Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-97

Multi-Word Expressions in English-Latvian SMT: Problems and Solutions

Inguna SKADIŅA[1]
University of Latvia, Institute of Mathematics and Computer Science

Abstract. Processing of multi-word expressions (MWEs) is well known 'pain in the neck' of human language technology researchers. The problem of MWE treatment affects almost any natural language processing task, including different levels of text analysis and automated translation. It is extremely complicated task for machine translation (MT), as it includes identification, alignment and translation. Many on-line machine translation systems translate MWEs as phrases, not as one complex unit. In this paper several experiments are presented where possible ways how statistical MT system could learn translations are investigated. Although there is no significant improvement achieved in automatic evaluation, manual inspection of translations revealed some improvement in fluency and adequacy of translations.

Keywords: multi-word expressions, statistical machine translation, Latvian language, alignment

1. Introduction

Processing of multi-word expressions (MWEs) is well known 'pain in the neck' [1] for human language technology researchers. The problem of MWE treatment affects almost any natural language processing task, including different levels of text analysis [2]. It is extremely complicated task for machine translation (MT), as it usually includes not only translation step, but also identification and alignment tasks. Many on-line machine translation systems translate MWEs as phrases that consist of separately translatable words, not as a complex unit. For example, an idiom *'raining cats and dogs*' is wrongly translated into Latvian as '*līst suņiem un kaķiem*', instead of '*līst kā pa Jāņiem*'. However, not only idiomatic expressions are translated incorrectly, but also well-known terms, e.g., '*part of speech*' is translated as '*daļa vārda*', instead of '*vārdšķira*'; or fixed phrases, such as '*in review*' in sentence '*the scheme in review is incompatible*' is translated inadequately as '*shēma pārskatā*' instead of '*apskatāmā shēma*'.

There have been several attempts to find the best way how to treat MWEs in the statistical machine translation (SMT). Most of them deal with widely used languages, such as English and Spanish [4], English and French [5], Chinese and English [6]. Only few studies dealing with some specific groups of MWEs (e.g. phrasal verbs, terminology) investigate automatic translation of MWEs into morphology rich free word order under-resourced language [7, 8, 9].

[1] Corresponding Author: Inguna Skadiņa, Raiņa bulv. 29, Riga, Latvia; e-mail:inguna.skadina@lumii.lv

In this paper several ways how MWEs could be obtained and how statistical MT system could learn translations of MWEs are investigated for English-Latvian language pair. Two approaches – pattern-based and statistical – are assessed. Although there is no significant improvement achieved in automatic evaluation, manual inspection of translations has demonstrated some improvement in fluency and adequacy of translations.

2. Related Work

In this work we follow definition of MWEs proposed by Baldwin and Kim [3], who defined MWEs as "lexical items that (a) can be decomposed in multiple lexemes and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". As MWEs include different types of lexical items, different approaches and methods are applied to MWE translation with the means of SMT. Among the most popular are single tokenization, adding MWE dictionaries to training data and MWE annotation in translation tables.

Most of the studies researching ways how automatically translate MWEs have been made for widely used languages. Lambert and Banchs [4] proposed a technique that allows to extract a bilingual multiword expression dictionary from a parallel corpus. Extracted MWE pairs are then detected in a training corpus and the source words and target words of each detected MWE are grouped in a unique "super-token" during training and are unjoined afterwards. The authors demonstrated that such approach allows to improve both alignment quality and translation accuracy.

Bouamor et al. [5] has analysed translation of MWEs (compounds, idiomatic expressions and collocations) in a French-English SMT task. For MWE extraction morphosyntactic patterns are applied on a POS tagged text. A vector space model is used for alignment. Authors analysed three methods, how to integrate automatically extracted bilingual MWEs in a phrase-based SMT system - retraining with MWEs as parallel corpus, MWE dictionary as the second phrase table and a specific feature in phrase table. In these experiments the best results were obtained with the retraining approach.

For domain adaptation where MWEs mainly represent terminological units several authors [6, 9] use automatically extracted bilingual domain dictionaries of multiword expressions as an additional resource during training. Three strategies which were already mentioned before - retraining SMT with MWEs as parallel sentence pairs, new feature in translation table for bilingual MWES, additional phrase table - are applied. The authors conclude that the usage of an additional feature to represent whether a bilingual phrase contains bilingual MWEs performs the best in most cases, while the other two strategies can also improve the quality of the SMT system, although not as much as the first one.

Where it concerns smaller languages Kordoni and Simova [7] have described phrasal verb translation with a help of dictionary in English-Bulgarian statistical machine translation. Both - automatic and human evaluations - showed that proposed integration strategies bring improvements in translation output.

