Enhancing Semantic Representations of Bilingual Word Embeddings with Syntactic Dependencies

Linli Xu, Wenjun Ouyang, Xiaoying Ren, Yang Wang and **Liang Jiang**
Anhui Province Key Laboratory of Big Data Analysis and Application
School of Computer Science and Technology, University of Science and Technology of China
linlixu@ustc.edu.cn, {oy01, wjren}@mail.ustc.edu.cn
angyan@ustc.edu.cn, jal@mail.ustc.edu.cn

Abstract

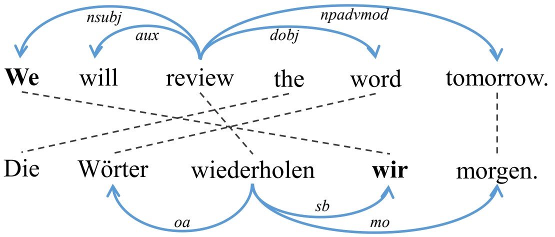
Cross-lingual representation is a technique that can both represent different languages in the same latent vector space and enable the knowledge transfer across languages. To learn such representations, most of existing works require parallel sentences with word-level alignments and assume that aligned words have similar Bag-of-Words (BoW) contexts. However, due to differences in grammar structures among different languages, the contexts of aligned words in different languages may appear at different positions of the sentences. To address this issue of different syntactics across different languages, we propose a model of bilingual word embedding integrating syntactic dependencies (DepBiWE) by producing dependency parse-trees which encode the accurate relative positions for the contexts of aligned words. In addition, a new method is proposed to learn bilingual word embeddings from dependency-based contexts and BoW contexts jointly. Extensive experimental results on a real world dataset clearly validate the superiority of the proposed model DepBiWE on various natural language processing (NLP) tasks.

1 Introduction

Distributed word representations, also known as word embeddings, have been extensively applied in various natural language processing (NLP) tasks. Different to the traditional representation of words as discrete and distinct symbols, distributed word representation embeds words into a low dimensional continuous vector space according to the distributional hypothesis stating that words with similar contexts have similar semantic meanings [Collobert and Weston, 2008; Mikolov *et al.*, 2013c; Levy *et al.*, 2015]. Specifically, with the successful applications of BoW (Bag-of-Words)-based methods [Mikolov *et al.*, 2013c] in the monolingual scenarios, including language modeling [Bengio *et al.*, 2003], text classification [Kim, 2014] and parsing [Socher *et al.*, 2013], studies have been conducted to extend the monolingual methods to cross-lingual scenarios, especially the tasks that require knowledge transfer from high-resource languages to low-resource languages, e.g., cross-lingual semantic analysis [Zhou *et al.*, 2015] and cross-lingual document classification [Klementiev *et al.*, 2012]. In principle, it is possible to map vocabularies of two or more languages into a shared vector space of cross-lingual representations because there exists a strong similarity between the vector spaces of different languages [Mikolov *et al.*, 2013b], with resembling semantic properties of word pairs across languages.

Various approaches have been proposed for cross-lingual word embeddings in the literature, which can be divided into three rough categories. The first uses a cross-lingual dictionary with translation between pairs of words in different languages [Gouws and Søgaard, 2015; Duong *et al.*, 2016]. The second group of methods are proposed based on sentence-aligned parallel corpus and can be applied to machine translation tasks [AP *et al.*, 2014; Gouws *et al.*, 2015; Shi *et al.*, 2015]. The third category leverages both sentence-level alignment and word-level alignment when learning word embeddings. Among them, a cross-lingual regularization is introduced in [Klementiev *et al.*, 2012] to pull the embedding vectors of aligned words closer. Word alignment is also leveraged in the model of CLC+WA [Shi *et al.*, 2015] by counting the co-occurrence using BoW contexts. In [Luong *et al.*, 2015], the skip-gram model [Mikolov *et al.*, 2013c] is extended to the cross-lingual setting by integrating the monolingual and cross-lingual objectives, and predicting the BoW contexts of both the target words and their aligned words.

The above methods are restricted in the sense that they assume that aligned words have similar contexts, while ignoring the difference of word orders across different languages. As a counter-example, Figure 1-**Top** shows a pair of sentences of different parse structures across different languages (English-German). As shown in Figure 1, given a window size of 1, the BoW context of the word *we*, which is *will*, is irrelevant to *wiederholen* or *morgen*, which are the BoW contexts of the word *wir* in German. With the increase of the window size, only an uncorrelated word *Wörter* is included. Figure 1 shows a contradiction to the bilingual distributional hypothesis, which is inherited from the monolingual counterpart, that words with similar contexts across different languages should have similar meanings.



