Bilingual Word Embeddings for Bilingual Terminology Extraction from Specialized Comparable Corpora

Amir Hazem[1] **Emmanuel Morin**[1]

[1] LS2N - UMR CNRS 6004, Université de Nantes, France
{amir.hazem,emmanuel.morin}@univ-nantes.fr

Abstract

Bilingual lexicon extraction from comparable corpora is constrained by the small amount of available data when dealing with specialized domains. This aspect penalizes the performance of distributional-based approaches, which is closely related to the reliability of word's cooccurrence counts extracted from comparable corpora. A solution to avoid this limitation is to associate external resources with the comparable corpus. Since bilingual word embeddings have recently shown efficient models for learning bilingual distributed representation of words, we explore different word embedding models and show how a general-domain comparable corpus can enrich a specialized comparable corpus via neural networks.

1 Introduction

Bilingual lexicon extraction from comparable corpora has shown substantial growth since the seminal work of Fung (1995) and Rapp (1995). Comparable corpora, which are comprised of texts sharing common features such as domain, genre, sampling period, etc. and without having a source text/target text relationship (McEnery and Xiao, 2007), are more abundant and reliable resources than parallel corpora. On the one hand, parallel corpora are difficult to obtain for language pairs not involving English. On the other hand, as parallel corpora are comprised of a pair of translated texts, the vocabulary appearing in the translated texts is highly influenced by the source texts. These problems are aggravated in specialized and technical domains.

Although it is easier to build large general-domain comparable corpora, specialized comparable corpora are often of modest size (around 1 million words) due to the difficulty to obtain many specialized documents in a language other than English. Consequently, word co-occurrence counts of the historical context-based projection approach, known as the *standard approach* (Fung, 1995; Rapp, 1995), dedicated to bilingual lexicon extraction from comparable corpora are unreliable in specialized domain. This problem persists with other paradigms such as Canonical Correlation Analysis (CCA) (Gaussier et al., 2004), Independent Component Analysis (ICA) (Hazem and Morin, 2012) and Bilingual Latent Dirichlet Allocation (BiLDA) (Vulić et al., 2011).

A solution to avoid this limitation and to increase the representativity of distributional representations is to associate external resources with the specialized comparable corpus. These resources can be lexical databases such as WordNet which allows the disambiguation of translations of polysemous words (Bouamor et al., 2013) or general-domain data to improve word co-occurrence counts of specialized comparable corpora (Hazem and Morin, 2016).

Our work is in this line and attempts to find out how a general-domain data can enrich a specialized comparable corpora to improve bilingual terminology extraction from specialized comparable corpora. Since bilingual word embeddings have recently provided efficient models for learning bilingual distributed representation of words from large general-domain data (Mikolov et al., 2013), we contrast different popular word embedding models for this task. In addition, we explore combinations of word embedding models as suggested by Garten et al. (2015) to improve distributed representations. We compare the results obtained with the state-of-the-art context-based projection approach. Our results show under which conditions the proposed model can compete with state-

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 685–693,
Taipei, Taiwan, November 27 – December 1, 2017

of-the art approaches. To the best of our knowledge, this is the first time that word embedding models have been used to extract bilingual lexicons from specialized comparable corpora.

The remainder of this paper is organized as follows. Section 2 presents the two state-of-the-art approaches used to extract bilingual lexicons from comparable corpora. Section 3 describes the two data combination approaches adapted to Skip-gram and CBOW models (Mikolov et al., 2013). Section 4 describes the different linguistic resources used in our experiments. Section 5 is then devoted to a large-scale evaluation of the different proposed methods. Finally, Section 6 presents our conclusion.

2 State-of-the-Art Approaches

In this section, we describe the two state-of-the-art approaches used to extract bilingual lexicons from comparable corpora.

These approaches are both based on monolingual lexical context analysis and relies on the distributional hypothesis (Harris, 1968) which postulates that a word and its translation tend to appear in the same lexical contexts. This is the hypothesis that tends to be reduced to the famous sentence of the British linguist J. R. Firth (1957, p. 11) who said: "*You shall know a word by the company it keeps.*" even if the context was related to collocates.

