Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction

Ivan Vulić and Marie-Francine Moens
Department of Computer Science
KU Leuven, Belgium
{ivan.vulic|marie-francine.moens}@cs.kuleuven.be

Abstract

We propose a simple yet effective approach to learning bilingual word embeddings (BWEs) from non-parallel document-aligned data (based on the omnipresent skip-gram model), and its application to bilingual lexicon induction (BLI). We demonstrate the utility of the induced BWEs in the BLI task by reporting on benchmarking BLI datasets for three language pairs: (1) We show that our BWE-based BLI models significantly outperform the MuPTM-based and context-counting models in this setting, and obtain the best reported BLI results for all three tested language pairs; (2) We also show that our BWE-based BLI models outperform other BLI models based on recently proposed BWEs that require parallel data for bilingual training.

1 Introduction

Dense real-valued vectors known as distributed representations of words or *word embeddings* (WEs) (Bengio et al., 2003; Collobert and Weston, 2008; Mikolov et al., 2013a; Pennington et al., 2014) have been introduced recently as part of neural network architectures for statistical language modeling. Recent studies (Levy and Goldberg, 2014; Levy et al., 2015) have showcased a direct link and comparable performance to "more traditional" distributional models (Turney and Pantel, 2010), but the skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013c) is still established as the state-of-the-art word representation model, due to its simplicity, fast training, as well as its solid and robust performance across a wide variety of semantic tasks (Baroni et al., 2014; Levy et al., 2015).

A natural extension of interest from monolingual to multilingual word embeddings has occurred recently (Klementiev et al., 2012; Zou et al., 2013; Mikolov et al., 2013b; Hermann and Blunsom, 2014a; Hermann and Blunsom, 2014b; Gouws et al., 2014; Chandar et al., 2014; Soyer et al., 2015; Luong et al., 2015). When operating in multilingual settings, it is highly desirable to learn embeddings for words denoting similar concepts that are very close in the *shared inter-lingual embedding space* (e.g., the representations for the English word *school* and the Spanish word *escuela* should be very similar). These shared interlingual embedding spaces may then be used in a myriad of multilingual natural language processing tasks, such as fundamental tasks of computing cross-lingual and multilingual semantic word similarity and *bilingual lexicon induction (BLI)*, etc. However, all these models critically require at least sentence-aligned parallel data and/or readily-available translation dictionaries to induce *bilingual word embeddings* (BWEs) that are consistent and closely aligned over languages in the same semantic space.

Contributions In this work, we alleviate the requirements: (1) We present the first model that is able to induce bilingual word embeddings from non-parallel data without any other readily available translation resources such as pre-given bilingual lexicons; (2) We demonstrate the utility of BWEs induced by this simple yet effective model in the BLI task from comparable Wikipedia data on benchmarking datasets for three language pairs (Vulić and Moens, 2013b). Our BLI model based on our novel BWEs significantly outperforms a series of strong baselines that reported previous best scores on these datasets in the same learning setting, as well as other BLI models based on recently proposed BWE induction models (Gouws et al., 2014; Chandar et al., 2014). The focus of the work is on learning lexicons from document-aligned comparable corpora (e.g., Wikipedia articles aligned through inter-wiki links).

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 719–725,
Beijing, China, July 26-31, 2015.

