English-Latvian SMT: knowledge or data?

Inguna Skadiņa
Institute of Mathematics and Computer Science, University of Latvia
Riga, Latvia
Inguna.Skadina@lumii.lv

Edgars Brālītis
Institute of Mathematics and Computer Science, University of Latvia
Riga, Latvia
Edgars.Bralitis@lumii.lv

Abstract

In cases when phrase-based statistical machine translation (SMT) is applied to languages with rather free word order and rich morphology, translated texts often are not fluent due to misused inflectional forms and wrong word order between phrases or even inside the phrase. One of possible solutions how to improve translation quality is to apply factored models. The paper presents work on English-Latvian phrase-based and factored SMT systems and, using evaluation results, demonstrates that although factored models seem more appropriate for highly inflected languages, they have rather small influence on translation results, while using phrase-model with more data better translation quality could be achieved.

1 Introduction

In the last decade statistical machine translation (SMT) has become one of the most popular approaches in the field of automated translation. SMT started with word-based models, but significant advances were made with the introduction of phrase-based models.

Statistical Machine Translation tries to generate translations on the basis of statistical models, with parameters derived from the analysis of bilingual text corpora. SMT approach is language independent, but it requires large bilingual corpora for training. If such corpora are available, good results can be achieved in translating texts of a similar kind. The main advantage of SMT approach is a possibility to build up the system in a relatively small period of time.

One of the prerequisites for classical SMT systems is availability of large parallel corpus which computer then uses in the training process. The lack of large parallel corpus is the main reason why experiments with SMT in Baltic countries have been started only recently, i.e., implementation of Estonian-English (Fishel et al., 2007) and English-Latvian (Skadiņa and Brālītis, 2007) SMT systems have been reported only in 2007.

Phrase-based models (Koehn et al., 2003) typically deals with words or phrases thus often generating wrong form if the text is translated into morphologically rich language. In factored translation models (Koehn and Hoang, 2007), the surface forms are augmented with factors, such as grammatical information and base form. Thus factored models usually improve machine translation performance for problems such as morphology, free word order, and sentence-level grammatical coherence. For instance, English-Czech factored SMT reached 27.04% BLEU for all morphological features and 27.45% BLEU for selected morphological features, in comparison to the baseline of 25.82% BLEU (Koehn and Hoang, 2007).

The paper presents application of factored approach to English-Latvian SMT and discusses evaluation results, demonstrating that simple factored models have no enough influence on translation quality, i.e., with phrase-based models and more data better results could be achieved as with factored models and less data.

2 English-Latvian factored translation model

Latvian language is typical representative of morphologically rich languages. Almost all open word classes, i.e., nouns, adjectives, numerals, pronouns, and verbs, are inflective.

Latvian nouns and pronouns have 6 cases in both singular and plural. Adjectives, numerals and participles have 6 cases in singular and plural, 2 genders and definite and indefinite form. In Latvian conjugation system there are two numbers, three persons and three tenses (present, future and

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 242–245

past tenses), both simple and compound and 5 moods. Moreover, inflected forms are highly ambiguous. Nouns in Latvian have 29 graphically different endings and only 13 of them are unambiguous, adjectives have 24 graphically different endings and half of them are ambiguous, verbs have 28 graphically different endings and only 17 of them are unambiguous. The most common ambiguity classes are feminine singular genitive vs. feminine plural nominative and masculine singular accusative vs. masculine plural genitive.

Initially the phrase-based model was built for JRC Acquis 2.2. corpus (Steinberger et al., 2006). Human analysis of translation results allowed us to conclude that one of the central problems, which make translation abstruse, is wrong inflectional form (Skadiņa and Brālītis, 2007). Selection of wrong inflectional form not only influences fluency of translation, but in complex sentences (as most of legal texts) makes translation abstruse. Therefore, to improve translation quality, factored SMT system which uses Latvian morphological analyzer was built (Figure 1).

