Improving SMT with Morphology Knowledge for Baltic Languages

Raivis SKADIŅŠ[a,b], Kārlis GOBA[a] and Valters ŠICS[a]
[a] *Tilde SIA, Latvia,* [b] *University of Latvia, Latvia*

Introduction

In the recent years, several machine translation systems have been built for the Baltic languages. Besides Google and Microsoft machine translation engines and research experiments with statistical MT for Latvian [1] and Lithuanian, there are both English-Latvian [2] and English-Lithuanian [3] rule-based MT systems available.

Both Latvian and Lithuanian are morphologically rich languages with quite free word order. In combination with the limited availability of parallel corpora for these languages, it poses a sparseness problem for phrase-based SMT. This research is a part of a project to build the best general-purpose phrase-based SMT using publicly available and proprietary corpora and tools. During the project we added language-specific knowledge to assess the possible improvement of translation quality.

This paper reports on implementation, as well as automatic and human evaluation of English-Latvian and Lithuanian-English statistical machine translation systems. Results of human evaluation show that integrating morphology knowledge into SMT gives significant improvement of translation quality compared to baseline SMT.

1. SMT resources and training

For training the SMT systems, both monolingual and bilingual sentence-aligned parallel corpora of substantial size are required. The corpus size largely determines the quality of translation, as has been shown both in case of multilingual SMT [4] and English-Latvian SMT [1].

For all of our trained SMT systems the parallel training corpus includes publicly available DGT-TM and OPUS corpora [5], as well as our proprietary localization corpus obtained from translation memories that were created during localization of software content, appliance user manuals and software help content. We additionally included word and phrase translations from bilingual dictionaries to increase word coverage. Total size of English-Latvian parallel data is 3.23 M sentence pairs and 2.71 M sentence pairs for English-Lithuanian.

Monolingual corpora were prepared from the corresponding monolingual part of parallel corpora, as well as news articles from Web for Latvian and European Parliament Proceedings and News Commentary[1] for English. Total size of Latvian monolingual corpus is 319 M words and 521 M words for English.

The baseline SMT models were trained on lowercased surface forms for source and target languages only. The SMT baseline models were trained for reference point to assess the relative improvement of additional data manipulation, factors, corpus size and language models. We used Moses SMT toolkit [6] for SMT system training and decoding, and extended the SMT system within the Moses framework by integrating morphology knowledge.

Both Latvian and Lithuanian belong to the class of inflected languages which are complex from the point of view of morphology. There are over 2000 different morphology tags for Latvian and Lithuanian. With Baltic languages as the target languages for SMT, the high inflectional variation of target language increases data sparseness at the boundaries of translated phrases, where a language model over surface forms might be inadequate to estimate the probability of target sentence reliably. Following the approach of English-Czech factored SMT [7] we introduced an additional language model over disambiguated morphologic tags in the English-Latvian system. The tags contain morphologic properties generated by a statistical morphology tagger. The order of the tag LM was increased to 7, as the tag data has significantly smaller vocabulary. The described system is later referred as ‘SMT tag’. As many small languages do not have part of speech tagger we made another English-Latvian experiment where we used inflectional suffixes instead of morphological tags. This system is later referred as ‘SMT suffix’.

[1] The monolingual training data from the Fourth Workshop on Statistical Machine Translation (http://www.statmt.org/wmt09/translation-task.html)

When translating from morphologically rich language, the SMT baseline system will not recognize the word forms that are not present in the training data. To reduce the data sparseness, we split each Lithuanian word into stem and an optional inflectional suffix which were treated as separate tokens during the training process. Suffixes were prefixed by a special symbol to avoid overlapping with stems. This system is later referred as 'SMT stem/suffix'.

Parallel data used to train all SMT systems mentioned before come from reliable sources and contain really parallel segments. We also used bigger but not so reliable parallel data automatically extracted from comparable web corpus. These parallel data are extracted from c.a. 159,000 comparable html and pdf documents crawled from the web (3.48 M sentences) and from 104 fiction books (0.66 M sentences). Extraction of parallel data from comparable corpus was done in Accurat project[2]. We used these automatically extracted data together with parallel data to train larger scale English-Latvian factored SMT system with morphological tags. This system is later referred as 'SMT tag+'.