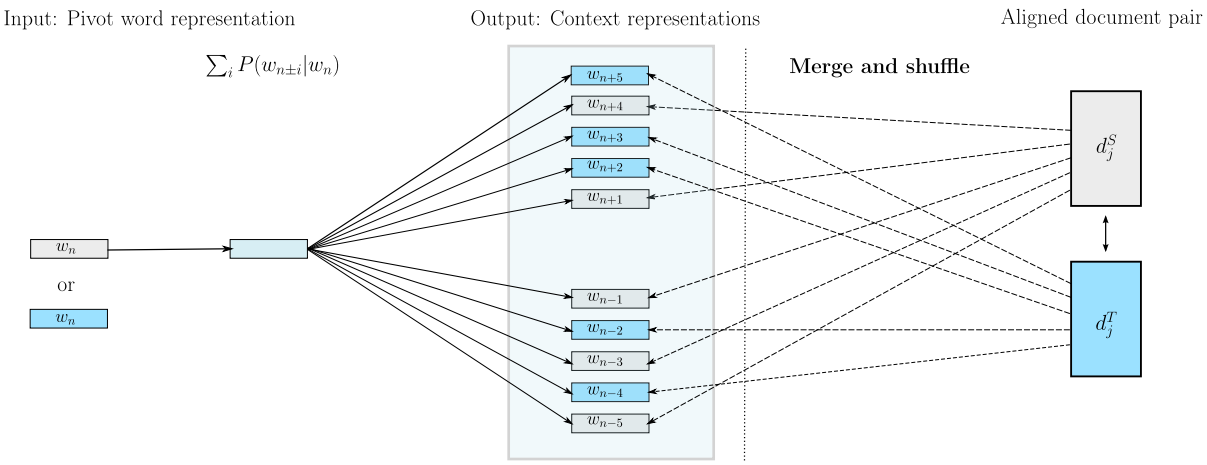


Figure 1: The architecture of our BWE Skip-Gram model for learning bilingual word embeddings from document-aligned comparable data. Source language words and documents are drawn as gray boxes, while target language words and documents are drawn as blue boxes. The right side of the figure (separated by a vertical dashed line) illustrates how a pseudo-bilingual document is constructed from a pair of two aligned documents; two documents are first merged, and then words in the pseudo-bilingual document are randomly shuffled to ensure that both source and target language words occur as context words.

2 Model Architecture

In the following architecture description, we assume that the reader is familiar with the main assumptions and training procedure of SGNS (Mikolov et al., 2013a; Mikolov et al., 2013c). We extend the SGNS model to work with bilingual document-aligned comparable data. An overview of our architecture for learning BWEs from such comparable data is given in fig. 1.

Let us assume that we possess a document-aligned comparable corpus which is defined as $\mathcal{C} = \{d_1, d_2, \ldots, d_N\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \ldots, (d_N^S, d_D^T)\}$, where $d_j = (d_j^S, d_j^T)$ denotes a pair of aligned documents in the source language L_S and the target language L_T, respectively, and N is the number of documents in the corpus. V^S and V^T are vocabularies associated with languages L_S and L_T. The goal is to learn word embeddings for all words in both V^S and V^T that will be semantically coherent and closely aligned over languages in a shared cross-lingual word embedding space.

In the first step, we *merge* two documents d_j^S and d_j^T from the aligned document pair d_j into a single "pseudo-bilingual" document d_j' and remove sentence boundaries. Following that, we *randomly shuffle* the newly constructed pseudo-bilingual document. The intuition behind this pre-training completely random shuffling step[1] (see fig. 1) is to assure that each word w, regardless of its actual language, obtains word collocates from both vocabularies. The idea of having bilingual contexts for each pivot word in each pseudo-bilingual document will steer the final model towards constructing a shared inter-lingual embedding space. Since the model depends on the alignment at the document level, in order to ensure the bilingual contexts instead of monolingual contexts, it is intuitive to assume that larger window sizes will lead to better bilingual embeddings. We test this hypothesis and the effect of window size in sect. 4.

The final model called BWE Skip-Gram (BWESG) then relies on the monolingual variant of skip-gram trained on the shuffled pseudo-bilingual documents.[2] The model learns word embeddings for source and target language words that are aligned over the d embedding dimensions and may be represented in the same shared cross-lingual embedding space. The BWESG-based representation of word w, regardless of its actual language, is then a d-dimensional vector: $\vec{w} = [f_1, \ldots, f_k, \ldots, f_d]$, where $f_k \in \mathbb{R}$ denotes the score for the k-th inter-lingual feature within the d-dimensional shared embedding space. Since all words share the embedding space, semantic similarity between words may be computed both

[1] In this paper, we investigate only the random shuffling procedure and show that the model is fairly robust to different outputs of the procedure if the window size is large enough. As one line of future work, we plan to investigate other, more systematic and deterministic shuffling algorithms.