The two approaches are known as distributional and distributed semantics (according to Hermann and Blunsom (2014)). The first one is based on vector space models while the second one is based on neural language models.

2.1 Context-Based Projection Approach

The historical context-based projection approach, known as the standard approach, has been studied by a number of researchers (Fung, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Morin et al., 2007; Prochasson and Fung, 2011; Bouamor et al., 2013; Morin and Hazem, 2016, among others). Its implementation can be carried out by applying the following steps:

1. For each word w of the source and the target languages, we build a context vector (resp. $\mathbf{s}$ and $\mathbf{t}$ for source and target languages) consisting in the measure of association of each word that appears in a short window of words around w. The association measures traditionally studied are Mutual Information, Log-likelihood, and the Discounted Odds-Ratio.

2. For a word i to be translated, its context vector $\mathbf{i}$ is projected from the source to the target language by translating each element of its context vector thanks to a bilingual seed lexicon.

3. The translated context vector $\bar{\mathbf{i}}$ is compared to each context vector $\mathbf{t}$ of the target language using a similarity measure such as Cosine or weighted Jaccard. The candidate translations are then ranked according to the scores of a given similarity measure.

This approach is very sensitive to the choice of parameters. We invite readers to consult the study of Laroche and Langlais (2010) in which the influence of parameters such as the size of the context, the choice of the association and similarity measures have been examined.

In order to improve the quality of bilingual terminology extraction from specialized comparable corpora, Hazem and Morin (2016) have proposed two ways to combine specialized comparable corpora with external resources. The hypothesis is that word co-occurrences learned from a large general-domain corpus for general words improve the characterisation of the specific vocabulary of the specialized corpus. The first adaptation called Global Standard Approach (GSA) is basic and consists in building the context vectors from a comparable corpus composed of the specialized and the general comparable corpora. The second adaptation called Selective Standard Approach (SSA) is more sophisticated and comprises: i) independently building context vectors of the specialized and general comparable corpus and then ii) merging under certain conditions, specialized and general context vectors of the words belonging to the specialized corpus when they appear in the general corpus.

2.2 Word Embedding Based Approach

Bilingual word embeddings has become a source of great interest in recent times (Mikolov et al., 2013; Vulić and Moens, 2013; Zou et al., 2013; Chandar et al., 2014; Gouws et al., 2014; Artetxe et al., 2016, among others). Mikolov et al. (2013) was the first to propose a method to learn a linear transformation from the source language to the

target language to improve the task of lexicon extraction from bilingual corpora.

During the training time of Mikolov's method, for all $\{x_i, z_i\}_{i=1}^{n}$ bilingual word pairs of the seed lexicon, the word embedding $x_i \in \mathbb{R}^{d_1}$ of word i in the source language and the word embedding $z_i \in \mathbb{R}^{d_2}$ of its translation in the target language are computed. A transformation matrix W such as Wx_i approximates z_i is then learned by the objective function as follows:

$$\min_{W} \sum_{i=1}^{n} \|Wx_i - z_i\|^2 \quad (1)$$

At prediction time, we can transfer the word embedding x for a word to be translated in the target language using the translation matrix such as $z = Wx$. The candidate translations are obtained by ranking the closest target words to z according to a similarity measure such as the Cosine measure.

Recently, Artetxe et al. (2016) presented an approach for learning bilingual mappings of word embeddings that preserves monolingual invariance using several meaningful and intuitive constraints related to other proposed methods (Faruqui and Dyer, 2014; Xing et al., 2015). These constraints are orthogonality, vectors length normalization for maximum cosine and mean centering for maximum covariance. Monolingual invariance tends to preserve the dot products after mapping, in order to avoid performance drop in monolingual tasks, while dimension-wise mean centering tends to insure that two randomly taken words would not be semantically related. This approach has shown meaningful improvements for both monolingual and bilingual tasks.

