-Estonian
Europarl corpus	+	–	+
DGT-TM corpus	+	+	+
JRC Acquis corpus	+	–	+
ECDC-TM corpus	+	–	+
EAC-TM corpus	–	–	+
News Commentary and News Crawl corpora	+	+	+
OPUS corpus			
EMEA corpus	+	+	+
ECB corpus	+	–	+
OpenSubtitles	+	–	+
EU Constitution	+	–	+
KDE documentation	+	+	+

Table 3. Corpora and dictionaries collected by Tilde and used to train the MT systems

Corpora	English-Latvian	English-Lithuanian	English-Estonian
Term dictionaries from eurotermbank.com	+	+	+
English-Latvian dictionary	+	–	–
Assistive technology term dictionary	+	–	+
English-Lithuanian dictionary	–	+	–
Translation memories from localization	+	+	+
EU BookShop corpus	+	–	+
BookMT corpus	+	+	+
Webscrape corpus	+	–	–
Monolingual WebNews corpus	+	+	+
ACCURAT Wikipedia corpus	–	–	+
Bible corpus	+	–	+
Parallel website corpus	+	–	+
RAPID corpus	+	–	+
National legislation corpora	+	–	+
Estonian Open Parallel Corpus	–	–	+

Different MT systems use different amounts of parallel data originating from EU documents. The latest systems (Latvian- English and Estonian-English-Estonian) include all available data from all releases of DGT-TM, Europarl and JRC-Acquis corpora, which is c.a. 5.5 M parallel sentences. The proportion of EU data to all data used in training is about 43 to 47%.

3. Evaluation

The BLEU metric [19] was used for the automatic evaluation using a balanced general domain evaluation corpus[5] that represents general domain data, which is a mixture of texts in different domains, representing the expected translation needs of a typical user.

[5] http://metashare.tilde.com/repository/browse/accurat-balanced-test-corpus-for-under-resourced-languages/7922fbd2a37611e3960f001dd8b71c19d96efef81e1948988b8a71b2d9d37937

It includes texts from fiction, business letters, IT texts, news, magazine articles, legal documents, popular science texts, manuals and EU legal texts. The evaluation corpus contains 512 parallel sentences in English, Estonian, Latvian and Lithuanian.

The summary of the automatic evaluation results in comparison with Google[6], Microsoft[7] and the University of Tartu[8] machine translation systems is presented in Figure 1.

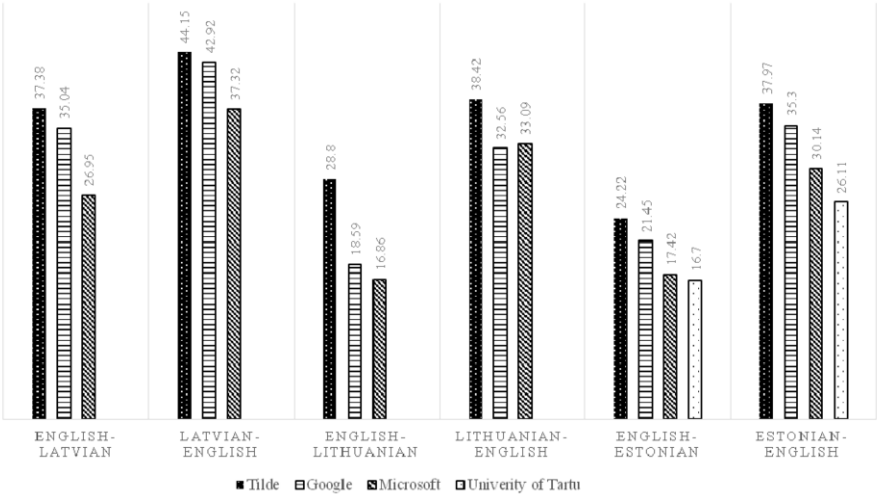

Figure 1. Our MT systems compared to Google, Microsoft and University of Tartu MT systems.

For human evaluation of the systems we used a ranking of translated sentences relative to each other. This is the official determinant of translation quality used in the Workshop on Statistical Machine Translation shared tasks [15]. Just as in our previous experiments [6], we ranked 2 MT systems and calculated how often evaluators preferred one system's translation to the other, and we calculated the confidence interval [20] to see the statistical relevance of the evaluation. The results of the human evaluation are given in Table 4, it shows that in all but one case the evaluators preferred the systems presented in this paper over other systems. Google's Lithuanian-English MT system was ranked better in human evaluation, although according to the automatic evaluation Tilde's MT system was better.

The English-Latvian MT system has been also evaluated in practical use for software localization where it helped to achieve 32.9% productivity increase [21].

[6] https://translate.google.com/

[7] http://www.bing.com/translator/

[8] http://masintolge.ut.ee/info/info.php?locale=en_US

Table 4. Manual evaluation results for 3 systems, balanced test corpus

MT System 1	MT System 2	System 1 preferred (%)	Confidence Interval
Tilde English – Latvian	Google	51.56	± 3.40
Tilde Latvian – English	Google	54.00	± 3.83
Tilde English – Lithuanian	Google	50.48	± 2.32
Tilde Lithuanian – English	Google	43.59	± 3.40
Tilde English – Estonian	Google	52.20	± 2.47
Tilde English – Estonian	University of Tartu	60.86	± 4.23
Tilde Estonian – English	Google	51.06	± 4.30

4. Conclusions

In this paper we have reported training and evaluation results of SMT systems for the languages of the Baltic States.

