# Analysis of MultiWord Expression translation errors in Statistical Machine Translation

**Natalia Klyueva**
Charles University in Prague
Faculty of Mathematics and Physics
kljueva@ufal.mff.cuni.cz

**Jeevanthi Liyanapathirana**
Copenhagen Business School
Denmark
jl.ibc@cbs.dk

## Abstract

In this paper, we analyse the usage of multiword expressions (MWE) in Statistical Machine Translation (SMT). We exploit the Moses SMT toolkit to train models for French-English and Czech-Russian language pairs. For each language pair, two models were built: a baseline model without additional MWE data and the model enhanced with information on MWE. For the French-English pair, we tried three methods of introducing the MWE data. For Czech-Russian pair, we used just one method – adding automatically extracted data as a parallel corpus.

## 1 Introduction

In this paper, we exploit a statistical machine translation (SMT) system, Moses, training it for two language pairs to explore how it cope with multiword expression translation in different languages. We will experiment with Czech-Russian, English-French language pairs to make sure that our conclusions are as language-independent as possible.

The problem of MWE in the area of SMT is a well-studied topic, next, we will name a few works that are most relevant to our work.

(Bouamor et al., 2012) described the way to extract an MWE bilingual lexicon lexicon from a parallel corpus and integrated this resource into an SMT system.

In the paper (Ghoneim and Diab, 2013) authors divided MWEs into several groups according to their parts of speech. They adopted the approaches to integrating MWE into SMT as described in (Carpuat and Diab, 2010): static (MWE on the source side are grouped with underscores) and dynamic (including MWE information straight into phrase tables) integration.

In our work, we will use three simpler methods of integrating MWE.

The paper is structured as follows. After the introduction, in Section 2, we present the notion and a basic classification of MWE. Next, we briefly describe the SMT system we are working with - Moses (Section 3). In Section 4 we present three methods to integrate MWE into SMT pipeline and test them for French-English language pair. In Section 5 we applied the most successful method from the previous experiment for Czech-Russian SMT, but MWE data we use here are sufficiently larger than for the previous experiment. Finally, we conclude in Section 6.

## 2 MultiWord Expressions

MWEs present a sequence of words with non-compositional meaning, they differ from language to language and are highly idiosyncratic. Even for the related languages we can not be sure if the structure of MWE is similar or not to say nothing about typologically different languages.

We can distinguish several types of the multiword expressions based on their part of speech and function in a sentence: noun multiword expressions, auxiliary multiword expressions, light verbs, idioms.

- Noun multiword expressions Multi-word expressions in our test sets are mainly named entities (NE) or belong to domain specific terminology (e.g. English-French : *military coup* – 'coup d'etat'. They generally contain a noun and some other part of speech. Those terms and NEs get translated properly if they were seen in the training data.

- Auxiliary multiword expressions present mainly multiword prepositions (e.g. English-French *with regard to* – 'en ce qui concerne' and SMT also does not have a problem to handle them properly because their co-

occurrence in the data is quite frequent and parts of an expression are not separated by other words.

- Light verb constructions (LVC) are generally formed by a verb and a noun where a verb does not bare its initial meaning, so that the whole construction takes the semantics of the noun. Some multiword verbs have identical component words in the languages (Czech: *hrát úlohu*, Russian: *igrat' rol'* – 'to play role', and some not (Czech: *dát smysl* – 'give sense' vs. Russian: *imet' smysl* – 'have sense'. Generally, multiword expressions are translated properly within SMT when an LVC presents an n-gram, but when a verb is separated from a noun, this LVC is often mistranslated.

- Idioms are MWEs that can include words of any part of speech and they generally bear a meaning that has very little to do with any component of MWE. Idiomatic constructions often present a challenge to MT systems because they might be equal in the languages (contain the same words), but that is not always the case. For example, the English idiom : *kick the bucket* will be translated into French as *casser sa pipe* (which is the literal meaning) in systems like Google Translate, whereas the real meaning or translation should be "mourir" , which means "to die" in English .

Multiword Expressions have a better chance to be handled properly within SMT than within Rule-Based MT if no explicit modeling of MWE was integrated into systems. If some MWE is frequently used in the training data or it is lexically fixed, it is more likely to be translated correctly.

## 3 SMT Moses

In our experiments, we exploited the toolkit Moses (Koehn et al., 2007), an open-source implementation of a phrase-based statistical translation system. **The Moses toolkit**[1] relies on and also includes several components for data preprocessing and MT evaluation, like GIZA++[2] involved in finding word alignment, the SRI Language Modeling or SRILM Toolkit,[3] implementation of model optimization (Minimum Error Rate Training, MERT) on a given development set of sentences.