[2] We were also experimenting with GloVe and CBOW, but they were falling behind SGNS on average.

monolingually and across languages. Given w, the most similar word cross-lingually should be its one-to-one translation, and we may use this intuition to induce one-to-one bilingual lexicons from comparable data.

In another interpretation, BWESG actually builds BWEs based on (pseudo-bilingual) document level co-occurrence. The window size parameter then just controls the amount of random data dropout. With larger windows, the model becomes prohibitively computationally expensive, but in sect. 4 we show that the BLI performance flattens out for "reasonably large" windows.

3 Experimental Setup

Training Data We use comparable Wikipedia data introduced in (Vulić and Moens, 2013a; Vulić and Moens, 2013b) available in three language pairs to induce bilingual word embeddings: (i) a collection of 13,696 Spanish-English Wikipedia article pairs (ES-EN), (ii) a collection of 18,898 Italian-English Wikipedia article pairs (IT-EN), and (iii) a collection of 7,612 Dutch-English Wikipedia article pairs (NL-EN). All corpora are theme-aligned comparable corpora, that is, the aligned document pairs discuss similar themes, but are in general not direct translations. Following prior work (Haghighi et al., 2008; Prochasson and Fung, 2011; Vulić and Moens, 2013b), we retain only nouns that occur at least 5 times in the corpus. Lemmatized word forms are recorded when available, and original forms otherwise. TreeTagger (Schmid, 1994) is used for POS tagging and lemmatization. After the preprocessing vocabularies comprise between 7,000 and 13,000 noun types for each language in each language pair. Exactly the same training data and vocabularies are used to induce bilingual lexicons with all other BLI models in comparison.

BWESG Training Setup We have trained the BWESG model with random shuffling on 10 random corpora shuffles for all three training corpora with the following parameters from the `word2vec` package (Mikolov et al., 2013c): stochastic gradient descent with a default learning rate of 0.025, negative sampling with 25 samples, and a subsampling rate of value $1e-4$. All models are trained for 15 epochs. We have varied the number of embedding dimensions: $d = 100, 200, 300$, and have also trained the model with $d = 40$ to be directly comparable to pre-trained state-of-the-art BWEs from (Gouws et al., 2014; Chandar et al., 2014). Moreover, in order to test the effect of window size on final results, we have varied the maximum window size cs from 4 to 60 in steps of 4.[3] Since cosine is used for all similarity computations in the BLI task, we call our new BLI model *BWESG+cos*.

Baseline BLI Models We compare BWESG+cos to a series of state-of-the-art BLI models from document-aligned comparable data:

(1) *BiLDA-BLI* - A BLI model that relies on the induction of latent cross-lingual topics (Mimno et al., 2009) by the bilingual LDA model and represents words as probability distributions over these topics (Vulić et al., 2011).

(2) *Assoc-BLI* - A BLI model that represents words as vectors of association norms (Roller and Schulte im Walde, 2013) over both vocabularies, where these norms are computed using a multilingual topic model (Vulić and Moens, 2013a).

(3) *PPMI+cos* - A standard distributional model for BLI relying on positive pointwise mutual information and cosine similarity (Bullinaria and Levy, 2007). The seed lexicon is bootstrapped using the method from (Peirsman and Padó, 2011; Vulić and Moens, 2013b).

All parameters of the baseline BLI models (i.e., topic models and their settings, the number of dimensions K, feature pruning values, window size) are set to their optimal values according to suggestions in prior work (Steyvers and Griffiths, 2007; Vulić and Moens, 2013a; Vulić and Moens, 2013b; Kiela and Clark, 2014). Due to space constraints, for (much) more details about the baselines we point to the relevant literature (Peirsman and Padó, 2011; Tamura et al., 2012; Vulić and Moens, 2013a; Vulić and Moens, 2013b).

Test Data For each language pair, we evaluate on standard 1,000 ground truth one-to-one translation pairs built for the three language pairs (ES/IT/NL-EN) (Vulić and Moens, 2013a; Vulić and Moens, 2013b). Translation direction is ES/IT/NL $\rightarrow$ EN.

