# A Comparison of Methods for Identifying the Translation of Words in a Comparable Corpus: Recipes and Limits

Laurent Jakubina and Philippe Langlais

DIRO
Université de Montréal
CP 6128 Succursale Centre-Ville
H3C3J7 Montréal, Québec, Canada
rali.iro.umontreal.ca

**Abstract.** Identifying translations in comparable corpora is a challenge that has attracted many researchers since a long time. It has applications in several applications including Machine Translation and Cross-lingual Information Retrieval. In this study we compare three state-of-the-art approaches for these tasks: the so-called context-based projection method, the projection of monolingual word embeddings, as well as a method dedicated to identify translations of rare words. We carefully explore the hyper-parameters of each method and measure their impact on the task of identifying the translation of English words in Wikipedia into French. Contrary to the standard practice, we designed a test case where we do not resort to heuristics in order to pre-select the target vocabulary among which to find translations, therefore pushing each method to its limit. We show that all the approaches we tested have a clear biased toward frequent words. In fact, the best approach we tested could identify the translation of a third of a set of frequent test words, while it could only translate around 10% of rare words.

## 1 Introduction

Extracting bilingual lexicons from comparable corpora has received a massive interest in the NLP community, mainly because parallel data is a scarce resource, especially for specific domains. More than twenty years ago, [5] and [20] described two methods for how this can be accomplished. While those studies differ in the nature of their underlying assumption, they both assume that the context of a word shares some properties with the context of its translation. In the case of [20] the assumption is that words in translation relation show similar co-occurrence patterns.

Many variants of [20] have been proposed since then. Some studies have for instance reported gains by considering syntactically motivated co-occurrences, either by the use of a parser [22] or by simple POS-based patterns [16]. Extensions of the approach in order to account for multiword expressions have also been proposed (*e.*g. [2]). Also, many have studied variants for extracting domain

specific lexicons; the medical domain being vastly studied, see for instance [1] and [15]. We refer the reader to [21] for an extensive overview of works conducted in this vein.

There is a trend of works on understanding the limits of this approach. See for instance the study of [11] in which a number of variants are being compared, or more recently the work of [9]. One important limitation of the context-based approach is its vulnerability to rare words, which has been demonstrated in [18]. The authors reported gains by predicting missing co-occurences thanks to co-occurences observed for similar words in the same language. In [19], the authors adapt the alignment technique of [6] initially proposed for parallel corpora by exploiting the document pairing structure of Wikipedia. The authors show that coupling this alignment technique with a classifier trained to recognize translation pairs, yield an impressive gain in performance for rare words.

Of course, many other approaches have been reported for mining translations in comparable corpora; again see [21] for an overview. A very attractive approach these days is to rely on so-called word embeddings trained with neural-networks thanks to gradient descent on large quantities of texts. In [13], Mikolov describes two efficient models for training such embeddings which are implemented in the popular `Word2Vec` toolkit. The same author also demonstrated that a mapping of word embeddings learnt independently for each language can be trained by making use of a seed bilingual lexicon [14].

In this study we compare three of the aforementioned approaches: the context-based approach [20], the word embedding approach of [14] and the approach of [19] dedicated to rare words. We investigate their hyper-parameters and test them on words with various properties. This comparison has been conducted on a very large scale setting, making use of the full English-French Wikipedia collection.

In the remainder of this article, we describe in Section 2 the approaches we tested. Our experimental protocol is presented in Section 3. Section 4 summarizes the best results we obtain with each approach, and report on the impact of their hyper-parameters on performances. In Section 5, we analyze the bias those approaches have toward some word properties. We conclude in Section 6.

## 2 Approaches

We implemented and tested three state-of-the-art alignment techniques. Two of them (`context` and `embedding`) makes use of a so-called bilingual seed lexicon, that is, pairs of words in translation relation; the other (`document`) is exploiting instead the structure of the comparable collection.

