Bilingual Contexts from Comparable Corpora to Mine for Translations of Collocations

Shiva Taslimipoor[1(✉)], Ruslan Mitkov[1], Gloria Corpas Pastor[2], and Afsaneh Fazly[3]

[1] Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK
{shiva.taslimi,r.mitkov}@wlv.ac.uk
[2] Univeristy of Malaga, Malaga, Spain
gcorpas@uma.es
[3] VerticalScope Inc., Toronto, Canada
afsaneh.fazly@gmail.com

Abstract. Due to the limited availability of parallel data in many languages, we propose a methodology that benefits from comparable corpora to find translation equivalents for collocations (as a specific type of difficult-to-translate multi-word expressions). Finding translations is known to be more difficult for collocations than for words. We propose a method based on bilingual context extraction and build a word (distributional) representation model drawing on these bilingual contexts (bilingual English-Spanish contexts in our case). We show that the bilingual context construction is effective for the task of translation equivalent learning and that our method outperforms a simplified distributional similarity baseline in finding translation equivalents.

Keywords: Collocations · Word vector representation
Distributional similarity · Comparable corpora

1 Introduction

Collocations are considered as one type of Multi-word Expressions (MWEs) [2, 7, 23]. While there are many studies on the automatic extraction of collocations from monolingual text [6, 19, 24], only a few have drawn on bilingual resources for the automatic treatment of collocations [3, 5, 15]. The need for representation of collocations in bilingual dictionaries is broadly discussed in [5]. To exemplify, collocations like *pay attention* and *pay homage*, require a different translation of the collocative verb in Spanish according to the base noun: *prestar/poner atención*, *rendir homenaje*.

Dealing with collocations, bilingually, is very interesting for two reasons: first, finding translation equivalents for these expressions is far from a resolved issue in Natural Language Processing (NLP); secondly, using bilingual corpora,

A. Gelbukh (Ed.): CICLing 2016, LNCS 9624, pp. 115–126, 2018.
https://doi.org/10.1007/978-3-319-75487-1_10

we can improve their identification especially for resource-poor languages. With regards to resource-poor languages, one approach that is indeed beneficial is to use comparable/non-parallel corpora. Although comparable corpora have been known to be helpful [14], their application to this task has been rather limited [9,21,26].

We propose an approach to find translation equivalents for collocations using comparable corpora. The idea is to use distributional similarity across bilingual corpora. By 'equivalent expressions' or 'equivalents' we refer to expressions which are translations of each other across languages. One of the premises in this methodology is that equivalent expressions are expected to appear in the same or similar contexts across languages.

Characterisation and comparison of context (distributional) vectors is known to be the standard approach to bilingual lexicon extraction from comparable corpora [4]. However, we aim to use such an approach to find translation equivalents for collocations. We benefit from a list of automatically aligned words to build bilingual contexts for our target expressions. We use the very recent word embedding approach [16] which employs the bilingual contexts to learn vector representations for words or expressions. Similar to [22], we use a strictly comparable corpora, in which the documents are paired to each other, to retrieve more relevant translations.

We focus on a particular type of collocations, namely those that are formed from a combination of a verb and a noun, e.g., *take part* in English, *formar parte* in Spanish. While the approach is language independent, in this particular study we seek to identify translation equivalents between Spanish and English collocations.

The remainder of this paper is organised as follows. The next section describes previous work addressing the task of bilingual translation equivalents extraction. In Sect. 3, we elaborate on the context similarity approach for identifying collocation translations. Section 4 includes the details of the data and the experiments which have been done for the task. We evaluate, report and discuss the results in Sect. 5 and we, finally, conclude in Sect. 6.

2 Related Work

The most common approach for extracting translation equivalents from parallel corpora is to use Statistical Machine Translation (SMT) [27]. Recently, several studies have suggested approaches for extracting *parallel segments* from comparable corpora for several different tasks, including bilingual lexicon construction [4,9,11,21], and sentence alignment for improving SMT [10,18,25]. Corpus-based distributional similarity has been used in a bilingual context to automatically discover translationally-equivalent *words* from comparable corpora [9,20,21]. It is not clear, however, whether a similar approach can be used for finding the translations of *multi-word collocations*.