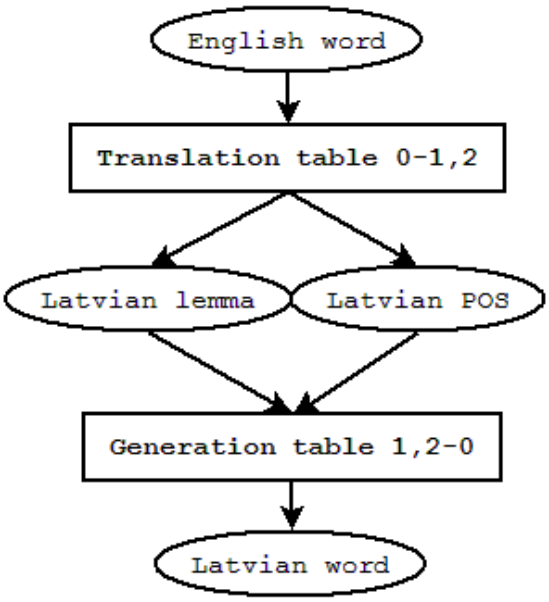

Figure 1. English-Latvian factored SMT

For Latvian language three factor model was chosen: inflected form (0), base form (or lemma) (1) and morphological tag (2). The translation process has been decompiled into the following steps:

1. English sentence has been translated into sequence of Latvian factors 1 and 2, using translation table 0-1,2
2. Sequence of Latvian factors 1 and 2 were translated into factor 0, using generation table 1,2-0

In addition three Latvian language models were implemented for each factor. All language models have the same weight during translation process.

The system was built using well known tools and techniques: after text normalization (texts were converted to lower-case, empty lines deleted, punctuation marks were separated from words) the GIZA++ tool (Och and Ney, 2003) was used for translation models. For Latvian language models SRI LM Toolkit (Stolcke, 2002) with recommended parameters (modified Kneser-Ney discounting and interpolation) were used. We used Latvian morphological analyzer by Paikens (2007) and Latvian tagger developed by Virza (unpublished work). For decoding Moses decoder (Koehn, 2004) was used.

3 Evaluation

For test purposes two test collections were created. For automatic evaluation sentences were selected randomly (1 from 1000) from JRC 3.0 corpus after omitting sentences from JRC2.2 corpus, and excluding sentences with possibly wrong alignment. As result text collection for automatic evaluation contains 843 sentences. For human evaluation 200 sentences were chosen from the test collection. Sentences which were included into test collections were deleted from JRC3.0 and JRC2.2 corpora before the training.

The evaluation was performed for four systems: phrase-based model built from JRC2.2 corpus, factored model built from JRC2.2 corpus, phrase-based model built from JRC3.0 corpus and factored model built from JRC3.0 corpus.

At first influence of different parameters, i.e., n-grams in language model, target language corpus, choice of decoder, on phrase-based models was evaluated (Table 1). As it is shown below the size of corpora has considerable influence on BLEU score (Papineni et al., 2002), while choice of decoder and number of n-grams in language model has relatively small influence on translation quality.

Phrase table data	Total number of words	Decoder	Language model		
			Order	Training data	
				JRC Acquis 2.2	JRC Acquis 3.0
JRC Acquis 2.2	EN – 9 932 536, LV – 8 129 497	Pharaoh	3	**26.20**	**29.91**
			5	23.91	26.43
JRC Acquis 2.2	EN – 9 932 536, LV – 8 129 497	Moses	3	26.37	31.82
			5	26.45	**32.41**
			7	**26.63**	32.37
JRC Acquis 3.0	EN -55 537 910, LV – 44 703 607	Moses	3	31.68	43.28
			5	**31.99**	44.93
			7	31.74	**44.97**

Table 1. Evaluation results (Bleu scores) for phrase-based models

While influence of size of training corpora on translation quality is obvious result, our main goal was to evaluate the influence of factored models on translation quality (Table 2). The first results show that it is possible to increase translation performance using factored models as it is in case of phrase-based model built form JRC Acquis 2.2 corpus and corresponding (same training data, language model order and other parameters) factored model. Factored model built from JRC3.0 Acquis corpus is slightly outperformed by corresponding phrase-based model.