### 2.1 Context-Based Projection (`context`)

In [20], each word of interest is represented by a so-called context vector (the words it co-occurs with). Source vectors are projected (or "translated") thanks to

a seed lexicon. Candidate translations are then identified by comparing projected contexts with target ones, thanks to a similarity measure such as cosine.

We conducted our experiments using the `eCVa` Toolkit[1] [9] which implements several association measures (for building up the context vectors) and similarities (for comparing vectors).

### 2.2 Document-based Alignment (`document`)

This approach relies on a pre-established pairing of comparable documents in the collection. Given such a collection, word translations are identified based on the assumption that source and target words should appear collection-wise in similar pairs of (comparable) documents.

The approach was initially proposed for handling the case of rare words for which co-occurrence vectors are very sparse. On a task of identifying the translation of medical terms, the authors showed the clear superiority of this approach over the context-based projection approach. In their paper, the authors show that coupling this approach to a classifier trained to recognize translation pairs is fruitful. We did not implement this second stage however, since the performance of the first stage was not judged high enough, as discussed later on.

### 2.3 Word Embedding Alignment (`embedding`)

Word Embeddings (continuous representations of words) has attracted many NLP researchers recently. In [14], the authors report on an approach where word embeddings trained mono-lingually are linearly mapped thanks to a projection matrix $W$ which coefficients are determined by gradient descent in order to minimize the distance between projected embeddings and target ones, thanks to a seed bilingual lexicon $\{(x_i, y_i)\}_{i\in[1,n]}$:

$$\min_W \sum_{i=1}^{n} ||W\hat{x_i} - \hat{y_i}||^2 \tag{1}$$

where $\hat{x_i}$ and $\hat{y_i}$ are the (monolingual) embeddings of the words $x_i$ and $y_i$ respectively. Given a term $x$ absent from the seed lexicon but for which an embedding $\hat{x}$ has been trained, translations are determined by selecting the words which target embeddings are the closest to the projected one, $W\hat{x}$, thanks to a distance (cosine in their case).

We implemented this approach by using the `Word2Vec`[2] toolkit [13] for training word embeddings. The linear mapping was trained thanks to the implementation described in [3].

[1] `http://rali.iro.umontreal.ca/rali/en/ecva-toolkit`
[2] `https://code.google.com/p/word2vec/`

## 3 Experimental Protocol

The approaches described in the previous section have been configured to produce a ranked list of (at most) 20 candidate French translations for a set of English test words. We measure their performance with accuracy at rank 1, 5 and 20; where accuracy at rank $i$ (TOP@$i$) is computed as the percentage of test words for which a reference translation is identified in the first $i$ candidates proposed.

Each approach used the exact same comparable collection, and possibly a seed lexicon. In the remainder, we describe all the resources we used.

### 3.1 Comparable Corpus

We downloaded the Wikipedia dump of June 2013 in both English and French. The English dump contains 3 539 093 articles, and the French one contains 1 334 116. The total number of documents paired by an inter-language link is 757 287. While some pairs of documents are likely parallel [17], most ones are only comparable [8]. The English vocabulary totalizes 7.3 million words (1.2 billion tokens); while the French vocabulary counts 3.6 million ones (330 million tokens).

We used all the collection without any particular cleaning, which departs from similar studies where heuristics are being used either to reduce the size of the collection or the list of candidate terms among which a translation is searched for. For instance, in [19] the authors built a comparable corpus of 20 169 document pairs and a target vocabulary of 128 831 words. Also, they concentrated on nouns only. This is far lower than the figures in our setting. While our choice brings some technical issues (computing context vectors for more than 3M words is for instance rather challenging), we feel it gives a better picture of the merit of the approaches we tested. This is further discussed in Section 4.

### 3.2 Test Sets

We built two test sets for evaluating our approaches; one named **1k-low** gathering 1 000 *rare* English words and their translations, where we defined rare words as those occurring at most 25 times in English Wikipedia; and **1k-high** gathering 1 000 words occurring more than 25 times. For the record, 6.8 million words (92%) in English Wikipedia occur less than 26 times.

