CONSISTENCY AND COMPLETENESS OF MULTIWORD EXPRESSIONS DURING TRANSLATION

Katerina Zdravkova[1]*, Aleksandar Petrovski*[2]*, Tomaž Erjavec*[3]
[1]Faculty of Computer Science and Engineering, University of Skopje, Macedonia
[2]Faculty of Informatics, Slavic University, Sveti Nikole, Macedonia
[3]Department of Intelligent Systems, Jožef Stefan Institute, Slovenia
e-mails: katerina.zdravkova@finki.ukim.mk; a.petrovski.sise@gmail.com; tomaz.erjavec@ijs.si

ABSTRACT

One of the crucial challenges of statistical machine translation is the lexical consistency of manually translated words and multiword expressions (MWEs) with multiple occurrences in the source language. In this paper, we present the degree of translation inconsistency and we introduce the index of translation completeness of fixed MWEs. The research was based on the recently developed system that intends to extract the entire candidate MWEs from Orwell's 1984 parallel corpora and to predict their translations between English, Macedonian, and Slovene.

1 INTRODUCTION

Since the early 1990s, traditional rule-based machine translation (MT) has been enhanced and replaced by the statistical MT [1]. The efficiency of, at that time rather revolutionary approach has been proved, and many tools and parallel corpora (many of them collected in http://www.statmt.org/) have been developed to enable an effective translation of written texts, no matter the languages involved it the process.

Statistical MT of MWEs can be successfully performed using non-hierarchical phrase-based SMT, which exploits only the continuous phrases [2]. In an absence of relevant parallel MWE corpora between English, Macedonian and Slovene, we decided to create an own system, which extracts all the candidate MWEs from sentence aligned corpora and then predicts their translations. The proposed system consisted of four complementary phases:

- *extraction* of all candidate continuous sequences of words that appear in each language at least twice,
- *syntactical filtering* of obtained candidates, using a predefined set of eligible syntactic expressions,
- prediction of *potential translation equivalents* from corresponding pairs of aligned sentences where MWEs appear, and
- cross-evaluation of candidate translations, interchanging the source and the target language.

In this paper we evaluate the efficiency of the system and try to determine the key causes of wrong expressions and inaccurate translation. The structure of the paper is the following: The analysis and research of the document-level consistency is presented in the second section. The typical examples of translation inconsistency and incompleteness are illustrated in the third section. The consistency index and the degree of translation completeness are introduced in the fourth section. Following the same section, the lexical consistency and the completeness of English to Macedonian and English to Slovene translation of Orwell's 1984 are calculated. The paper concludes with the ideas that might improve the quality of document-based statistical MT.

2 ANALYSIS OF PREVIOUS RESEARCH

Multiword expressions, which are defined as combinations or strings of words without a unique syntactic or semantic property, are among the crucial obstacles of machine translation [4]. Many lexicons include significant amounts of MWEs, including lists of phrasal verbs (*think of / мисли на / misliti na*), nominal multiwords (*dark-haired girl / темнокоса девојка / temnolaso dekle*), pronouns (*almost nothing / скоро ништо / skoraj ničesar*), adverbs (*during his childhood / за време на неговото детство / med njegovim otroštvom*) and other phrases.

Multiword expressions are extremely frequent. Jackendoff estimated that they appeared in the speaker's lexicon with a comparable frequency with the simple words [5]. In addition, they are very heterogeneous [4]. Even when MWEs are restricted to fixed strings, their treatment in MT is one of the most challenging NLP tasks [6]. In order to become useful for further MT research, MWEs should be extracted out of a parallel and aligned corpus. Their automatic identification and acquisition have been exhaustively researched by many authors. Most of the proposed techniques identify MWEs using different statistical measures, for example, the mutual information, permutation frequency, and Pearson's chi-square [4]. Statistical measures can be extended with various additional information concerning the word alignment [4, 8].

The detection of missing lexical entries for MWEs based on error mining methods and maximum entropy model was recommended by Zhang et al [7]. Apart from proposing their approach, they list the ten most frequent and least frequent MWEs using Google search engine. Statistical properties were also efficient during the extraction of non-compositional compounds [4].

Once extracted from parallel aligned corpora, MWEs can undergo through the translation process. The typical recent SMT tools, such as Moses (http://www.statmt.org/moses/) are phrase-based models [8]. Moses used the Bayes rule to initially calculate the probability for translating a foreign sentence into English. The same approach was very soon implemented for many other languages, including Macedonian [9]. Numerous experiments have shown that Moses performs much better that word-based models, and more significantly, it appeared that the use of syntax doesn't lead to better performance.

