Journal of Artificial Intelligence Research 55 (2016) 953-994 Submitted 09/15; published 04/16

Bilingual Distributed Word Representations from Document-Aligned Comparable Data

Ivan Vulić IV250@CAM.AC.UK
University of Cambridge
Department of Theoretical and Applied Linguistics
9 West Road, CB3 9DP, Cambridge, UK

Marie-Francine Moens MARIE-FRANCINE.MOENS@CS.KULEUVEN.BE
KU Leuven
Department of Computer Science
Celestijnenlaan 200A, 3001 Heverlee, Belgium

Abstract

We propose a new model for learning bilingual word representations from non-parallel document-aligned data. Following the recent advances in word representation learning, our model learns dense real-valued word vectors, that is, bilingual word embeddings (BWEs). Unlike prior work on inducing BWEs which heavily relied on parallel sentence-aligned corpora and/or readily available translation resources such as dictionaries, the article reveals that BWEs may be learned solely on the basis of document-aligned comparable data without any additional lexical resources nor syntactic information. We present a comparison of our approach with previous state-of-the-art models for learning bilingual word representations from comparable data that rely on the framework of multilingual probabilistic topic modeling (MuPTM), as well as with distributional local context-counting models. We demonstrate the utility of the induced BWEs in two semantic tasks: (1) bilingual lexicon extraction, (2) suggesting word translations in context for polysemous words. Our simple yet effective BWE-based models significantly outperform the MuPTM-based and context-counting representation models from comparable data as well as prior BWE-based models, and acquire the best reported results on both tasks for all three tested language pairs.

1. Introduction

A huge body of work in distributional semantics and word representation learning almost exclusively revolves around the *distributional hypothesis* (Harris, 1954) - an idea which states that similar words occur in similar contexts. All current corpus-based approaches to semantics rely on the contextual evidence in one way or another. Roughly speaking, word representations are typically learned using these two families of distributional context-based models: (1) global matrix factorization models such as latent semantic analysis (LSA) (Landauer & Dumais, 1997) or generative probabilistic models such as latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003), which model the word co-occurrence at the document or paragraph level; or (2) local context window models that represent words as sparse high-dimensional context vectors, and model the word co-occurrence at the level of selected neighboring words (Turney & Pantel, 2010), or generative probabilistic models that learn the probability distribution of a vocabulary word in the context window as a latent variable (Deschacht & Moens, 2009; Deschacht, De Belder, & Moens, 2012).

On the other hand, dense real-valued vectors known as distributed representations of words or *word embeddings* (WEs) (e.g., Bengio, Ducharme, Vincent, & Janvin, 2003; Collobert & Weston, 2008; Mikolov, Chen, Corrado, & Dean, 2013a; Pennington, Socher, & Manning, 2014) have been introduced recently, first as part of neural network based architectures for statistical language modeling. WEs serve as richer and more coherent word representations than the ones obtained by the aforementioned traditional distributional semantic models, with illustrative comparative studies available in the recently published relevant work (e.g., Mikolov, Yih, & Zweig, 2013d; Baroni, Dinu, & Kruszewski, 2014; Levy, Goldberg, & Dagan, 2015).

A natural extension of interest from monolingual to multilingual word embeddings has occurred recently (e.g., Klementiev, Titov, & Bhattarai, 2012; Hermann & Blunsom, 2014b). When operating in multilingual settings, it is highly desirable to learn embeddings for words denoting similar concepts that are very close in the *shared bilingual embedding space* (e.g., the representations for the English word *school* and the Spanish word *escuela* should be very similar). These BWEs may then be used in a myriad of multilingual natural language processing tasks and beyond, such as fundamental tasks leaning on such bilingual meaning representations, e.g., computing cross-lingual and multilingual semantic word similarity and extracting bilingual word lexicons using the induced bilingual embedding space (see Figure 1). However, all these models critically require (at least) sentence-aligned parallel data and readily-available translation dictionaries to induce *bilingual word embeddings* (BWEs) that are consistent and closely aligned over different languages.

1.1 Contributions

To the best of our knowledge, this article presents the first work to showcase that bilingual word embeddings may be induced directly on the basis of comparable data without any additional bilingual resources such as sentence-aligned parallel data or translation dictionaries. The focus is on document-aligned comparable corpora (e.g., Wikipedia articles aligned through inter-wiki links, news texts discussing the same theme).

Our new bilingual embedding learning model makes use of *pseudo-bilingual documents* constructed by merging the content of two coupled documents from a document pair, where we propose and evaluate two different strategies on how to construct such pseudo-bilingual documents: (1) *merge and randomly shuffle* strategy which randomly permutes words from both languages in each pseudo-bilingual document, and (2) *length-ratio shuffle* strategy, a deterministic method that retains monolingual word order while intermingling the words cross-lingually. These additional pre-training shuffling strategies ensure that both source language words and target language words occur in the contexts of each source and target language word. A monolingual model such as *skip-gram with negative sampling (SGNS)* from the `word2vec` package (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013c) is then trained on these "shuffled" pseudo-bilingual documents. By this procedure, we steer semantically similar words from different languages towards similar representations in the shared bilingual embedding space, and effectively use available bilingual contexts instead of monolingual ones. The model treats documents as bags-of-words (i.e., it does not include any syntactic information) and does not even rely on any sentence boundary information.

In summary, the main contributions of this article are:

(1) We present BWE Skip-Gram (BWESG), the first model that induces bilingual word embeddings directly from document-aligned non-parallel data. We test and evaluate two main variants of the model based on the pre-training shuffling step. The main strength of the presented model lies in its favourable trade-off between simplicity and effectiveness.
(2) We provide a qualitative and quantitative analysis of the model. We draw analogies and comparisons with prior work on inducing word representations from the same data type: document-aligned comparable corpora (e.g., models relying on the multilingual probabilistic topic modeling framework (MuPTM)).
(3) We demonstrate the utility of induced BWEs *at the word type level* in the task of bilingual lexicon extraction (BLE) from Wikipedia data for three language pairs. A BLE model based on our BWEs significantly outperforms MuPTM-based and context-counting BLE models, and acquires the best reported scores on the benchmarking BLE datasets.
(4) We demonstrate the utility of induced BWEs *at the word token level* in the task of suggesting word translations in context (SWTC) (Vulić & Moens, 2014) for the same three language pairs. A SWTC model based on our BWEs again significantly outscores the best scoring MuPTM-based SWTC models in the same setting without any use of parallel data and translation dictionaries, and again acquires the best reported results on the benchmarking SWTC datasets.
(5) We also present a comparison with state-of-the-art BWE induction models (Mikolov, Le, & Sutskever, 2013b; Hermann & Blunsom, 2014b; Gouws, Bengio, & Corrado, 2015) in BLE and SWTC. Results reveal that our simple yet effective approach is on-par with or outperforms other BWE induction models that rely on parallel data or readily available dictionaries to learn shared bilingual embedding spaces. In addition, preliminary experiments with BWESG on parallel Europarl data demonstrate that the model is also useful when trained on sentence-aligned data, reaching the performance of benchmarking BWE induction models from parallel data (e.g., Hermann & Blunsom, 2014b).

2. Related Work

In this section we further motivate why we opt for building a model for inducing bilingual word embeddings from comparable document-aligned data. For a clearer overview, we have split related work into three broad clusters: (1) monolingual word embeddings, (2) bilingual word embeddings, and (3) bilingual word representations from document-aligned data.

2.1 Monolingual Word Embeddings

The idea of representing words as continuous real-valued vectors dates way back to mid-80s (Rumelhart, Hinton, & Williams, 1986; Elman, 1990). The idea met its resurgence a decade ago (Bengio et al., 2003), where a neural language model learns word embeddings as part of a neural network architecture for statistical language modeling. This work inspired other approaches that learn word embeddings within the neural-network language modeling framework (Collobert & Weston, 2008; Collobert, Weston, Bottou, Karlen, Kavukcuoglu, & Kuksa, 2011). Word embeddings are tailored to capture semantics and encode a continuous

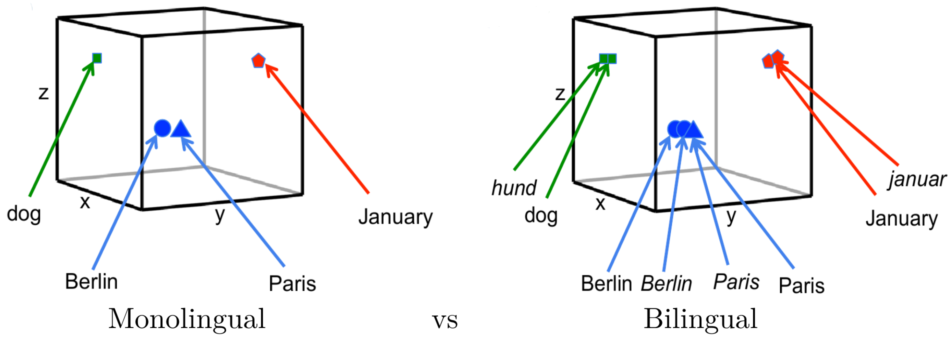


Figure 1: A toy 3D shared bilingual embedding space from Gouws et al. (2015): While in monolingual spaces words with similar meanings should have similar representations, in bilingual spaces words in two different languages with similar meanings should have similar representations (both mono- and cross-lingually).

notion of semantic similarity (as opposed to semantically poorer discrete representations), necessary to share information between words and other text units.

Recently, the skip-gram and continuous bag-of-words (CBOW) model of Mikolov et al. (2013a, 2013c) revealed that the full neural-network structure is not needed at all to learn high-quality word embeddings (with extremely decreased training times compared to the full-fledged neural network models, see Mikolov et al., 2013a for the full analysis of complexity of the models). These models are in fact simple single-layered architectures, where the objective is to predict a word's context given the word itself (skip-gram) or predict a word given its context (CBOW). Similar models called vector log-bilinear models were recently proposed (Mnih & Kavukcuoglu, 2013). Other models inspired by skip-gram and CBOW are GloVe (Global Vectors for Word Representation) (Pennington et al., 2014), which combines local and global contexts of a word into a unified model, and a model which relies on dependency-based contexts instead of simpler word-based contexts (Levy & Goldberg, 2014a), and new models are steadily emerging (e.g., Lebret & Collobert, 2014; Lu, Wang, Bansal, Gimpel, & Livescu, 2015; Stratos, Collins, & Hsu, 2015; Trask, Gilmore, & Russell, 2015; Liu, Jiang, Wei, Ling, & Hu, 2015).

An interesting finding has been discussed recently (Levy & Goldberg, 2014b): the popular skip-gram model with negative sampling (SGNS) (Goldberg & Levy, 2014) is simply a model which implicitly factorizes a word-context matrix, with its cells containing pointwise mutual information (PMI) scores of the respective word and context pairs, shifted by a global constant. In other words, the SGNS performs exactly the same thing as traditional distributional models (i.e., context counting plus context weighting and/or dimensionality reduction), with a slight improvement in performance with SGNS (Baroni et al., 2014; Levy et al., 2015).

All these low-dimensional vectors, besides improving computational efficiency, lead to better generalizations, even allowing to generalize over the vocabularies observed in labelled data, and hence partially alleviating the ubiquitous problem of data sparsity. Their utility has been validated and proven in various semantic tasks such as semantic word similarity, synonymy detection or word analogy solving (Mikolov et al., 2013d; Baroni et al., 2014; Pennington et al., 2014). Moreover, word embeddings have been proven to serve as useful unsupervised features for plenty of downstream NLP tasks such as named entity recognition, chunking, semantic role labeling, part-of-speech tagging, parsing, selectional preferences (Turian, Ratinov, & Bengio, 2010; Collobert et al., 2011; Chen & Manning, 2014).

Due to its simplicity, as well as its efficacy and consequent popularity in various tasks (Mikolov et al., 2013c; Levy & Goldberg, 2014b), with a clear advantage on similarity tasks when compared to traditional models from distributional semantics (Levy et al., 2015) in this article we will focus on the adaptation of SGNS (Mikolov et al., 2013c). In Section 3, we provide a very brief overview of the model, and then follow up with our new bilingual model which is based on SGNS.

2.2 Bilingual Word Embeddings

Bilingual word representations could serve as an useful source knowledge for problems in cross-lingual information retrieval (Levow, Oard, & Resnik, 2005; Vulić, De Smet, & Moens, 2013), statistical machine translation (Wu, Wang, & Zong, 2008), document classification (Ni, Sun, Hu, & Chen, 2011; Klementiev et al., 2012; Hermann & Blunsom, 2014b; Chandar, Lauly, Larochelle, Khapra, Ravindran, Raykar, & Saha, 2014; Vulić, De Smet, Tang, & Moens, 2015), bilingual lexicon extraction (Tamura, Watanabe, & Sumita, 2012; Vulić & Moens, 2013a), or knowledge transfer and annotation projection from resource-rich to resource-poor languages for a myriad of NLP tasks such as dependency parsing, POS tagging, semantic role labeling or selectional preferences (Yarowsky & Ngai, 2001; Padó & Lapata, 2009; Peirsman & Padó, 2010; Das & Petrov, 2011; Täckström, Das, Petrov, McDonald, & Nivre, 2013; Ganchev & Das, 2013; Tiedemann, Agić, & Nivre, 2014; Xiao & Guo, 2014). Other interesting application domains are machine translation (e.g., Zou, Socher, Cer, & Manning, 2013; Wu, Dong, Hu, Yu, He, Wu, Wang, & Liu, 2014; Zhang, Liu, Li, Zhou, & Zong, 2014) and cross-lingual information retrieval (e.g., Vulić & Moens, 2015). Moreover, by making the transition from monolingual to bilingual settings and building a *shared bilingual embedding space* (see again Figure 1 for an illustrative example), one is able to extend or rather generalize semantic tasks such as semantic similarity computation, synonymy detection or word analogy computation across languages. Following the success in monolingual settings, a body of recent work on word representation learning has therefore focused on learning bilingual word embeddings (BWEs).