WORDS	BoW contexts	DEP contexts
we	*will*, review	review
wir	*Wörter*, wiederholen, *morgen*	wiederholen
word	*will*, review, *tomorrow*	review
Wörter	wiederholen, *wir*	wiederholen

Figure 1: An example of BoW (Bag-of-Words) contexts (window size=2) and DEP (dependency) contexts extracted for bilingual models. **Top**: the word alignments of an example parallel sentence pair (English-German) with the corresponding dependency parse-trees. **Bottom**: the BoW contexts and DEP contexts extracted for words *we* (*wir*) and *word* (*Wörter*) in two sentences.

To address the above issue, we consider different syntactic structures in parallel sentences across languages as important clues when obtaining the contexts of aligned words in the bilingual model. Specifically, we propose a dependency-based bilingual word embedding model (DepBiWE) where a pair of dependency parse-trees [De Marneffe and Manning, 2008] are produced to capture the syntactic contexts of aligned words across languages. This is illustrated in Figure 1, the context of word *we* according to the dependency parse-tree is *review*, which is similar to *wiederholen* in the parallel sentence. In addition, we further introduce a regularization term (DepBiWE+R) to enhance the quality of cross-lingual embeddings by pulling the representations of similar words close to each other in the phrase-level semantic space. Furthermore, considering that building dependency parse-trees can be expensive on large-scale corpus, we propose a cross learning method which integrates BoW contexts as an unsupervised supplementary context information and learns the word representations based on the dependency and BoW contexts jointly. By integrating the BoW-based topical contexts and the dependency-based syntactic contexts, the cross-lingual performance can been effectively improved while both semantics and syntactics are preserved in the learned bilingual word embeddings.

To evaluate the quality of the embeddings learned by the proposed bilingual word embedding model, experimental investigation is conducted on monolingual word similarity, cross-lingual word similarity and cross-lingual dictionary induction tasks, which demonstrates significant improvements over the state-of-the-art. We further apply the proposed method to the task of cross-lingual document classification on real-world datasets to justify the practical effectiveness of the DepBiWE model by exploiting syntactic dependencies.

The main contributions of this paper are:

- We consider different syntactic structures in parallel sentences across different languages. By obtaining the contexts of aligned words across different languages with dependency parse-trees, a novel bilingual word embedding model is designed.
- We propose a new cross learning method which learns the word representations based on the dependency contexts and BoW contexts jointly.
- The proposed bilingual embedding model achieves a significant improvement over the state-of-the-art methods.

2 Related Work

2.1 Dependency-based Word Embeddings

Traditional neural word embedding models transform the word representation problem into a word prediction task [Collobert and Weston, 2008; Mikolov *et al.*, 2013a; 2013c]. However, most of word representation techniques rely on linear BoW contexts. Recently, [Levy and Goldberg, 2014] first propose a dependency-based word embedding model, which captures the dependency-based word syntactic contexts instead. Sequentially, linear BoW contexts and dependency paths are integrated in [Yin *et al.*, 2016] for aspect term extraction. On the other hand, for the bilingual word representation problem, syntactic dependencies can provide more important clues due to different word orders across languages. In this paper, we incorporate syntactic structures in the parallel corpus to better encode the semantic and syntactic information in bilingual word embeddings.

2.2 Bilingual Word Embeddings

Existing methods of bilingual word representations can be grouped into three categories: *monolingual mapping*, *monolingual adaption* and *cross-lingual training*.

In *monolingual mapping*, word embeddings are first trained on each monolingual corpora independently [Mikolov *et al.*, 2013c], and then a transformation matrix is learned which maps word representations from one language to another language. Among them, [Mikolov *et al.*, 2013b] utilize a set of meaning-equivalent pairs to learn the linear mapping; while canonical correlation analysis (CCA) is employed in [Faruqui and Dyer, 2014] to project words from two languages to a shared bilingual embedding space. On the other hand, *monolingual adaption* jointly optimizes the monolingual objectives of each language, with a cross-lingual objective to enforce the bilingual constraint [Zou *et al.*, 2013].

Unlike the schemes above which fix representations on either one or both languages, *cross-lingual training* learns bilingual word embeddings from a parallel corpus by optimizing a cross-lingual objective that encourages embeddings of similar words from different languages to be close to each other in a common vector space. Representatives include the methods that train cross-lingual word embeddings using a bilingual dictionary with pairs of translations between words in different languages. Alternatively, supervised information of sentence-level alignments can be introduced instead of word-level alignments [AP *et al.*, 2014; Gouws *et al.*, 2015; Shi *et al.*, 2015]. Word-level and sentence-level alignments can be further combined. As an example, a multi-task learning framework is proposed in [Klementiev *et al.*, 2012]; while in BiSkip [Luong *et al.*, 2015], the skip-gram model [Mikolov *et al.*, 2013c] is extended to bilingual scenarios where separate contexts of aligned word pairs are jointly predicted.