## 4 English-French SMT

This section will describe the experiments we conducted in translating multiword expressions from French to English. The subsections will explain the process of extracting multiword expressions, the word alignment procedure, and the integration of the extracted information for the statistical machine translation system.

### 4.1 Multiword Expression extraction

The first step of our experiments was to extract monolingual multiword expressions from a corpus. Choosing the proper multiword expressions was quite tricky, depending on the available resources.

We used a method of extracting multiword expressions using a linguistic rule based approach. We determined some of the most common types of linguistic rules which would effectively constitute in a multiword expression (e.g. Noun-Adj, Adj-Noun, Noun-Noun). Altogether, we defined 10 rules. Once the rules are determined, we use these linguistic rules to extract the potential multiword expression from the corpora.

Once the potential candidates for multiword expressions are extracted, the most frequent candidates in the extracted set were considered potential candidates to be used in training the machine translation system. In order to remove the irrelevant candidates in the process, we conducted a simple approach : if an MWE is included inside another, having the same frequency, we remove the one smaller in size. If not, we keep both.

We conducted this experiment to extract the multiword expression candidates in the French side of the corpus.

### 4.2 Word level alignment

Once the potential MWE s are extracted, the next step is to find the potential translations in English for them. For this purpose, we used the GIZA++ alignment toolkit. A parallel corpus (which included the MWEs we extracted) was trained, and

[1] http://www.statmt.org/moses/
[2] http://www.fjoch.com/GIZA++.html
[3] http://www.speech.sri.com/projects/srilm/

the alignments for the extracted MWEs were extracted out of the alignment output.

This way, the parallel MWE pairs were extracted out of the corpus. The next step was to incorporate that knowledge into a machine translation system.

### 4.3 Integrating information into Moses system

In order to integrate the above mentioned MWE pairs to the system, we conducted three different approaches.

#### 4.3.1 Adding MWE pairs into training data

The first approach was based on simply including the extracted MWE pairs to the SMT system. This way, the extracted MWEs were considered as more training data.

#### 4.3.2 Adding MWE pairs into the phrase table

In this approach, we made use of the phrase table which is created in the Moses SMT system. We inserted the extracted MWE pairs as phrase pairs in the lexical table which is generated while training the MT system. The probability for the lexical phrase pair (which is, here, a MWE pair) is set to 1.

#### 4.3.3 Integrating features into Moses decoder

In the third approach, we inserted a simple feature to the Moses feature file, and used it for the MERT training. This feature simply mentions whether the phrase pair in concern is a multiword expression or not.

### 4.4 Experiments

As mentioned earlier, we use Moses as our statistical machine translation system. In order to extract the linguistics features, we used Stanford parser, and the TreeTagger toolkit. Plus, to generate the alignment model (to extract the MWE pairs), we used GIZA++ toolkit.

To conduct this experiment, we extracted 50 potential MWE candidates. Then, we conduct the above mentioned approaches for English to French data sets. We consider the Europarl parallel corpus for French to English for this purpose.

Table 1 shows the dataset we used for training the statistical machine translation system.

Table 2 below shows the BLEU scores we got for a test set of 10000 sentences , which include the MWEs we extracted. The baseline approach depicts the normal BLEU score we get for the parallel corpus, and the next three lines demonstrate the BLEU score we obtained using each of the approaches mentioned in section 4.3.

| | French | English |
|---|---|---|
| **Sentences** | 32000 | 33000 |
| **Words** | 120000 | 150000 |

Table 1: Europarl corpus : French to English . The statistics show the number of words and sentences in the corpus in each side

| **Method** | BLEU |
|---|---|
| Baseline System | 21.67 |
| Adding MWE pairs into training data | 21.88 |
| Adding MWE pairs into the phrase table | 21.68 |
| Integrating features into Moses decoder | 19.2 |

Table 2: BLEU Scores for each approach

Table 2 shows that two approaches we conducted slightly increase the BLEU score. However, the approach of integrating features into Moses decoder degrades the performance. This gives a positive potential to the fact that incorporating MWE s to the SMT system in different manners can effectively increase the BLEU score.

It should be also mentioned that it is quite difficult to evaluate the efficiency of our proposed approaches by incorporating a significantly small number of MWEs, e.g. 50. Also, the alignment models can also give some amount of noise in their alignments, so the extracted MWE pairs are not 100% accurate. These reasons might have contributed to the fact of having a relatively low increase in BLEU score.