The current research on inducing BWEs critically relies on sentence-aligned parallel data or readily available bilingual lexicons to achieve the coherence of representations across languages (e.g., to build similar representations for similar concepts in different languages such as *January-januari*, *dog-hund* or *sky-hemel*). We may cluster the current work in three different groups: (1) the models that rely on hard word alignments obtained from parallel data to constrain the learning of BWEs (Klementiev et al., 2012; Zou et al., 2013; Wu et al., 2014); (2) the models that use the alignment of parallel data at the sentence level (Kočiský, Hermann, & Blunsom, 2014; Hermann & Blunsom, 2014a, 2014b; Chandar et al., 2014; Shi, Liu, Liu, & Sun, 2015; Gouws et al., 2015); (3) the models that critically require readily available bilingual lexicons (Mikolov et al., 2013b; Faruqui & Dyer, 2014; Xiao & Guo, 2014). The main disadvantage of all these models is the limited availability of parallel data and bilingual lexicons, resources which are scarce and/or domain-restricted for plenty of language pairs. In this work, we significantly alleviate the requirements: unlike prior work, we show that BWEs may be induced solely on the basis of document-aligned comparable data without any additional need for parallel data or bilingual lexicons. Note that (in theory) the work of Hermann and Blunsom (2014b), and Chandar et al. (2014) may also be extended to the same setting with document-aligned data, as these two models originally rely on sentence embeddings computed as aggregations over their single word embeddings plus sentence alignments. In this work, by testing and comparing to the BiCVM model of Hermann and Blunsom, we show that these models do not work well in practice after replacing the very strong bilingual signal coded in parallel sentences with the noisy bilingual signal given by document alignments and non-parallel data.

2.3 Bilingual Word Representations from Document-Aligned Data

Prior work on inducing bilingual word representations in the early days followed the tradition of window-based context-counting distributional models (Rapp, 1999; Gaussier, Renders, Matveeva, Goutte, & Déjean, 2004; Laroche & Langlais, 2010) and it again required a bilingual lexicon as a critical resource. In order to tackle this issue, recent work relies on the supervision-lighter framework of multilingual probabilistic topic modeling (MuPTM) (Mimno, Wallach, Naradowsky, Smith, & McCallum, 2009; Boyd-Graber & Blei, 2009; De Smet & Moens, 2009; Ni, Sun, Hu, & Chen, 2009; Zhang, Mei, & Zhai, 2010; Fukumasu, Eguchi, & Xing, 2012) or other similar models for latent structure induction (Haghighi, Liang, Berg-Kirkpatrick, & Klein, 2008; Daumé III & Jagarlamudi, 2011).

Words in this setting are represented as real-valued vectors with conditional topic probability scores $P(z_k|w_i)$, regardless of their actual language. Topics z_k are in fact latent inter-lingual concepts discovered directly from multilingual comparable data using a multilingual topic model such as bilingual LDA. We discuss the MuPTM-based representations in more detail in Section 4.1.

MuPTM-based bilingual word representations induced from comparable data have demonstrated its utility in tasks such as cross-lingual semantic similarity computation and bilingual lexicon extraction (Vulić, De Smet, & Moens, 2011; Liu, Duh, & Matsumoto, 2013) and suggesting word translations in context (Vulić & Moens, 2014). In this work, we compare the state-of-the-art MuPTM-based word representations induced from the same type of comparable corpora with BWEs learned by our new model in these two semantic tasks.

Another recent model (Søgaard, Agić, Martínez Alonso, Plank, Bohnet, & Johannsen, 2015) is also able to learn from document-aligned data. It is a count-based model which builds binary word vectors denoting the occurrence of each word in each document pair. Dimensionality reduction is then applied post-hoc on the induced sparse vectors. Since the links between documents are known, the model is able to learn cross-lingual correspondences between words and, consequently, bilingual word representations. Exactly the same idea was already introduced as a baseline model by Vulić et al. (2011), where TF-IDF weights were used instead of binary indices, and no dimensionality reduction was applied post-hoc. The model of Vulić et al. was surpassed by baseline models from document-aligned data briefly discussed in Section 4.1, while the model of Søgaard et al. obtains results that are very similar to the BWE baselines compared against in this work (described in Section 4.2).

3. BWESG: Model Architecture

Our new bilingual model is an extension of SGNS to bilingual settings with document-aligned comparable training data. This section describes the underlying SGNS and two variants of our SGNS-based BWE induction model.

3.1 Skip-Gram with Negative Sampling (SGNS)

Our departure point is the log-linear SGNS of Mikolov et al. (2013c) as implemented in the `word2vec` package.[1] The SGNS model learns word embeddings (WEs) in a similar way to neural language models (Bengio et al., 2003; Collobert & Weston, 2008), but without a non-linear hidden layer.

In the monolingual setting, we assume one language L with vocabulary V, and a corpus of words $w \in V$, along with their contexts $c \in V^c$, where V^c is the context vocabulary. Contexts for each word w_n are typically neighboring words in a context window of size cs (i.e., $w_{n-cs}, \ldots, w_{n-1}, w_{n+1}, \ldots, w_{n+cs}$), so effectively it holds $V^c \equiv V$.[2]

Each word type $w \in V$ is associated with a vector $\vec{w} \in \mathbb{R}^d$ (its pivot word representation or pivot word embedding, see Figure 2), and a vector $\vec{w_c} \in \mathbb{R}^d$ (its context embedding). d is the dimensionality of the WE vectors, which, as a model input parameter, has to be set in advance before the training procedure commences. The entries in these vectors are latent, and treated as parameters θ to be learned by the model. In short, the idea of the skip-gram model is to scan through the corpus (which is typically unannotated, Mikolov et al., 2013a) *word by word* in turn (i.e., these are the pivot words), and learn from the pairs *(word, context word)*. The learning goal is to maximize the ability of predicting context words for each pivot word in the corpus. Let $ob = 1$ denote that the pair of words (w, v) is observed in the corpus and thus belongs to the training set D. The probability of $(w, v) \in D$ is defined by the softmax function:

$$P(ob = 1|w, v, \theta) = \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v_c})} \tag{1}$$

Each word token w in the corpus is treated in turn as the pivot and all pairs of word tokens $(w, w \pm 1)$,...,$(w, w \pm t(cs))$ are appended to D, where $t(cs)$ is an integer sampled from a uniform distribution on $\{1, \ldots, cs\}$.[3] The global training objective J is then to maximize the probabilities that all pairs from D are indeed observed in the corpus:

$$J = \arg\max_{\theta} \sum_{(w,v) \in D} \log \frac{1}{1 + \exp(-\vec{w} \cdot \vec{v_c})} \tag{2}$$

where θ are the parameters of the model, that is, pivot and context word embeddings which have to be learned. One may see that this objective function has a trivial solution by setting

[1] `https://code.google.com/p/word2vec/`

[2] Testing other options for context selection such as dependency-based contexts (Levy & Goldberg, 2014a) is beyond the scope of this work, and it was shown that these contexts may not lead to any gains in the final WEs (Kiela & Bottou, 2014).

[3] The original skip-gram model utilizes dynamic window sizes, where cs denotes the maximum window size. Moreover, the model takes into account sentence boundaries in context selection, that is, it selects as context words only words occurring in the same sentence as the pivot word.

$\vec{w} = \vec{v_c}$, and $\vec{w} \cdot \vec{v_c} = Val$, where Val is a large enough number (Goldberg & Levy, 2014). In order to prevent this trivial training scenario, the *negative sampling* procedure comes into the picture (Collobert & Weston, 2008; Mikolov et al., 2013c).

In short, the idea behind negative sampling is to present the model with a set D' of artificially created or sampled "negative pivot-context" word pairs (w, v'), which by assumption serve as negative examples, that is, they do not occur as observed/positive *(word, context)* pairs in the training corpus. The model then has to adjust the parameters θ in such a way to also maximize the probability that these negative pairs will not occur in the corpus. While the interested reader may find further details about the negative sampling procedure, and the new exact objective function along with its derivation elsewhere (Levy & Goldberg, 2014b), for illustrative purposes and simplicity, here we present the approximative objective function with negative sampling by Goldberg and Levy:

$$J = \arg\max_{\theta} \sum_{(w,v)\in D} \log \frac{1}{1+\exp(-\vec{w}\cdot\vec{v_c})} + \sum_{(w,v')\in D'} \log \frac{1}{1+\exp(\vec{w}\cdot\vec{v_c'})} \tag{3}$$

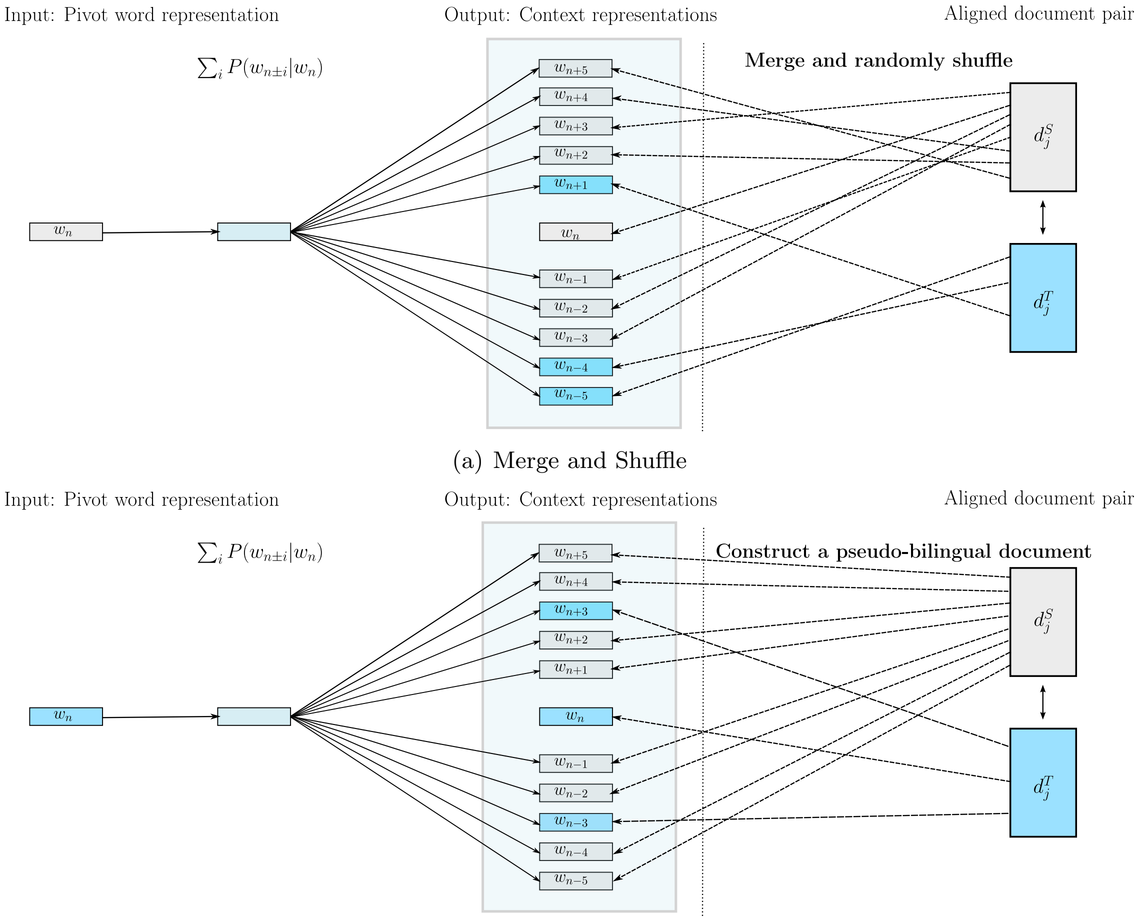
The free parameters θ are updated using stochastic gradient descent and backpropagation, with learning rate typically controlled by Adagrad (Duchi, Hazan, & Singer, 2011) or with a global linearly decreasing learning rate. By optimizing the objective from eq. (3), the model incrementally pushes observed pivot WEs towards context WEs of their collocates in the corpus. In the words of distributional hypothesis - after training, words that occur in similar contexts should end up having similar word embeddings. In other words, to link the terminology of distributional hypothesis and the modeling assumptions of SGNS - words that predict similar contexts end up having similar word embeddings.

3.2 Final Model - BWESG: BWE Skip-Gram

In the next step, we propose a novel method that extends SGNS to work with bilingual document-aligned comparable data. Let us assume that we possess a document-aligned comparable corpus, defined as $\mathcal{C} = \{d_1, d_2, \ldots, d_N\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \ldots, (d_N^S, d_N^T)\}$. $d_j = (d_j^S, d_j^T)$ 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 pairs in the corpus. V^S and V^T are vocabularies associated with languages L_S and L_T. The goal is to learn a shared bilingual embedding space given the data (Figure 1) and document alignments as the only bilingual signal during training. We present two strategies that, coupled with SGNS, lead to such shared bilingual spaces. An overview of the architecture for learning BWEs from document-aligned comparable data with the two strategies is given in Figures 2(a) and 2(b).