Other work has focused on learning bilingual word representations without word-to-word alignments of comparable corpora. Chandar et al. (2014) and Gouws et al. (2014) use multilingual word embeddings based on sentence-aligned parallel data whereas Vulić and Moens (2015, 2016) propose a model to induce bilingual word embeddings directly from document-aligned non-parallel data. Theses works are based on sentence- or document-aligned of general-domain comparable corpora and are outside the scope of this study. It is unlikely, not to say impossible, to find this type of alignment in a specialized comparable corpus.

3 Data Combination Using Neural Networks

Recently, Hazem and Morin (2016) have shown that using external data drastically improves the performance of the traditional distributional-based approach for the task of bilingual lexicon extraction from specialized comparable corpora. Mikolov et al. (2013) have also shown that distributed vector representations over large corpora in a continuous space model capture many linguistic regularities and key aspects of words. Based on these findings, we pursue the preceding works and propose different ways to combine specialized and general domain data using neural network models. We adapt the two data combination approaches proposed in Hazem and Morin (2016) (see Section 2.1) using Skip-gram and CBOW models (Mikolov et al., 2013). Inspired by the work of Garten et al. (2015) which studied different combinations of distributed vector representations for word analogy task, we also propose different Skip-gram and CBOW models combinations over specialized and general domain data.

3.1 Global Data Combination Using Neural Network Models

This approach can be seen as similar to the GSA approach (Hazem and Morin, 2016), the difference is that instead of using the context-based projection approach to build context vectors, we use the distributed Skip-gram or Continuous Bag-of-Words (CBOW) models proposed in Mikolov et al. (2013). Given a specialized and a general domain corpus, we create a new corpus which is the combination of both. We then learn a Skip-gram model (respectively a CBOW model) using this new generated corpus. We denote this approach by GSG for the global[1] Skip-gram model and GCBOW for the global CBOW model.

After combining the two corpora, the steps for extracting bilingual lexicons are as follows:

1. We first build a CBOW (respectively a Skip-gram) model for source and target languages.

2. Then, we apply bilingual mapping (Artetxe et al., 2016) between the source and the target CBOW models (respectively the Skip-gram models). The mapping step needs a bilingual dictionary to compute the mapping ma-

[1]The term *global* refers to a global combination of data without any specific criterion.

trix. We used a dictionary subset of the 5,000 more frequent translation pairs. Different sizes of the seed dictionary have been studied and discussed in Jakubina and Langlais (2016). It should be noted that in our experiments, no great impact has been observed when varying the size of the dictionary subset.

3. For each word to be translated, we compute a Cosine similarity between its mapped embedding vector and the embedding vectors of all the target words.

4. Finally, we rank the candidates according to their similarity score.

3.2 Specific Data Combination Using Neural Network Models

This approach is in the line of the SSA (Hazem and Morin, 2016) approach but the idea is not exactly the same. Similarly to them we build two separate representations. One learned from a specialized domain corpus and the second learned from a general-domain corpus, but unlike them we concatenate the distributed models while they merge [2] distributional context vectors.

Our goal is to capture the two word characterisations thanks to CBOW/Skipgram models. One is issued from the specialized domain and the other one from the general domain, to finally combine both representations in the perspective of obtaining a better word representation. Our approach is as follows:

1. We first build a CBOW (respectively a Skip-gram) model for both specialized and general domain source and target languages.

2. Then, we concatenate source CBOW vectors (respectively Skip-gram vectors) of the specialized and the general domain data. We apply the same process for specialized and general-domain target data.

3. We apply bilingual mapping (Artetxe et al., 2016) between the source and target concatenated vectors.

4. For each word to be translated, we compute a Cosine similarity between its mapped embedding vector and the embedding vectors of all the target words.

5. Finally, we rank the candidates according to their similarity score.

3.3 Combining Distributed Representations

We follow the findings of Garten et al. (2015) where they have shown substantial improvements on a standard word analogy task, combining distributed vector representations (more specifically, vectors concatenation). They compared their hybrid methods and have shown their advantages especially when training data is limited, which is the main problem in the task of extracting bilingual terminology from specialized comparable corpora.