NLP systems that need to translate collocations often use pre-existing lexicons of collocation translations [15]. However, such lexicons do not provide translations of all collocations, as new combinations are created and used on a daily

basis. Thus, it is important to develop a method that can automatically find translation equivalents for multi-word collocations. Bouamor et al. [3] use distributional models to align MWEs to improve the performance of a machine translation system. However, their method relies on sentence-aligned (parallel) corpora. Rapp and Sharoff [22] also investigate the use of word co-occurrence patterns across languages to extract translations of single and multi-word terms. Like [11] they avoid using a large initial bilingual dictionary. While their approach delivers good results in finding the translations of single words, they do not report good results for MWEs. Even for single words their results only cover words that are salient words (keywords) according to their frequency patterns.

We also 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 [16]. Our results suggest that similarities measured using word embeddings are more meaningful and lead to better translations.

3 Distributional Similarity Across Languages

According to the distributional similarity hypothesis, terms that are translation equivalents 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 *poner en marcha* co-occuring with words, such as *problema*, *decisión* and *mercado*, and the potential English translation of it, *to launch*, co-occurring with the translations of the Spanish context words, i.e., *concern*, *decision*, *market*, respectively.

Distributional similarity has been widely used to find pairs (words or terms) that are semantically similar; however, the applications have mainly focused on similar pairs within a single language. We use an extended version of a state-of-the-art distributional similarity method to identify translation equivalents for collocations. Specifically, we define context in a bilingual space by pairing words from the two languages that we know are translations of each other. Note that we do not rely on a clean bilingual lexicon. Instead, we take the word pairs from a noisy bilingual lexicon, which is automatically learned by using a word alignment tool.[1]

3.1 Word Vector Representation

To represent words using context vectors, we use the `word2vec` method proposed by Mikolov et al. [16]. The method employs the patterns of word co-occurrences within a small window to predict similarities among words. The idea is to represent each word as a dense vector (a.k.a. word embeddings) derived by various

[1] We use the lexicon built by applying GIZA++ on the Spanish–English portion of the Europarl.

training methods, which in turn have been inspired by neural-network language modelling [13]. The new word embedding approach uses a neural network to learn low-dimensional word vectors from raw (monolingual) text. The standard implementation of `word2vec` constructs bag-of-words contexts for all single-word terms that appear in a training corpus. We adapt the model to our task of finding translation equivalents for multi-word collocations, by: (i) treating sequences of words as single units/terms, and (ii) defining bilingual contexts by drawing on a core set of known translation pairs. To do this we use the generalised word embedding approach proposed by [13] that allows us to define bilingual contexts. Although the generalised version of `word2vec` was originally used to extract dependency-based word embeddings [13], we can easily adapt it to our specific task of vector construction for multi-word collocations using bilingual contexts.

3.2 Bilingual Phrase Vector Representation

In standard `word2vec`, using a window of size k around a target word w, $2k$ context words are produced: the k words before and the k words after w. We base our context extraction on this standard, with the difference that we extract only specific words rather than all the words in the context window. Our favourable context words come from a bilingual dictionary of words. Specifically, we focus on nouns as the most important components of meaning, and use a core lexicon of paired English–Spanish nouns as our bilingual context terms. The generalised `word2vec` model (called `word2vecf`)[2] can then be trained on these pairs, resulting in the vectors of the two languages to be defined over the same space (of paired English–Spanish nouns), and to be comparable.

3.3 Translation Equivalent Extraction

Given a target collocation s from the source language (e.g., Spanish), our goal is to find the best translation equivalent in the target language (e.g., English). First, we identify a set of candidate translations for s, from a Spanish–English comparable corpora that we automatically build by pairing documents from the two languages. Next, we rank these candidates according to their semantic similarity to the target collocation. The following subsections explain these two steps in more detail.