3.2.1 Merge and Shuffle

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'. Following that, we randomly *shuffle* the newly constructed pseudo-bilingual document. A shuffle is a (random) permutation of the word tokens given in two different languages forming the pseudo-bilingual document. The pre-training shuffling step (see Figure 2(a)) assures that each word w, regardless of its actual language, obtains word collocates from both vocabularies. The idea of obtaining bilingual contexts for each pivot word in each pseudo-bilingual document will steer the


Input: Pivot word representation
Output: Context representations
Aligned document pair
$\sum_i P(w_{n\pm i}|w_n)$
Merge and randomly shuffle
w_{n+5}
w_{n+4}
w_{n+3}
w_{n+2}
w_{n+1}
w_n
w_{n-1}
w_{n-2}
w_{n-3}
w_{n-4}
w_{n-5}
d_j^S
d_j^T
Construct a pseudo-bilingual document

(a) Merge and Shuffle

(b) Length-Ratio Shuffle

Figure 2: The architecture of our BWE Skip-Gram (BWESG) model for learning bilingual word embeddings from document-aligned comparable data with two different pre-training strategies: (1) non-deterministic *merge and shuffle*, (2) deterministic *length-ratio shuffle*. 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 figures (separated by vertical dashed lines) illustrates how a pseudo-bilingual document is constructed from a pair of two aligned documents.

final model towards constructing a shared bilingual 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 Section 7.3. In another interpretation, since the model relies only on (pseudo-bilingual) document level co-occurrence, the window size parameter then just controls the amount of random data dropout, that is, the number of positive document-level training examples. The locality feature of SGNS is not preserved due to the shuffling procedure.

3.2.2 Length-Ratio Shuffle

The non-deterministic and uncontrollable nature of the *merge and shuffle* procedure opens up a possibility of accidentally obtaining "bad shuffles" that will result in sub-optimal word representations. Therefore, we also propose a *deterministic strategy* for building pseudo-bilingual documents suitable for bilingual training. Source and target language words are inserted into an (initially empty) pseudo-bilingual document in turn based on the ratio of document lengths, with word order preserved. Document lengths are measured in terms of word tokens, and let us denote them as m_S and m_T for an aligned document pair (d_j^S, d_j^T). Let us assume, without loss of generality, that $m_S \geq m_T$. The procedure then proceeds as follows (if $m_T > m_S$ the procedure proceeds in an analogous manner with the roles of d_j^S and d_j^T reversed):

1. Pseudo-bilingual document d'_j is empty: $d'_j = \{\}$.
2. Compute the ratio: $R = \lfloor \frac{m_S}{m_T} \rfloor$.
3. Scan through aligned documents d_S and d_T simultaneously and (3.1) append R word tokens from d_j^S into d'_j; then (3.2) append 1 word token from d_j^T. Repeat steps 3.1 and 3.2 until all word tokens from d_j^T have been inserted into d'_j.
4. Insert remaining $m_S \mod m_T$ word tokens from d_j^S into d'_j.

Using a simple example, assume that we have an English (EN) document $\{Frodo, Sam, orcs, goblins, Mordor, ring\}$ and a Spanish (ES) document $\{anillo, orcos, mago\}$: the pseudo-bilingual document would be formed by inserting 1 Spanish word after 2 English words (as the length ratio is 6:3 = 2:1). The final pseudo-bilingual document is:
$\{Frodo_{EN}, Sam_{EN}, anillo_{ES}, orcs_{EN}, goblins_{EN}, orcos_{ES}, Mordor_{EN}, ring_{EN}, mago_{ES}\}$.

In another interpretation, the *length-ratio shuffle* strategy constructs a single permutation/shuffle of the pseudo-bilingual document controlled by the word order in two aligned documents as well as their length ratio. As before, the model relies on pseudo-bilingual document level co-occurrence, and the window size parameter controls the amount of (now non-random) data dropout. A difference lies in the fact that this procedure now keeps word order intact monolingually while constructing a pseudo-bilingual document.

The final BWE Skip-gram (BWESG) model then relies on the monolingual variant of SGNS (or any other monolingual WE induction model) trained on these shuffled/permuted pseudo-bilingual documents (using any of the proposed strategies).[4] The model learns word embeddings for source and target language words aligned over the d shared embedding dimensions. The BWESG-based representation of word w, regardless of its actual language, is then a d-dimensional vector: $\vec{w} = [f_1, \ldots, f_k, \ldots, f_d]$. $f_k \in \mathbb{R}$ denotes the value of the k-th shared inter-lingual feature within the d-dimensional shared bilingual embedding space. Since all words share the embedding space, semantic similarity between words may be computed both monolingually and across languages. We will extensively use this property in our evaluation tasks.

[4] We were also experimenting with GloVe and CBOW, but they were falling short of SGNS on average.

4. Baseline Representation Models

We quickly navigate through other approaches to bilingual word representation learning from document-aligned comparable data. The set of models in comparison may be roughly clustered into two main groups: (Group I) "pre-BWE" baseline representation models from document-aligned data, (Group II) benchmarking BWE induction models that were not originally developed for learning from document-aligned comparable data. While it is essential to compare the BWESG model with other frameworks for learning representations from document-aligned data (Group I), it is also crucial to detect main strengths of the BWESG model when compared to other approaches in the BWE learning framework which can also be adjusted to learn from document-aligned data (Group II).

4.1 Group I: Baseline Representation Models from Document-Aligned Data

We briefly describe three benchmarking Group I models.

4.1.1 Basic-MuPTM

The early approaches (e.g., Dumais, Landauer, & Littman, 1996; Carbonell, Yang, Frederking, Brown, Geng, Lee, Frederking, E, Geng, & Yang, 1997) tried to mine topical structure from document-aligned comparable texts using a monolingual topic model (e.g., LSA or LDA) trained on pseudo-bilingual documents with the target document simply appended to its source language counterpart, and then used the discovered latent topical structure as a shared semantic space in which both words and documents from two languages may be represented in a uniform way.

More recent work on multilingual probabilistic topic modeling (MuPTM) (Mimno et al., 2009; De Smet & Moens, 2009; Vulić et al., 2011) showed that word representations of higher quality may be built if a multilingual topic model such as bilingual LDA (BiLDA) is trained jointly on document-aligned comparable corpora by retaining the structure of the corpus intact (i.e., there is no need to construct pseudo-bilingual documents).

MuPTM discovers the latent structure of the observed data in the form of K latent cross-lingual topics $z_1, \ldots, z_K$ which optimally describe the generation of observed data. Extracting latent cross-lingual topics actually implies learning per-document topic distributions for each document in the corpus (probability scores $P(z_k|d_j)$), and discovering language-specific representations of these topics given by per-topic word distributions in each language (probability scores $P(w_i^S|z_k)$ and $P(w_i^T|z_k)$). Latent cross-lingual topics are in fact distributions over vocabulary words, and have their language-specific representation in each language. Per-document topic distributions and per-topic word distributions are obtained after training the topic model on multilingual data. The representation of some word $w \in V^S$ (or in an analogous manner $w \in V^T$) is then a K-dimensional vector:
$\vec{w} = [P(z_1|w), \ldots, P(z_k|w), \ldots, P(z_K|w)]$.

We call this representation model (RM) *Basic-MuPTM (BMu)*. Since the number of topics, that is, the number of vector dimensions K is typically high (Dinu & Lapata, 2010; Vulić et al., 2011), additional feature pruning (Reisinger & Mooney, 2010) may be employed in order to retain only the most descriptive dimensions in the MuPTM-based representation, which was shown to improve the performance on several semantic tasks (e.g., BLE or SWTC) (Vulić & Moens, 2013a; Vulić et al., 2015).

A multilingual topic model is typically trained by Gibbs sampling (Geman & Geman, 1984; Steyvers & Griffiths, 2007; Vulić et al., 2015). Similar to the SGNS/BWESG training procedure, Gibbs sampling for MuPTM/BiLDA also scans the training corpus word by word, and then cyclically updates topic assignments for each word token. However, unlike BWESG which uses only a subset of document-level training examples, Gibbs sampling for MuPTM uses all words from the source language document as well as all words from its coupled target language document to influence the topic assignment for the pivot word. The BWESG design relying on data dropout leads to decreased training times and computation costs to obtain final representations compared to Basic-MuPTM.

4.1.2 Association-MuPTM

Another representation is also based on the MuPTM framework: it contains association scores $P(w_a|w)$ for each $w, w_a \in V^S \cup V^T$ (Vulić & Moens, 2013a) as dimensions of real-valued word vectors. These association scores are computed as $P(w_a|w) = \sum_{k=1}^{K} P(w_a|z_k)$ $P(z_k|w)$ (Griffiths, Steyvers, & Tenenbaum, 2007), and the word vector is a $(|V^S| + |V^T|)$-dimensional vector: $\vec{w} = [P(w_1^S|w), \ldots, P(w_{|V^S|}^S|w), P(w_1^T|w), \ldots, P(w_{|V^T|}^T|w)]$. As with Basic-MuPTM, the original word representation may also be pruned post-hoc. We call this representation model *Association-MuPTM (AMu)*. Since this approach relies on the MuPTM training plus additional $|V^S| \cdot |V^T|$ computations to estimate association scores, the cost of obtaining Association-MuPTM representations is even higher than for Basic-MuPTM, but it leads to more robust word representations for the BLE task (Vulić & Moens, 2013a). While both Basic-MuPTM and Association-MuPTM produce high-dimensional real-valued vectors with plenty of near-zero dimensions (the number of dimensions is typically measured in thousands) which have to be pruned afterwards with the pruning parameter often set ad-hoc, BWESG produces lower-dimensional dense real-valued vectors, and no additional post-hoc feature pruning is required for BWESG.

4.1.3 Traditional-PPMI

A traditional approach to building bilingual word representations in (cross-lingual) distributional semantics is to compute weighted co-occurrence scores (e.g., using PMI, TF-IDF) between pivot words and their context words in a window of predefined size, plus an external bilingual lexicon to align context words/dimensions across languages (Gaussier et al., 2004; Laroche & Langlais, 2010). A weighting function (WeF), which is a standard choice in distributional semantics and yields optimal or near-optimal results over a group of semantic tasks (Bullinaria & Levy, 2007), is the smoothed positive pointwise mutual information statistic (Pantel & Lin, 2002; Turney & Pantel, 2010). Furthermore, in order to induce context words without the need for a readily available lexicon, we employ the bootstrapping procedure of Peirsman and Padó (2011), and Vulić and Moens (2013b). This representation model is called *Traditional-PPMI (TPPMI)*. The word representation is an R-dimensional vector: $\vec{w} = [sc_1(w, c_1), \ldots, sc_k(w, c_k), \ldots, sc_K(w, c_K)]$. The dimensions of the vector space are K one-to-one word translation pairs $c_k = (c_k^S, c_k^T)$, and $sc_k(w, c_k)$ is the weighted co-occurrence score of the pivot word w and the k-th context feature, where one computes the co-occurrence score using c_k^S if $w \in V^S$, or c_k^T if $w \in V^T$.

Vector dimensions $c_k = (c_k^S, c_k^T)$ in the Traditional-PPMI representation and similar models with other WeFs are typically the most frequent and reliable translation pairs in the corpus. As opposed to BWESG, the obtained word vectors are again high-dimensional (typically thousands of dimensions) sparse real-valued vectors. In addition, traditional-PPMI is a purely local distributional model deriving distributional context knowledge from narrow context windows (typically 3-10 surrounding words, e.g., Laroche & Langlais, 2010). A bootstrapping approach (Vulić & Moens, 2013b) which we use to induce the Traditional-PPMI representation starts from an automatically learned seed lexicon of one-to-one translation pairs obtained using some other model (e.g., Basic-MuPTM or Association-MuPTM), and then gradually detects new dimensions of the shared bilingual semantic space. We refer the interested reader to the relevant literature (Vulić & Moens, 2013b) for more details.

4.2 Group II: BWE Induction Models Adjusted to Document-Aligned Data

We now provide a quick overview of three representative benchmarking BWE models that learn from different types of bilingual and monolingual data.

4.2.1 BiCVM

Hermann and Blunsom (2014b) introduced a model called BiCVM (Bilingual Compositional Vector Model) that learns bilingual word embeddings from a sentence-aligned parallel corpus $\mathcal{C} = \{s_1, s_2, \ldots, s_N\} = \{(s_1^S, s_1^T), (s_2^S, s_2^T), \ldots, (s_N^S, s_N^T)\}$.[5] $s_j = (s_j^S, s_j^T)$ now denotes a pair of aligned sentences. The model assumes that the aligned sentences have the same meaning, which implies that their sentence representations should be similar. Assume two functions f and g which map sentences given in the source and language respectively to their semantic representations in $\mathbb{R}^d$, where d is again the representation dimensionality. The energy of the model given two sentences $(s_j^S, s_j^T) \in \mathcal{C}$ is then defined as: $E(s_j^S, s_j^T) = ||f(s_j^S) - g(s_j^T)||$. The goal is to minimize E for all semantically equivalent sentences (i.e., aligned sentences) in the corpus. In order to prevent the model from degenerating, they use a noise-contrastive large-margin update which ensures that the representations of non-aligned sentences observe a certain margin from each other. For every pair of parallel sentences (s_j^S, s_j^T), they sample a number of additional negative sentence pairs (s_j^S, n_{neg}^T) from the corpus (i.e., the sampled pairs are not observed as positive pairs in $\mathcal{C}$). These noise samples are used in formulating the hinge loss as follows: $E(s_j^S, s_j^T) = \max(mrg + \Delta E(s_j^S, s_j^T, n_{neg}^T), 0)$, where mrg is the margin, and $\Delta E(s_j^S, s_j^T, n_{neg}^T) = E(s_j^S, s_j^T) - E(s_j^S, n_{neg}^T)$. The loss is minimized for every pair of parallel sentences in the corpus with $L2$-regularization on the model parameters. The number of noise samples per each positive pair is a hyper-parameter of the model. A semantic signal is propagated from aligned sentences back to the individual words to obtain bilingual word embeddings. While the BiCVM model was originally built for sentence-aligned parallel data, exactly the same idea may be applied to document-aligned non-parallel data. In this paper, we test its ability to learn from noisier comparable data. The BWESG

[5] A very similar (but more expensive) model which also learns from parallel sentence-aligned data was also introduced by Chandar et al. (2014).

model is compared against BiCVM when inducing BWEs from both data types: comparable and parallel.