The reference translations where collected by crossing the vocabulary of French Wikipedia with a large in-house bilingual lexicon. Half of the test words have only one reference translation, the remainder having an average of 3 translations. It should be clear that each approach we tested could potentially identify the translations of each test word, and therefore have a perfect recall.

### 3.3 Seed Bilingual Lexicon

The `context` and `embedding` approaches both require a seed bilingual lexicon. We used the part of our in-house lexicon not used for compiling the test sets

aforementioned. For the `embedding` approach, we followed the advice of [3] and compiled lexicons of size up to 5 000 entries.[3] More precisely, we prepared three seed lexicons: **2k-low** which gathers 2 000 entries involving rare English words (words occurring at most 25 times); **5k-high** gathering 5 000 entries whose English words are not rare, and **5k-rand** which gathers 5 000 entries randomly picked. For the `context` approach, we used 107 799 words of our in-house dictionary not belonging to the test material.

## 4 Results and Recipes

In this section, we report on the best performance we obtained with the approaches we experimented with. We further explore the impact of their hyperparameters on performance.

### 4.1 Overall performance

The performance of each approach on the two test sets are reported in Table 1, where we only report the best variant of each approach according to TOP@1. This table calls for some comments. First, we observe a huge performance drop of the approaches when asked to translate rare words. While `context` and `embedding` perform (roughly) equally well on frequent words, with an accuracy at rank 1 of around 20%, and 45% at rank 20, both approaches on the **1k-low** test set could translate correctly only 2% of the test words in the first place; the `embedding` approach being the less impacted at rank 20 with 12% of test words being correctly translated. What comes to a disappointment is the poor performance of the `document` approach which was specifically designed to handle rare words. We come back to this issue later on.

| | **1k-low** | | | **1k-high** | | |
|---|---|---|---|---|---|---|
| | TOP@1 | TOP@5 | TOP@20 | TOP@1 | TOP@5 | TOP@20 |
| `embedding` | **2,2** | **6,1** | **11,9** | **21,7** | **34,2** | **44,9** |
| `context` | 2,0 | 4,3 | 7,6 | 19,0 | 32,7 | 44,3 |
| `document` | 0,7 | 2,3 | 5,0 | *10,0* | *19.0* | *24.0* |
| *oracle* | 4,6 | 10,5 | 19,0 | 31,8 | 46,8 | 57,6 |

Table 1: Performance of the different approaches discussed in Section 2 on our two test sets. The best variant according to TOP@1 is reported for each approach. On **1k-high**, we could only compute at the time of writing the performance of the `document` approach on a subset of 100 entries.

It is interesting to note that approaches are complementary as evaluated by an oracle which picks one of the three candidate lists produced for each term.

[3] The authors tried larger lexicons without success.

This is the results reported in the last line of Table 1. On rare words, we observe almost twice the performance of individual approaches, while on **1k-high**, an absolute gain of 10 to 15 in TOP@1, 5 *and* 20 is observed. This said, we observe that no more than 57% (resp. 19%) of the test words in **1k-high** (resp. **1k-low**) could be translated in the top-20 positions, which is disappointing.

Examples of outputs produced by our best configurations are reported in Figure 1. While it is rather difficult to pinpoint why many test words were not translated, we observe tendencies. First, we notice a "thesaurus effect", that is, candidates are often related to the words being translated, without being translations, as *aromatisé* (aromatized) proposed for the English word *donut*. Some errors are simply due to morphological variations and could have been counted correct, as *pathologique* produced instead of *pathologiquement*. We also observe a few cases where candidates are acceptable, but simply not sanctioned by our reference.