In our case, word embedding models lead to three different ways of concatenation. The first one is a CBOW model concatenation between the specialized and the general domain data. The second one is a Skip-gram model concatenation and the third one is a concatenation of both CBOW and Skip-gram models.

If for instance we have a 100 dimension specialized CBOW model and a 200 dimension general domain CBOW model. The concatenation will lead to a resulting 300 dimension CBOW model. If we also have a 100 dimension specialized Skip-gram model and a 200 dimension general domain Skip-gram model. The concatenation will lead to a resulting 300 dimension Skip-gram model. Finally, if we concatenate the CBOW and Skip-gram models, this will result in a 600 dimension combined model. This final concatenation process allows to take advantage of both CBOW and Skip-gram models and to learn a mapping matrix of the combined models. To our knowledge this is the first attempt to first encode CBOW⌒CBOW[3], Skip-gram⌒Skip-gram and CBOW⌒Skip-gram models before learning a bilingual mapping matrix.

4 Data and Resources

In this section, we briefly outline the different textual resources used for our experiments: the comparable corpora, the bilingual dictionary and the terminology reference list. These textual resources are a subset of the data used in Hazem and Morin (2016).

[2] The merging process can be seen as a simple vector addition.

[3] CBOW⌒CBOW stands for the concatenation of two CBOW vectors issued from two different training datasets.

4.1 Comparable corpora

The specialized comparable corpus consists of scientific papers collected from the Elsevier website[4]. The scientific papers were taken from the medical domain within the sub-domain of "breast cancer". The breast cancer comparable corpus (BC) is composed of 103 English documents and 130 French documents.

The four general-domain comparable corpora are of different types and sizes often used in multiple evaluation campaigns such as WMT. News commentary corpus consists of political and economic commentary crawled from the web (NC), Europarl corpus is a parallel corpus extracted from the proceedings of the European Parliament (EP7), JRC acquis corpus is a collection of legislative European Union documents (JRC) and Common Crawl corpus (CC) which encompasses over petabytes of web crawled data collected over seven years. It should be noted that we do not take advantage of the parallel information encoded in the parallel corpora.

Table 1 shows the number of content words (# content words) for each corpus.

Comparable corpus	# content words	
	FR	EN
BC	8,221	79.07
NC	5.7M	4.7M
EP7	61.8M	55.7M
JRC	70.3M	64.2M
CC	91.3M	81.1M

Table 1: Characteristics of the specialized comparable corpus and the external data.

The documents were pre-processed through basic linguistic steps such as tokenization, part-of-speech tagging and lemmatization using the TTC TermSuite[5] tool that applies the same method to several languages including English and French. Finally, the function words were removed thanks to a stopword list and the hapax [6] were discarded.

4.2 Bilingual Dictionary

The bilingual dictionary is the French/English dictionary ELRA-M0033[7] (243,539 entries). This resource is a general language dictionary which contains only a few terms related to the medical domain.

4.3 Gold Standard

The bilingual terminology reference list required to evaluate the quality of bilingual terminology extraction from comparable corpora has been derived from the UMLS[8] meta-thesaurus. The terminology reference list is composed of 248 single word pairs for which each word appears at least 5 times in the comparable corpus. This list is of a standard size compared to other works such as Chiao and Zweigenbaum (2002): 95 single words, Morin et al. (2007): 100 single words and Bouamor et al. (2013): 125 and 79 single words.

5 Experiments and Results

The first piece of work comparing methods for identifying translation pairs in comparable corpora was presented in Jakubina and Langlais (2016). However the evaluation was conducted on Wikipedia, which is a general domain corpus. In our case, we are interested in specialized domains where there is a lack of specialized data. Our experiments aim at exploring word embeddings performance in specialized comparable corpora, which is to our knowledge, the first attempt at tackling this problem. Moreover, and as it has been pointed out in Mikolov et al. (2013), applications to low resource domains is a very interesting topic where there is still much to be explored.