4.2.2 Mikolov's Mapping

Another collection of BWE induction models (Mikolov et al., 2013b; Faruqui & Dyer, 2014; Dinu, Lazaridou, & Baroni, 2015; Lazaridou, Dinu, & Baroni, 2015) assumes the following setup: first, two monolingual embedding spaces, $\mathbb{R}^{dim_S}$ and $\mathbb{R}^{dim_T}$, are induced separately in each of the two languages using a standard monolingual WE model such as SGNS (Mikolov et al., 2013a, 2013c). dim_S and dim_T denote the dimensionality of monolingual embedding spaces in the source and target language respectively. The bilingual signal is provided in the form of word translation pairs (x_i, y_i), where $x_i \in V^S$, $y_i \in V^T$, and $\vec{x_i} \in \mathbb{R}^{dim_S}$, $\vec{y_i} \in \mathbb{R}^{dim_T}$. Training is cast as a multivariate regression problem: it implies learning a function that maps the source language vectors from the training data to their corresponding target language vectors. A standard approach (Mikolov et al., 2013b; Dinu et al., 2015) is to assume a linear map $\mathbf{W} \in \mathbb{R}^{dim_S \times dim_T}$, where a L_2-regularized least-squares error objective (i.e., ridge regression) is used to learn the map $\mathbf{W}$: it is learned by solving the following optimization problem (typically by stochastic gradient descent):
$\min_{\mathbf{W} \in \mathbb{R}^{dim_S \times dim_T}} ||\mathbf{X}\mathbf{W} - \mathbf{Y}||_F^2 + \lambda ||\mathbf{W}||_F^2$.

$\mathbf{X}$ and $\mathbf{Y}$ are matrices obtained through the respective concatenation of source language and target language vectors from training pairs. Once the linear map $\mathbf{W}$ is estimated, any previously unseen source language word vector $\vec{x_u}$ may be straightforwardly mapped into the target language embedding space $\mathbb{R}^{dim_T}$ as $\mathbf{W}\vec{x_u}$. After mapping all vectors $\vec{x}$, $x \in V^S$, the target embedding space $\mathbb{R}^{dim_T}$ in fact serves as a bilingual embedding space (Figure 1).

Although the main strength of the model is its ability to learn embeddings on larger monolingual training sets, the model may also be adjusted to the setting where the only training data are document-aligned comparable data as follows: (1) Automatically learn a seed lexicon or reliable one-to-one translation pairs from document-aligned data using a bootstrapping approach from Vulić and Moens (2013b), (2) Train two separate monolingual embedding spaces on two separated halves of the document-aligned data set (i.e., using only source language documents and only target language documents), (3) Learn the mapping between the two spaces using the pairs from Step 1.

4.2.3 BilBOWA

Another collection of BWE induction models jointly optimizes two monolingual objectives, with the cross-lingual objective acting as a cross-lingual regularizer during training (Klementiev et al., 2012; Gouws et al., 2015; Soyer, Stenetorp, & Aizawa, 2015). The idea behind joint training may be summarized by the simplified formulation (Luong, Pham, & Manning, 2015): $\gamma(Mono_S + Mono_T) + \delta Bi$.

The monolingual objectives $Mono_S$ and $Mono_T$ ensure that similar words in each language are assigned similar embeddings and aim to capture the semantic structure of each language, whereas the cross-lingual objective Bi ensures that similar words across languages are assigned similar embeddings, and ties the two monolingual spaces together into a bilingual space. Parameters γ and δ govern the influence of the monolingual and bilingual components.[6] The bilingual signal for these models, now acting as the cross-lingual regularizer during the joint training, is provided in sentence-aligned parallel data. Although they use the same data sources, the models differ in the choice of monolingual and cross-lingual objectives. In this work, we opt for the BilBOWA model of Gouws et al. (2015) as the representative model to be included in the comparisons, due to its previous solid performance and robustness in the BLE task, its reduced complexity reflected in fast computations on massive datasets, as well as its public availability. In short, the BilBOWA model combines SGNS for the monolingual objectives together with the cross-lingual objective that minimizes the L_2-loss between the bag-of-word vectors of parallel sentences. For more details about the exact training procedure, we refer the interested reader to the Gouws et al.'s work.

Again, although the main strength of the model is its ability to learn embeddings on larger monolingual training sets, the model may also be adjusted to the setting with document- or sentence-aligned data by: (1) using two halves of the aligned corpus for separate monolingual training, (2) using the alignment signal for bilingual training.

5. From Word Representations to Semantic Word Similarity

Assume now that we have induced bilingual word representations, regardless of the chosen RM. Given two words w_i and w_j, irrespective to their actual language, we may compute the degree of their semantic similarity by applying a *similarity function* (SF) on their vector representations $\overrightarrow{w_i}$ and $\overrightarrow{w_j}$: $sim(w_i, w_j) = SF(\overrightarrow{w_i}, \overrightarrow{w_j})$. Different choices (or rather families of) SFs are cosine, the Kullback-Leibler or the Jensen-Shannon divergence, the Hellinger distance, the Jaccard index, etc. (Lee, 1999; Cha, 2007), and different RMs typically require different SFs to produce optimal or near-optimal results over various semantic tasks. When working with word embeddings, a standard choice for SF is cosine similarity (*cos*) (Mikolov et al., 2013c), which is also a typical choice in traditional distributional models (Bullinaria & Levy, 2007). The similarity is then computed as follows:

$$sim(w_i, w_j) = cos(w_i, w_j) = \frac{\overrightarrow{w_i} \cdot \overrightarrow{w_j}}{|\overrightarrow{w_i}| \cdot |\overrightarrow{w_j}|} \tag{4}$$

On the other hand, a good choice for SF when working with probabilistic RMs such as Basic-MuPTM and Association-MuPTM RS is the Hellinger distance (Pollard, 2001; Cha, 2007; Kazama, Saeger, Kuroda, Murata, & Torisawa, 2010), which displays excellent results in the BLE task (Vulić & Moens, 2013a). The similarity between words w_i and w_j using the Hellinger distance is computed as follows:

$$sim(w_i, w_j) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{K} \left(\sqrt{P(f'_k|w_i)} - \sqrt{P(f'_k|w_j)} \right)^2} \tag{5}$$

Note that the Hellinger distance is applicable only if word representations are probability distributions, which is the case for Basic-MuPTM and Association-MuPTM. $P(f'_k|w_i)$ de-

[6] Setting $\gamma = 0$ reduces the model to the setting similar to BiCVM (Hermann & Blunsom, 2014b). $\gamma = 1$ results in the models of Klementiev et al. (2012), Gouws et al. (2015), and Soyer et al. (2015).

notes the probability score for the k-th dimension (f'_k) in the vector representation with Basic-MuPTM or Association-MuPTM.[7]

For each word w_i, we can build a *ranked list* $RL(w_i)$ which consists of all other words w_j ranked according to their respective semantic similarity scores $sim(w_i, w_j)$. Additionally, we label the ranked list $RL(w_i)$ that is pruned at position M as $RL_M(w_i)$. Since we may retain language labels for words when training in multilingual settings (e.g., language labels are marked by different colors in Figure 2), we may compute: (1) *monolingual similarity*, e.g., given $w_i \in V^S$, we retain only $w_j \in V^S$ in the ranked list (analogous for $w_i \in V^T$), (2) *cross-lingual similarity* (CLSS), e.g., given $w_i \in V^S$, we retain only $w_j \in V^T$, and (3) *multilingual similarity*, where we retain all words $w_j \in V^S \cup V^T$. When computing CLSS for w_i, the most similar word cross-lingually is called the cross-lingual *nearest neighbor*.

We will employ the models of context-insensitive CLSS at the word type level to extract bilingual lexicons from document-aligned or sentence-aligned data, and to compare all representation models in the BLE task in Section 7.

5.1 Context Sensitive Models of (Cross-Lingual) Semantic Similarity

The context-insensitive models of semantic similarity provide ranked lists of semantically similar words *invariably* or *in isolation*, and they operate at the level of word types. They do not explicitly encode different word senses. In practice, it means that, given a sentence *"The coach of his team was not satisfied with the game yesterday."*, these context-insensitive CLSS models are not able to detect that the Spanish word *entrenador* is more similar to the polysemous English word *coach* in the context of this sentence than the Spanish word *autocar*, although *autocar* is listed as the most semantically similar word to *coach* globally/invariably without any observed context. In another example, while the Spanish words *partido*, *encuentro*, *cerilla* or *correspondencia* are all highly similar to another ambiguous English word *match* when observed in isolation, given the Spanish sentence *"She was unable to find a match in her pocket to light up a cigarette."*, it is clear that the strength of cross-lingual semantic similarity should change in context as only *cerilla* exhibits a strong cross-lingual semantic similarity to *match* within this particular sentential context.

The goal now is to build BWE-based models of cross-lingual semantic similarity in context, similar to context-aware CLSS models proposed by Vulić and Moens (2014). Two key questions are: (i) How to provide BWE-based representations beyond word level to represent the context of a word token?; (ii) How to use the contextual knowledge in a context-sensitive model of semantic similarity?

Following Vulić and Moens (2014), given a word token w in context (e.g., a window of words, a sentence, a paragraph, or a document), we build its context set or rather context bag $Con(w) = \{cw_1, \ldots, cw_r\}$ by harvesting r neighboring words in the chosen context scope (e.g., the context bag may comprise all content-bearing words in the same sentence as the pivot word token, the so-called *sentential context*). In order to present the context $Con(w)$ in the d-dimensional embedding space, we need to apply a model of *semantic composition* to learn its d-dimensional vector representation $\overrightarrow{Con(w)}$.

[7] Prior work has shown that the results for Basic-MuPTM and Association-MuPTM are slightly higher when cosine is replaced with the Hellinger distance. Therefore, in this particular case we have opted for the Hellinger distance to report a more competitive baseline.

Formally, given word w, we may specify the vector representation of the context bag $Con(w)$ as the d-dimensional vector/embedding:

$$\overrightarrow{Con(w)} = \overrightarrow{cw_1} \star \overrightarrow{cw_2} \star \ldots \star \overrightarrow{cw_r} \tag{6}$$

where $\overrightarrow{cw_1}, \ldots, \overrightarrow{cw_r}$ are d-dimensional WEs learned from the data, and $\star$ is a compositional vector operator such as addition, point-wise multiplication, tensor product, etc.

A plethora of models for semantic composition have been proposed in the relevant literature, differing in their choice of vector operators, input structures and required knowledge (Mitchell & Lapata, 2008; Baroni & Zamparelli, 2010; Rudolph & Giesbrecht, 2010; Socher, Huval, Manning, & Ng, 2012; Blacoe & Lapata, 2012; Clarke, 2012; Hermann & Blunsom, 2014b; Milajevs, Kartsaklis, Sadrzadeh, & Purver, 2014), to name only a few. In this work, driven by the observed linear linguistic regularities in the embedding spaces (Mikolov et al., 2013d), we opt for simple *addition* (denoted by $+$) from Mitchell and Lapata (2008) as the compositional operator, due to its simplicity, the ease of applicability on bag-of-words contexts, and its relatively solid performance in various compositional tasks (Mitchell & Lapata, 2008; Milajevs et al., 2014). The d-dimensional embedding $\overrightarrow{Con(w)}$ is then:

$$\overrightarrow{Con(w)} = \overrightarrow{cw_1} + \overrightarrow{cw_2} + \ldots + \overrightarrow{cw_r} \tag{7}$$

If we use any BWE-based RM, we may compute the context-sensitive semantic similarity score $sim(w_i, t_j, Con(w_i))$ between t_j and w_i given its context $Con(w_i)$ in the shared bilingual embedding space as follows:

$$sim(w_i, t_j, Con(w_i)) = SF(\overrightarrow{w_i'}, \overrightarrow{t_j}) \tag{8}$$

$t_j \in V^T$ is any target language word, and $\overrightarrow{t_j}$ its word representation, while $\overrightarrow{w_i'}$ is the new "contextualized" vector representation for w_i modulated by its context $Con(w_i)$, that is, its context-aware representation. Vulić and Moens (2014) introduced a linear interpolation of two d-dimensional vectors as a plausible solution for the modulation/contextualization. The modulation of representation for w_i is computed as follows:

$$\overrightarrow{w_i'} = (1 - \lambda) \cdot \overrightarrow{w_i} + \lambda \cdot \overrightarrow{Con(w_i)} \tag{9}$$

where $\overrightarrow{w_i}$ is the word embedding for w_i computed at the word type level, $\overrightarrow{Con(w_i)}$ is the embedding for the context bag computed using eq. (7), and λ is an interpolation parameter. Another set of similar models that can yield context-sensitive similarity computations has been proposed very recently, and has displayed very competitive results regardless of its simplicity (Melamud, Levy, & Dagan, 2015). Here, we present two best scoring context-sensitive models which we adapt to the bilingual setting:

$$\textsc{Add-Melamud:} \quad sim(w_i, t_j, Con(w_i)) = \frac{SF(w_i, t_j) + \sum_{cw_i \in Con(w_i)} SF(cw_i, t_j)}{|Con(w_i)| + 1}$$

$$\textsc{Mult-Melamud:} \; sim(w_i, t_j, Con(w_i)) = \sqrt[|Con(w_i)|+1]{SF(w_i, t_j) \cdot \prod_{cw_i \in Con(w_i)} SF(cw_i, t_j)}$$

Note that for the Mult model one has to avoid negative values, so a simple shift to an all-positives interval is required, e.g., the shifted cosine score becomes $cos'(x,y) = \frac{cos(x,y)+1}{2}$. Unlike the models of Vulić and Moens, these two models do not aggregate single word representations into one vector that represents the context, but compute similarity scores separately with each word from the context. For more details regarding the models, we refer the interested reader to the original Melamud et al.'s work .