## 5 Czech-Russian SMT

Our second experiment with Czech-Russian language pair includes only one method of introducing MWE. We will exploit the simplest method described in Section 4.3.1 - adding MWE lexicon as a parallel corpus and retraining the system on the enhanced data.

### 5.1 Baseline SMT

We trained a baseline system on data coming from news domain[4] and from the domain of fiction[5]. Europarl corpus does not include version in Russian, so we can not add parallel data from this resource. The data for training a language model for the target language - Russian - were compiled from various online resources, see (Bílek, 2014) for details. Table 3 presents the statistics of the training data.

| corpus | sentences |
|---|---|
| **news** | 93432 |
| **fiction** | 148810 |
| total | 242242 |

Table 3: Size of training data

### 5.2 MWE from wikipedia headlines

We used a list of names and phrases from Wikipedia headlines for the pair Czech-Russian. The headlines were automatically extracted from the wikipedia dumps in XML (`https://dumps.wikimedia.org/`). The headlines were not necessarily multiword expressions, but for the sake of our experiment, we extracted only MWEs. Following is the example of several entities from the list:

Dějiny Říma История Рима
Higašijama Император Хигасияма
Štika obecná Щука
Vánoční stromek Новогодняя ёлка
Křižák obecný Крестовик обыкновенный
Gaius Licinius Macer Гай Лициний Макр
Ryzec pravý Рыжик настоящий
Mealyho automat Автомат Мили
Kočka bažinná Камышовый кот
Švýcarská hymna Гимн Швейцарии
Zápach z úst Галитоз
Leon V. Arménský Лев V Армянин
Dopravna Раздельный пункт
Krevní plazma Плазма крови

Figure 1: Czech-Russian MWEs from Wikipedia headlines

The automatically extracted data are not very clean; there are no light verb constructions and hardly any idioms, mostly they are Named Entities. Total number of MWE pairs extracted from the Wikipedia is 87354.

[4] `http://ufal.mff.cuni.cz/umc/cer/`

[5] Czech-Russian side of Intercorp, `https://ucnk.ff.cuni.cz/intercorp/`, not an open-source

### 5.3 Results of the experiment

Using the factored configuration of Moses, we ran two experiments:

- the baseline with models trained on data without the Wikipedia headlines
- model trained on data including the headlines

Table 4 demonstrates the difference in performance between the baseline system and the system trained on data with additional MWE resource. In addition to BLEU, we calculated the number of out-of-vocabulary (OOV) words - searching for Latin characters in the translation output (Czech words left untranslated by Moses).

| | BLEU | OOV |
|---|---|---|
| **Baseline system** | 17,23% | 1216 |
| **With MWE** | 17,90% | 1011 |

Table 4: BLEU score and OOV rate for SMT trained on data with and without MWE resource

The BLEU score in the second experiment was slightly better than in the baseline, but, evidently, this improvement is insignificant. The number of out-of-vocabulary words decreased by 205 individual tokens. This may be attributed to the positive effect of adding new data.

### 5.4 Examples of improved MWE

We examined the list of OOV words in the output from the two experiments. Among those 205 words that were recognized and translated in the second experiment, there were MWEs from the added resource, such as *Carlo Ancelotti*, *Amschel Rothschild*, *alt soprán* etc. The following MWEs were not translated or mistranslated in baseline, but were translated correctly according to the added data in the improved setup: *Higgsův boson* – 'Bozon Higgsa' (Higgs boson), *Velký hadronový urychlovač* – 'Bol'shoy adronniy collajder' (LHC), *Praní špinavých peněz* – 'Otmivanie deneg' (money laundering) etc.

## 6 Conclusion

In this work, we presented experiments with integrating MWEs into SMT for the two language pairs - French-English and Czech-Russian. We tested three methods of including MWE information into SMT. It turned out that for the concrete language pair (French-English) and the concrete

MWE list the method of introducing MWE as additional parallel data scored better than other methods. We adopted this method for the pair Czech-Russian and added an automatically extracted resource. In both cases, the increase in BLEU score was very little, but this often happens when improving concerns one concrete linguistic issue.

## Acknowledgments

This work has been using language resources distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

## References

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In *LREC*, pages 674–679.

Karel Bílek. 2014. A Comparison of Methods of Czech-to-Russian Machine Translation. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Marine Carpuat and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 242–245. Association for Computational Linguistics.

Mahmoud Ghoneim and Mona T. Diab. 2013. Multiword Expressions in the Context of Statistical Machine Translation. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1181–1187.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.