Evaluation Metrics Since we can build a one-to-one bilingual lexicon by harvesting one-to-one translation pairs, the lexicon qualiy is best reflected in the Acc_1 score, that is, the number of source language (ES/IT/NL) words w_i^S from ground truth translation pairs for which the top ranked word cross-lingually is the correct trans-

[3] We will make all our BWESG BWEs available at: `http://people.cs.kuleuven.be/~ivan.vulic/`

Spanish-English (ES-EN)			Italian-English (IT-EN)			Dutch-English (NL-EN)		
(1) **reina**	(2) **reina**	(3) **reina**	(1) **madre**	(2) **madre**	(3) **madre**	(1) **schilder**	(2) **schilder**	(3) **schilder**
(Spanish)	(English)	(Combined)	(Italian)	(English)	(Combined)	(Dutch)	(English)	(Combined)
rey	*queen(+)*	*queen(+)*	padre	*mother(+)*	*mother(+)*	kunstschilder	*painter(+)*	*painter(+)*
trono	*heir*	rey	moglie	*father*	padre	schilderij	*painting*	kunstschilder
monarca	*throne*	trono	sorella	*sister*	moglie	kunstenaar	*portrait*	*painting*
heredero	*king*	*heir*	figlia	*wife*	*father*	olieverf	*artist*	schilderij
matrimonio	*royal*	*throne*	figlio	*daughter*	sorella	olieverfschilderij	*canvas*	kunstenaar
hijo	*reign*	monarca	fratello	*son*	figlia	schilderen	*impressionist*	*portrait*
reino	*succession*	heredero	casa	*friend*	figlio	frans	*cubism*	olieverf
reinado	*princess*	*king*	amico	*childhood*	*sister*	nederlands	*art*	olieverfschilderij
regencia	*marriage*	matrimonio	marito	*family*	fratello	componist	*poet*	schilderen
duque	*prince*	*royal*	donna	*cousin*	*wife*	beeldhouwer	*drawing*	*artist*

Table 1: Example lists of top 10 semantically similar words for all 3 language pairs obtained using BWESG+cos; $d = 200, cs = 48$; (col 1.) only source language words (ES/IT/NL) are listed while target language words are skipped (monolingual similarity); (2) only target language words (EN) are listed (cross-lingual similarity); (3) words from both languages are listed (multilingual similarity). EN words are given in italic. The correct one-to-one translation for each source word is marked by (+).

lation in the other language (EN) according to the ground truth over the total number of ground truth translation pairs (=*1000*) (Gaussier et al., 2004; Tamura et al., 2012; Vulić and Moens, 2013b).

4 Results and Discussion

Exp 0: Qualitative Analysis Tab. 1 displays top 10 semantically similar words monolingually, across-languages and combined/multilingually for one ES, IT and NL word. The BWESG+cos model is able to find semantically coherent lists of words for all three directions of similarity (i.e., monolingual, cross-lingual, multilingual). In the combined (multilingual) ranked lists, words from both languages are represented as top similar words. This initial qualitative analysis already demonstrates the ability of BWESG to induce a shared cross-lingual embedding space using only document alignments as bilingual signals.

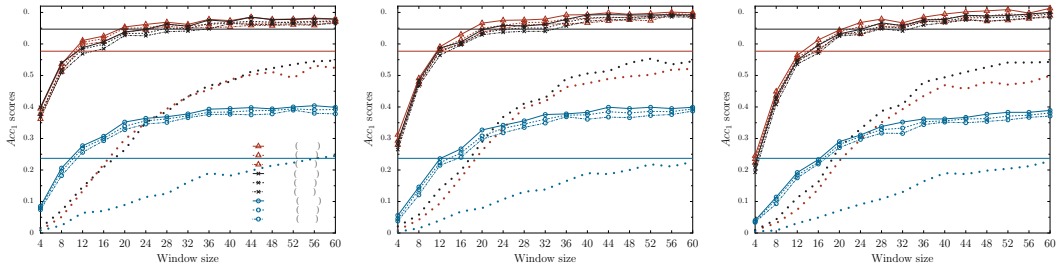
Exp I: BWESG+cos vs. Baseline Models In the first experiment, we test whether our BWESG+cos BLI model produces better results than the baseline BLI models which obtain current state-of-the-art results for BLI from comparable data on these test sets. Tab. 2 summarizes the BLI results.