We will employ the models of context-sensitive CLSS at the word token level to compare all representation models in the task of suggesting word translations in context in Section 8.

6. Training Setup

In this section, we provide an insight into training data and experimental setup for our BWESG model and all baseline models.

6.1 Training Data

To induce bilingual word embeddings as well as to be directly comparable with baseline representations from prior work, we use a dataset comprising a subset of comparable Wikipedia data available in three language pairs (Vulić & Moens, 2013b, 2014)[8]: (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 of each other. To be directly comparable to prior work in the two evaluation tasks (Vulić & Moens, 2013b, 2014), 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 steps vocabularies comprise between 7,000 and 13,000 noun types for each language in each language pair, and the training corpora are quite small: ranging from approximately 1.5M tokens for NL-EN to 4M for ES-EN. Exactly the same training data and vocabularies are used to train all representation models in comparison (both from Group I and Group II, see Section 4).

We also demonstrate that it is simple and straightforward to train BWESG on parallel sentence-aligned data using the same modeling principles. For that purpose, we use Europarl.v7 (Koehn, 2005) for all three language pairs obtained from the OPUS website (Tiedemann, 2012).[9] As the only preprocessing step, we retain only words occurring at least 5 times in the corpus. Each corpus contains approximately 2M parallel sentences, vocabularies are by an order of magnitude larger than from the smaller Wikipedia data (i.e., varying from 45K EN word types to 75K NL word types), and the corpora sizes are approximately 120M tokens. Data statistics of the two data sources, Wikipedia vs Europarl, are provided in Table 1. The statistics reveal the different nature of the two corpora, with significantly more variance and noise reported for the Wikipedia data.

[8] Available online: `people.cs.kuleuven.be/~ivan.vulic/software/`

[9] `http://opus.lingfil.uu.se/`

Corpus:	**Wikipedia**			**Europarl**		
Pair:	ES-EN	IT-EN	NL-EN	ES-EN	IT-EN	NL-EN
Average length (OTHER)	111	84	51	29	29	27
Average length (EN)	174	154	129	28	29	27
Average length difference	127	125	102	3	4	4

Table 1: Training data statistics: Non-parallel document-aligned Wikipedia vs parallel sentence-aligned Europarl for all three language pairs. OTHER = ES, IT or NL. Lengths are measured in word tokens. Averages are rounded to the closest integer.

6.2 Trained BWESG Models

To test the effect of random shuffling in the *merge and shuffle* BWESG strategy, we have trained the BWESG model with 10 random corpora shuffles for all three training corpora. We also train BWESG with the *length-ratio shuffle* strategy. All parameters are set to default suggested parameters for SGNS from the `word2vec` package: stochastic gradient descent (SGD) with a linearly decreasing global learning rate of 0.025, 25 negative samples, subsampling rate $1e-4$, and 15 epochs.

We have varied the number of dimensions $d = 100, 200, 300$. We have also trained BWESG with $d = 40$ to be directly comparable to readily available sets of BWEs from prior work (Chandar et al., 2014). Moreover, to test the effect of window size on the final results, i.e., the number of positives used for training, we have varied the maximum window size cs from 4 to 60 in steps of 4.[10]

We will make our pre-training and training code for BWESG publicly available, along with all BWESG-based bilingual word embeddings for the three language pairs at: `http://liir.cs.kuleuven.be/software.php`.

6.3 Baseline Representations: Group I

All parameters of the baseline representation models (i.e., topic models and their settings, the number of dimensions K, the values for feature pruning, window size, weighting and similarity functions) were optimized in prior work. Therefore, the settings are adopted directly from previous work (Griffiths et al., 2007; Bullinaria & Levy, 2007; Dinu & Lapata, 2010; Vulić & Moens, 2013a, 2013b; Kiela & Clark, 2014), and we encourage the interested reader to check the details and exact parameter setup in the relevant literature. We provide only a short overview here.

For Basic-MuPTM and Association-MuPTM, as in the work of Vulić and Moens (2013a), a bilingual latent Dirichlet allocation (BiLDA) model was trained with $K = 2000$ topics and the standard values for hyper-parameters: $\alpha = 50/K$, $\beta = 0.01$ (Steyvers & Griffiths, 2007). Post-hoc semantic space pruning was employed with the pruning parameter set to 200 for Basic-MuPTM and to 2000 for Association-MuPTM. We refer the reader to the relevant paper for more details.

For Traditional-PPMI, as in the work of Vulić and Moens (2013b), a seed lexicon was automatically obtained by bootstrapping from the initial seed lexicon of reliable pairs stem-

[10] We remind the reader that we slightly abuse terminology here, as the BWESG windows do not include the locality component any more.

ming from the Association-MuPTM model (with the same parameters for Association-MuPTM as listed above). The window size was fixed to 6 in both directions. We again refer the reader to the paper for more details.

6.4 Baseline Representations: Group II

All baseline BWE models were trained with the same number of dimensions as BWESG: $d = 100, 200, 300$. Other model-specific parameters were taken as suggested in prior work.

For BiCVM, we use the tool released by the authors.[11] We train an additive model, with hinge loss margin $mrg = d$ as in the original paper, batch size of 50, and noise parameter of 10. All models were trained with 200 iterations.

For Mikolov, we train two monolingual SGNS models using the original `word2vec` package, SGD with a global learning rate of 0.025, 25 negative samples, subsampling rate $1e-4$, and 15 epochs. The seed lexicon required to learn the mapping between two monolingual spaces is exactly the same as for Traditional-PPMI.

For BilBOWA, we use SGD with a global learning rate 0.15 for training[12], 25 negative samples, subsampling rate $1e-4$, and 15 epochs. For BilBOWA and Mikolov, we vary the window size the same way as in BWESG.

6.5 Similarity Functions

Unless stated otherwise, a similarity function used in all similarity computations with all RMs is cosine (*cos*). The only exceptions are Basic-MuPTM and Association-MuPTM where the Hellinger distance (HD) was used since it consistently outperformed cosine for these two RM types in prior work (see Footnote 7).

6.6 A Roadmap to Experiments

In the first experiment, we quickly visually inspect the obtained lists of semantically similar words using the BWESG bilingual representation model. Following that, we compare BWESG-based models for bilingual lexicon extraction (BLE) and suggesting word translations in context (SWTC) against both groups of baseline models discussed in Section 4. The experiments and results for the BLE task are presented in Section 7, while the experiments and results for SWTC are presented in Section 8.

7. Evaluation Task I: Bilingual Lexicon Extraction

One may employ the context-insensitive CLSS models from Section 5 to extract bilingual lexicons automatically from data.

[11] `https://github.com/karlmoritz/bicvm`

[12] Suggestions for parameter values received through personal correspondence with the authors. The software is available online: `https://github.com/gouwsmeister/bilbowa`

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	portret	*canvas*	kunstenaar
hijo	*reign*	monarca	fratello	*son*	figlia	schilderen	*brush*	*portrait*
reino	*succession*	heredero	casa	*friend*	figlio	frans	*cubism*	olieverf
reinado	*princess*	*king*	amico	*childhood*	*sister*	nederlands	*art*	portret
regencia	*marriage*	matrimonio	marito	*family*	fratello	componist	*poet*	schilderen
duque	*prince*	*royal*	donna	*cousin*	*wife*	beeldhouwer	*drawing*	*artist*

Table 2: Example lists of top 10 semantically similar words for all 3 language pairs obtained using BWESG (length-ratio shuffle); $d = 200, cs = 48$; (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). The correct one-to-one translation is marked by (+).

7.1 Task Description

By harvesting cross-lingual nearest neighbors, one is able to build a bilingual lexicon of one-to-one translation pairs (w_i^S, w_j^T). We test the validity of our BWEs and baseline representations in the BLE task.

7.2 Experimental Setup

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) by Vulić and Moens (2013a, 2013b). Translation direction is ES/IT/NL → EN. The data is available online.[13]

Evaluation Metrics Since we can build a one-to-one bilingual lexicon by harvesting one-to-one translation pairs, the lexicon quality 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 translation 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 & Moens, 2013b). Similar trends are observed within a more lenient setting with Acc_5 and Acc_{10} scores, but we omit these results for clarity and the fact that the actual BLE performance is best reflected in Acc_1.

[13] `http://people.cs.kuleuven.be/ ivan.vulic/software/`

Spanish-English (ES-EN)				**Italian-English (IT-EN)**			
BWESG	BMu	AMu	TPPMI	BWESG	BMu	AMu	TPPMI
cebolla	**cebolla**	**cebolla**	**cebolla**	**golfo**	**golfo**	**golfo**	**golfo**
onion(+)	dessert	dessert	sauce	gulf(+)	whale	coast	coast
dish	salad	walnut	cheese	coast	dolphin	isthmus	sea
marinade	nut	salad	garlic	coastline	coast	coastline	island
cuisine	walnut	nut	salad	bay	suborder	fjord	bay
soup	rice	hazelnut	chili	island	cadmium	ferry	lagoon
sauce	toast	porridge	onion(+)	peninsula	ferry	monsoon	harbour
cheese	porridge	rice	cuisine	settlement	monsoon	mainland	beach
coriander	paddy	marinade	flavor	shore	fjord	seaside	shore
vegetable	tuber	toast	bread	tourism	isthmus	isle	river
tortilla	potato	paddy	dish	ferry	mainland	suborder	lake

Table 3: Example lists of top 10 semantically similar words for ES-EN and IT-EN, obtained using BWESG (length-ratio, $d = 200, cs = 48$), and the three representation models from Group I. The correct translation is marked by (+).

7.3 Results and Discussion

Table 2 displays top 10 semantically similar words monolingually, across-languages and combined/multilingually for one ES, IT and NL word, while Table 4 shows the first set of BLE results.

7.3.1 Experiment 0: Qualitative Analysis and Comparison

BWESG 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 bilingual embedding space using only document alignments as bilingual signals.[14]

In another brief analysis, we qualitatively compare the cross-lingual ranked lists acquired by BWESG with the other three baseline CLSS/BLE models from Group I. The lists for one ES word and one IT word are presented in Table 3. For the two example words, BWESG is the only model which is able to rank the actual correct translations as nearest cross-lingual neighbors. It is already symptomatic that the word *gulf*, which is the correct translation for *golfo*, does not occur in the ranked list $RL_{10}(golfo)$ at all in case of the three baseline models. We will soon quantitatively confirm this initial suspicion, and demonstrate that BWESG is superior to the three baseline models in the BLE task.

As an aside, Table 3 also clearly reveals the difficulty of judging the quality of models for computing semantic similarity/relatedness solely based on the observed output of the models. The lists $RL_{10}(cebolla)$ and $RL_{10}(golfo)$ appear significantly different across all

[14] We also conducted a small experiment on solving word analogies using monolingual English embedding spaces, and then we repeated the experiment with the same vocabulary and bilingual English-Spanish/Italian/Dutch embedding spaces. The results follow the findings of Faruqui and Dyer (2014), where only slight (and often insignificant) fluctuations for SGNS vectors were reported (e.g., the fluctuations are $< 1\%$ on average in our experiments) when moving from monolingual to bilingual embedding spaces. We may conclude that the linguistic regularities established for monolingual embedding spaces (Mikolov et al., 2013d) induced by SGNS also hold in bilingual embedding spaces induced by BWESG.

Pair:	**ES-EN**			**IT-EN**			**NL-EN**		
BWESG	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
Merge and Shuffle									
cs:16,MIN	0.607	0.600	0.577	0.585	0.597	0.571	0.293	0.244	0.219
cs:16,AVG	0.617	0.613	0.596	0.599	0.601	0.583	0.300	0.254	0.224
cs:16,MAX	0.625	0.630	0.613	0.607	0.606	0.596	0.307	0.267	0.233
cs:48,MIN	0.658	0.676	0.672	0.662	0.677	0.672	0.378	0.366	0.354
cs:48,AVG	0.665	0.685	0.688	0.669	0.683	0.683	0.389	0.381	0.363
cs:48,MAX	0.675	0.694	**0.705**	0.677	**0.692**	0.689	0.394	0.395	0.377
BWESG	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
Length-Ratio									
cs:16	0.627	0.610	0.602	0.613	0.614	0.595	0.303	0.275	0.237
cs:48	**0.678**	**0.701**	0.703	**0.679**	0.689	**0.692**	**0.397**	**0.396**	**0.382**
BWESG	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
No Shuffling									
cs:16	0.218	0.176	0.139	0.209	0.198	0.162	0.070	0.068	0.049
cs:48	0.511	0.497	0.480	0.523	0.540	0.526	0.214	0.198	0.197
BMu	0.441	0.441	0.441	0.575	0.575	0.575	0.237	0.237	0.237
AMu	0.518	0.518	0.518	0.618	0.618	0.618	0.236	0.236	0.236
TPPMI	0.577	0.577	0.577	0.647	0.647	0.647	0.206	0.206	0.206

Table 4: BLE performance in terms of Acc_1 scores for all tested BLE models for Spanish-English, Italian-English and Dutch-English with all bilingual word representations learned from document-aligned Wikipedia data. For BWESG with *merge and shuffle* we report maximum (MAX), minimum (MIN) and average (AVG) scores over 10 random corpora shuffles. Highest scores per column are in bold.