For English-Latvian machine translation, Deksne et al. [8] propose to use special dictionary of MWEs and include MWE processing step in a rule-based machine translation. Pinnis and Skadiņš [9] analyse terminology translation problem for a

narrow domain English-Latvian SMT system. They report transformation of translation model into term-aware phrase tables as the most successful approach.

3. Pattern-Based Approach for MWE Identification

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

3.1. Data and tools

For experiments the DGT-TM corpus [10] of legal documents containing about 1.63 million unique parallel English-Latvian sentence pairs was used. Although the corpus does not contain idiomatic expressions, it contains a lot of terminological units, light verb constructions and named entities that needs to be treated as MWEs.

Before training 1000 sentences were randomly selected for tuning and another 1000 sentences were randomly selected for tests. Test and tuning data were removed from a training data before building SMT systems with Moses toolkit [11] and tuning with MERT [12]. BLEU metrics [13] was used for automatic evaluation.

3.2. Identification, extraction and alignment of multi-word expressions

To create MWE aware MT system, the MWEs needs to be integrated into a MT system. In most cases there is no bilingual MWE dictionary available. Thus at first such dictionary needs to be created.

The first step is to identify and to extract monolingual MWE candidates. The *mwetoolkit* [14] was used for MWE candidate extraction. The *mwetoolkit* allows to define morpho-syntactic patterns that are then applied for MWE candidate extraction. Due to the rich morphology of the Latvian language and to limit overgeneration more patterns were created for the Latvian language (in total 210) as for the English language (in total 57). Most of the patterns used for this task describe different noun phrases.

Extracted list of MWE candidates contains all strings of words that correspond to the patterns, thus the association measures are usually applied to extract most reliable candidates. The Dice's coefficient was used as association measure in these experiments.

Then the MPAligner tool [15] was applied for MWE alignment and creation of bilingual MWE dictionary. MPAligner aligns possible translation equivalents and assigns a reliability score to each pair. Most reliable candidate pairs that had the confidence score above 0.7 were kept for the further experiments. Two sets of MWEs were used to create bilingual MWE dictionary: (1) all extracted candidates and (2) the top 200 thousand candidates that were filtered out by calculating Dice's coefficient. From the first set 55 363 MWE pairs where kept, while from the second set only 4437 pairs were left. So small amount of MWE pairs can be explained by morphological richness of the Latvian language. The further experiments were performed with the dictionary of 55 363 MWE pairs.

3.3. Integration of MWE dictionary into SMT system

Three methods how to integrate MWE dictionary were investigated: (1) MWE candidate pairs were added to the parallel corpus and the SMT system was retrained; (2) two translation tables were used – the first translation table was extracted from parallel corpus and the second translation table was created from MWE pairs using reliability scores assigned by MPAligner; and (3) a binary feature was included in translation table to indicate presence of MWEs [16].

Table 1 provides summary of automatic evaluation results of different approaches. The automatic evaluation results are very close to the baseline and falls into confidence interval. Similarly to Bouamor [5], the most successful approach was adding MWEs to parallel data (+0.14 BLEU), two translation tables and introduction of additional feature that indicates MWE (reported as most successful by Pinnis and Skadiņš [9]) lead to smaller improvements.

Table 1. Application of different strategies for integration of bilingual MWE dictionary into SMT system.

Method	**BLEU without tuning**	**BLEU with MERT tuning**
Baseline	45.83	46.35
Baseline + MWE as training data	**45.95**	**46.49**
Two translation tables	45.76	**46.46**
Additional feature	45.53	**46.40**

As automatic evaluation results were close to the baseline, manual analysis of translations to find main differences between baseline system and the system with MWEs was performed with iBleu tool [17] for the best system (Baseline + MWE as training data). The manual inspection of translations showed that in some cases the improved SMT system provides more precise translation, i.e., improves fluency and adequacy. Figure 1 illustrates a case where the improved SMT system translates term ‘short-term toxicity test’ correctly, while translation of the baseline system is incorrect.

English: fish , short-term toxicity test on embryo and sac-fry stages
Human: īstermiņa toksicitātes tests zivīm embrija un dzeltenummaisa attīstības posmos
Baseline: īslaicīgas iedarbības toksicitātes tests zivju embriju un pieņu / ikru attīstības stadijas
MWE SMT: īstermiņa toksicitātes tests zivīm embrija un dzeltenummaisa attīstības posmos

Figure 1. Example of term translation by baseline and improved system.

However, we also found that current solution, namely creation of bilingual MWE dictionary, is limited to the linguistic patterns defined for MWE candidate identification and dictionary used for MWE alignment.

4. Application of Lexical Association Measures for MWE Identification

In the second experiment monolingual MWE candidates were extracted using lexical association measure and then filtered according to their frequency and cost of the association measure. The main hypothesis which we wanted to investigate was that such approach allows to identify frequently used MWEs which are hard to recognize with the linguistic patterns.