Candidate Extraction. To extract candidate translations for a collocation, we examine a set of automatically paired *comparable documents* from the two languages. Specifically, for each collocation s, we examine all target language documents that are paired to the source language documents containing s. We take a set of frequent unigrams, bigrams, and trigrams (which are verb combinations) appearing in these documents as candidate translations for s.[3] The details of pairing documents in comparable corpora is explained in Sect. 4.1.

[2] The software is available in the websites of the authors of [13].

[3] We set the frequency threshold to 10 in our experiments.

Ranking Candidates Using Cross-Lingual Similarity. We construct a cross-lingual vector representation for each collocation s, and for each of its candidate translations, drawing on our proposed approach for defining a cross-lingual semantic space (see Sect. 3 above). The winning candidate is the one that has the highest similarity to the collocation s.

4 Experimental Setup

4.1 Corpus

We use a corpus of comparable English–Spanish documents that we build from various news sources on the Web, as explained below.

Collecting News Documents from the Web. News texts are rich sources of shared content, and hence have commonly been used to construct comparable corpora [1,8,17]. To build our corpus of comparable English–Spanish documents, we collect news feeds from a variety of news sources, including the ABC news,[4] Yahoo news,[5] CNN news,[6] Sport news,[7] and Euronews[8] in both Spanish and English languages. We focus on documents from July to December 2015. We use a tool from the ACCURAT project[9] to extract comparable documents from the news texts [1].

Computing Document Comparability. ACCURAT also comes with a tool, called `DictMetric`, which is designed to measure the comparability levels of document pairs via cosine similarity [26]. The tool is specifically proposed to provide a data for extracting parallel segments with high performance. To measure the comparability of two documents in different languages, one language get translated to the other. The tool translates non-English texts into English by using lexical mapping from the available GIZA++ based bilingual dictionaries. Since the proportion of overlapped lexical information in two documents is the key factor in measuring their comparability, the tool converts the texts into index vectors and then computes the comparability score of document pairs by applying cosine similarity measure on the index vectors.

Using the ACCURAT toolkit, we compute the comparability of all pairs of Spanish and English documents. We extract the pairs with the comparability score (cosine similarity) of higher than 0.45 as aligned comparable documents. This result in 16,436 English documents (with around 11 million word tokens) and 11,468 Spanish documents (with around 6 million word tokens).

[4] http://www.abc.es and http://www.abc.net.au.
[5] http://es.noticias.yahoo.com and http://uk.news.yahoo.com.
[6] http://cnnespanol.cnn.com and http://cnn.com.
[7] http://www.sport.es/es and http://www.sport-english.com/en.
[8] http://es.euronews.com and http://euronews.net.
[9] http://www.accurat-project.eu.

Each English document is paired to at least one Spanish document; equally, there is at least one paired English document for every Spanish document.[10]

4.2 Experimental Expressions

Our methodology is to use bilingual word vector representation to find translations for collocations across comparable corpora. To report the results, we focus on 9 highly-frequent verbs in English and 6 in Spanish. These verbs tend to frequently combine with many different nouns in their direct object positions to form multi-word collocations. The verbs are: *take*, *have*, *make*, *give*, *get*, *find*, *pay*, *lose* in English, and *tener*, *dar*, *hacer*, *formar*, *tomar*, *poner* in Spanish. We extract all occurrences of these verbs followed with a noun, from the whole News corpora, focusing only on those combinations that have a frequency higher than 10. This process results in 1,007 English Verb+Noun collocations, and 930 Spanish Verb+Noun collocations, which are annotated by two human annotators as being semantically coherent collocations, or arbitrary sequences of words. We measure inter-annotator agreement using the Kappa score: Kappa is 0.67 for English expressions, and 0.61 for Spanish expressions. Among these candidate expressions, only 162 English expressions and 187 Spanish expressions occur with frequency higher than 9 in our paired comparable documents. We run the experiments only on these expressions.