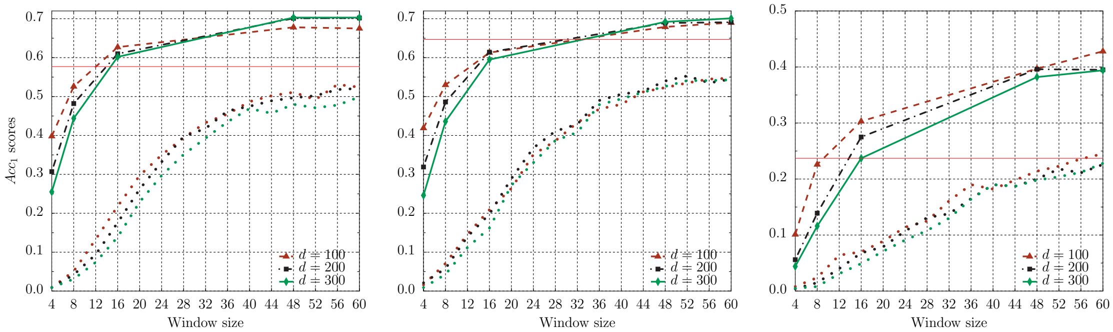
four models, yet all these lists contain words which appear semantically related to the source word. Therefore, we require a more systematic quantitative task-oriented comparison of induced word representations.

7.3.2 Experiment I: BWESG vs Group I

Table 4 shows the first set of results in the BLE task: we report scores with two different BWESG strategies as well as with a BWESG model which does not shuffle pseudo-bilingual documents. The previous best reported Acc_1 scores with baseline representations for the same training+test combination are also reported in the table. By zooming into the table multiple times, we summarize the most important findings.

BWESG vs Baseline Representations The results clearly reveal the superior performance of the BWESG model for BLE which relies on our new framework for inducing bilingual word embeddings from document-aligned comparable data over other BLE models relying on previously used bilingual word representations from the same type of training data. The increase in Acc_1 scores over the best scoring baseline models is 22.2% for ES-EN, 7% for IT-EN and 67.5% for NL-EN.

BWESG Shuffling Strategy Although both BWESG strategies display results that are above established baselines, there is a clear advantage to the *length-ratio shuffle* strategy, which displays a solid and robust performance across a variety of parameters and all three language pairs. Another advantage of that strategy is the fact that it has a deterministic outcome and does not suffer from "sub-optimal" random shuffles. In summary, we suggest



(a) Spanish-English (b) Italian-English (c) Dutch-English

Figure 3: Acc_1 scores in the BLE task with BWESG *length-ratio shuffle* for all 3 language pairs, and varying values for parameters cs and d. Solid (red) horizontal lines denote the highest baseline Acc_1 scores for each language pair. Thicker dotted lines refer to BWESG without shuffling.

using the *length-ratio shuffle* strategy in future work, and along the same line we opt for that strategy in all further experiments.

The results also reveal that shuffling is universally useful, as BWESG without shuffling relies largely on monolingual contexts and cannot reach the performance of BWESG with shuffling. A partial remedy for the problem is to train BWESG with more document-level training pairs (i.e., by increasing the window size), but that leads to prohibitively expensive models, and nonetheless BWESG without shuffling with larger cs-s still falls short of BWESG with both shuffling strategies (see also Figures 3(a)-3(c)).

Window Size: Number of Training Pairs The results confirm the intuition that larger window sizes, i.e., more training examples lead to better results in the BLE task. For all embedding dimensions d-s, BWESG exhibits a superior performance for $cs = 48$ than for $cs = 16$, and the performance with $cs = 48$ and $cs = 60$ seems relatively stable: intuitively, more training pairs leads to a slightly better BLE performance, but the curve slowly flattens out (Figures 3(a)-3(c)). This finding reveals that even a coarse tuning of these parameters might lead to optimal or near-optimal scores for BLE with BWESG.

Differences across Language Pairs A lower increase in Acc_1 scores for IT-EN is attributed to the fact that the test set for IT-EN comprises IT words with occurrence frequencies above 200 in the training data (Vulić & Moens, 2013a), while the other two test sets comprise randomly sampled words covering all frequency spectra. As expected, all models in comparison are able to effectively utilize distributional signals for higher-frequency words, but BWESG still displays the best performance, and these improvements in Acc_1 scores are statistically significant (using McNemar's statistical significance test, $p < 0.05$).[15]

Further, the lowest overall scores for all models in comparison are observed for NL-EN. We attribute it to using less training data for NL-EN when compared to ES-EN and IT-EN (i.e., training corpora for ES-EN and IT-EN are almost triple the size of training corpora for NL-EN). However, we observe that the increase obtained by BWESG is even more prominent in this setting with limited training data. The lower results of TPPMI compared

[15] McNemar's significance test is very common in the NLP literature, especially when Acc_1 scores are reported. It utilizes the standard 2×2 contingency table, and may be observed as a paired version of the more common chi-square test. The reader is referred to the original work of McNemar (1947).

Pair:	**ES-EN**			**IT-EN**			**NL-EN**		
BWESG	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
Length-Ratio									
cs:48	**0.678**	**0.701**	**0.703**	**0.679**	**0.689**	**0.692**	**0.397**	**0.396**	**0.382**
Mikolov									
cs:4	0.187	0.151	0.282	0.368	0.382	0.533	0.042	0.068	0.120
cs:8	0.305	0.306	0.420	0.462	0.518	0.582	0.076	0.095	0.145
cs:16	0.344	0.396	0.486	0.472	0.539	0.602	0.117	0.161	0.184
cs:48	0.311	0.375	0.477	0.458	0.536	0.591	0.132	0.178	0.202
cs:60	0.324	0.389	0.479	0.460	0.538	0.597	0.151	0.180	0.209
BiCVM									
iterations:200	0.342	0.384	0.403	0.309	0.366	0.377	0.068	0.084	0.083

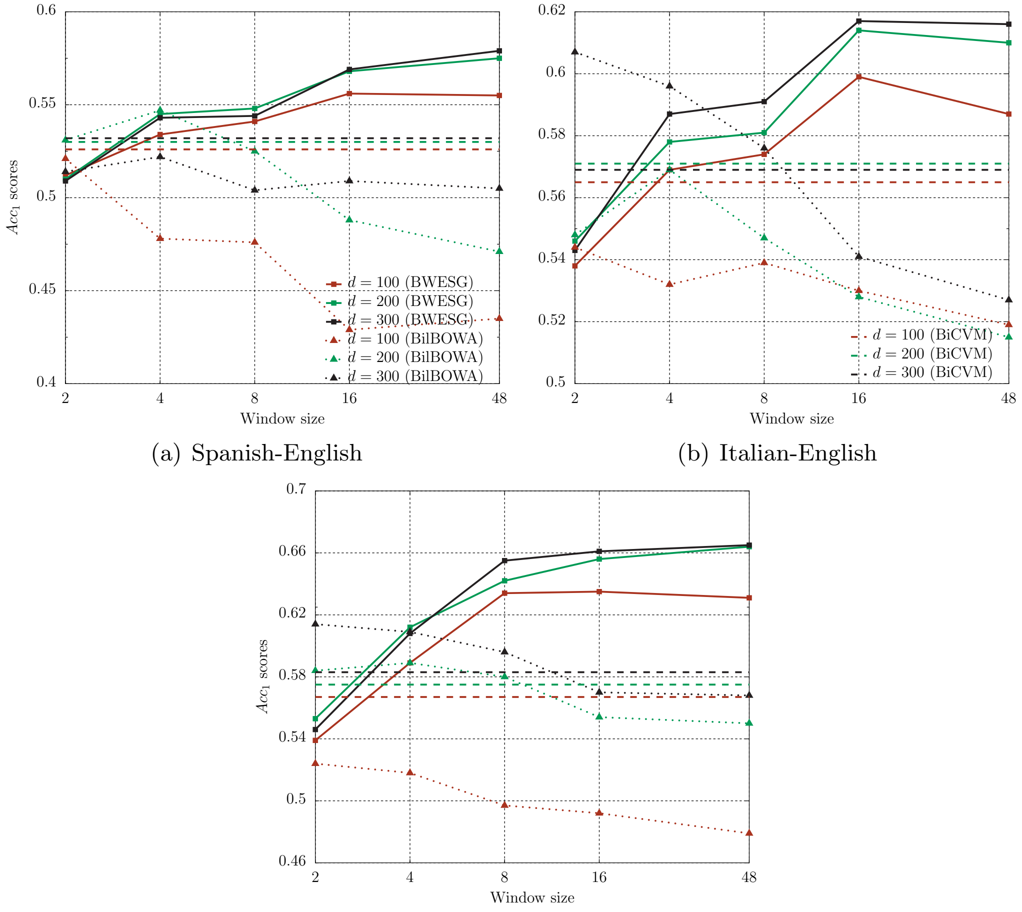
Table 5: BLE results: Comparison of BWESG with (1) the BWE induction model of Mikolov et al. (2013b) relying on SGNS, (2) BiCVM: the BWE induction model of Hermann and Blunsom (2014b) initially developed for parallel sentence-aligned data. All models were trained on the same document-aligned training Wikipedia data with exactly the same vocabularies.

to other two baseline models are also attributed to the overall lower quality and size of NL-EN training data, which is then reflected in a lower quality of seed lexicons necessary to start the bootstrapping procedure from Vulić and Moens (2013b).

Computational Complexity BWESG trained with larger values for d and cs yields richer semantic representations, but also naturally leads to increased training times. However, due to a lightweight design of the supporting SGNS, the times are by the order of magnitude lower than the training times for Basic-MuPTM or Association-MuPTM. Typically, several hours are needed to train BWESG with $d = 300$ and $cs \approx 48 - 60$, whereas it takes two to three days to train a bilingual topic model with $K = 2000$ on the same training set using the multi-threaded architectures on 10 Intel(R) Xeon(R) CPU E5-2667 2.90GHz processors. The BWESG model scales as expected (i.e., training time increases linearly with the window size with all other parameters being equal), and enjoys all the advantages (training time-wise and memory-wise) of the original `word2vec` package. A logical explanation for the behaviour follows from the interpretation of SGNS provided by Levy and Goldberg (2014a), e.g., using a window size of 48 instead of a window size 16 basically means using 3 times more positive examples for training (e.g., approximately 15 minutes is needed to train 300-dimensional ES-EN BWESG embeddings with $cs = 16$ using the Wikipedia data as opposed to 46 minutes with $cs = 48$, measured again on 10 Intel(R) Xeon(R) processors).

7.3.3 Experiment II: BWESG vs Other BWE Induction Models (Group II)

All further experiments are conducted using BWESG with the *length-ratio shuffle* strategy. Note that again all models in comparison use exactly the same data sources and vocabularies as BWESG and Group I models from the previous section. The results with BiCVM and the Mikolov model are summarized in Table 5: the comparison reveals a clear and prominent advantage for the BWESG model given the same data and training setup. We do not report absolute scores of the BilBOWA model in this setup as they were much lower than the other two baseline models. The BiCVM model, although in theory fit to learn from



(a) Spanish-English

(b) Italian-English

(c) Dutch-English

Figure 4: Comparison of BWESG (solid curves) with two other models that rely on parallel training data: (1) BilBOWA (dotted curves), (2) BiCVM: the BWE induction modelinitially developed for parallel sentence-aligned data (dashed horizontal lines). All models were trained on the same sentence-aligned training Europarl data with exactly the same vocabularies. BLE is performed over the same search space for all models. x axes are in log scale.

document-aligned data, is unable to compete with BWESG when learning BWEs from the noisier setting with non-parallel data.

We also present a preliminary study where we compare BWSESG and Group II models in the setup with parallel sentence-aligned data. Results are summarized in Figures 4(a)-4(c).[16] The preliminary results clearly demonstrate that BWESG is able to learn BWEs from parallel data without the slightest change in its modeling principles. While the BilBOWA model displays better results for lower values of the *cs* parameter, to our own surprise, the

[16] Note that the absolute scores are not directly comparable to the BLE scores when the model is trained on Wikipedia data (Tables 4 and 5) due to different training data, different preprocessing steps and vocabularies. Different vocabularies also result in different BLE search spaces and coverages of the test sets (e.g., some very common Spanish nouns from the test set such as *nadador (swimmer)* or *colmillo (tusk)* are not observed in Europarl due to the domain shift).

BWESG model is comparable to or even better than the baseline models with larger window sizes. The BiCVM model, which implicitly utilizes the entire sentence span in training also outperforms BWESG with smaller windows, but BWESG again performs significantly better with larger windows. The BWESG performance flattens out quicker than with the Wikipedia data (compare the results with $cs = 16$ and $cs = 48$), which is easily explained by the decreased length of aligned items as provided in Table 1 (i.e., sentences vs documents).

For English-Spanish, we can also compare BWESG to pre-trained 40-dimensional embeddings of Chandar et al. (2014), as their embeddings were also induced on the same Europarl data. While their model's Acc_1 score is 0.432 for $d = 40$, BWESG obtains Acc_1 scores of 0.502 ($d = 40$, $cs = 8$), 0.535 ($d = 40$, $cs = 16$) or 0.529 ($d = 40$, $cs = 48$).

8. Evaluation Task II: Suggesting Word Translations in Context

In another task, we test the ability of BWEs to produce context-sensitive semantic similarity modeling (see Section 5.1), which in turn may be used to solve the task of suggesting word translations in context (SWTC) proposed recently (Vulić & Moens, 2014). The goal now is to build BWESG-based models for SWTC given the sentential context, similar as in the prior work. We show that our new BWESG-based SWTC models outperform the best SWTC models of Vulić and Moens, as well as other SWTC models which rely on the baseline word representations discussed in Section 4.