The evaluation results show that the presented MT systems slightly outperform MT systems created by global MT developers Google and Microsoft in both automatic and human evaluations. It shows that it is possible to achieve and exceed the quality level set by Google and Microsoft even for general domain MT.

Our results show that big amount of high quality training data is very important to build competitive general domain MT systems, and it is possible to collect a significant amount of training data. The most important sources of MT training data are:

- Publicly available parallel and monolingual corpora;
- Multilingual websites, books and other sources of parallel and comparable texts that can be crawled and aligned.

The systems presented are available as a free online service at http://translate.tilde.com, they are also included in software packages Tildes Birojs/Biuras. English-Latvian has been tested in practical use for software localization where it helped to achieve a productivity increase for the English-Latvian language pair.

The reported methods can also be applied to build MT systems for other under-resourced languages.

We are planning to continue our work to build ever better general domain MT systems for the languages of the Baltic States by (i) collecting new parallel and monolingual data, (ii) cleaning collected data, and (iii) continuously retraining MT systems using all the collected corpora. The other promising way for improvements is integrating more language pair specific linguistic knowledge in statistical MT.

Acknowledgements

The research leading to these results has received funding from the research project "2.6. Multilingual Machine Translation" of EU Structural funds, contract No. L-KC-11-0003 signed between ICT Competence Centre (www.itkc.lv) and Investment and Development Agency of Latvia.

References

[1] Koehn, P., Birch, A., & Steinberger, R. (2009). 462 Machine Translation Systems for Europe, Proceedings of MT Summit XII.

[2] Rehm, G. & Uszkoreit, H., editors. (2012). META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg etc. 32 volumes on 31 European languages.

[3] Och, F. J. (2005). Statistical Machine Translation: Foundations and Recent Advances. Tutorial at the Tenth Machine Translation Summit. Phuket, Thailand.

[4] Vasiļjevs, A., Skadiņš, R., & Tiedemann, J. (2012). LetsMT!: Cloud-Based Platform for Do-It-Yourself Machine Translation. In Proceedings of the ACL 2012 System Demonstrations (pp. 43–48). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P12-3008

[5] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, In Proceedings of the ACL 2007 Demo and Poster Sessions (pp. 177-180). Prague.

[6] Skadiņš, R., Goba, K., & Šics, V. (2010). Improving SMT for Baltic Languages with Factored Models. In Proceedings of the Fourth International Conference Baltic HLT 2010, Frontiers in Artificial Intelligence and Applications, Vol. 2192 (pp. 125–132). Riga: IOS Press.

[7] Deksne, D., & Skadiņš, R. (2012). Data Pre-Processing to Train a Better Lithuanian-English MT System. In A. Tavast, K. Muischnek, & M. Koit (Eds.), Frontiers in Artificial Intelligence and Applications, Volume 247: Human Language Technologies – The Baltic Perspective (pp. 36–41). IOS Press. doi:10.3233/978-1-61499-133-5-36

[8] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Conference Proceedings: the tenth Machine Translation Summit. Phuket, Thailand: AAMT, pp. 79-86

[9] Steinberger, R., Eisele, A., Klocek, S., Pilos, S., & Schlüter, P. (2012). DGT-TM: A freely Available Translation Memory in 22 Languages. Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012). Istanbul, Turkey, pp. 454-459.

[10] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D. (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'2006), pp. 24-26. Genoa, Italy.

[11] ECDC-TM. (2012). Retrieved from http://ipsc.jrc.ec.europa.eu/?id=782

[12] EAC-TM. (2012). Retrieved from http://ipsc.jrc.ec.europa.eu/?id=784

[13] Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)

[14] Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V) (pp. 237-248). Amsterdam/Philadelphia: John Benjamins.

[15] Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation (pp. 1-44). Sofia, Bulgaria: Association for Computational Linguistics.

[16] Skadiņš, R., Tiedemann, J., Rozis, R., & Deksne, D. (2014). Billions of Parallel Words for Free: Building and Using the EU Bookshop Corpus. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 1850–1855). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/846_Paper.pdf

[17] Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A., & Babych, B. (2012). ACCURAT toolkit for multi-level alignment and information extraction from comparable corpora. In Proceedings of System Demonstrations Track of ACL 2012 (pp. 91–96). Retrieved from http://dl.acm.org/citation.cfm?id=2390470.2390486

[18] Moore, R.C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora. In Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users. London, UK: Springer-Verlag, pp. 135-144.

[19] Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics. : ACL

[20] Wallis, S.A. (2013). Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. Journal of Quantitative Linguistics 20:3, 178-208. DOI:10.1080/09296174.2013.799918

[21] Skadiņš, R., Pinnis, M., Vasiļjevs, A., Skadiņa, I., & Hudik, T. (2014). Application of Machine Translation in Localization into Low-Resourced Languages. In M. Tadić, P. Koehn, J. Roturier, & A. Way (Eds.), Proceedings of the 17th Annual Conference of the European Association for Machine Translation EAMT2014 (pp. 209–216). Dubrovnik: European Association for Machine Translation. Retrieved from http://hnk.ffzg.hr/eamt2014/EAMT2014_proceedings.pdf