4.3 Vector Construction

Recall that to construct vectors for our English and Spanish expressions, we need a seed list of paired context words (a.k.a., the bilingual context pairs). For this purpose, we use a subset of the word alignments resulting from applying GIZA++ on the English–Spanish Europarl parallel corpus [12]. Specifically, we only consider pairs of frequent nouns that have an alignment probability of higher than 0.2, where frequent nouns in a language are those that appear 50 times or more in Europarl. As a result we have a list of 4,700 bilingual contexts.

For learning the vectors, we use the following corpora to extract word co-occurrence statistics: the monolingual English and Spanish components from the Europarl, and the English and Spanish components of our News corpora. We index all the English and Spanish *verb combinations* (unigrams, bigrams, trigrams) according to their occurrences with the context word pairs. Specifically, from the window of 10 words around a target expression, we capture any word that exists in our bilingual context pairs (focusing on the relevant language given the language of the target expression). The `word2vecf` software is then used to train vectors on the indexed corpus. We then apply our methodology to find translations for collocations in both directions: Spanish to English, and English to Spanish.

[10] The comparable corpora that we prepared is available on https://github.com/shivaat/EnEsCC.

Note that we focus on finding translations for Verb+Noun combinations. We assume that for most such expressions, the translation equivalent is either a Verb (unigram), a Verb+Noun (bigram), or a Verb+Noun with an intervening word, such as a determiner or an adjective (trigram). We thus consider as our candidate translations all unigram Verbs, bigram Verb+Noun combinations, and trigram Verb+Noun combinations with an intervening word. For every expression from the source language (e.g., Spanish), our goal is to find the five most cross-lingually similar Verb or Verb + Noun combination in the target language (e.g., English).

5 Evaluation and Results

Baseline. We implement a simple distributional similarity approach as our baseline. Given two expressions (from the two languages), we measure their similarity by comparing their corresponding sets of (bilingual) context pairs (using a context window of size 10). We use the Jaccard similarity coefficient to measure similarity. The baseline uses our comparable corpora to find translation candidates for each expression, but relies on the above simple similarity to rank these candidates.

Using Loosely Comparable Corpora. We also perform experiments to investigate the advantage of using comparable corpora with high level of similarity for finding the candidate translations of an expression. To do so, we add noisy alignments to our accurately-aligned documents. Specifically, for each source-language (e.g., Spanish) document, paired with several highly-similar target-language (e.g., English) documents, we align an extra set of 2,000 randomly selected target-language documents.[11] This process results in a larger but noisy corpus of comparable documents. Our goal here is to understand whether using a larger set of documents that may contain more candidate translations is helpful, despite the noise. That is, we intend to understand whether a method like `word2vec` is sufficiently robust to noise, and hence capable of finding good translations from documents that are not perfectly aligned. If that is the case, then we can avoid the rather expensive process of building highly-accurate comparable corpora. We apply both the baseline and our proposed approach (the one that uses `word2vec`) to this noisy data, and compare the results with those on the smaller corpora with the more accurately aligned documents.

Results and Discussion. We ask a human expert to rate the top-ranked translations produced by each of the methods for each expression. We ask the expert to give a rating of 1 if there is at least one good translation in the top-5-ranked list; otherwise, the list is given a rating of 0. We also have 25% of the resulted translation lists annotated by a second annotator. The inter-annotator

[11] Note that we add noise in both Spanish–English and English–Spanish directions.

agreement in terms of Kappa is 0.80 both for finding translations for Spanish expressions and for finding translations for English expressions.

Note that we use a similarity measure to rank the candidate translations of each expression. By using different threshold values for this similarity, we get ranked lists of varying sizes. The higher this threshold, the smaller the number of the resulting translation candidates, and hence the higher the number of expressions for which we may not have any good translations. In other words, we can trade off accuracy (precision) for coverage (recall). We thus set the similarity thresholds to different values in order to measure accuracy for varying degrees of coverage (from around 10% to around 80%). Doing so gives us a better understanding of the overall performance of each method.