In this section, we compare different word embedding representations for the extraction of bilingual terms from specialized comparable corpora. We contrast Skip-gram and CBOW models as well as different ways of combining them over specialized and general domain corpora.

5.1 Word2vec

For word2vec, we used as settings a window size of 10, negative sampling of 5, sampling of 1e-3 and training over 15 iterations. We applied both Skip-gram and CBOW models[9] to create vectors of 100 dimensions. We used hierarchical softmax for training the Skip-gram model. Other settings were assessed but on average the chosen ones tend to give the best results on our data.

[4] www.elsevier.com

[5] code.google.com/p/ttc-project

[6] Tokens that appear only once in the corpus.

[7] catalog.elra.info/product_info.php?products_id=666

[8] www.nlm.nih.gov/research/umls

[9] To train word embedding models we used the gensim toolkit (Rehurek and Sojka, 2010)

Corpus	*CBOW*	*SG*	*Concat*
BC	17.1	12.8	20.8
NC	33.9	31.2	33.6
EP7	42.3	40.8	43.1
JRC	40.3	40.5	43.4
CC	60.9	56.0	**61.0**
BC ∪ NC	42.9	37.7	46.3
BC ∪ EP7	47.2	49.0	53.3
BC ∪ JRC	49.9	46.5	53.0
BC ∪ CC	67.7	63.2	**68.4**
BC⌢(BC ∪ NC)	45.5	30.7	48.1
BC⌢(BC ∪ EP7)	51.6	35.7	53.8
BC⌢(BC ∪ JRC)	53.7	36.3	56.1
BC⌢(BC ∪ CC)	70.7	40.2	**70.9**

Table 2: Results (MAP %) of word2vec using the Skip-gram model (noted SG), the Continuous Bag of Words model (noted CBOW) and the concatenation of both models (noted Concat). The window size was set to 10 and the vector size to 100.

5.2 Bilingual Mappings of Word Embeddings

For mapping words of the source language to the target language we follow the method presented in Artetxe et al. (2016) where they presented an efficient exact method to learn the optimal linear transformation that gives the best results in translation induction. While we contrasted different configurations of there framework, we only present the best results. We used the orthogonal mapping with length normalization and mean centering of vectors[10].

5.3 Results

Table 2 shows the results of word2vec according to different configurations. The first column represents the corpora that have been used to train word2vec models. The first bloc lines compares the performance of word2vec models on the specialized breast cancer corpus (BC) and the four external data that are: news commentary (NC), EuroParl (EP7), JRC acquis (JRC) and common crawl (CC). The second bloc lines compares the models trained on the combination of BC with each external corpus. For instance BC ∪ EP7 consists of training word2vec on the combination of BC and EP7. Finally, the third bloc lines shows the concatenation of two models, the first one trained on the specialized corpus (BC) and the second one trained on the combination of BC with a given external corpus (BC ∪ EP7 for instance). This is noted by BC⌢(BC∪EP7) for the concatenation (represented by ⌢) of vectors of BC with vectors of BC ∪ EP7. The first column of the second and third blocs shows CBOW concatenation while the second shows Skip-gram concatenation. The third column shows CBOW and Skip-gram concatenation. Concatenate BC with BC ∪ EP7 instead of BC with EP7 (noted BC⌢EP7) for instance, insures the presence of all specialized domain words in both models.

The first observation that can be seen from Table 2 is that the results are very low when using the specialized BC corpus only (the maximum obtained MAP is 20.8% for the Concat model). The second observation is that using external data only gives better results. The best obtained MAP score is 61.0% when using CC corpus. This is certainly due to the presence of medical terms in the general domain corpora which have been crawled from the web and can contain scientific pages. Over the first bloc lines results, we can see that CBOW model outperforms Skip-gram model in most cases which is not surprising in the sense that CBOW aims at characterising a word according to its context while Skip-gram characterises a context according to a given word. From this point CBOW is more appropriate for our task. However, combining both CBOW and Skip-gram as shown by the Concat model, always improves the MAP scores. This is an important finding that shows that both models are complementary.