2. Results and conclusions

We used BLEU [8] metric for automatic evaluation. The summary of automatic evaluation results is presented in Table 1.

Table 1. Automatic evaluation BLEU scores

System	Language pair	BLEU
Google[3]	English-Latvian	32.9
SMT baseline	English-Latvian	24.8
SMT suffix	English-Latvian	25.3
SMT tag	English-Latvian	25.6
SMT tag+	English-Latvian	33.0
Google	Lithuanian-English	29.5
SMT baseline	Lithuanian-English	28.3
SMT stem/suffix	Lithuanian-English	28.0

For Lithuanian-English system we also measured the out of vocabulary (OOV) rate on both per-word and per-sentence basis (Table 2). The per-word OOV rate is the percentage of untranslated words in the output text, and the per-sentence OOV rate is the percentage of sentences that contain at least one untranslated word.

Table 2. OOV rates for Lithuanian-English

System	Language pair	OOV, Words	OOV, Sentences
SMT baseline	Lithuanian-English	3.31%	39.8%
SMT stem/suffix	Lithuanian-English	2.17%	27.3%

Table 3. Manual evaluation results. Comparison of two systems

System1	System2	Language pair	p	ci
SMT tag	SMT baseline	English-Latvian	58.67 %	±4.98 %
Google	SMT tag	English-Latvian	55.73 %	±6.01 %
SMT stem/suffix	SMT baseline	Lithuanian-English	52.32 %	±4.14 %

The best systems where compared to baseline systems and the best English-Latvian factored system to Google SMT system. Manual evaluation was done comparing two systems. As a result of such comparison we get a percentage showing how often evaluators preferred one system over the other and a confidence interval [9]. Results of manual evaluation are given in Table 3.

By development of factored EN-LV SMT models we expected to improve human assessment of quality by targeting local word agreement and inter-phrase consistency. Human evaluation shows a clear preference for factored SMT over the baseline SMT, which operates only with the surface forms. However, automated metric scores show only slight improvement on test corpus (BLEU 24.8% vs 23.8%). Although using of suffixes instead of morphological tags did not give as good improvement it still gives better results as the baseline system.

By developing of the LT-EN SMT Stem/suffix model we expected to increase overall translation quality by reduction of untranslated words. The BLEU score slightly decreased (BLEU 28.0% vs 28.3%), however the OOV rate differs significantly. Human evaluation results suggest that users prefer lower OOV rate despite slight reduction in overall translation quality in terms of BLEU score.

[2] European Union Seventh Framework Programme (FP7/2007-2013) project, www.accurat-project.eu

[3] Google Translate (http://translate.google.com/) as of July 2010

Acknowledgements

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347 and this research was partially supported by the European Social Fund (ESF) activity No. 1.1.2.1.2 "Support to doctor's studies", project No. 2009/0138/1DP/1.1.2.1.2/09/IPIA/VIAA/004.

References

[1] I. Skadiņa, E. Brālītis, English-Latvian SMT: knowledge or data, in *Proceedings of the 17th Nordic Conference on Computational Linguistics NODALIDA*, Odense, Denmark, NEALT Proceedings Series, Vol. 4 (2009), 242–245., 2009.

[2] R. Skadiņš, I. Skadiņa, D. Deksne, T. Gornostay, English/Russian-Latvian Machine Translation System, in *Proceedings of HLT'2007*, Kaunas, Lithuania, 2007.

[3] E. Rimkute, J. Kovalevskaite, Linguistic Evaluation of the First English-Lithuanian Machine Translation System, in *Proceedings of HLT'2007*, Kaunas, Lithuania, 2007.

[4] P. Koehn, J.F. Och, D. Marcu, Statistical Phrase-Based Translation, in *Proceedings of HLT/NAACL 2003*, 2003.

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

[6] P. Koehn, M. Federico, B. Cowan, R. Zens, C. Duer, O. Bojar, A. Constantin, E. Herbst, Moses: Open Source Toolkit for Statistical Machine Translation, in *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177-180, Prague, 2007.

[7] O. Bojar, D. Mareček, V. Novák et al., English-Czech MT in 2008, in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, Association for Computational Linguistics, 2009

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

[9] R. Skadiņš, K. Goba and V. Šics, Improving SMT for Baltic Languages with Factored Models, *in Proceedings of the Fourth International Conference Baltic HLT 2010*, Riga, 2010