8.1 Task Description

Given an occurrence of a polysemous word $w_i \in V^S$ and the context of that occurrence, the SWTC task is to choose the correct translation in the target language L_T of that particular occurrence of w_i from the given set $\mathcal{TC}(w_i) = \{t_1, \ldots, t_{tq}\}$, $\mathcal{TC}(w_i) \subseteq V^T$, of its tq possible translations/meanings. We may refer to $\mathcal{TC}(w_i)$ as an *inventory of translation candidates* for w_i. The task of *suggesting word translations in context* (SWTC) may be interpreted as ranking the tq translation candidates with respect to the observed local context $Con(w_i)$ of the occurrence of the word w_i. The best scoring translation candidate according to the scores $sim(w_i, t_j, Con(w_i))$ (see Section 5.1) in the ranked list is then the correct translation for that particular occurrence of w_i observing its local context $Con(w_i)$.

8.2 Experimental Setup

Test Data We use the SWTC test set introduced recently (Vulić & Moens, 2014). The test set comprises 15 polysemous nouns in three languages (ES, IT and NL) along with sets of their translation candidates (i.e., sets $\mathcal{TC}$). For each polysemous noun, the test sets provide 24 sentences extracted from Wikipedia which illustrate different senses and translations of the pivot polysemous noun, accompanied by the annotated correct translation for each sentence. It yields 360 test sentences for each language pair (and 1080 test sentences in total). An additional set of 100 IT sentences (5 other polysemous IT nouns plus 20 sentences for each noun) is used as a development set to tune the parameter λ (see Section 5.1) for all language pairs and all models in comparison. In summary, the final aim may be formulated as follows: For each polysemous word w_i in ES/IT/NL, the goal is to suggest its correct translation in English given its sentential context.

Evaluation Metrics Since the task is to present a list of possible translations to a SWTC model, and then let the model decide a single most likely translation given the word and its sentential context, we measure the performance again as *Top 1* accuracy (Acc_1).

8.3 Results and Discussion

We again compare against Group I and Group II models. Note that the Group I models held previously best reported SWTC scores when training on the Wikipedia data that we also use in this work.

8.3.1 Experiment I: BWESG vs Group I

Models in Comparison (1) *BWESG+add*. RM: BWESG. SF: cos. Composition: addition. $\lambda = 1.0$. The value for λ suggests that only context is used to disambiguate the meaning of a polysemous word and to guess its most likely translation in context.[17]
(2) *BMu+HD+S*. RM: BasicMuPTM. SF: Hellinger distance. Composition: Smoothed-Fusion[18] of Vulić and Moens (2014). $\lambda = 0.9$.
(3) *BMu+Cue+S*. RM: BasicMuPTM. SF: Cue or Association measure (Steyvers & Griffiths, 2007; Vulić & Moens, 2013a). Composition: Smoothed-Fusion. $\lambda = 0.9$. The Cue similarity is tailored for probabilistic models and computed as the association score $P(t_i|w_i') = \sum_{k=1}^{K} P(t_i|z_k)P(z_k|w_i')$, where z_k denotes k-th latent feature, and $P(z_k|w_i')$ denotes the modulated probability score obtained by smoothing the probabilistic representations of w_i and its context $Con(w_i)$.
(4) *TPPMI+add*. RM: Traditional-PPMI. SF: cos. Composition: addition. $\lambda = 0.9$.

Again, all parameters of the baseline representation models are adopted directly from prior work where they were optimized on development sets comprising additional 100 sentences (Vulić & Moens, 2014). In addition, BMu+HD+S and BMu+Cue+S also rely on the procedure of context sorting and pruning (Vulić & Moens, 2014), where the idea is to retain only context words which are most semantically similar to the given pivot polysemous word, and then use them in computations. The procedure, however, produces significant gains only for probabilistic models (BMu+HD+S and BMu+Cue+S), and therefore, we employ it only for these models. BMu+HD+S and BMu+Cue+S with context sorting and pruning were the best scoring models in the introductory SWTC paper of Vulić and Moens and currently produce state-of-the-art SWTC results on these test sets.[19]

Table 6 summarizes the results and comparison with Group I models on the SWTC task. NO-CONTEXT refers to the context-insensitive majority baseline (i.e., always choosing the most semantically similar translation candidate obtained by BWESG at the word type level, without taking into account any context information).

[17] We have also experimented with the context-sensitive CLSS models proposed by Melamud et al. (2015), but we do not report the actual scores as this model, although displaying a similar relative ranking of different representation models, was consistently outperformed by the models of Vulić and Moens (2014) in our evaluation runs: ≈0.75-0.80 vs ≈0.60-0.65 for the models of Melamud et al. (2015).

[18] In short, Smoothed-Fusion is a probabilistic variant of the context-sensitive modeling idea presented by equations (7)-(9). For more details, check the work of Vulić and Moens (2014).

[19] We omit results for the Association-MuPTM RM since SWTC models based on Association-MuPTM were consistently outperformed by SWTC models based on Basic-MuPTM across different settings.

Pair:	**ES-EN**			**IT-EN**			**NL-EN**		
BWESG+add	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
Length-Ratio									
cs:16	**0.794***	**0.767***	0.752*	**0.817***	0.789	0.794	0.778*	0.769*	0.767*
cs:48	0.752*	0.758*	**0.764***	0.814*	**0.831***	**0.814***	**0.797***	**0.789***	**0.775***
BWESG+add	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
No Shuffling									
cs:16	0.717	0.717	0.694	0.747	0.728	0.728	0.722	0.686	0.678
cs:48	0.731	0.692	0.686	0.775	0.778	0.758	0.739	0.733	0.719
NO-CONTEXT	0.406	0.406	0.406	0.408	0.408	0.408	0.433	0.433	0.433
BMu+HD+S	0.664	0.664	0.664	0.731	0.731	0.731	0.669	0.669	0.669
BMu+Cue+S	0.703	0.703	0.703	0.761	0.761	0.761	0.712	0.712	0.712
TPPMI+add	0.619	0.619	0.619	0.706	0.706	0.706	0.614	0.614	0.614

Table 6: A comparison of SWTC models for Spanish-English, Italian-English and Dutch-English with all bilingual word representations learned from document-aligned Wikipedia data. The asterisk (*) denotes statistically significant improvements of BWESG+add over the strongest baseline according to a McNemar's statistical significance test ($p < 0.05$). Highest scores per column are in bold.

BWESG vs Baseline Representations The results reveal that BWESG outperforms baseline bilingual word representations from Group I also in the SWTC task. The improvements are prominent for all reported values of parameters d and cs, and are often statistically significant even when compared to the strongest baseline (which is the fine-tuned BMu+Cue+S model with context sorting and pruning for all three language pairs from Vulić & Moens, 2014). The increase in Acc_1 scores over the strongest baseline is 12.9% for ES-EN, 11.9% for IT-EN, and 12.4% for NL-EN. The obtained results surpass previous state-of-the-art scores and are currently the best reported results on the SWTC datasets when using non-parallel data to learn semantic representations.

BWESG Shuffling Strategy Although BWESG without shuffling (due to a reduced complexity of the SWTC task compared to BLE) already displays encouraging results, there is again a clear advantage to the *length-ratio shuffle* strategy, which displays an excellent performance for all three language pairs. In simple words, shuffling is again useful.

Dimensionality and Number of Training Pairs Unlike in the BLE task, the highest Acc_1 scores on average are obtained by using lower-dimensional word embeddings (i.e., $d = 100$). The phenomenon may be attributed to the effect of semantic composition and the reduced complexity of the SWTC task compared to the BLE task. First, although enlarging the dimensionality of embeddings leads to an increased semantic expressiveness within

Senses:	**2 senses**	**3 senses**	**4 senses**
Model	Acc_1	Acc_1	Acc_1
BMu+Cue+S	0.827	0.619	0.417
BWESG+add	**0.834**	**0.804**	**0.583**

Table 7: A comparison of the best scoring baseline model BMu+Cue+S and the best scoring BWESG+add model over different clusters of words (2-sense, 3-sense and 4-sense words) for Spanish-English.

the shared bilingual embedding space, it may be harmful when working with composition models, since the simple additive model of semantic composition may produce more erroneous dimensions when constructing higher-dimensional context embeddings out of single word embeddings. Second, due to its design, the SWTC task requires coarser-grained representations than BLE. While in the BLE task the goal is to detect a translation of a word from a vocabulary which typically spans (tens of) thousands of words, in the SWTC task the goal is to detect the most likely translation of a word given its sentential context, but from a small closed vocabulary of 2-4 possible translations from the translation inventory. Therefore, it is highly likely that even low-dimensional embeddings are sufficient to produce plausible rankings for the SWTC task, while at the same time, they are not sufficient and expressive enough to find correct translations in BLE. More training pairs (i.e., larger windows) still yield better results on average in the SWTC task. In summary, the choice of representation granularity is dependent on the actual task, which consequently leads to the conclusion that optimal values for d and cs are largely task-specific (compare also results in Table 4 and Table 6).

Testing Polysemy In order to test whether the gain in performance for BWESG+add is derived mostly from the effective handling of the easiest set of words, that is, bisemous words (polysemous words with only 2 translation candidates), we have performed an additional experiment, where we have measured Acc_1 scores separately for words with 2, 3, and 4 different senses. Results indicate that the performance gain comes mostly from gains on trisemous and tetrasemous words, while the scores on bisemous words are comparable. Table 7 shows Acc_1 over different clusters of words for ES-EN, and similar scoring patterns are observed for IT-EN and NL-EN.

Differences across Language Pairs Due to the reduced complexity of SWTC, we may also observe relatively higher results for NL-EN when compared to ES-EN and IT-EN, as opposed to their relative performance in the BLE task, where the scores for NL-EN are much lower than scores for ES-EN and IT-EN. Since SWTC is a less difficult task which requires coarse-grained representations, even limited amounts of training data may be sufficient to learn word embeddings which are useful for the specific task. This finding is in line with the recent work of Gouws and Søgaard (2015).

8.3.2 Experiment II: BWESG vs. Other BWE Induction Models (Group II)

We again test other BWE induction models in the SWTC task, using the same training setup and sets of embeddings as introduced in Section 7.3.3 for the BLE task. The representations were now plugged in the context-sensitive CLSS modeling framework from Section 5.1, and the optimization of parameters for SWTC has been conducted in the same manner as for BWESG. The results with the Mikolov model and BiCVM are summarized in Table 8. The results with BilBOWA are very similar to BiCVM, so we do not report it for brevity.

BWESG outperforms other BWE induction models in the SWTC task and further confirms its utility in cross-lingual semantic modeling. The model of Mikolov et al. (2013b) constitutes a stronger baseline: Good results in the SWTC task with this model are an interesting finding *per se*. While the model is not competitive with BWESG and other baseline representations models from document-aligned data in a more difficult BLE task when using noisy one-to-one translation pairs, its performance on the less complex SWTC

Pair:	**ES-EN**			**IT-EN**			**NL-EN**		
BWESG+add **Length-Ratio**	d=100	d=200	d=300	d=100	d=200	d=300	d=100	d=200	d=300
cs:16	**0.794**	**0.767**	0.752	**0.817**	0.789	0.794	0.778	0.769	0.767
cs:48	0.752	0.758	**0.764**	0.814	**0.831**	**0.814**	**0.797**	**0.789**	**0.775**
Mikolov									
cs:4	0.742	0.739	0.725	0.733	0.706	0.692	0.692	0.700	0.700
cs:8	0.767	0.750	0.747	0.767	0.747	0.744	0.694	0.697	0.672
cs:16	0.769	0.744	0.747	0.758	0.755	0.758	0.725	0.700	0.689
cs:48	0.678	0.642	0.669	0.714	0.714	0.747	0.725	0.711	0.708
cs:60	0.636	0.658	0.656	0.725	0.725	0.742	0.722	0.728	0.722
BiCVM									
iterations:200	0.547	0.567	0.539	0.636	0.664	0.642	0.586	0.567	0.581

Table 8: SWTC results: Comparison of BWESG with (1) the BWE induction model of Mikolov et al. (2013b) relying on SGNS, (2) BiCVM: the BWE induction model of Hermann and Blunsom (2014b) initially developed for parallel sentence-aligned data. All models were trained on the same document-aligned training Wikipedia data with exactly the same vocabularies.

task with a reduced search space is solid even when the model relies on the imperfect set of translation pairs to learn the mapping between two monolingual embedding spaces.

8.3.3 Further Discussion

By analyzing the influence of pre-training shuffling on the results in two different evaluation tasks, we may safely establish its utility when inducing bilingual word embeddings using the BWESG model. While we have already presented two shuffling strategies in this work, one line of future work will investigate different possibilities of "blending in" words from two different vocabularies into pseudo-bilingual documents in a more structured and systematic manner. For instance, one approach to generating pseudo-training sentences for learning from textual and perceptual modalities has been recently introduced (Hill & Korhonen, 2014). However, it is not straightforward how to extend this approach to the generation of pseudo-bilingual training documents.

Another idea in the same vein is to build artificial training data of higher-quality starting from noisy comparable data by: (1) computing semantically similar words monolingually and across-languages from the noisy data, (2) retaining only highly reliable pairs of similar words using an automatic selection procedure (Vulić & Moens, 2012), (3) building pseudo-bilingual documents using only reliable context word pairs. In other words, the questions is: Is it possible to choose positive training pairs more systematically to reduce the noise stemming from non-parallel data? The construction of such artificial training data and training on such data would then proceed in a bootstrapping fashion, and the model should be able to steadily reduce noise inherently present in comparable data. The idea of "improving corpus comparability" was only touched upon in previous work (Li & Gaussier, 2010; Li, Gaussier, & Aizawa, 2011).

While the entire framework proposed in this article is in theory completely language pair agnostic as it does not make any language pair dependent modeling assumptions, we acknowledge the fact that all three language pairs comprise languages coming from the same phylum, that is, the Indo-European language family. Future extensions also include porting the framework to other more distant language pairs that do not share the same roots nor the same alphabet (e.g., English-Chinese/Hindi/Arabic), and for which benchmarking test sets are still scarce for a variety of semantic tasks (e.g., SWTC) (Camacho-Collados, Pilehvar, & Navigli, 2015). We believe that larger window sizes may solve difficulties with different word orderings (e.g., for Chinese-English).