3 Bilingual Word Embeddings with Syntactic Dependencies

We consider the task of learning bilingual word representations from two languages l_1 and l_2. Specifically, the goal is to learn word embedding matrices of the vocabularies in two languages, and the training data consists of two monolingual corpus and a parallel corpus with word alignments. Let W^{l_i} $(i = 1, 2)$ be the vocabulary of language l_i and $\mathbf{W}^{l_i} \in \mathbb{R}^{|W^{l_i}| \times d}$ be the corresponding word embedding matrix, where d is dimensionality of the word embedding vector. We further denote the vocabulary of contexts in language l_i as C^{l_i}, with the corresponding context embedding matrix $\mathbf{C}^{l_i} \in \mathbb{R}^{|C^{l_i}| \times d}$. The embedding vectors of a word w and a context c are represented by $\mathbf{w}$ and $\mathbf{c}$ respectively.

To enhance the quality of bilingual word embeddings across different languages, we leverage a general objective which consists of the monolingual components from each language, as well as the cross-lingual component. The joint objective can be formulated as:

$$L = \alpha(L^{l_1}_{\text{mono}} + L^{l_2}_{\text{mono}}) + \beta L_{\text{cross}}, \tag{1}$$

where $L^{l_1}_{\text{mono}}$ and $L^{l_2}_{\text{mono}}$ denote the monolingual objectives derived from the languages l_1 and l_2 respectively, while L_{cross} corresponds to the cross-lingual objective which is used to map the two monolingual word embeddings across languages into a common vector space. α and β are hyperparameters that balance the importance of the monolingual terms and the cross-lingual term.

3.1 Dependency-based Monolingual Objectives

Due to different syntactic structures across different languages, syntactic dependencies can provide more important clues when designing the monolingual objectives in the task of mapping two vocabularies into a shared semantic space. To formulate the monolingual objective for each language, we follow the paradigm of [Levy and Goldberg, 2014], which generalizes the skip-gram principle [Mikolov *et al.*, 2013c] by extending from BoW contexts to dependency contexts.

Formally, for each language, after parsing a given sentence $s = \{w_1, w_2, \ldots, w_{|s|}\}$, we derive the corresponding DEP contexts. Specifically, for a target word w_i with modifiers $m_1, m_2, ..., m_k$ and a head h, we construct the DEP contexts in the form of $(m_1, dr_1), ..., (m_k, dr_k)$, (h, dr_h^{-1}), where dr denotes the type of dependency relation (e.g. *amod*, *nsubj* or *root*) between the head and the modifier, while dr^{-1} is the inverse relation. For example, in Figure 1, the dependency relation between *we* and *review* is *nsubj*. After deriving the DEP contexts, we consider a word-context pair (w, c) where c is the DEP context of w and utilize the target word w to predict its DEP context. The dependency-based monolingual objective function for language l_i can be defined as:

$$L^{l_i}_{\text{mono}} = \sum_{(w,c)\in D^{l_i}_m} L^{l_i}_{w,c} \tag{2}$$

$$L^{l_i}_{w,c} = -\log\sigma(\mathbf{w}^\top \mathbf{c}) - \sum_{(w,c')\in \text{NEG}(w,c)} \log\sigma(-\mathbf{w}^\top \mathbf{c}') \tag{3}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{R}^d$ denote the vector representations of the target word w and DEP context c respectively. $D^{l_i}_m$ is the set of dependency-based word-context pairs in language l_i, $\text{NEG}(w, c)$ is a set of negative sampled word-context pairs for (w, c), and σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$.

3.2 Dependency-based Cross-lingual Objectives

We proceed to the cross-lingual objective in this subsection. The fundamental principle here is that similar words from different languages should have similar embeddings. To achieve that, the dependency-based cross-lingual objective is designed to pull the two monolingual word embeddings into a common vector space by enforcing the words with similar dependency-based cross-lingual contexts across languages to be projected into similar embeddings. In addition, we propose a regularization term as another cross-lingual objective to minimize the distance between similar dependency-phrases in the phrase-level semantic space.

Dependency-based cross-lingual contexts

For the first cross-lingual objective, we generalize the monolingual distributional hypothesis of word embeddings to the bilingual setting, i.e., the words with similar contexts across languages have similar semantic meanings.

Based on the DEP contexts, it is natural to further define the dependency-based cross-lingual contexts which capture more similar contexts of aligned words in the parallel sentence pairs. Concretely, we formulate the dependency-based objective in the bilingual setting with cross-lingual prediction tasks from l_1 to l_2 and from l_2 to l_1 which is similar to the dependency-based monolingual objectives.