For MWE extraction the *Collocate* tool [18] was used. Different association measures were investigated and the *log-likehood* score was selected for this series of experiments. Different thresholds based on frequency and association measure were applied for MWE filtering. In addition, invalid phrases (e.g. phrases ending with conjunction, preposition followed by determiner, etc.) and strings that include numbers were filtered out using regular expressions.

After MWE candidate extraction the manual inspection and comparison of the most frequent MWEs in Latvian and English was performed. This manual inspection confirmed our hypothesis that there are frequently used collocations, e.g. '*saskaņā ar'* (*in accordance with*), '*attiecībā uz'* (*regarding to*), *'pamatojoties uz'* (*based on*) for which application of pattern based approach could lead to overgeneration. The comparison revealed that most frequent MWEs in English and Latvian are different. Thus it was decided not to create a bilingual MWE dictionary, but treat the MWEs as single units (tokens were concatenated with underscore) during the training of translation and language model.

Three SMT systems were built using this approach. Number of MWEs, applied filters and automatic evaluation results for each system are summarized in Table 2. Similarly to the previous series of experiments the BLEU scores are close to the baseline.

At first all MWE candidates with frequency above 3 were included into training data. The automatic evaluation with BLEU metric show +0.5 BLEU improvement by this system over the baseline. Our hypothesis was that data are noisy, therefore two additional training sets using different frequency filters were created and corresponding SMT systems were built. As Table 2 shows the best result is achieved by the smallest set of MWE candidates before MERT applied. However, after MERT application, the best result is achieved by system with largest set of MWEs.

Table 2. Results of automatic evaluation.

System	**Number of collocations**		**BLEU (without tuning)**	**BLEU (with MERT tuning)**
	English	**Latvian**		
Baseline			**45.83**	**46.35**
S1: Minimal frequency >3	1087932	795063	**45.88**	**46.86**
S2: Frequency and cost >9	1074112	556695	**46.75**	44.57
S3: Frequency for Latvian >4, freq. for English >9	98843	88943	**46.96**	45.13

Similarly to the previous set of experiments, manual inspection of obtained results was performed allowing to notice some improvements in translation adequacy.

English: legislative procedure ongoing
Human: notiek likumdošanas procedūra
Baseline: uzsākta leģislatīvā procedūra
Improved SMT: notiek likumdošanas procedūra

Figure 2. Influence of alignment and translation

Figure 2 illustrates improvement of alignment and translation due to the treatment of compound nominal 'legislative procedure' as a single unit: both translations of this term - 'likumdošanas' and 'leģislatīvā' - are correct, while correct translation of verb 'ongoing' is 'notiek', while 'uzsākta' corresponds to translation of 'is started'.

Figure 3 illustrates the case where both – baseline and MWE aware – translations are almost perfect, however different. One of differences is translation of phrasal verb 'draw up', which is translated by human and improved MT system as 'veic', while translation of the baseline system is 'izstrādā' (*develop*, also could be used in this case).

English: the agency shall draw up analytical accounts of its revenue and expenditure.
Human: aģentūra veic analītisku ieņēmumu un izdevumu uzskaiti.
Baseline: aģentūra izstrādā analītisko uzskaiti par tā ieņēmumiem un izdevumiem.
Improved SMT: aģentūra veic analītisku ieņēmumu un izdevumu uzskaiti.

Figure 3. Differences in translation between baseline and improved SMT system (automatically extracted collocation candidates are underlined)

Figure 4 illustrates the case when phrasal verb 'concludes that' is correctly translated by both systems, while collocation 'in review' is correctly translated as 'apskatāmā' by improved system, while translation 'atbalsta' (*support*) by the baseline system could be considered as partly correct. The interesting case is different correct translations of phrase 'is incompatible' ('par nesaderīgu' by baseline system; 'nav saderīga' by improved system) which demonstrates influence of collocation concatenation during training.

English: the commission concludes that the scheme in review is incompatible with the common market.
Human: tādēļ komisija secina, ka apskatāmā nodokļu shēma nav saderīga ar kopējo tirgu .
Baseline: komisija secina, ka šī atbalsta shēma ir uzskatāma par nesaderīgu ar kopējo tirgu.
Improved SMT: komisija secina, ka apskatāmā shēma nav saderīga ar kopējo tirgu .

Figure 4. Differences in translation between baseline and improved SMT system (automatically extracted collocation candidates are underlined)

Finally, output of different MT systems illustrating translation of MWE 'set out' is shown in Figure 5

English: are aggressive as set out in articles 8 and 9.
Human: tā ir agresīva, kā izklāstīts 8. un 9. pantā.
Baseline: ir agresīvi, kā noteikts 8. un 9. pantu.
S1: ir agresīvas, kā izklāstīts 8 . un 9. pantu.
S2: agresīva, kā norādīts 8. un 9. pantā.
S3 : ir agresīvas, kā norādīts 8 . un 9 . pantā .