9. Conclusions and Future Work

We have proposed and described Bilingual Word Embeddings Skip-Gram (BWESG), a simple yet effective bilingual word representation learning model which is able to induce bilingual word embeddings solely on the basis of document-aligned comparable data. BWESG is based on the omnipresent skip-gram with negative sampling (SGNS). We have presented two ways to build pseudo-bilingual documents on which a monolingual SGNS (or any monolingual WE induction model) may be trained to produce shared bilingual embedding spaces. The BWESG model does not make any language-pair dependent assumptions nor requires language-pair specific external resources such as bilingual lexicons, predefined category/ontology knowledge or parallel data. We have showed that the model may be trained on non-parallel and parallel data without any changes in modeling principles, which, complemented with its simplicity and lightweight design makes it potentially very useful as a tool for researchers in machine translation and information retrieval.

We have employed induced BWEs in two semantic tasks: (1) bilingual lexicon extraction (BLE), and (2) suggesting word translations in context (SWTC). Our new BWESG-based BLE and SWTC models outperform previous state-of-the-art models for BLE and SWTC from document-aligned comparable data and related BWE induction models (Mikolov et al., 2013b; Chandar et al., 2014; Gouws et al., 2015). The findings in this article follow the recently published surveys from Baroni et al. (2014), and Levy et al. (2015) regarding a solid and robust performance of neural word representations/word embeddings in semantic tasks: our new BWESG-based models for BLE and SWTC significantly outscore previous state-of-the-art distributional approaches on both tasks across different parameter settings. Even more encouraging is the fact that these new state-of-the-art results are attained using default parameter settings for the BWESG model as suggested in the `word2vec` package without any development set. Further (finer) tuning of model parameters in future work may lead to higher-quality bilingual embedding spaces.

Several straightforward lines of future research have already been tackled in Section 7 and Section 8. For instance, the current *length-ratio* shuffling strategy may be replaced by a more advanced shuffling method in future work. Moreover, BWEs induced by BWESG may be used in other semantic tasks besides the ones discussed in this work, and it would be interesting to experiment with other types of context aggregation and selection beyond the bag-of-words assumption, such as dependency-based contexts (Levy & Goldberg, 2014a), or other objective functions during training in the same vein as proposed by Levy and Goldberg (2014b). Similar to the evolution in multilingual probabilistic topic modeling, another path of future work may lead to investigating bilingual models for learning BWEs which will be able to jointly learn from separate documents in aligned document pairs, without the need to construct pseudo-bilingual documents.

A natural step in the text representation learning research is to extend the focus from single word representations to composite phrase, sentence and document representations (Hermann & Blunsom, 2013; Kalchbrenner, Grefenstette, & Blunsom, 2014; Le & Mikolov, 2014; Soyer et al., 2015; Kiros, Zhu, Salakhutdinov, Zemel, Torralba, Urtasun, & Fidler, 2015; Hill, Cho, Korhonen, & Bengio, 2016). In this article, we have relied on a simple composition model based on vector addition, and have shown that this model performs excellent in the SWTC task. However, in the long run this model is not by any means sufficient to effectively capture all complex compositional phenomena in the data. Several models which aim to learn sentence and document embeddings have been proposed recently, but they critically rely on sentence-aligned parallel data. It is yet to be seen how to build structured multilingual phrase, sentence and document embeddings solely on the basis of comparable data. Such low-cost multilingual embeddings beyond the word level extracted from comparable data may find its application in a variety of tasks such as statistical machine translation (Mikolov et al., 2013b; Zou et al., 2013; Zhang et al., 2014; Wu et al., 2014), semantic tasks such as multilingual semantic textual similarity (Agirre, Banea, Cardie, Cer, Diab, Gonzalez-Agirre, Guo, Mihalcea, Rigau, & Wiebe, 2014), cross-lingual information retrieval (Vulić et al., 2013; Vulić & Moens, 2015) or cross-lingual document classification (Klementiev et al., 2012; Hermann & Blunsom, 2014b; Chandar et al., 2014).

In another future research path, 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 learning from large unaligned multilingual datasets (Mikolov et al., 2013b; Al-Rfou, Perozzi, & Skiena, 2013). In the long run, this idea may lead to large-scale learning models from huge amounts of multilingual data without any requirement for parallel data or manually built bilingual lexicons.

Acknowledgments

This work was done while Ivan Vulić was a postdoctoral researcher at Department of Computer Science, KU Leuven supported by the PDM Kort fellowship (PDMK/14/117). The work was also supported by the SCATE project (IWT-SBO 130041) and the ERC Consolidator Grant LEXICAL: Lexical Acquisition Across Languages (648909). We would also like to thank the anonymous reviews for their helpful suggestions that helped us to greatly improve the presentation of our work.

References

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., & Wiebe, J. (2014). SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SEMEVAL)*, pp. 81–91. Association for Computational Linguistics.

Al-Rfou, R., Perozzi, B., & Skiena, S. (2013). Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL)*, pp. 183–192.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 238–247.

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1183–1193.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 546–556.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J., & Blei, D. M. (2009). Multilingual topic models for unaligned text. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 75–82.

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

Camacho-Collados, J., Pilehvar, M. T., & Navigli, R. (2015). A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1–7.

Carbonell, J. G., Yang, J. G., Frederking, R. E., Brown, R. D., Geng, Y., Lee, D., Frederking, Y., E, R., Geng, R. D., & Yang, Y. (1997). Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 708–714.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, *1*(4), 300–307.

Chandar, S., Lauly, S., Larochelle, H., Khapra, M. M., Ravindran, B., Raykar, V. C., & Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 1853–1861.

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 740–750.

Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, *38*(1), 41–71.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 160–167.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Das, D., & Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 600–609.

Daumé III, H., & Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 407–412.

De Smet, W., & Moens, M.-F. (2009). Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM@CIKM)*, pp. 57–64.

Deschacht, K., De Belder, J., & Moens, M.-F. (2012). The latent words language model. *Computer Speech & Language*, *26*(5), 384–409.

Deschacht, K., & Moens, M.-F. (2009). Semi-supervised semantic role labeling using the latent words language model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 21–29.

Dinu, G., & Lapata, M. (2010). Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1162–1172.

Dinu, G., Lazaridou, A., & Baroni, M. (2015). Improving zero-shot learning by mitigating the hubness problem. In *ICLR Workshop Papers*.

Duchi, J. C., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, *12*, 2121–2159.

Dumais, S. T., Landauer, T. K., & Littman, M. (1996). Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *Proceedings of the SIGIR Workshop on Cross-Linguistic Information Retrieval*, pp. 16–23.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Faruqui, M., & Dyer, C. (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)*, pp. 462–471.

Fukumasu, K., Eguchi, K., & Xing, E. P. (2012). Symmetric correspondence topic models for multilingual text analysis. In *Proceedings of the 25th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 1295–1303.

Ganchev, K., & Das, D. (2013). Cross-lingual discriminative learning of sequence models with posterior regularization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1996–2006.

Gaussier, É., Renders, J.-M., Matveeva, I., Goutte, C., & Déjean, H. (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)*, pp. 526–533.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(6), 721–741.

Goldberg, Y., & Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *CoRR*, *abs/1402.3722*.

Gouws, S., Bengio, Y., & Corrado, G. (2015). BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 748–756.

Gouws, S., & Søgaard, A. (2015). Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1386–1390.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 771–779.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hermann, K. M., & Blunsom, P. (2013). The role of syntax in vector space models of compositional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 894–904.

Hermann, K. M., & Blunsom, P. (2014a). Multilingual distributed representations without word alignment. In *Proceedings of the 2014 International Conference on Learning Representations (ICLR)*.

Hermann, K. M., & Blunsom, P. (2014b). Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 58–68.

Hill, F., Cho, K., Korhonen, A., & Bengio, Y. (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the ACL*, *4*, 17–30.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 255–265.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 655–665.

Kazama, J., Saeger, S. D., Kuroda, K., Murata, M., & Torisawa, K. (2010). A Bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 247–256.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 36–45.

Kiela, D., & Clark, S. (2014). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pp. 21–30.

Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R. S., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-thought vectors. In *Proceedings of the 28th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*.

Klementiev, A., Titov, I., & Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pp. 1459–1474.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT SUMMIT)*, pp. 79–86.

Kočiský, T., Hermann, K. M., & Blunsom, P. (2014). Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 224–229.

Landauer, T. K., & Dumais, S. T. (1997). Solutions to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Laroche, A., & Langlais, P. (2010). Revisiting context-based projection methods for term-translation spotting in comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 617–625.

Lazaridou, A., Dinu, G., & Baroni, M. (2015). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, pp. 270–280.

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pp. 1188–1196.

Lebret, R., & Collobert, R. (2014). Word embeddings through Hellinger PCA. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 482–490.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 25–32.

Levow, G.-A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management*, *41*(3), 523–547.

Levy, O., & Goldberg, Y. (2014a). Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 302–308.

Levy, O., & Goldberg, Y. (2014b). Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 2177–2185.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, *3*, 211–225.

Li, B., & Gaussier, É. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pp. 644–652.

Li, B., Gaussier, É., & Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 473–478.

Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., & Hu, Y. (2015). Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1501–1511.

Liu, X., Duh, K., & Matsumoto, Y. (2013). Topic models + word alignment = a flexible framework for extracting bilingual dictionary from comparable corpus. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*, pp. 212–221.

Lu, A., Wang, W., Bansal, M., Gimpel, K., & Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 250–256.

Luong, T., Pham, H., & Manning, C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 151–159.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157.

Melamud, O., Levy, O., & Dagan, I. (2015). A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pp. 1–7.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of the 2013 International Conference on Learning Representations (ICLR): Workshop Papers*.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *CoRR*, *abs/1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 3111–3119.

Mikolov, T., Yih, W., & Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 746–751.

Milajevs, D., Kartsaklis, D., Sadrzadeh, M., & Purver, M. (2014). Evaluating neural word representations in tensor-based compositional settings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 708–719.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 880–889.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 236–244.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pp. 2265–2273.

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2009). Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International World Wide Web Conference (WWW)*, pp. 1155–1156.

Ni, X., Sun, J.-T., Hu, J., & Chen, Z. (2011). Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of the 4th International Conference on Web Search and Web Data Mining (WSDM)*, pp. 375–384.

Padó, S., & Lapata, M. (2009). Cross-lingual annotation projection for semantic roles. *Journal of Artificial Intelligence Research*, *36*, 307–340.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 613–619.

Peirsman, Y., & Padó, S. (2010). Cross-lingual induction of selectional preferences with bilingual vector spaces. In *Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 921–929.

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

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability.* Cambridge University Press.

Rapp, R. (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)*, pp. 519–526.

Reisinger, J., & Mooney, R. J. (2010). A mixture model with sharing for lexical semantics. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1182.

Rudolph, S., & Giesbrecht, E. (2010). Compositional matrix-space models of language. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 907–916.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

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

Shi, T., Liu, Z., Liu, Y., & Sun, M. (2015). Learning cross-lingual word embeddings via matrix co-factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 567–572.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1201–1211.

Søgaard, A., Agić, v., Martínez Alonso, H., Plank, B., Bohnet, B., & Johannsen, A. (2015). Inverted indexing for cross-lingual nlp. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1713–1722.

Soyer, H., Stenetorp, P., & Aizawa, A. (2015). Leveraging monolingual data for crosslingual compositional word representations. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR).*

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of Latent Semantic Analysis*, *427*(7), 424–440.

Stratos, K., Collins, M., & Hsu, D. (2015). Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 1282–1291.

Täckström, O., Das, D., Petrov, S., McDonald, R., & Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the ACL*, *1*, 1–12.

Tamura, A., Watanabe, T., & Sumita, E. (2012). Bilingual lexicon extraction from comparable corpora using label propagation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 24–36.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, pp. 2214–2218.

Tiedemann, J., Agić, Z., & Nivre, J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, pp. 130–140.

Trask, A., Gilmore, D., & Russell, M. (2015). Modeling order in neural word embeddings at scale. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2266–2275.

Turian, J. P., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 384–394.

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

Vulić, I., De Smet, W., & Moens, M.-F. (2011). Identifying word translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pp. 479–484.

Vulić, I., De Smet, W., & Moens, M.-F. (2013). Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, *16*(3), 331–368.

Vulić, I., De Smet, W., Tang, J., & Moens, M. (2015). Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management*, *51*(1), 111–147.

Vulić, I., & Moens, M.-F. (2012). Detecting highly confident word translations from comparable corpora without any prior knowledge. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 449–459.

Vulić, I., & Moens, M.-F. (2013a). Cross-lingual semantic similarity of words as the similarity of their semantic word responses. In *Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 106–116.

Vulić, I., & Moens, M.-F. (2013b). 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)*, pp. 1613–1624.

Vulić, I., & Moens, M.-F. (2014). Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data.

In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 349–362.

Vulić, I., & Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 363–372.

Wu, H., Dong, D., Hu, X., Yu, D., He, W., Wu, H., Wang, H., & Liu, T. (2014). Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 142–146.

Wu, H., Wang, H., & Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pp. 993–1000.

Xiao, M., & Guo, Y. (2014). Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL)*, pp. 119–129.

Yarowsky, D., & Ngai, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 200–207.

Zhang, D., Mei, Q., & Zhai, C. (2010). Cross-lingual latent topic extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1128–1137.

Zhang, J., Liu, S., Li, M., Zhou, M., & Zong, C. (2014). Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 111–121.

Zou, W. Y., Socher, R., Cer, D., & Manning, C. D. (2013). Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1393–1398.