For a parallel sentence pair (s^{l_1}, s^{l_2}) with word alignments and the dependency parse-tree pair as illustrated in Figure 1-**Top**, given an alignment link between a word $w^{l_1}_i$ in language l_1 and a word $w^{l_2}_j$ in language l_2, we define the dependency-based cross-lingual contexts of $w^{l_1}_i$ as the DEP contexts of $w^{l_2}_j$ in s^{l_2} and vice versa. As a comparison to the BoW-based paradigm where the contexts of aligned words may appear at different positions in the parallel sentence pair, the dependency parse-tree pairs enable the contexts of aligned words to be also aligned across languages despite of different positions. For example, as shown in Figure 1-**Bottom**, the BoW cross-lingual contexts of aligned words *we* (*wir*) contain a similar context *wiederholen* (*review*), but the word *will* does not correspond to *Wörter* or *morgen*; meanwhile, the dependency-based cross-lingual contexts of aligned words *we* (*wir*) are the same, which are *wiederholen* (*review*). Furthermore, syntactic dependencies are both more inclusive and more focused than BoW [Levy and Goldberg, 2014], as a dependency-based model can filter out some contexts which are within windows but not directly related to the target words.

Based on the dependency-based cross-lingual contexts and monolingual objectives formulated in Equation (2), we define a joint learning objective with the *cross-lingual training* principle as follows,

$$L = \alpha \sum_{(w,c)\in D_m} L_{w,c} + \beta \sum_{(w,c)\in D_{bi}} L_{w,c} \tag{4}$$

where D_m is a union set of co-occurring dependency-based word-context pairs in two languages: $D_m = D^{l_1}_m \cup D^{l_2}_m$, and D_{bi} is the set of pairs of words and the corresponding dependency-based cross-lingual contexts. $L_{w,c}$ is defined in Equation (3). Bilingual word embeddings can be learned by optimizing the dependency-based joint learning objective above, and the algorithm is called DepBiWE.

Cross-lingual phrase-level regularization

We can further augment the DepBiWE model and enhance the quality of cross-lingual embeddings by making full use of word alignment information in the parallel corpus with a cross-lingual regularization in terms of phrase-level semantic similarities.

In the dependency parse-tree of the parallel sentence pairs, we define the dependency-phrase p as a word pair $(w, w_{dr^{-1}})$, where $w_{dr^{-1}}$ is the head of w and dr^{-1} denotes their inverse dependency relation. The representation of a dependency-phrase $\mathbf{p}$ can be represented as the sum of two word vectors, i.e., $\mathbf{p} = \mathbf{w} + \mathbf{w}_{dr^{-1}}$. By incorporating the phrase-level semantic information, we encourage the representations of similar dependency-phrases to be close, as we can derive the aligned dependency-phrases from the aligned words in the parallel sentence pairs. For example, (*review*, *word*) and (*wiederholen*, *Wörter*) in Figure 1 are aligned as dependency-phrases. The more dependency-phrase pairs are identified in the parallel corpus, the closer the embeddings for the two dependency-phrases will be pushed together. By minimizing the distance between aligned dependency-phrases, the auxiliary cross-lingual regularization term can be written as:

$$L_R = \gamma_R \sum_{(p_i^{l_1}, p_j^{l_2}) \in D_p} ||\mathbf{p}_i^{l_1} - \mathbf{p}_j^{l_2}||^2, \qquad (5)$$

where D_p is a set of aligned dependency-phrase pairs extracted from the parallel corpus. The regularization term is combined with the joint objective in Equation (4) to learn bilingual word embeddings (DepBiWE+R), where γ_R is a tradeoff parameter to control the contribution of the phrase-level regularization term.

3.3 Integration of Semantic Spaces

Dependency parse-trees can be regarded as the supervised information from corpus which is valuable yet expensive to obtain, and only applies to small-scale data. This prohibits the dependency-based bilingual word embedding model from being applied to large-scale corpus. On the other hand, the quality of the parsers affects the performance of dependency-based embedding methods. Fortunately, the BoW-based embeddings learned from large-scale monolingual corpus can be incorporated as unsupervised information without parsers, which can be combined with the supervised dependency-based embeddings via joint learning and make the bilingual word embedding model more robust to parsing error.

Specifically, the dependency-based bilingual embedding matrix $\mathbf{W}_s$ learned with supervised dependency parse-tree information and the BoW-based monolingual embedding matrix $\mathbf{W}_u$ learned from large-scale unsupervised data represent two different semantic vector spaces respectively. To integrate the two different semantic spaces for the better word representations, we propose a joint learning scheme to encourage the model to learn similar representations in both $\mathbf{W}_s$ and $\mathbf{W}_u$. Two corresponding context matrices $\mathbf{C}_s$ and $\mathbf{C}_u$ are learned simultaneously by optimizing the joint objective,

l_1-l_2	#S	#l_1-W	#l_2-W	#l_1-V	#l_2-V
en-de	1.9M	55M	52M	40k	50k
en-fr	2.0M	50M	51M	40k	50k
en-es	1.9M	49M	51M	40k	50k

Table 1: The size of the parallel corpus of three language pairs after preprocessing the data. #S denotes the number of sentence pairs, and #l_i-W represents the number of tokens of the parallel corpus in language l_i, while #l_i-V is the vocabulary size.

$$L_C = (L_{w_u,c_u} + L_{w_s,c_s}) + \gamma_C (L_{w_u,c_s} + L_{w_s,c_u}) \qquad (6)$$

where w_u and w_s denote two different representations of the same target word w, while c_u and c_s correspond to the BoW context and the DEP context of the target word respectively. L_{w_u,c_u} and L_{w_s,c_s} are the loss functions corresponding to BoW-based and dependency-based bilingual embedding learning respectively, while L_{w_u,c_s} and L_{w_s,c_u} are the loss functions integrating the supervised dependency-based embeddings and the BoW-based embeddings learned from large-scale monolingual corpus, which encourage the model to learn similar representations in both $\mathbf{W}_s$ and $\mathbf{W}_u$. γ_C is a tradeoff parameter of the integrated model DepBoW.