Caseli and her collaborators combined phrase-based and word-based model, creating the first alignment based MWE extraction method [4]. For each language, they created an output of the aligner or the tagger along with the target words that were aligned to them. Inspired by this project, we suggest a new, slightly less rigorous approach [10]. Its intention is to identify all the MWEs appearing in the multilingual sentence aligned Multext-East corpus [11]. The effectiveness of the system will be illustrated with the examples of aligned English to Macedonian and Slovene translation of 968 multiword expressions existing in the English original of Orwell's novel 1984.

The extraction process in these two projects passed through a pre-processing phase, which produced parallel, sentence aligned and PoS tagged multilingual corpora [4, 12]. Furthermore, some MWEs were word-aligned to be associated with semantics [4]. False positive examples were syntactically filtered using patterns or syntactic constraints. Many inadequate candidates were further eliminated using the cross-evaluation mentioned in the introduction of this paper [11]. In our system, we eliminated the syntactically ineligible MWEs using different patterns [10]. In many occasions, the filtering process using the cross-evaluation offered a very good result.

The implementation of mutual cross-evaluation among English, Macedonian and Slovene revealed that in many occasions:

- manual translator of Orwell's 1984 was either inconsistent or had "an artistic freedom",
- inflectional paradigms, which are richer in the Slavic languages can influence the translation,
- the context in which the same target MWE appeared can also influence its translation.

As a result, many MWEs were translated with an MWE that is shorter than the real target, up to the extreme not to be translated at all. Partial incompleteness or the entire absence of the target MWEs were the main drawback of our system.

3 EXAMPLES OF INACCURATE TRANSLATIONS

The English version of Orwell's 1984, which serves as a base for the parallel corpus contains 6701 sentences and 104302 words. Macedonian translation consisted of 6712 sentences with 98846 words. The amount of continuous word sequences with multiple occurrences in both languages exceeded 15000. The English candidate MWEs were matched with the translated Macedonian MWEs. As a result, the extraction phase ended up with 968 English MWEs, the majority of which produced a target Macedonian MWE [10].

Due to the abundance of Slovene nominal inflections, the amount of omitted Slovene translations was higher. Here are some typical examples that explain the deficiency of the statistical machine translation without a morphological extension we created. The cross-evaluation, which reverted the source and the target language revealed that in some occasions two different English MWEs were translated with the same MWE. This can be treated as a revert inconsistency. Table 1. presents several cases of inconsistencies across three languages. The omitted parts of the most acceptable translations in the corresponding language are presented in the parentheses. The MWEs in bold are the starting points for the translation.

Language	English	Macedonian	Slovene
Multiword expression	**the seconds were ticking by**	секундите минуваа (отчукувајќи)	sekunde so tiktakale mimo
Mac 1: секундите минуваа отчукувајќи ...			
Mac 2: секундите минуваа бескрајно долги ...			
Multiword expression	**almost on a level with**	речиси на исто (со)	no translation
Mac 1: ... речиси на исто ниво со ...			
Mac 2: ... речиси на исто рамниште со ...			
Slov 1: ... skoraj na ravni z ...			
Slov 2: ... skoraj v isti višini z ...			
Multiword expression	the first thing	(прва работа што мора) да ја сфатиш е	**prva stvar ki jo moraš ...**
Eng 1: the first thing for you to understand ...			
Eng 2: the first thing you must realize ...			

Table 1: *Incompleteness due to lexical inconsistency*

Slavic incomplete or missing translations of English due to inflections are presented in the Table 2. The parentheses in the Macedonian example are given to describe the MWE.

Language	English	Macedonian	Slovene
Multiword expression	**smell of her hair**	(мирисот) на нејзината коса	vonj njenih las
Mac 1: ... (пријатниот) мирис на нејзината коса			
Mac 2: мирисот на нејзината коса			
Multiword expression	**ideologically neutral**	идеолошки неутрален	ideološko nevtralen/na
Slov 1: ... (nobena beseda ... ni bila) ideološko nevtralna			
Slov 2: ... (predmet govora ni bil) ideološko nevtralen			
Multiword expression	**against us**	против нас	po robu (proti nam)
Slov 1: ... (nikdar ne) postavi po robu			
Slov 2: ... (in se nam) postavila po robu			

Table 2: *Incompleteness due to inflections*

In many occasions, the context was the crucial reason of the pruned or missing translations. It is worth mentioning that there were several examples with shorter multiword expression even in the English original, as presented in the Table 3. They are a result of the reverse order of source and target extraction due to cross-evaluation.

In order to distinguish the importance of the context, the original source contexts in parallel with the target ones is also specified. Although the absence of Slovene translation in the last example is mainly due to the inflections, it is presented here, because the different grammatical cases (genitive in *prepisovalne ekipe* and locative in *prepisovalni ekipi*) themselves also arise from the context.

