Translation of Multi-word Expressions into Under-resourced Language: Case of English-Latvian Statistical Machine Translation

Inguna Skadiņa
Institute of Mathematics and Computer Science,
University of Latvia

WG3

1 Introduction

One of the most complex problems in processing of multiword expressions (MWEs) is automated translation, as it requires not only identification of MWEs, but also finding correct translation equivalents. Most of research on MWE translation has been made for widely used languages, e.g., Bouamor et al. (2012) has analysed translation of MWEs in a French-English statistical machine translation (SMT) task. Where it concerns smaller languages, Kordoni and Simova (2014) described phrasal verb translation with a help of dictionary in English-Bulgarian SMT task.

For English-Latvian machine translation, Deksne et al. (2008) propose to use special dictionary of MWEs in a rule-based machine translation. Pinnis and Skadiņš (2012) analyse terminology translation problem for narrow domain English-Latvian SMT system. They report transformation of translation model phrase tables into term-aware phrase tables as most successful approach.

The aim of this study is to assess applicability of different MWE extraction and translation approaches when translation is performed into morphology rich under-resourced language Latvian. Two series of experiments were performed: using pattern-based methods for MWE identification and using statistical methods for MWE identification. Both experiments have shown some positive results.

2 Experiments with linguistically motivated MWEs

In first series of experiments monolingual MWE candidates were identified using linguistic patterns and then aligned to extract possible translation pairs. The extracted MWE candidate pairs were then integrated into SMT system using three different approaches.

The *JRC Acquis* corpus (Stenberger et al. 2006) containing about 1.47 million unique parallel sentence pairs was used for these experiments. The development corpus (1134 sentences) and the test corpus (1599 sentences) was extracted from parallel corpus by selecting random sentences before training.

We applied *mwetoolkit*[1] (Ramisch, 2015) for MWE identification and annotation in the corpus. 210 morpho-syntactic patterns for Latvian and 57 patterns for English were created for this task. The extracted initial MWE candidate lists contained 610 thousand English candidates and 4,46 million Latvian candidates. Such a big difference can be explained by rich morphology of the Latvian language. The candidates then were filtered by calculating Dice's coefficient and keeping top 100 thousand.

In a next step the bilingual MWE dictionary was created with *MPAligner* toolkit[2] (Pinnis, 2013). *MPAligner* at first extracts all possible translations of MWEs and then selects those that are above specified threshold. Initially the toolkit extracted 230 thousand candidate pairs (including duplicates). After filtering, 41,689 pairs were kept for our experiments.

Three possible ways how to integrate MWEs into the SMT system were investigated. At first extracted MWE pairs were added to the parallel corpus. In the second experiment the extracted MWE pairs were used to build a second translation table for the SMT system. Finally, the MWE data were integrated in the SMT system by adding a new feature in a translation table. For training and translation *Moses toolkit* (Keohn et al., 2007) with default settings was used. KenLM (Heafield et al., 2013) was used to create 5-gram language model and MERT was applied for tuning (Och, 2003).

[1] http://mwetoolkit.sourceforge.net/
[2] https://github.com/pmarcis/mp-aligner

Table 1 provides summary of automatic evaluation results. The best results are achieved by adding MWEs as second translation table, while adding MWEs to parallel data gives only small improvement. Finally, introduction of new feature for MWEs did not lead to improvements. We found that the improved SMT system provides more precise translation, i.e., improves fluency and adequacy.

Method	**BLEU (no tuning)**	**BLEU (after tuning)**
Baseline	**56.00**	62.40
Baseline + MWE training data	55.98	**62.44**
Two translation tables	55.76	**62.55**
Additional feature	55.85	62.27

Table 1: Results of automatic evaluation.