4 Experiments

4.1 Data and Setup

We train our dependency-based bilingual models for the English-German (en-de), English-French (en-fr) and English-Spanish (en-es) language pairs on the Europarl v7 parallel corpus[1] [Koehn, 2005]. To preprocess the dataset, we lowercase and tokenize all words and select the top words according to their term frequencies in the training corpus. The words with low frequencies for all languages are mapped to <unk>. The statistics of the parallel corpus for all language pairs are summarized in Table 1.

In our experiments, the Europarl corpus is used for both monolingual training and bilingual training. Parameters for bilingual embedding learning are set as suggested in BiSkip [Luong *et al.*, 2015] and fixed for all experiments. The subsampling rate, negative sampling size are set to $1e$-4 and 30 respectively; the default learning rate of Stochastic Gradient Decent (SGD) is set to 0.025 and gradually decreases to $2.5e$-6 when training is finished. The dimensionality of all embedding vectors d is set to 200, and experiments are run for 10 epochs. We set the monolingual weight α and bilingual weight β in Equation (4) to 1.0 and 4.0 respectively, with the regularization weight γ_R =0.1. Word alignments are obtained with FastAlign [Dyer *et al.*, 2013], and a python library spaCy[2] is employed to produce the dependency parse-trees for all languages in the parallel corpus for the dependency-based models.

We compare our proposed bilingual word embedding models based on syntactic dependencies with baselines including SGNS [Mikolov *et al.*, 2013c] and DepWE [Levy

[1] http://www.statmt.org/europarl/

[2] https://spacy.io/docs/usage/dependency-parse

Models	Monolingual Word Similarity			CLWS	CLDI	
	SemLex	RW	SCWS	semeval2017	Accuracy	MRR
SGNS	31.7	42.2	49.2	-	-	-
DepWE	33.8	41.9	49.5	-	-	-
CLC-WA	22.5	27.2	36.0	35.3	64.9	52.6
CLC+WA	23.5	25.6	35.2	33.8	62.7	51.8
BiSkip (MA)	33.6	48.1	47.6	55.3	79.9	64.0
BiSkip (UA)	32.8	47.5	46.9	53.1	78.8	63.7
DepBiWE (MA)	36.9	46.0	52.8	52.7	79.7	64.8
DepBiWE	37.4	47.1	**53.4**	56.4	81.9	**65.8**
DepBiWE+R	**38.0**	**48.9**	52.3	**60.4**	**82.7**	65.7

Table 2: The results of various models on both the monolingual (monolingual word similarity) and cross-lingual (cross-lingual word similarity and cross-lingual dictionary induction) evaluation tasks on language pair en-de. DepBiWE (MA), DepBiWE and DepBiWE+R are our proposed methods. The best performance for each dataset and evaluation task is in **bold**.

and Goldberg, 2014] for monolingual word embeddings, as well as four cross-lingual word embedding models: cross-lingual matrix co-factorization without (CLC-WA) and with word alignments (CLC+WA) [Shi *et al.*, 2015]; the bilingual skip-gram model exploiting unsupervised word alignments (BiSkip (UA)) or assuming monotonic word alignments (BiSkip (MA)) [Luong *et al.*, 2015]. Notice that both CLC+WA and BiSkip (UA) employ word alignments generated by the FastAlign [Dyer *et al.*, 2013] software which is the same to our models, while DepBiWE (MA) resembles BiSkip (MA) by assuming monotonic word alignments. All algorithms are trained on the Europarl corpus, and we fix the window size to 5 for all the BoW-based methods.

4.2 Evaluation

We evaluate the quality of the induced cross-lingual word embeddings in this section. First, we measure the performance of the learned embeddings monolingually in terms of word similarities in a single language on standard similarity datasets. Next, we evaluate the similarity of nearby pairs of words from two languages in the embedding space on the tasks of cross-lingual word similarity and cross-lingual dictionary induction. We further demonstrate the effectiveness of the proposed models by feeding the learned embeddings to a practical NLP task of cross-lingual document classification.

