WG3: Resource-poor Translation of Multiword Expressions

Shiva Taslimipoor and Ruslan Mitkov
Wolverhampton University

Rationale and Related Work

The most obvious resources such as parallel corpora and dictionaries do not cover all interpretation and translation of multiword expressions (MWE); they are scarce and in many cases not available at all. In order to benefit from the wider availability of comparable corpora, in this project we propose a novel 'semantic similarity' methodology to extract and translate multiword units from comparable corpora. This methodology does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. While in this particular study we have covered English and Spanish, the methodology is not restricted to any particular pair of languages. We focus on a particular subclass of MWEs: verb-noun expressions (collocations) such as *take advantage*, *make sense* and *prestar atención* (meaning *pay attention*).

The cross-lingual analysis of MWEs and automatic extraction of their translation equivalents is still an under-researched topic [1, 2]. The need for representation of collocations as a type of MWEs in bilingual dictionaries is broadly discussed in [2]. NLP systems that need to translate collocations often use pre-existing lexicons of collocation translations [4], which do not provide translations of all collocations, as new combinations are created and used on a daily basis. Bouamor et al. [1] use distributional models to align MWEs in order to improve the performance of machine translation systems. However, their method relies exclusively on a sentence-aligned corpus. Rapp and Sharoff [7] also investigate the use of the patterns of word-co-occurrence across languages to extract MWE translations. While their approach delivers good results in establishing translations of single words, they do not report good results for MWEs.

Based on the promising results of using corpus-based distributional similarity in a bilingual context to discover translationally-equivalent words [6], we use context similarity to automatically extract translations for a set of experimental collocations in English and Spanish. However, we define the contexts bilingually and we draw on word embeddings for learning vector representations for our target expressions [5]. Our results suggest that similarities modelled via word embeddings are more meaningful and lead to better translations.

Outline of the methodology

The methodology pursued in this study employs distributional similarity across bilingual corpora. A fundamental premise is that equivalent expressions may share common concepts in their contexts. These shared concepts are in turn expressed by words/terms that are translation equivalents in the two languages. For example, we might expect to see the Spanish expression *tener lugar* co-occuring with words, such as *terremoto*, *disturbio*, *seminario*, *elecciones*, *debate*, and the potential English translation of it, *take place*, co-occurring with the translations of the Spanish context words, i.e., *earthquake*, *riot*, *seminar*, *elections*, *debate*, respectively.

We use an extended version of a state-of-the-art distributional similarity method to identify translation equivalents for collocations, which is known as word2vec [5]. The underlying idea is to represent words as dense vectors (also referred to as word embeddings) which are learned by neural networks [3]. This increasingly popular model is adapted to this study by (i) regarding sequences of words as single units and (ii) defining bilingual contexts as set of translation pairs which are

obtained by an automatically learned Machine Translation system. We adapt the generalised version of word2vec (word2vecf) [3] to the task of vector construction for multi-word collocations using the bilingual contexts. To extract candidate translations for a MWE, we examine a set of automatically paired comparable documents from the two languages. Specifically, for each MWE *vn*, we examine all target language documents that are paired to the source language documents containing *vn*. We take a set of frequent verb unigrams, verb + noun bigrams, and verb + noun with an intervening word trigrams (which are verb combinations) appearing in these documents as candidate translations for *vn*. The winning candidate is the one that has the highest similarity to the MWE *vn*. We use a corpus of comparable English-Spanish documents that we build from various news sources on the Web and we pair the documents in comparable corpora by using the ACCURAT toolkit.[1] The details of the methodology is explained in [8].

Experiments and Evaluation

A native speaker was asked to review and rate the top-ranked translations identified by each of the methods for each expression. The annotator was instructed to give a score of 1 if there was at least one correct translation in the top-ranked list, and a score of 0, otherwise. A simple distributional similarity approach based on the Jaccard similarity coefficient was implemented as a baseline. Given two expressions from two different languages, their similarity was computed by comparing their corresponding sets of bilingual context pairs within a window of 10 words from the comparable corpora. We also perform experiments to investigate the robustness of the approach while adding noise to the corpora.

As aforementioned, similarity measures were used to rank the candidate translations for each expression. Experiments with different similarity thresholds were conducted, which as expected, reflected the trade-off between coverage (recall) and accuracy (precision). Table 1 shows accuracy and coverage values for the preliminary results on finding translations of the Spanish expressions.

Table 1: The accuracy of the baseline compared to the word2vec approach in extracting translations of Spanish Expressions.

	coverage	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
using paired CC	baseline	82%	55%	24%	22%	18%	16%	12%
	word2vec	50%	46%	40%	36%	34%	32%	33%
using CC + noise	baseline	78%	50%	24%	18%	14%	13%	8%
	word2vec	44%	45%	38%	**37%**	30%	**33%**	32%

As can be seen in the first row of the table, the baseline accuracy is high when we limit the method with a very low coverage, but drops down quickly as we increase coverage. Compared to the baseline, the word2vec approach is more stable across the different degrees of coverage: in fact, the performance of word2vec drops only slightly when we move from coverage of 30% to almost 80%. Importantly, even for a very high degree of coverage (i.e., 70%-80%) word2vec performs much better than the baseline in terms of accuracy. Furthermore, we observe that almost in all coverages the performance of the baseline approach drops by using the noisy paired documents, while the word2vec appears to be more stable on the two different corpora.

[1] http://www.accurat-project.eu

Conclusions

We have proposed a method for extracting cross-lingual contexts from comparable corpora, which we have then used to build embedding-based vector representations for multi-word collocations using a state-of-the-art technique (word2vec). We use these vectors to find translation equivalents for Verb+Noun combinations between Spanish and English. We show that our approach outperforms a simple distributional similarity baseline. We should note, however, that the results should be regarded as preliminary. Future experiments will focus on improving the performance further as follows. First, we intend to compile and employ larger corpora of comparable documents, in order to increase the coverage and also the accuracy by providing more context. Secondly, syntactic structure of MWEs can be added to the word2vec approach to draw on the grammatical dependencies of context which is expected to model better vector representations.

References

[1] Bouamor, D., Semmar, N., and Zweigenbaum, P. (2012). Identifying bilingual multi-word expressions for statistical machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[2] Corpas Pastor, G. (In press). Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues. In *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches (eds. S. Torner y E. Bernal). Series "Theoretical Developments in Hispanic Linguistics (ed. J. Gutiérrez-Rexach)*, pages 139–160, Chicago, IL: Ohio State University Press.

[3] Levy, O. and Goldberg, Y. (2014). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

[4] Mendoza Rivera, O., Mitkov, R., and Corpas Pastor, G. (2013). A flexible framework for collocation retrieval and translation from parallel and comparable corpora. In *Workshop on Multi-word Units in Machine Translation and Translation Technology*.

[5] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[6] 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.

[7] Rapp, R. and Sharoff, S. (2014). Extracting multiword translations from aligned comparable documents. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014*, pages 83–91, Gothenburg, Sweden.

[8] Taslimipoor, S., Mitkov, R., Corpas Pastor, G., and Fazly, A. (2016). Bilingual contexts from comparable corpora to mine for translations of collocations. In *Proceedings of the Seventeenth International Conference on Intelligent Text Processing and Computational Linguistics (to be appeared)*, CICLing '16, Konya, Turkey.