| | | | | |
|---|---|---|---|---|
| donut | | beigne | | |
| `context` | - | aromatisé (0.05) | donut (0.05) | beignet (0.04) |
| `embedding` | - | liper (0.54) | babalous (0.53) | savonnettes (0.52) |
| brilliantly | | brillamment | | |
| `context` | - | imaginatif (0.05) | captivant (0.05) | rusé (0.05) |
| `embedding` | - | éclatant (0.69) | pathétique (0.67) | émouvant (0.66) |
| gentle | | doucet, | doux, | délicat, |
| `context` | - | enjoué (0.05) | serviable (0.05) | affable (0.04) |
| `embedding` | - | colérique (0.76) | enjoué (0.75) | espiègle (0.75) |
| pathologically | | pathologiquement | | |
| `context` | - | cordonale (0.05) | pathologique (0.05) | diagnostiqué (0.05) |
| `embedding` | - | psychosexuel (0.60) | psychoaffectif (0.60) | piloérection (0.59) |

**Fig. 1.** Top-3 candidates produced by the two best approaches for a few test words.

We compared a number of variants of each approach in order to better understand the figures reported. We summarize the main outcomes of our investigations in the following subsections.

### 4.2 Recipe for the `context` approach

We ran over 50 variants of the context-based approach, varying some meta-parameters, the influence of which is summarized in the sequel.

Each word in a context vector can be weighted by the strength of its relationship with the word being translated. We tested 4 main association measures and found PMI (point-wise mutual information) and discontinuous odds-ratio

(see [4]) to be the best ones. Other popular association measures such as log-likelihood ratios drastically underperforms on **1k-high** (TOP@20 of 7.8 compared to 44.3 for PMI).

Context words are typically picked within a window centered on the word to translate. While the optimal window size somehow varies with the association measure considered, we found the best results for a window size of 7 (3 words before and after) for the **1k-high** test set, and a much larger window size (31) for the **1k-low** test set. For rare words, context vectors are very sparse, therefore, increasing the window size leads to better performance.

Last, we observed a huge boost in performance at projection time by including in the projected vector source words unknown from our seed lexicon. Without doing this, the best configuration we tested decreased in TOP@20 from 44.3 to 31.7. Our explanation for this unexpected gain is that some of those words are proper names or acronyms which presence in the context vector might help to discriminate translations.

### 4.3 About the `document` approach

Since this approach does not deliver competitive results (see Table 1) we did not investigate many configurations as we did for the other two approaches we tested. One reason for the disappointing results we observed compared to the gains reported in [19] might be the very different nature of the datasets used in our experiments. As a matter of fact, we used the entire English-French Wikipedia collection while in [19] they only selected a set of 20 000 document pairs. Also our target vocabulary contains almost 3 millions words while theirs is gathering 120k nouns.

We conducted a sanity check where we randomly selected from our target vocabulary a subset of 120k words to which we added the reference translations of our test words (so that the aligner could identify them). On the **1k-low** test set, this led to an increased performance with a TOP@1 of 4.9 (compared to 0.7) and a TOP@20 of 20.2 (compared to 5.0). Those results are actually very much in line with the ones reported by the authors, suggesting that the approach does not scale well to large datasets.

### 4.4 Recipe for the `embedding` approach

We trained 130 configurations varying meta-parameters of the approach. In the following, we summarize our main observations.

First of all, the configurations which perform the best on **1k-high** and **1k-low** are different. On the former test set, our best configuration consists in training embeddings with the *cbow* model and the *negative sampling* (10 samples). The best window size we observed is 11, and increasing the dimensionality of the embeddings increases performance steadily. The largest dimensionality for which we managed to train a model is 200.[4] Those findings confirm observations

[4] We ran our training on nodes of a cluster that can accommodate up to 64 Gb of memory.

made by [3] for frequent terms. In this work the authors managed to trained a model with embeddings of size 300. On their task, the best model trained performs a TOP@1 of 30 while on our task, the configuration described achieved a TOP@1 of 22. In [13] the author also report a TOP@1 of around 30 for embeddings of size 1000. Recall that in our case, the target vocabulary size at test time is around 3 million words, while for instance in [13] it is in the order of a hundred thousands (depending of the language pair considered), which might account for some differences in performance.