As the most striking finding, the results reveal superior performance of the BWESG-cos model for BLI which relies on our new framework for inducing bilingual word embeddings over other BLI models relying on previously used bilingual word representations. The relative increase in Acc_1 scores over the best scoring baseline BLI models from comparable data is 19.4% for the ES-EN pair, 6.1% for IT-EN (significant at $p < 0.05$ using McNemar's test) and 65.4% for NL-EN. For large enough values for cs ($cs \geq 20$) (see also fig. 2(a)-2(c)), almost all BWESG+cos models for all language pairs outperform the highest baseline results. We may also observe that the performance of BWESG+cos is fairly stable for all models with larger values for cs ($cs \geq 20$). This finding reveals that even a coarse tuning of these parameters might lead to optimal or near-optimal scores in the BLI task with BWESG+cos.

Pair:	**ES-EN**	**IT-EN**	**NL-EN**
Model	Acc_1	Acc_1	Acc_1
BiLDA-BLI	0.441	0.575	0.237
Assoc-BLI	0.518	0.618	0.236
PPMI+cos	0.577	0.647	0.206
BWESG+cos			
d:100,cs:16	0.617	0.599	0.300
d:100,cs:48	0.667	0.669	0.389
d:200,cs:16	0.613	0.601	0.254
d:200,cs:48	0.685	**0.683**	**0.392**
d:300,cs:16	0.596	0.583	0.224
d:300,cs:48	**0.689**	**0.683**	0.363
d: 40,cs:16	0.558	0.533	0.266
d: 40,cs:48	0.578	0.595	0.308
CHANDAR	0.432	-	-
GOUWS	0.516	0.557	0.575

Table 2: BLI performance for all tested BLI models for ES/IT/NL-EN, with all bilingual word representations except CHANDAR and GOUWS learned from comparable Wikipedia data. The scores for BWESG+cos are computed as post-hoc averages over 10 random shuffles.

Exp II: Shuffling and Window Size Since our BWESG model relies on the pre-training random shuffling procedure, we also test whether the shuffling has significant or rather minor impact on the induction of BWEs and final BLI scores. Therefore, in fig. 2, we present maximum, minimum, and average Acc_1 scores for all three language pairs obtained using 10 different random corpora shuffles with $d = 100, 200, 300$ and varying val-



(a) $d = 100$ (b) $d = 200$ (c) $d = 300$

Figure 2: Maximum (MAX), minimum (MIN) and average (AVG) Acc_1 scores with BWESG+cos in the BLI task over 10 different random corpora shuffles for all 3 language pairs, and varying values for parameters cs and d. Solid horizontal lines denote the highest baseline Acc_1 scores for each language pair. NOS (thicker dotted lines) refers to BWESG+cos without random shuffling.

ues for cs. Results reveal that random shuffling affects the overall BLI scores, but the variance of results is minimal and often highly insignificant. It is important to mark that even the minimum Acc_1 scores over these 10 different random shuffles are typically higher than the previous state-of-the-art baseline scores for large enough values for d and cs (compare the results in tab. 2 and fig. 2(a)-2(c)). A comparison with the BWESG model without shuffling (NOS on fig. 2) reveals that shuffling is useful even for larger cs-s.

Exp III: BWESG+cos vs. BWE-Based BLI We also compare our BWESG BLI model with two other models that are most similar to ours in spirit, as they also induce shared cross-lingual word embedding spaces (Chandar et al., 2014; Gouws et al., 2014), proven superior to or on a par with the BLI model from (Mikolov et al., 2013b). We use their pre-trained BWEs (obtained from the authors) and report the BLI scores in tab. 2. To make the comparison fair, we search for translations over the same vocabulary as with all other models. The results clearly reveal that, although both other BWE models critically rely on parallel Europarl data for training, and Gouws et al. (2014) in addition train on entire monolingual Wikipedias in both languages, our simple BWE induction model trained on much smaller amounts of document-aligned non-parallel data produces significantly higher BLI scores for IT-EN and ES-EN with sufficiently large windows.