Table 1 shows accuracy and coverage values for finding translations of the Spanish expressions; Table 2 gives the results for English expressions. Note that we show the results for both the baseline and the word2vec method, using both corpora of comparable documents: the (smaller and less noisy) corpus of highly-comparable documents (referred to as paird CC), and the larger and noisy corpus (referred to as CC + noise).

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%

Table 2. Comparing the accuracy of the baseline with the word2vec approach in extracting translations of English expressions.

	Coverage	10–20%	20–30%	30–40%	40–50%	50–60%	60–70%	70–80%
Using paired CC	Baseline	79%	52%	46%	35%	26%	22%	18%
	word2vec	39%	37%	34%	36%	34%	29%	31%
Using CC + noise	Baseline	70%	50%	24%	22%	18%	12%	13%
	word2vec	38%	34%	31%	**39%**	**39%**	**32%**	31%

As can be seen in the first rows of both tables, the baseline accuracy/precision is high when we limit the method with a very low coverage/recall, but drops down quickly as we increase coverage. Note that when coverage is low, many expressions do not have any translation equivalents. But those that do have candidates, have a few accurate ones, and hence it is easy for a simple method such as the baseline to pick the best.

Compared to the baseline, the word2vec approach is more stable across the different degrees of coverage for both translation directions: in fact, the performance of word2vec drops only slightly when we move from a 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 (33% compared to 12% for Spanish-to-English, and 31% versus 18% for English-to-Spanish).

Next, we compare the results using the two corpora. Investigating the baseline approach over the two corpora, we observe that almost in all coverages the performance of the baseline approach drops by using the noisy paired documents. This can be seen in both Tables 1 and 2 for both directions of Spanish to English and English to Spanish translations. Then we compare the results of word2vec: Interestingly, the performance of word2vec is reasonably close on the two different corpora, even though the CC + noise has a much higher degree of noise. The better accuracies of word2vec in some cases when we use the larger noisy corpora are shown in bold. This is an interesting result, suggesting that even using a large but noisy corpus of comparable documents, we can find reasonable translations for multiword collocations by relying on a robust and accurate method such as word2vec.

Semantically Coherent Collocations. Our experimental Verb+Noun combinations (that we try to find translations for) include a range of expressions, from frequent collocations (*get things*), to multi-word verbal units (*make reference*), to more idiomatic expressions (*take place*). It is thus interesting to find out whether the performance of our method differs on these different types of expressions. For this, we take a subset of expressions from each language that has been annotated as a semantically-coherent MWE by two annotators. This selection process results in 80 Spanish and 101 English expressions. Table 3 shows accuracy of the word2vec method for both Spanish and English subsets when coverage is set to around 80% (using the cleaner comparable corpora for finding candidates). The results show that, for both languages, accuracy improves when we focus on these subsets (48% versus 33% for Spanish expressions, and 44% versus 31% for English).

Table 3. The accuracy of the word2vec approach in extracting translations of multiword collocations from comparable corpora.

	Accuracy	
	Spanish	English
word2vec approach	48%	44%

6 Conclusions and Future Work

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 also show that, in contrast to the distributional similarity baseline, the word2vec approach is less vulnerable to noise in the corpus (in terms of comparability of the aligned documents).

Future experiments will focus on improving the results further as follows: First, preparing larger corpora of comparable documents, in order to increase the coverage and also the accuracy by providing more context. Secondly, we can take into account expressions that have more than one intervening word between the Verb and the Noun components (both for our experimental collocations, and for the translation candidates). Third, syntactic structure can be added to the word2vec approach to draw on the grammatical dependencies of context and hence form better vector representations (as suggested in [13]).

Acknowledgments. This work has been partially supported by the LATEST (Ref: 327197-FP7-PEOPLE-2012-IEF) project. The authors would like to express their gratitude to Anna de Santis and Lorena Gomez for their annotation work.