SMT	BLEU score
JRC Acquis 2.2. phrase-based	26.37
JRC Acquis 2.2. factored	28.96
JRC Acquis 3.0 phrase-based	43.28
JRC Acquis 3.0 factored	42.98

Table 2. Influence of factored model on translation quality

Although JRC Acquis 2.2. corpus is almost five times smaller than JRC Acquis 3.0 corpus, it is sufficient for translation dictionary of EU legislation domain: in test corpus of 200 sentences and 5313 running words in Latvian reference translation, only 33 words have been left without translation, in 9 cases word was not translated by all SMT systems, thus only in 24 cases English word was not in JRC Acquis 2.2. translation model.

To compare automatic evaluation results with human intuition, the simple human evaluation was performed. The evaluator compared translations of four systems: phrase-based model built from JRC2.2 corpus, factored model built from JRC2.2 corpus, phrase-based model built from JRC3.0 corpus and factored model built from JRC3.0 corpus, by answering two questions for each sentence in test collection:

1. Which translation is better?
2. Is translation understandable easily?

Evaluator may select several translations in case the output of systems is similar. Evaluation results are summarized in Table 3.

	Chosen as the best (or one of best)	Easily understandable translations
JRC Acquis 2.2 phrase-based	20	12
JRC Acquis 2.2 factored	42	18
JRC Acquis 3.0 phrase-based	57	30
JRC Acquis 3.0 factored	74	28
All	71	15

Table 3. Results of human evaluation

The human evaluation showed the similar tendency – the size of training corpus has great influence on translation performance. 58 translations (29%) generated by systems trained on JRC Acquis 3.0 corpus are evaluated as understandable, while for systems trained on JRC Acquis 2.2 only 30 translations (15%) are evaluated as understandable. In 71 cases (35.5%) human evaluator has classified all translations as equal in translation quality; however, most of them are not easily understandable.

4 Conclusions

The paper presents first results of English-Latvian factored SMT systems showing that at current stage, better results could be achieved with more data as by intelligence, i.e., factored models.

We plan to make deeper and more precise human evaluation of current systems for further elaborations. We plan to research reasons why factored models have not demonstrated sufficient improvements in translation quality, especially for system trained on large (JRC Acquis 3.0) corpus and research possibilities to elaborate factored models.

Recent versions of SMT systems presented here are available at eksperimenti.ailab.lv/smt.

Acknowledgments

The work presented was supported by Latvian Council of Science through projects Evaluation of Statistical Machine Translation Methods for English-Latvian Translation system (2005-2008) and Application of Factored Methods in English-Latvian Statistical Machine Translation System (2009-2012). We would also like to thank reviewers for useful comments.

References

Fishel Mark, Kaalep Heiki-Jaan, Muischnek Kadri. 2007. Estonian-English Statistical Machine Translation: the First Results. Nivre J., Kaalep H., Muischnek K., Koit M. (eds.) *Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007*. Tartu, 278–283.

Koehn Philipp. 2004. Pharaoh: a beam search decoder for statistical machine translation. In: *6th Conference of the Association for Machine Translation in the Americas, AMTA*, Lecture Notes in Computer Science. Springer.

Koehn Philipp, Och Franz Josef, and Marcu Daniel. 2003. Statistical phrase based translation. *Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL).* Edmonton, Canada, pp. 48-54.

Koehn Philipp and Hieu Hoang. 2007. Factored Translation Models. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, pp. 868–876.

Och Franz Josef, Ney Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51.

Papineni Kishore, Roukos Salim, Ward Tood, Zhu Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, Pennsylvania, pp. 311-318.

Paikens Pēteris. Lexicon-Based Morphological Analysis of Latvian Language.2007. *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, Kaunas, pp. 235–240.

Skadiņa Inguna, Brālītis Edgars. 2007. Experimental Statistical Machine Translation System for Latvian. *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, Kaunas, 2007, pp. 281-286.

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

Stolcke Andreas. 2002. SRILM - an extensible language modeling toolkit, *ICSLP-2002*, 901-904.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206