Monolingual word similarity

We start with evaluating the semantic quality of the learned embeddings in terms of monolingual word similarity [Iacobacci *et al.*, 2015] on the following three datasets: SemLex (999 pairs), RareWord (RW) (2034 pairs), SCWS (1762 pairs). Each dataset contains tuples in the form of (w_1, w_2, s), where s denotes the semantic similarity score between w_1 and w_2 rated by humans. The evaluation is based on Spearman's rank correlation coefficient between semantic similarity scores and cosine similarity scores given by the word representations.

Table 2 shows the performance of the proposed models compared to various cross-lingual embedding baselines in terms of monolingual word similarity. We also compare our methods with monolingual embedding models SGNS [Mikolov *et al.*, 2013c] and DepWE [Levy and Goldberg, 2014] trained on the English corpus from the language pair en-de. From Table 2 one can observe that, on the three monolingual word similarity datasets, the proposed DepBiWE models achieve a better monolingual performance by bilingual training with more corpus from another language. In addition, the DepBiWE models outperform the bilingual baselines, which justifies the quality of the embeddings learned by DepBiWE. By adding a cross-lingual phrase-level regularization, DepBiWE+R effectively improves the quality of monolingual word representations for English. Also notice that the bilingual word embedding baselines with word alignments CLC+WA and BiSkip (UA) achieve little improvement compared with models without word alignments in terms of the monolingual performance, which implies the information of word alignment is not effectively exploited in these models.

Cross-lingual performance

We proceed to evaluate the cross-lingual performance of the dependency-based DepBiWE models with tasks of cross-lingual word similarity (CLWS) and cross-lingual dictionary induction (CLDI).

For the cross-lingual word similarity task, each word pair consists of two words from different languages, which is similar to the monolingual word similarity task, and we employ a cross-lingual word similarity dataset[3] (semeval2017 with 914 pairs) of the en-de language pair proposed by [Camacho-Collados *et al.*, 2015]. The task of cross-lingual dictionary induction [Upadhyay *et al.*, 2016] evaluates the quality of cross-lingual word embeddings by detecting word pairs from two languages that are semantically similar. We use similar settings as in [Upadhyay *et al.*, 2016], and generate the gold dictionary using the Open Multilingual WordNet data released by [Bond and Foster, 2013], which includes synset alignments across 26 languages. We delete words from each synset with frequency less than 1000 in the vocabulary for each language. In this way, a gold dictionary of 1340 word pairs is generated from the aligned synsets for the en-de language pair. Given the entries (w^{l_1}, w^{l_2}) in the gold dictionary, where w^{l_1} and w^{l_2} are the lemmas and have the same meaning, we calculate the accuracy that w^{l_2} appears in the top-10 words which are nearest to w^{l_1} in the semantic space.

[3] http://alt.qcri.org/semeval2017/task2/index.php?id=task-details

Models	Data	en→de	de→en
MT baseline	Europarl	68.1	67.4
Majority class	-	46.8	46.8
Klementiev et al.	Europarl+RCV	77.6	71.1
BiCVM	Europarl	83.7	71.4
BilBOWA	Europarl+RCV	86.5	75.0
BiSkip (256d)	Europarl	88.4	**80.3**
CLC+WA	Europarl+RCV	90.0	75.0
DepBiWE	Europarl	85.8	74.7
DepBiWE+R	Europarl	**90.8**	79.2

Table 3: Accuracy of cross-lingual document classification (in percent) on language pair en-de. Data denotes the corpus used for training bilingual word embedding models. The proposed models DepBiWE and DepBiWE+R are compared to baselines. The best scores are in **bold**.

We also evaluate the quality of cross-lingual embeddings by mean reciprocal rank (MRR).

The results of various algorithms on the CLWS and CLDI evaluation tasks on the language pair en-de are summarized in Table 2, where one can observe that the dependency-based DepBiWE models achieve superior scores compared with prior works including CLC-WA, CLC+WA and BiSkip (with alignments or without alignments). As a comparison, the difference of word orders in parallel sentences degrades the performance of CLC+WA and BiSkip with word alignments, while the DepBiWE models capture more similar contexts satisfying the distributional hypothesis in the bilingual setting.

Compared to DepBiWE (MA), DepBiWE achieves better results, which indicates the information of word alignment is effectively exploited. The cross-lingual performance of the proposed methodology is further improved in the regularized model DepBiWE+R, which justifies the regularization that two similar dependency-phrases in different languages are close to each other in the embedding space.

Cross-lingual document classification

In this section, the quality of the learned cross-lingual embeddings are evaluated via the cross-lingual document classification (CLDC) task on the language pair en-de which tests the semantic transfer of information across languages as introduced in [Klementiev *et al.*, 2012]. In the CLDC task, for the language pair (l_1, l_2), a classifier is trained using labeled documents in language l_1 and then applied to classify documents in language l_2, and vice-versa. A document is represented as an idf weighted sum of the embedding vectors of all its tokens, while idf weights of words are computed using all documents from that language in RCV1/RCV2. The training and test data are sourced from the Reuters RCV1/RCV2 multilingual corpus [Lewis *et al.*, 2004] and are assigned to only one of four topics: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). For the classification experiments, 15,000 documents for each language are selected randomly from the RCV1/RCV2 corpus, in which 5,000 documents are used as the test data and a subset with varying sizes between 100 and 10,000 of the remainder serves as the training data. Meanwhile, we keep 1,000 documents as the development set for hyper-parameter tuning. A multi-class document classifier is trained for 10 epochs with an averaged perceptron algorithm [Klementiev *et al.*, 2012], and we use the accuracy to evaluate the performance of the knowledge transfer.