For the **1k-low** test set, the best configuration we found consists in training a *skip-gram* model with *hierarchical softmax*, and a window size of 21 (10 words on both sides of each word), and an embedding dimensionality of 250 (the largest value we could afford on our computer). Again this confirms tendencies observed in [3] for the case of translating unfrequent words. Also, [14] observed that the *skip-gram* model and the *hierarchical softmax* training algorithm are both preferable when translating unfrequent words.

Regarding the influence of the seed lexicon, we observed that using the largest one is preferable. This confirms the findings in [3] that a lexicon of 5k is optimal for training the mapping of embeddings. For the low frequency test set, we further observed that using a seed lexicon of rare words (**5k-rand**) is better. By doing so, we could improve TOP@1 of 1 absolute point and TOP@20 of 3 points. On the **1k-high** test set, the best performance is obtained with **5k-high** (TOP@1 of 44.9%), then **5k-rand** (TOP@1 of 40.5%) and **2k-low** (TOP@1 of 10.3%).

## 5 Analysis

In the previous sections, we analyzed the impact of the hyper-parameters of each approach on performance. In this section, we analyze more precisely the results of the best configuration of each aligner in terms of a few properties of our test set. We believe such an analysis useful for comparison purposes. Also, in order to foster reproducibility, we are happy to share the test sets as well as the seed lexicons we used in this study.[5]

### 5.1 Frequency

We already observed a clear bias of the approaches we tested toward frequency. Figure 2 reports the performance of the best configuration of each approach when translating test words which frequency in English Wikipedia do not exceed a given threshold. For instance, we observe that on the subset of test words which frequency is 10 or less, the best approach according to TOP@1 (`embedding`) achieves a score of 8.76%. The frequency biais is clearly observable, and even for rather large frequency thresholds. Both `context` and `embedding` compete across frequencies, but if not too frequent test words have to be translated, and if a shortlist is what matters (TOP@20), then `context` might be the good approach to go with.

[5] Downloadable at `http://rali.iro.umontreal.ca/rali/en/ecva-toolkit`.

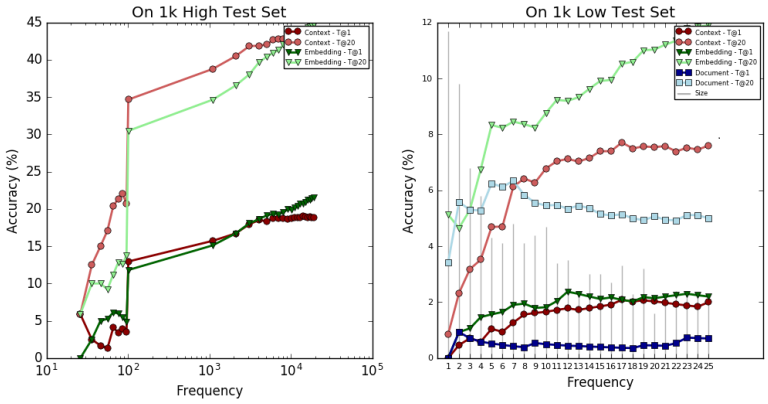


**Fig. 2.** TOP@1 and TOP@20 of the best configurations of each approach as a function of the frequency of the test words in English Wikipedia. Read the text for more.

### 5.2 String Similarity

It is rather difficult in the relevant literature to get a clear sense of the intrinsic difficulty of the translation task being tackled. In particular, the similarity of test words and their reference translation is almost never reported, while for some language pairs, it is a relevant information that is even used as a feature in some approaches (*e.g.*, [11]). Figure 3 illustrates the performance of our best configurations as a function of the edit-distance between test words and their reference translation.[6]

As we expected, we observe overall a decrease of performance for words which reference translation is dissimilar. On words translated verbatim, TOP@1 performance is as high as 79.3% for `context` and 61.8% for `embedding`, while both approaches compare as the edit-distance augments. On rare words, `embedding` seems to be less sensitive to edit-distance.

