Evaluation of the system for extraction of multi-word expressions and prediction of their translations

WG2 and WG3

Katerina Zdravkova
University Sts Cyril and Methodius
Faculty of Computer Science and Engineering
Skopje, Macedonia
katerina.zdravkova@finki.ukim.mk

Aleksandar Petrovski
International Slavic University
Faculty of Informatics
Sveti Nikole, Macedonia
a.petrovski.sise@gmail.com

Abstract

During PARSEME meeting in Athens, we proposed an unsupervised learning system intended to extract all the candidate multiword expressions from sentence aligned parallel corpora and to predict their translations. The system was created using the parallel corpora of Orwell's 1984, which is a part of Multext-East project. In this paper we evaluate the efficiency of the system and try to determine the major drawbacks leading to wrong expressions and inaccurate translation. They will be illustrated with the examples of a bilingual translation of Orwell's 1984.

1 Introduction

In the recent years, one of the most explored NLP challenges were multiword expressions (MWEs), mainly because, as Caseli et al. (2009) claim "the methods and techniques developed for the treatment of simplex words are not necessarily suitable for them". Furthermore, their automatic prediction and extraction from corpora is far from being trivial. As a rule, MWEs models implement manually created lists of annotated candidates, such as those which were examined by de Caseli (2009) and Farahmand (2014).

The system for MWEs extraction and prediction of their translations presented earlier on (Zdravkova, 2014) is an ad hoc system that implements annotation in the post processing phase in order to eliminate those candidate MWEs which contain additional lexemes, for example the strings, "in his present position *he*", "his glass *was*", or "*of* the orators of the party".

2 Extraction and syntactical filtering

The system consisted of four complementary phases. The presumption of the extraction phase was that any string appearing in the text at least twice in the monolingual corpus is a candidate MWE. It ended up with more that 10000 candidate expressions, most of them parts of longer expressions. The elimination of sub-expressions led to almost 3500 candidate MWEs in each language, most of them useless: *тоа би ја / toa bi ja*, instead of *тоа би ја усреќило / toa bi ja usrekjilo* = that will make her happy; *рече тој со / reche toj so*, instead of *рече тој со недоверба / reche toj so nedoverba* = he said with mistrust; *зелка и на / zelka i na*, which is completely meaningless.

Syntactical filtering using monolingual annotated dictionaries restricted the candidate MWEs to a limited set of eligible expressions. For example, the implementation of rules for multiword nouns suggested by Laporte (2008) created less than 500 phrases, some of them inflections of the same phrase: *атомската бомба / atomskata bomb* = the atomic bomb, *атомски бомби / atomski bombi* = atomic bombs, *обичен човек / obichen chovek* = an ordinary man, *обичните луѓе / obichnite lugje* = the ordinary men or the ordinary people, *шаховска табла / shahovska tabla* = a chess board, *шаховската табла / shahovskata tabla* = the chess board, etc. Since the filtering phase seemed to be too restrictive, we decided to proceed towards the translation phase by skipping its results, using all candidate MWEs no matter their actual validity.

3 Translation and cross evaluation

The basic hypothesis we based the translation on is the following: If a candidate MWE exists in the source language at least twice, even within one sentence, than its translation will be paired with exactly the same amount of target MWEs existing in the aligned sentences. In such case, the translated MWE is the intersection of all the repeated expressions existing in the target language.

The cross evaluation phase matched the candidate translations from the target language with the candidate MWEs, when the target language was used as a source language for the extraction phase. It eliminated many irrelevant strings without implementing any syntactic filtering.

4 Evaluation of the results

The implementation of this approach using the English original and its Macedonian manual translation led to 968 English candidate MWEs and their translations. Such a small amount enabled profound manual evaluation of the results, which will be discussed further in more details. The extraction precision of the system was rather low, particularly many candidate MWEs ended with an excessive lexeme, such as: almost on a level *with*, *of* human life, *a* little behind, in some cases *they*, his glass *was*, etc. The amount of erroneously extended MWEs was 196, which together with the 22 meaningless strings lead to 22.52% inaccuracy.

We noticed several problems during the translation process. The first occurred due to the existence of two candidate MWEs in two mutually aligned sentences. They produced two incomplete, thus inaccurate translations. For example, the phrase "a comb and a piece of toilet paper" was translated with two of its constituents: *чешел и / cheshel i* = "comb and", and *парче тоалетна хартија / parche toaletna hartija* = "a piece of paper"; "guilty of the crimes they were charged with" as *виновни за / vinovni za* = "guilty of" and *за кои беа обвинети / za koi bea obvineti* = they were charged with, where the noun *криминал / kriminal* = crimes was omitted from both parts. Due to inconsistent translation, multiword nouns "thieves bandits" got two translations: *растурачи на дрога / rasturachi na droga* = "drug dillers" and *крадци бандити / kradci banditi* = "thieves bandits". If it can be tolerated, the translation of the phrase "their hands crossed on their knees" with *затворениците седеа / zatvorenicite sedea* = "the prisoners sat", and *неподвижно со / nepodvizhno so* = "immobile with" is completely incorrect. These examples are the extreme ones. The typical incomplete MWEs due to inconsistent translation were found in 102 cases, such as: "with the tips of his fingers", "true feelings towards big brother" or "sweet summer air".

5 Conclusions

The proposed system based on a very small parallel and sentence aligned corpus proved that the proposed approach can be useful for further extraction and translation of MWEs even without a profound syntactic analysis. In some occasions, even the erroneously extended MWEs were accurately translated. It appeared that the major misleading for the system were inconsistent human translations. In the last few years, consistent translations have been thoroughly examined, and the Herfindahl-Hirschman Index (HHI) measure was adapted by Itagaki (2007) to express the consistency index. We propose the index of completeness as its upgrading, to express the degree of correct translation of a complete MWE.

References

Caseli, H. D., Ramisch, C., Nunes, M. D. V. and Villavicencio, A., 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44 (1-2): 59-77.

Farahmand, M., & Martins, R. (2014). A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features. *EACL 2014*.

Itagaki, M., Aikawa, T., & He, X. (2007). Automatic validation of terminology translation consistency with statistical method. Proceedings of MT summit XI, 269-274.

Laporte, E., Nakamura, T., & Voyatzi, S. (2008). A French corpus annotated for multiword nouns. In Proceedings of the Language Resources and Evaluation Conference. Workshop Towards a Shared Task on Multiword Expressions: 27-30.

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