<table>
<tr><td>Language</td><td>English</td><td>Macedonian</td><td>Slovene</td></tr>
<tr><td>Multiword expression</td><td><b>for more than half an hour</b></td><td>(за) повеќе од половина час</td><td>za več kot pol ure</td></tr>
<tr><td colspan="4">Eng 1: ... and never for more than half an hour at a time
Mac 1: ... и никогаш повеќе од половина час</td></tr>
<tr><td colspan="4">Eng 2: ... to turn off the telescreen for more than half an hour
Mac 2: ... да го држат исклучен телекранот повеќе од половина час</td></tr>
<tr><td>Multiword expression</td><td><b>definitive edition</b></td><td>дефинитивното издание</td><td>no translation (dokončna izdaja)</td></tr>
<tr><td colspan="4">Eng 1: ... (the eleventh edition is the) definitive edition ...
Slov 1: ... (enajsta izdaja je) dokončna</td></tr>
<tr><td colspan="4">Eng 2: ... (we were producing a) definitive edition ...
Slov 2: ... (pripravljali smo) končno izdajo</td></tr>
<tr><td>Multiword expression</td><td>(in) the rewrite squad</td><td><b>во одделот за препишување</b></td><td>no translation (prepisovalna ekipa)</td></tr>
<tr><td colspan="4">Mac 1: ... до завршните работи во одделот за препишување
Eng 1: ... (down to final touching-up by) the rewrite squad
Slov 1: ... (pa do končne obdelave) prepisovalne ekipe</td></tr>
<tr><td colspan="4">Mac 2: ... никогаш не работев во одделот за препишување
Slov 2: ... (nikdar nisem bila v) prepisovalni ekipi
Eng 2: (i was never in) the rewrite squad</td></tr>
</table>

Table 3: *Incompleteness due to the context*

In the next section, the inconsistency index, which was proposed by Itagaki et al. [3] will be introduced and calculated for those MWEs that existed in all the three languages. In parallel with the inconsistency index, we also propose the degree of incompleteness, which is the direct consequence of the inconsistent translation.

4 CONSISTENCY AND COMPLETENESS OF MULTIWORD EXPRESSIONS

Human translators usually work with very large translation units. Without a large list of own translated phrases, or an automated translation tool, the possibility to inconsistently generate the translation is high.

In 2007, Itagaki, Aikawa and He decided to devise an index to assess the terminology translation consistently [4]. They discovered that the estimation could be effectively done using the Herfindahl-Hirschman Index (*HHI*), which was previously used to measure the market concentration. The index is calculated as:

$$HHI = \sum_{i=1}^{n} S_i^2$$

where *S* is the ratio of each translation (*i*) to the total number of translations (*n*) within a product. To simplify the definition, whenever one word is translated with *n* different words, each one with a frequency S_i, in such case, the consistency of the translation is the sum of squared frequencies within the document [4, 13].

HHI is applicable to multiword expressions, replacing the single words to lexical units. For example, the multiword expression *the dark-haired girl*, which appears twice in the source language was uniquely translated to Macedonian (*темнокосата девојка*) and Slovene (*temnolaso dekle*), so its consistency is 1. The English MWE *during his childhood* also appeared twice, with two Macedonian translations: *за време на неговото детство* and *во текот на неговото детство*, and a unique Slovene translation *med njegovim otroštvom*. The consistency of the Macedonian translation is $0.5^2 + 0.5^2 = 0.25$., while the Slovene consistency is 1. The translation of the phrases that appear in the English original more than once was always perfectly consistent (*yes said Winston* / *да рече Винстон* / *da je rekel Winston*; *how many fingers Winston* / *колку прсти Винстоне* / *koliko prstov Winston*).

By adopting the consistency index of lexems to lexical units, i.e. to multiword expressions, we also propose to calculate their relative consistency as a ratio between *HHI* and the cardinality of the set of all multiword expressions appearing in the target corpus at least twice:

$$RC = \frac{HHI}{|MWE|}$$

In the Macedonian version, 48 out of 968 English MWEs had no translation due to inconsistent translation, or the translation consisted of only one word, which was excluded from the MWE corpus. Further 127 were partially inconsistent, thus the consistency index was 836.75, or relatively 86.44%.

The translation to Slovene had a relative consistency of 80.40%, due to 162 MWEs without a translation, and 91 with partial inconsistency, or a total consistency index of 778.25.

The examples presented in the tables above indicate that the key outcome of human inconsistency used as a source in the statistical machine translation systems is the incompleteness of generated target expressions. To measure the degree of incompleteness of MWE translations, we propose the index of completeness *DG* of a single MWE calculated as:

$$DG = \frac{length(generated\ MWE)}{length(complete\ MWE)}$$

For example, the English expression *almost on a level with* is translated with *речиси на исто* instead of *речиси на исто ниво со*. Its completeness is 0.6. But, whenever the short MWE is not a subset of the complete MWE, such as the translation of *against us* to Slovene, which was *po robu* (see Table 2.), in such case the completeness is 0. This estimation can be done after a manual inspection of the translated MWEs.