Figure 5. Comparison of output from different MWE aware SMT systems

5. Conclusion

In this paper we presented two series of experiments with MWE extraction and integration in a phrase-based SMT system. In the first series of experiments we extracted MWEs using patterns and automatically created a bilingual MWE dictionary, which was then integrated into SMT system in three different ways. The automatic evaluation results are very close to the baseline. The most successful approach was adding MWEs to parallel data, second translation table and introduction of new feature for MWEs lead to smaller improvements. Manual inspection of obtained results showed some improvement in fluency and adequacy of obtained translations.

In the second series of experiments we used statistical association measures to extract MWE candidates that were then integrated into SMT systems. This approach achieved +0.5 BLEU improvement over the baseline for the best system. Similarly to the first series of experiments increase of adequacy of translations has been noticed during manual inspection of obtained results.

We see obtained results as a baseline for the next experiments where we plan to combine both approaches to improve fluency and adequacy of translations.

Acknowledgements

The research was supported by Grant 271/2012 from the Latvian Council of Science. This work has been supported by the IC1207 COST Action PARSEME as part of its scientific program. I also would like to thank Mārcis Pinnis, Matīss Rikters and Raivis Skadiņš for their valuable comments and support.

References

[1] I.A. Sag, T. Baldwin, F. Bond, A. Copestake, D. Flickinger. *Multiword expressions: a pain in the neck for nlp.* Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, CICLing'02, Springer-Verlag, 2002.

[2] A. Savary, M. Sailer, Y. Parmentier, M. Rosner, V. Rosén, A. Przepiórkowski, C. Krstev, V. Vincze, B. Wójtowicz, G. Smørdal Losnegaard, C. Parra, E. J. Waszczuk, M. Constant, P. Osenova, F. Sangati. *PARSEME – PARSing and Multiword Expressions within a European multilingual network.* Proceeding of the 7th Language & Technology Conference (LTC 2015), 2015.

[3] T. Baldwin and S.N. Kim. *Multiword Expressions*. Handbook of Natural Language Processing, 2010.

[4] P. Lambert and R. Banchs. *Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation.* Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context., 2006.

[5] D. Bouamor, N. Semmar and P. Zweigenbaum. *Identifying bilingual multi-word expressions for statistical machine translation.* Proceedings of LREC 2012, Eigth International Conference on Language Resources and Evaluation, 2012.
[6] 1.Z. Ren, Y. Lu, Q. Liu, and Y. Huang. *Improving statistical machine translation using domain bilingual multiword expressions.* Proceedings of the Workshop on Multiword Expressions : Identification, Interpretation, Disambiguation and Applications, 2009, 47–57.
[7] V. Kordoni, I. Simova. *Multiword Expressions in Machine Translation.* LREC 2014, Nineth International Conference on Language Re-sources and Evaluation, 2014.
[8] D. Deksne, R. Skadins and I. Skadina. *Dictionary of Multiword Expressions for Translation into Highly Inflected Languages.* Proceedings of the International Conference on Language Resources and Evaluation LREC 2008, 2008.
[9] M. Pinnis, R.Skadiņš. *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, 2012, 176-184.
[10] R. Steinberger, A. Eisele, S. Klocek, S. Pilos and P.Schlüter, *DGT-TM: A freely Available Translation Memory in 22 Languages.* Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), 2012.
[11] P. Koehn, H. Hoang, A. Birch, Ch. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Ch. Moran, R. Zens, Ch. Dyer, O. Bojar, A. Constantin, E. Herbst. *Moses: Open Source Toolkit for Statistical Machine Translation*, ACL 2007, demonstration session, 2007.
[12] F.F. Och. *Minimum Error Rate Training in Statistical Machine Translation.* ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003.
[13] K. Papineni, S. Roukos, T. Ward, W. Zhu. *BLEU: a method for automatic evaluation of machine translation.* ACL-2002: 40th Annual meeting of the Association for Computa-tional Linguistics, 2002.
[14] C. Ramisch, *Multiword Expressions Acquisition: A Generic and Open Framework.* Theory and Applications of Natural Language Processing series XIV, Springer, 2015.
[15] M. Pinnis. *Context Independent Term Mapper for European Languages*. Recent Advances in Natural Language Processing, 2013.
[16] A. Bisazza, N. Ruiz, and M. Federico. *Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation*. International Workshop on Spoken Language Translation (IWSLT), 2011.
[17] N. Madnani. *iBLEU: Interactively Debug-ging & Scoring Statistical Machine Translation Systems*. Proceedings of the Fifth IEEE Interna-tional Conference on Semantic Computing, 2011.
[18] M. Barlow. Collocate 1.0: Locating collocations and terminology, 2004.