Table 3 summarizes the results of various algorithms on the CLDC task. Among the baselines, *MT* translates the target documents into the source language using a statistical machine translation (SMT) system. *Majority Class* is a system where test documents are simply classified into the class with the most training samples. We also compare our dependency-based models to the state-of-the-art bilingual word embedding methods including [Klementiev *et al.*, 2012], BiCVM [Hermann and Blunsom, 2014], BilBOWA [Gouws *et al.*, 2015], bilingual skip-gram (BiSkip) [Luong *et al.*, 2015] and a matrix co-factorization framework with word alignments (CLC+WA) [Shi *et al.*, 2015]. As shown in Table 3, the DepBiWE models achieve competitive performance compared with the state-of-the-art. In the meantime, the regularization term further improves the accuracy in DepBiWE+R, which achieves the best performance (90.8%) in en→de while the model is only trained on the Europarl corpus.

4.3 Influence of Word Order

To verify the effectiveness of the dependency-based model on the language pairs with different grammatical structures, we evaluate the BiSkip (UA) and DepBiWE+R models on two additional language pairs (en-fr and en-es) with similar parameter setup as used in Section 4.1. To achieve that, we quantify the difference in word order of three language pairs (en-de, en-fr and en-es) and compare the performance of the BoW-based method and the dependency-based method with various evaluation tasks on them. The results of the monolingual word similarity and CLDI tasks are shown in Table 4. Notice that in the CLDI task, the gold dictionaries for en-fr and en-es are generated similarly to that of en-de, the sizes of which are 616 for en-fr and 1394 for en-es respectively. The difference in word order is quantified by computing the probability of alignment of word's BoW context in the parallel corpus, and is summarized in the first column of Table 4.

From Table 4, it can be shown that with increasing grammatical difference in a language pair, the performance of both methods on various tasks degrades in general, which is not surprising. Further, by comparing the average improvement of our model over the BoW-based method for each evaluation task on three language pairs in Table 4, one can observe that the more difference there exists in a language pair, the greater improvement the proposed model achieves. This justifies the advantage of exploiting different syntactic structures across different languages when learning bilingual word embeddings, which is more evident for language pairs with greater grammatical difference.

4.4 Efficacy of Integrating Semantic Spaces

Here we investigate the results of joint embedding learning by integrating supervised dependency-based embeddings with unsupervised BoW-based embeddings on the language pair en-de. For supervised dependency-based embeddings, we

l_1-l_2 (Difference)	Models	Monolingual Word Similarity			CLDI		AvgImprovement
		SemLex	RW	SCWS	Accuracy	MRR	
en-de (59.8%)	BiSkip (UA)	32.8	47.5	46.9	78.8	63.7	**3.58** ↑
	DepBiWE+R	38.0	48.9	52.3	82.7	65.7	
en-fr (63.2%)	BiSkip (UA)	34.1	48.5	48.5	81.6	68.2	**2.46** ↑
	DepBiWE+R	36.4	51.4	51.3	83.6	70.7	
en-es (67.8%)	BiSkip (UA)	33.1	45.4	48.7	75.9	62.3	**2.06** ↑
	DepBiWE+R	35.9	47.2	50.6	78.5	63.5	

Table 4: The performance of various evaluation tasks on three language pairs. AvgImprovement represents the average improvement of the DepBiWE+R over the BiSkip (UA) across all evaluation tasks.

Models	CLWS	CLDI	en→de	de→en
DepBiWE	56.4	81.9	85.8	74.7
DepBiWE+WV	40.8	63.7	64.5	43.5
DepBoW ($\mathbf{W}_u$)	50.4	79.4	86.2	75.2
DepBoW ($\mathbf{W}_s$)	59.8	**82.5**	87.3	**78.6**
DepBoW ($\mathbf{W}_u + \mathbf{W}_s$)	**60.3**	82.2	**91.4**	77.8

Table 5: Performance on the tasks of CLWS, top-10 accuracy in CLDI and CLDC with different word representations learned in the integration model DepBoW and our fundamental model DepBiWE. The best scores are in **bold**.

use the Europarl corpus with dependency parse-trees, while BoW vectors are trained on the Europarl corpus for unsupervised BoW-based embeddings, utilizing pre-trained word vectors which are learned from the large-scale Wikipedia corpus [Raganato *et al.*, 2016] from two languages independently. In our experiments, we fine-tune the BoW-based embedding vectors with a learning rate of 2.5e-5, and γ_C in Equation (6) is set to 0.3.