However, the results for NL-EN with all BLI models from comparable data from tab. 2 are significantly lower than with the GOUWS BWEs. We attribute it to using less (and clearly insufficient) document-aligned training data for NL-EN (i.e., training corpora for ES-EN and IT-EN are almost double or triple the size of training corpora for NL-EN, see sect. 3).

5 Conclusions and Future Work

We have proposed Bilingual Word Embeddings Skip-Gram (BWESG), a simple yet effective model that is able to learn bilingual word embeddings solely on the basis of document-aligned comparable data. We have demonstrated its utility in the task of bilingual lexicon induction from such comparable data, where our new BWESG-based BLI model outperforms state-of-the-art models for BLI from document-aligned comparable data and related BWE induction models.

The low-cost BWEs may be used in other (semantic) tasks besides the ones discussed here, and it would be interesting to experiment with other types of context aggregation and selection beyond random shuffling, and other objective functions. Preliminary studies also demonstrate the utility of the BWEs in monolingual and cross-lingual information retrieval (Vulić and Moens, 2015).

Finally, we may use the knowledge of BWEs obtained by BWESG from document-aligned data to learn bilingual correspondences (e.g., word translation pairs or lists of semantically similar words across languages) which may in turn be used for representation learning from large unaligned multilingual datasets as proposed in (Haghighi et al., 2008; Mikolov et al., 2013b; Vulić and Moens, 2013b). In the long run, this idea may lead to large-scale fully data-driven representation learning models from huge amounts of multilingual data without any "pre-requirement" for parallel data or manually built lexicons.

Acknowledgments

We would like to thank the reviewers for their insightful comments and suggestions. This research has been carried out in the frameworks of the SCATE project (IWT-SBO 130041) and the PARIS project (IWT-SBO 110067).

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, pages 238–247.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

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 *NIPS*, pages 1853–1861.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167.

Éric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Hervé Déjean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *ACL*, pages 526–533.

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

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, pages 771–779.

Karl Moritz Hermann and Phil Blunsom. 2014a. Multilingual distributed representations without word alignment. In *ICLR*.

Karl Moritz Hermann and Phil Blunsom. 2014b. Multilingual models for compositional distributed semantics. In *ACL*, pages 58–68.

Douwe Kiela and Stephen Clark. 2014. A systematic study of semantic vector space model parameters. In *CVSC Workshop at EACL*, pages 21–30.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *COLING*, pages 1459–1474.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *NIPS*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, 3:211–225.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR Workshop Papers*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

David Mimno, Hanna Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *EMNLP*, pages 880–889.

Yves Peirsman and Sebastian Padó. 2011. Semantic relations in bilingual lexicons. *ACM Transactions on Speech and Language Processing*, 8(2):article 3.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *ACL*, pages 1327–1335.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *EMNLP*, pages 1146–1157.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Hubert Soyer, Pontus Stenetorp, and Akiko Aizawa. 2015. Leveraging monolingual data for crosslingual compositional word representations. In *ICLR*.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of Latent Semantic Analysis*, 427(7):424–440.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2012. Bilingual lexicon extraction from comparable corpora using label propagation. In *EMNLP*, pages 24–36.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artifical Intelligence Research*, 37(1):141–188.

Ivan Vulić and Marie-Francine Moens. 2013a. Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *NAACL*, pages 106–116.

Ivan Vulić and Marie-Francine Moens. 2013b. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *EMNLP*, pages 1613–1624.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *SIGIR,* to appear.

Ivan Vulić, Wim De Smet, and Marie-Francine Moens. 2011. Identifying word translations from comparable corpora using latent topic models. In *ACL*, pages 479–484.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398.