We also define a combined completeness *CC* of all *m* MWEs extracted from the source corpus as:

$$CC = \frac{1}{m}\sum_{j=1}^{m} S_j^2 DG_j^2$$

The combined completeness of the MWE *almost on a level with* is 0.25 * 0.36 = 0.09. The combined translation of a consistent translations is 1.

Due to the higher consistency, Macedonian translations had a higher combined completeness of 83.42%, compared to 74.97% for Slovene translations.

5 CONCLUSIONS AND FURTHER RESEARCH

The proper identification of MWEs that appear multiple times in the parallel sentence aligned corpora offers an opportunity to improve the quality of statistical machine translation.

In the research presented in this paper, we tried to define a framework for effective treatment of lexical units across languages. It passed through four complementary phases presented in the introduction of the paper. In order to measure the correctness of MWE extraction process, as well as the translation prediction, we measured the consistency and completeness of generated translations of MWEs existing in the small parallel Multext-East corpus. We intend to implement the same approach to measure the same parameters in the raw material obtained when Moses SMT toolkit, which was implemented over SETimes corpus [9].

In order to improve the quality of the created translation system, we will first incorporate MWE lexical entries, which are currently created for the Macedonian language [14]. They will consist of fixed MWE lexical entries used in the current stage of the system, and extended with semi-fixed and flexible MWEs. We will also intend to study the lexical cohesion, and extend the document-level translation to a larger collection. Inspired by Ben et al., the final goal in this direction will be the integration of the model into a hierarchical phrase-based SMT system [15]. Most current SMT systems translate sentences individually, assuming that the sentences in a text are independent [16]. A further extension of the system is directed towards the extraction of the common knowledge about multiword expressions out of a continuous context and its incorporation into a translation system capable to competently deal with them.

References

[1] P. F. Brown et al. A statistical approach to machine translation, *Computational linguistics 16.2*, pp. 79-85, 1990.

[2] M. Galley, C. D. Manning. Accurate non-hierarchical phrase-based translation, *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 966-974, 2010.

[3] M. Itagaki, T. Aikawa, X. He. Automatic Validation of Terminology Translation Consistency with Statistical Method, *Proceedings of MT summit XI, 269-274*, pp. 269-274, 2007.

[4] H. Caseli, C. Ramish, M. Nunes, A. Villavicencio. Alignment based extraction of multiword expressions, *Language Resources & Evaluation 44*, pp. 59-77, 2010.

[5] R. Jackendoff. 'Twistin' the night away, *Language 73*, pp. 534-559, 1997.

[6] J. Tiedemann. To cache or not to cache, Experiments with Adaptive Models in Statistical Machine Translation, *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pp. 189-194, 2010.

[7] Y. Zhang, V. Kordoni, A. Villavicencio, M. Idiart. Automated Multiword Expression Prediction for Grammar Engineering, *Proceedings of the 5th Workshop on Important Unresolved Matters*, pp. 44-52, 2006.

[8] P. Koehn, F. J. Och, D. Marcu. Statistical phrase-based models, *Proceedings of NAACL 2003*, pp. 48-54, 2003

[9] M. Stolikj, K. Zdravkova. Resources for Machine Translation of the Macedonian Language, *online Proceedings of ICT Innovations 2009*.

[10] K. Zdravkova, A. Petrovski. System for extraction of potential multi-word expressions and prediction of their translations from a multilingual corpus, *PARSEME 2nd general meeting*, poster 43, 2014.

[11] T. Erjavec. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages, *Language Resources and Evaluation*, Vol. 46 / 1, pp. 131-142, 2012.

[12] V. Vojnovski, S. Džeroski, T. Erjavec. Learning PoS tagging from a tagged Macedonian text corpus, Proceedings of SiKDD 2005, Ljubljana, Slovenia, pp. 199-202, 2005.

[13] L. Guillou. Analysing Lexical Consistency in Translation, *Proceedings of DiscoMT*, Sofia, Bulgaria, pp. 10-18, 2013.

[14] A. Petrovski, K. Zdravkova. How to create a MWE lexical entry?. *PARSEME 3rd general meeting*, poster 8, group B, 2014.

[15] G. Ben, D. Xiong, Z. Teng, Y. Lu, Q. Liu. Bilingual Lexical Cohesion Trigger Model for Document-Level Machine Translation, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 382-386, 2013.

[16] S. Stymne, J. Tiedemann, C. Hardmeier, J. Nivre. Statistical Machine Translation with Readability Constraints, *Proceedings of the 19th Nordic Conference of Computational Linguistics*, pp. 375-474, 2013.