According to the second bloc lines results, we

[10]These parameters have also shown the best results in Artetxe et al. (2016).

Word2vec + Mapping	BC_{concat}	CC_{concat}
$Unconstrained + Original$	11.4	48.0
$Orthogonal + Original$	19.5	69.9
$Unconstrained + Unit$	11.8	50.4
$Orthogonal + Unit$	19.0	70.1
$Unconstrained + Unit + Center$	12.3	54.0
$Orthogonal + Unit + Center$	**20.8**	**70.9**

Table 3: Results (MAP %) of word2vec using different mapping techniques.

	BC	NC	EP7	JRC	CC
SA	27.0	45.3	48.5	52.0	75.5
GSA	-	**58.9**	58.3	61.7	80.2
SSA	-	**58.9**	**60.8**	**66.6**	**82.3**
$GCBOW$	17.1	42.9	47.2	49.9	67.7
GSG	12.8	37.7	49.2	46.5	63.2
$GCBOW + GSG$	20.8	46.3	53.3	53.0	68.4
$SCBOW$	-	45.5	51.6	53.7	70.7
SSG	-	30.7	35.7	36.3	40.2
$SCBOW + SSG$	-	48.1	53.8	56.1	70.9

Table 4: Results (MAP %) of the *Standard Approach* (SA), the *Global Standard Approach* (GSA) and the *Selective Standard Approach* (SSA) for the breast cancer corpus (BC) using the different external data (the improvements indicate a significance at the 0.001 level using the Student t-test).

observe that combining the specialized bilingual corpus (BC) with external data always improves the results. This can be noticed for the BC $\cup$ CC corpus where we increase the MAP score for CBOW from 60.9% to 67.7%, the MAP score for Skip-gram from 56% to 63.2% and the MAP score for the Concat approach from 61% to 68.4%. These are also interesting results which coincide with the observations of Hazem and Morin (2016). Hence, combining specific and general domain corpora before applying any model always benefits the task of bilingual terminology extraction.

Finally, for the third bloc lines where we combine the model issued from the first bloc (the BC model) and models issued from the second bloc lines (BC$\cup$EP7 for instance), two observations need to be pointed out. The first one is that for CBOW model concatenation we still get improvements as we can see for CC where we gain 3 points (from 67.7% to 70.7% of MAP score). The second surprising observation is that Skip-gram concatenation decreases the results. This drop may suggest that Skip-gram concatenation is not appropriate for this configuration. However the concatenation of CBOW and Skip-gram models (the Concat approach) still improves the results as we can see for BC$^\frown$(BC $\cup$ JRC) where we move from 53% to 56.1% of MAP score and for BC$^\frown$(BC $\cup$ CC) where we increase the MAP score from 68.4% to 70.9%. For this last result the gain is not that important compared to the CBOW$^\frown$CBOW model that obtains a MAP score of 70.7%.

In Table 3 we report a comparison of different mappings as studied in Artetxe et al. (2016). We contrast unconstrained and orthogonal mappings using original as well as length normalization (noted $unit$) and mean centering (noted $Center$). As we can see, the best results are those obtained using orthogonal mapping with length normalization and mean centering. We reach the same conclusions as Artetxe et al. (2016).

In Table 4 we present the results from Hazem and Morin (2016) of the standard approach (noted SA) using only specialized comparable corpora (BC) or using only external data (NC, EP7, JRC and CC), in addition to the two adapted standard approaches (noted GSA and SSA) using the combination of each specialized comparable corpus with each corpus of the external data. We also

report the best results of each of the three word embedding methods that we introduced earlier in Table 2.

Even if we obtained improvements over our different embedding models, we are still below the standard approach as seen in Table 4. The lack of specialized data may partially explain these lower results as well as the multiple tuning parameters of CBOW and Skip-gram models.