Figure 1 illustrates a case where the improved SMT system translates MWE correctly, while translation of the baseline system is incorrect.

```
English: the hazard represented by biological agents
Reference: draudiem , ko rada bioloģiskie aģenti
Baseline: draudiem , ar bioloģiskiem aģentiem
Improved: draudiem , ko rada bioloģiskie aģenti
```

Figure 1. MWE translation with baseline and improved system.

3 Experiments with MWE candidates extracted using association measures

The manual analysis of the first experiment results revealed several issues of the chosen approach. At first, patterns allow to recognize only MWEs which are defined by patterns. Secondary, many short MWEs are already included in a phrase table. Thus in a second experiment we concentrated on MWEs that could be extracted using association measures.

We used *DGT-TM corpus* (Stenberger et al., 2012) containing about 1.63 million unique parallel sentence pairs for these experiments. The tuning set contained 2000 randomly selected segments and the test corpus contained 1000 randomly selected sentences that were selected before training. For collocation extraction we used *Collocate* tool (Barlow, 2004) and applied the Log Likelihood score. As the most frequent collocations were different for each language, we did not align collocations, but we treated them as single unit during the training. Different thresholds (minimal frequency and cost) were applied. Automatic evaluation results of these experiments are summarized in Table 2.

System	**Number of collocations**		**BLEU** (no tuning)	**BLEU** (after tuning)
	English	**Latvian**		
Baseline			45.83	46.35
Minimal frequency >3	1,087,932	795,063	43.87	44.78
Frequency and cost >9	1,074,112	556,695	45.05	43.00
Frequency for Latvian >4, frequency for English >9	98,843	88,943	45.36	43.72

Table 2:Results of automatic evaluation.

Similarly, to previous set of experiments the BLEU scores are close to the baseline. However, in these experiments they never exceed the baseline. We also performed manual inspection of obtained results and noticed some improvement in adequacy, as it is illustrated in *Figure 2*.

```
English: declaration by bulgaria
Reference: bulgārijas deklarācija
Baseline: deklarācija , ko bulgārija
SMT with MWEs: bulgārjas republikas deklarācija
```

Figure 2. Example of translation.

4 Conclusion

We presented two series of experiments with MWE extraction and integration in a phrase-based SMT system. In the first series of experiments the best automatic evaluation results were achieved using two phrase tables. In the second series of experiments none of them exceeded the baseline in terms of BLEU scores. In both cases manual inspection of obtained results showed improvement of fluency and adequacy. We see these results as a baseline for the next experiments where we plan to combine both approaches to improve fluency and adequacy of translations.

References

Michael Barlow. Collocate 1.0: Locating collocations and terminology. 2004.

Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. *LREC 2012, Eigth International Conference on Language Resources and Evaluation*, Istanbul, Turkey, ELRA.

Daiga Deksne, Raivis Skadins and Inguna Skadina. 2008. Dictionary of Multiword Expressions for Translation into Highly Inflected Languages. *Proceedings of the International Conference on Language Resources and Evaluation LREC 2008*, Marrakech, Morocco.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Es-timation. ACL, Sofia, Bulgaria, 4—9 August, 2013.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation, *ACL 2007, demonstration session.*

Valia Kordoni, Iliana Simova. 2014. Multiword Expressions in Machine Translation. *LREC 2014, Nineth International Conference on Language Resources and Evaluation.*

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.*

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics.*

Mārcis Pinnis, Raivis Skadiņš. 2012. MT Adaptation for Under-Resourced Domains – What Works and What Not. Human Language Technologies – The Baltic Perspective. *Proceedings of the Fifth International Conference Baltic HLT 2012*, 176-184, Tartu, Estonia.

Mārcis Pinnis. 2013. Context Independent Term Mapper for European Languages. *Recent Advances in Natural Language Processing,* Hissar, Bulgaria.

Carlos Ramisch, 2015. *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing series XIV, Springer.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, 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, Italy.

Steinberger Ralf, Andreas Eisele, Szymon Klocek, Spyridon Pilos & Patrick Schlüter (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, 21-27 May 2012.