USING CROSS-LINGUAL CONTEXTS TO EXTRACT TRANSLATION EQUIVALENTS FOR MULTIWORD EXPRESSIONS FROM PARALLEL CORPORA

Shiva Taslimipoor
University of Wolverhampton
shiva.taslimi@wlv.ac.uk

Ruslan Mitkov
University of Wolverhampton
r.mitkov@wlv.ac.uk

Gloria Corpas Pastor
University of Malaga
gcorpas@uma.es

Within the Natural Language Processing and Computational Linguistic communities, multiword expressions (MWEs) are notoriously well-known to be difficult to translate (Sag et al. 2002). Word for word correspondences cannot be helpful in establishing the translation of MWEs. There is also no comprehensive dictionary featuring all MWEs as putting together all possible MWEs and their variants would not be a feasible exercise.

There is a large body of work describing different properties of various MWEs (Fazly 2007, Baldwin and Kim 2010). However the cross-lingual analysis of these expressions and automatic extraction of their translation equivalents is still an under-researched topic (Bouamor et al. 2012). Contrastive and cross-lingual studies based on parallel and comparable bilingual corpora can benefit from statistical analysis of the various categories of such expressions (Colson 2008, Corpas Pastor 2013). Based on the exploitation of parallel corpora to investigate MWEs, we have implemented a bilingual corpus-based approach to find translation equivalents for MWEs in two languages (Spanish and English in this case). It is worth pointing out that corpus-based distributional similarity has already offered promising results in the discovery of translationally equivalent words/terms in a bilingual scenario (Pekar et al. 2006). According to the distributional similarity premise, translation equivalent terms share common words in their contexts.

In this paper, we apply distributional similarity in a bilingual scenario to extract the English expressions deemed to be 'most similar' to Spanish MWEs. More specifically, we use Vector Space Models with bilingual contexts to find the similarities between Spanish and English MWEs.

As bilingual contexts, we exploit a list of all non-polysemous English nouns and their translations in Spanish from an on-line English-Spanish dictionary. For every Spanish expression, we extract its vector representing the pattern/distribution of the co-occurrences of the list of the prepared context nouns within a particular window around that expression. Once the vectors for all English expressions are retrieved, we then compute the vector similarity between the Spanish and the English expressions. The most similar sequence of words to each Spanish expression is proposed as its translation in English. Examples of automatically extracted translation equivalents using our methodology include examples like *dar la bienvenida* (a alg.) = *to welcome* sb. and *base de datos = database*. The final version of the paper will feature a section on the evaluation results of the developed methodology.

This method can be beneficial (and enhance the performance) of not only machine translation systems, but also offer new opportunities to cross-lingual studies on MWEs based on their occurrences in parallel corpora. This methodology could also assist lexicographers when deciding which MWEs should be listed in bilingual dictionaries as well as speed up the semi-automatic compilation of such dictionaries. In line with previous research on bilingual term extraction from parallel corpora (Ha et al., 2008), in this paper we also make use of the extraction of MWEs in one language to boost the extraction performance in the other.

References

BALDWIN, T. AND KIM, S. N. (2010). "Multiword expressions". *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, eds., CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

BOUAMOR, D., SEMMAR, N., AND ZWEIGENBAUM, P. (2012). "Identifying bilingual multi-word expressions for statistical machine translation". *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),* Istanbul, Turkey.

COLSON, J. P. (2008). "Cross-linguistic phraseological studies: An overview". *Phraseology: An interdisciplinary perspective*, S. Granger and F. Meunier, eds., John Benjamins Publishing Company, Amsterdam/ Philadelphia: Meunier, John Benjamins Publishing Company.

CORPAS PASTOR, GLORIA: "Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas". *In: Inés Olza / Elvira Manero (eds.): Fraseopragmática*, Berlin: Frank & Timme, 335-373.

FAZLY, A. (2007). "Automatic acquisition of lexical knowledge about multiword predicates". Ph.D. thesis, Department of Computer Science, University of Toronto, Department of Computer Science, University of Toronto.

HA, L. A., FERNANDEZ, G., MITKOV, R., AND PASTOR, G. C. (2008). "Mutual bilingual terminology extraction". *Proceedings of the International Conference on Language Resources*

and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco, http://www.lrec-conf.org/proceedings/lrec2008/summaries/463.html .

PEKAR, V., MITKOV, R., BLAGOEV, D., AND MULLONI, A. (2006). "Finding translations for low-frequency words in comparable corpora". *Machine Translation*, 20(4), 247-266.

SAG, I. A., BALDWIN, T., BOND, F., COPESTAKE, A. A., AND FLICKINGER, D. (2002). "Multiword expressions: A pain in the neck for nlp". *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing,* CICLing '02, London, UK, UK, Springer-Verlag, 1-15, <http://dl.acm.org/citation.cfm?id=647344.724004>.