6 Conclusion

In this paper, we have proposed and contrasted different data combinations using neural networks for bilingual terminology extraction from specialized comparable corpora. We have shown under which conditions external resources as well as Skip-gram and CBOW models can be jointly used to improve the performance of bilingual terms extraction. If the results are encouraging, we were unable to compete with the results of the historical context-based projection approach. However, our findings may suggest a starting point of applying word embeddings and the multiple proposed variants to specialized domains as well as to other tasks. We hope this work will help to lead the way in exploring low resource domains such as specialized comparable corpora.

Acknowledgments

The research leading to these results has received funding from the French National Research Agency under grant ANR-17-CE23-0001 ADDICTE (Distributional analysis in specialized domain).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Dhouha Bouamor, Nasredine Semmar, and Pierre Zweigenbaum. 2013. Context vector disambiguation for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 759–764, Sofia, Bulgaria.

A. P. Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas C. Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS'14)*, pages 1853–1861, Montreal, Quebec, Canada.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING'02)*, pages 1208–1212, Tapei, Taiwan.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL'14)*, pages 462–471, Gothenburg, Sweden.

John Rupert Firth. 1957. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32. Blackwell, Oxford.

Pascale Fung. 1995. Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In *Proceedings of the 3rd Annual Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Cambridge, MA, USA.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98)*, pages 1–17, Langhorne, PA, USA.

Justin Garten, Kenji Sagae, Volkan Ustun, and Morteza Dehghani. 2015. Combining distributed vector representations for words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 95–101, Denver, CO, USA.

Eric. Gaussier, Jean-Michel Renders, Irena. Matveeva, Cyril. Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, pages 526–533, Barcelona, Spain.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *CoRR*, abs/1410.2455.

Z. S. Harris. 1968. *Mathematical Structures of Language*. Interscience Publishers.

Amir Hazem and Emmanuel Morin. 2012. Adaptive dictionary for bilingual lexicon extraction from comparable corpora. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 288–292, Istanbul, Turkey.

Amir Hazem and Emmanuel Morin. 2016. Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 3401–3411, Osaka, Japan.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 58–68, Baltimore, Maryland.

Laurent Jakubina and Phillippe Langlais. 2016. A comparison of methods for identifying the translation of words in a comparable corpus: Recipes and limits. *Computación y Sistemas*, 20(3):449–458.

Audrey Laroche and Philippe Langlais. 2010. Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 617–625, Beijing, China.

Anthony McEnery and Zhonghua Xiao. 2007. Parallel and comparable corpora: What are they up to? In Gunilla Anderman and Margaret Rogers, editors, *Incorporating Corpora: Translation and the Linguist*, Multilingual Matters, chapter 2, pages 18–31. Clevedon, UK.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Emmanuel Morin, Béatrice Daille, Koichi Takeuchi, and Kyo Kageura. 2007. Bilingual Terminology Mining – Using Brain, not brawn comparable corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 664–671, Prague, Czech Republic.

Emmanuel Morin and Amir Hazem. 2016. Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. *Natural Language Engineering*, 22(4):575–601.

Emmanuel Prochasson and Pascale Fung. 2011. Rare Word Translation Extraction from Aligned Comparable Documents. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 1327–1335, Portland, OR, USA.

Reinhard Rapp. 1995. Identify Word Translations in Non-Parallel Texts. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'95)*, pages 320–322, Boston, MA, USA.

Reinhard Rapp. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pages 519–526, College Park, MD, USA.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, pages 479–484, Portland, OR, USA.

Ivan Vulić and Marie-Francine Moens. 2013. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1613–1624, Seattle, WA, USA.

Ivan Vulić and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15)*, pages 719–725, Beijing, China.

Ivan Vulić and Marie-Francine Moens. 2016. Bilingual distributed word representations from document-aligned comparable data. *Journal of Artificial Intelligence Research (JAIR)*, 55(1):953–994.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL'15)*, pages 1006–1011, Denver, CO, USA.

Will Y. Zou, Richard Socher, Daniel M. Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1393–1398, Seattle, WA, USA.