### 5.3 Medical terms

Medical term translation is the subject of many investigations (*e.g.* [15, 7, 10] to mention just a few). In order to measure if this very task is easier than translating any kind of word, we filtered our test words with an in-house list of 22 773 medical terms. We found only 22 medical terms in **1k-low** and 80 in **1k-high**. Although those figures are definitely not representative, we computed the performance of our best configurations on those subsets. The results are reported in Table 2. On frequent words, the gain in performance is especially marked for the `context` approach. We also note that the `document` approach seems to perform rather well on unfrequent medical terms, which is exactly the

[6] For easing the interpretation, we only considered test words having only one reference translation.

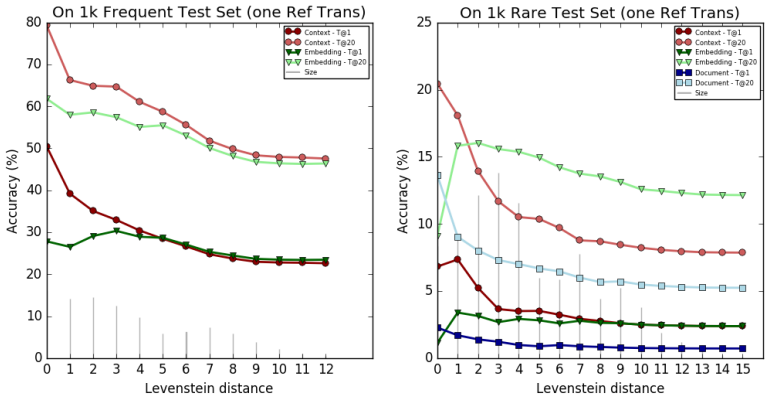


**Fig. 3.** TOP@1 and TOP@20 of the best configurations as a function of edit-distance between test words and their reference translation. Read the text for more.

setting studied in [19]. One possible explanation for this positive difference is that medical terms, at least in English Wikipedia tend to be rather frequent and their translation into French have an average edit-distance which is lower than for other types of words, two factors we have shown to impact performance positively.

| | **1k-low** | | **1k-high** | |
|---|---|---|---|---|
| | TOP@1 | TOP@20 | TOP@1 | TOP@20 |
| `embedding` | 4.5 (+2.7) | 13.6 (+1.7) | 27.5 (+5.8) | 53.7 (+8.8) |
| `context` | 0.0 (-2.0) | 4.5 (-3.1) | 48.7 (+29.7) | 72.5 (+28.3) |
| `document` | 4.5 (+3.8) | 22.7 (+17.7) | — | — |

Table 2: Performance (accuracy in %) on medical test words. Figures in parenthesis are absolute gains over the performance measured over the full test set.

## 6 Conclusion

In this study, we compared three approaches for identifying translations in a comparable corpus, and studied extensively how their hyper-parameters impact performance. We tested those approaches without reducing (somehow arbitrarily) the size of the target vocabulary among which to choose candidate translations. We also analyzed a number of properties of the test sets that we feel are worth reporting on when conducting such a task; among which the distribution of test words according to their frequency in the comparable collection and the distribution of their string similarity to their reference translation.

Among the observations we made, we noticed that the good old context-based projection approach [20] when appropriately configured competes with the more recent neural-network one, especially when translating frequent words. This observation echoes the observation made in [12] that carefully tuned count-based distributional methods are no worse than trained word-embeddings. This said, in our experiments, the embedding approach revealed itself as the method of choice overall. We also observed that the approach of [19] designed specifically for handling rare words, while being good at translating medical terms had a harder time translating other types of (unfrequent) words. Definitely, translating rare words is a challenge that deserves further investigations, especially since unfrequent words are pervasive.