References

1. Aker, A., Kanoulas, E., Gaizauskas, R.: A light way to collect comparable corpora from the web. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012) (2012)
2. Bannard, C.: A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions, pp. 1–8. Association for Computational Linguistics (2007)
3. Bouamor, D., Semmar, N., Zweigenbaum, P.: Identifying bilingual multi-word expressions for statistical machine translation. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey. European Language Resources Association (ELRA) (2012)
4. Bouamor, D., Semmar, N., Zweigenbaum, P.: Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Short Papers, vol. 2, pp. 759–764. Association for Computational Linguistics (2013)
5. Pastor, G.C.: Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues. In: Torner, S., Bernal, E. (eds.) Collocations and Other Lexical Combinations in Spanish: Theoretical and Applied Approaches, pp. 173–199. Routledge, Abingdon (2017)
6. Evert, S.: The statistics of word cooccurrences : word pairs and collocations. Ph.D. thesis, Universität Stuttgart, Holzgartenstr. 16, 70174 Stuttgart (2005)
7. Fazly, A.: Automatic acquisition of lexical knowledge about multiword predicates. Ph.D. thesis, Department of Computer Science, University of Toronto (2007)

8. Fung, P.: A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In: Farwell, D., Gerber, L., Hovy, E. (eds.) AMTA 1998. LNCS (LNAI), vol. 1529, pp. 1–17. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-49478-2_1
9. Fung, P., McKeown, K.: Finding terminology translations from non-parallel corpora. In: Proceedings of the 5th Annual Workshop on Very Large Corpora, pp. 192–202 (1997)
10. Ion, R.: PEXACC: a parallel sentence mining algorithm from comparable corpora. In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012) (2012)
11. Ismail, A., Manandhar, S.: Bilingual lexicon extraction from comparable corpora using in-domain terms. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 481–489. Association for Computational Linguistics (2010)
12. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Conference Proceedings: The Tenth Machine Translation Summit, Phuket, Thailand, pp. 79–86 (2005)
13. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, Short Papers, vol. 2, pp. 302–308. Association for Computational Linguistics (2014)
14. McEnery, A., Xiao, R.: Parallel and comparable corpora: what is happening. In: Incorporating Corpora: The Linguist and the Translator, pp. 18–31 (2007)
15. Mendoza Rivera, O., Mitkov, R., Corpas Pastor, G.: 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 (2013)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
17. Munteanu, D.S., Marcu, D.: Improving machine translation performance by exploiting non-parallel corpora. Comput. Linguist. **31**(4), 477–504 (2005)
18. Pal, S., Pakray, P., Naskar, S.K.: Automatic building and using parallel resources for SMT from comparable corpora. In: Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL, pp. 48–57 (2014)
19. Pecina, P.: An extensive empirical study of collocation extraction methods. In: Proceedings of the ACL Student Research Workshop, ACLstudent 2005, Stroudsburg, PA, USA, pp. 13–18. Association for Computational Linguistics (2005)
20. Pekar, V., Mitkov, R., Blagoev, D., Mulloni, A.: Finding translations for low-frequency words in comparable corpora. Mach. Transl. **20**(4), 247–266 (2006)
21. Rapp, R.: Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 519–526. Association for Computational Linguistics (1999)
22. Rapp, R., Sharoff, S.: Extracting multiword translations from aligned comparable documents. In: Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, Gothenburg, Sweden, pp. 83–91 (2014)
23. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: a pain in the neck for NLP. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 1–15. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45715-1_1

24. Smadja, F.: Retrieving collocations from text: Xtract. Comput. Linguist. **19**, 143–177 (1993)
25. Smith, J.R., Quirk, C., Toutanova, K.: Extracting parallel sentences from comparable corpora using document level alignment. In: Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2010), pp. 403–411 (2010)
26. Su, F., Babych, B.: Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. In: Proceedings of the Joint Workshop on Exploiting Synergies Between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL 2012, Stroudsburg, PA, USA, pp. 10–19. Association for Computational Linguistics (2012)
27. Tiedemann, J.: Extraction of translation equivalents from parallel corpora. In: Proceedings of the 11th Nordic Conference on Computational Linguistics, pp. 120–128 (1998)