Table 5 shows the results of various algorithms on three cross-lingual evaluation tasks. In the comparison, DepBiWE is the fundamental model based only on dependencies; DepBiWE+WV incorporates the dependency-based bilingual embeddings and the pre-trained BoW word vectors, which are learned independently, simply by summing them up; DepBoW ($\mathbf{W}_u$) and DepBoW ($\mathbf{W}_s$) represent the semantic spaces encoded by the matrices $\mathbf{W}_u$ and $\mathbf{W}_s$ respectively, which are learned from the integration model in Equation (6), and DepBow ($\mathbf{W}_u + \mathbf{W}_s$) integrates the two embedding matrices by summing them up. As can be observed, compared with DepBiWE, DepBoW ($\mathbf{W}_s$) and DepBoW ($\mathbf{W}_u + \mathbf{W}_s$) achieve superior results when the dependency-based and BoW-based semantic spaces are integrated, which justifies that the integration of different semantic spaces can effectively exploit monolingual word vectors to improve the quality of bilingual word embeddings.

5 Conclusion

In this paper, we propose a bilingual word embedding framework by exploiting syntactic dependencies (DepBiWE). We consider different syntactic structures across different languages by building dependency parse-trees to capture the syntactic contexts of aligned words in parallel sentences. We further introduce a regularization term based on the phrase-level semantic similarities. In addition, considering that building dependency parse-trees can be expensive on large-scale corpus, we propose a cross learning method to integrate the dependency-based embeddings with BoW-based embeddings learned from large-scale monolingual corpus. Extensive experiments are conducted to validate the superiority of the proposed framework over the state-of-the-art on various natural language processing tasks.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 61673364 and No. 61672487), the Fundamental Research Funds for the Central Universities (WK2150110008), and Jiangsu Natural Science Foundation (No.BK20151239). We thank Ivan Titov and Alexandre Klementiev for providing their evaluation package. We also appreciate the valuable comments from the anonymous reviewers.

References

[AP *et al.*, 2014] Sarath Chandar AP, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853–1861, 2014.

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb):1137–1155, 2003.

[Bond and Foster, 2013] Francis Bond and Ryan Foster. Linking and extending an open multilingual wordnet. In *ACL (1)*, pages 1352–1362, 2013.

[Camacho-Collados *et al.*, 2015] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *ACL (2)*, pages 1–7, 2015.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

[De Marneffe and Manning, 2008] Marie-Catherine De Marneffe and Christopher D Manning. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

[Duong *et al.*, 2016] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Learning crosslingual word embeddings without bilingual corpora. *arXiv preprint arXiv:1606.09403*, 2016.

[Dyer *et al.*, 2013] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics, 2013.

[Faruqui and Dyer, 2014] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. Association for Computational Linguistics, 2014.

[Gouws and Søgaard, 2015] Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *HLT-NAACL*, pages 1386–1390, 2015.

[Gouws *et al.*, 2015] Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 748–756, 2015.

[Hermann and Blunsom, 2014] Karl Moritz Hermann and Phil Blunsom. Multilingual models for compositional distributed semantics. *arXiv preprint arXiv:1404.4641*, 2014.

[Iacobacci *et al.*, 2015] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *The Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 86–91, 2015.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Klementiev *et al.*, 2012] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. Inducing crosslingual distributed representations of words. 2012.

[Koehn, 2005] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

[Levy and Goldberg, 2014] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL (2)*, pages 302–308, 2014.

[Levy *et al.*, 2015] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[Lewis *et al.*, 2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5(Apr):361–397, 2004.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D Manning. Bilingual word representations with monolingual quality in mind. In *HLT-NAACL*, pages 151–159, 2015.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[Mikolov *et al.*, 2013c] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Raganato *et al.*, 2016] Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. Automatic construction and evaluation of a large semantically enriched wikipedia. In *International Joint Conference on Artificial Intelligence*, pages 2894–2900, 2016.

[Shi *et al.*, 2015] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL (2)*, pages 567–572, 2015.

[Socher *et al.*, 2013] Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. Parsing with compositional vector grammars. In *ACL (1)*, pages 455–465, 2013.

[Upadhyay *et al.*, 2016] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Roth Dan. Cross-lingual models of word embeddings: An empirical comparison. In *Meeting of the Association for Computational Linguistics*, pages 1661–1670, 2016.

[Yin *et al.*, 2016] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. Unsupervised word and dependency path embeddings for aspect term extraction. In *International Joint Conference on Artificial Intelligence*, pages 2979–2985, 2016.

[Zhou *et al.*, 2015] Guangyou Zhou, Tingting He, Jun Zhao, and Wensheng Wu. A subspace learning framework for cross-lingual sentiment classification with partial parallel data. In *International Joint Conference on Artificial Intelligence*, pages 1426–1433, 2015.

[Zou *et al.*, 2013] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1398, 2013.