We also provide evidence that the approaches we tested are complementary and that combining their outputs should be fruitful. Similarly, since a given approach typically shows different performance depending on the properties of test words (their frequency, their nature), it is also likely that combining different variants of the same approach should lead to better performance. This is left as a future work.

## Acknowledgments

This work has been funded by the Quebec funding agency *Fonds de Recherche Nature et Technologies* (FRQNT).

## References

1. Chiao, Y.C., Sta, J.D., Zweigenbaum, P.: A novel approach to improve word translations extraction from non-parallel, comparable corpora. In: Proceedings of the International Joint Conference on Natural Language Processing, Hainan, China. AFNLP (2004)
2. Daille, B., Morin, E.: Effective Compositional Model for Lexical Alignment. In: Third International Joint Conference on Natural Language Processing. pp. 95–102 (2008)
3. Dinu, G., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. ResearchGate (Dec 2014)
4. Evert, S.: The statistics of word cooccurrences. Ph.D. thesis, Dissertation, Stuttgart University (2005)
5. Fung, P.: Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In: Third Workshop on Very Large Corpora. pp. 173–183 (1995)
6. Gale, W.A., Church, K.W.: Identifying Word Correspondences in Parallel Texts. In: HLT. vol. 91, pp. 152–157. Citeseer (1991)
7. Hazem, A., Morin, E.: Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora. In: LREC. pp. 288–292 (2012)
8. Hovy, E., Navigli, R., Ponzetto, S.P.: Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence 194, 2–27 (Jan 2013)

9. Jakubina, L., Langlais, P.: Projective methods for mining missing translations in dbpedia. In: Proceedings of the Eighth Workshop on Building and Using Comparable Corpora. pp. 23–31. Association for Computational Linguistics, Beijing, China (July 2015)
10. Kontonatsios, G., Korkontzelos, I., Tsujii, J., Ananiadou, S.: Combining String and Context Similarity for Bilingual Term Alignment from Comparable Corpora. In: EMNLP. pp. 1701–1712 (2014)
11. Laroche, A., Langlais, P.: Revisiting Context-based Projection Methods for Term-translation Spotting in Comparable Corpora. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 617–625. COLING '10, Association for Computational Linguistics (2010)
12. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics 3, 211–225 (2015)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
14. Mikolov, T., Le, Q.V., Sutskever, I.: Exploiting similarities among languages for machine translation (2013)
15. Morin, E., Prochasson, E.: Bilingual lexicon extraction from comparable corpora enhanced with parallel corpora. In: Proceedings of the 4th workshop on building and using comparable corpora: comparable corpora and the web. pp. 27–34. Association for Computational Linguistics (2011)
16. Otero, P.G.: Learning bilingual lexicons from comparable english and spanish corpora. Proceedings of MT Summit xI pp. 191–198 (2007)
17. Patry, A., Langlais, P.: Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. In: Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web. pp. 87–95. BUCC '11, Association for Computational Linguistics (2011)
18. Pekar, V., Mitkov, R., Blagoev, D., Mulloni, A.: Finding translations for low-frequency words in comparable corpora. Machine Translation 20(4), 247–266 (Mar 2006)
19. Prochasson, E., Fung, P.: Rare Word Translation Extraction from Aligned Comparable Documents. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1327–1335. HLT '11, Association for Computational Linguistics (2011)
20. Rapp, R.: Identifying Word Translations in Non-parallel Texts. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. pp. 320–322. ACL '95, Association for Computational Linguistics (1995)
21. Sharoff, S., Rapp, R., Zweigenbaum, P.: Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In: Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. (eds.) Building and Using Comparable Corpora, pp. 1–17. Springer Berlin Heidelberg (Jan 2013)
22. Yu, K., Tsujii, J.: Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. In: 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers. pp. 121–124 (2009)