Machine Translation (2021) 35:551–569
https://doi.org/10.1007/s10590-021-09283-z

Jointly learning bilingual word embeddings and alignments

Zhenqiao Song[1] · **Xiaoqing Zheng**[1] · **Xuanjing Huang**[1]

Received: 7 January 2021 / Accepted: 18 October 2021 / Published online: 1 November 2021

Abstract

Learning bilingual word embeddings can be much easier if the parallel corpora are available with their words well aligned explicitly. However, in most cases, the parallel corpora only provide a set of pairs that are semantically equivalent to each other at sentence level. While algorithms have been proposed to obtain word alignments, good alignments are still hard to achieve. In this study, we propose Bilingual word embeddings with soft alignment (BWESA) to learn bilingual word representations from the parallel corpora without explicit word-level alignment information. At the same time, this method learns to make 'soft' alignments between words by approximating a distribution for each word in a sentence to estimate how likely the word is aligned to the words in the parallel translation. Unlike previous methods that typically make use of a predetermined word alignment, our learning strategy makes similar words—properly chosen by the continuously improving word alignment—become closer in the shared vector space during the training process. This study is among the first to learn bilingual word alignments and embeddings in a joint manner. The proposed method was evaluated on two cross-lingual tasks (cross-lingual document classification and word translation) and achieved state-of-the-art or comparable results on all the tasks considered.

Keywords Bilingual word embeddings · Word alignment · Weakly-supervised learning · Cross-lingual document classification · Word translation

✉ Xiaoqing Zheng
zhengxq@fudan.edu.cn

Zhenqiao Song
zqsong17@fudan.edu.cn

Xuanjing Huang
xjhuang@fudan.edu.cn

[1] School of Computer Science, Fudan University, 825 Zhangheng Road, Shanghai, China

1 Introduction

Recently, learning word representations (also known as word embeddings) of natural languages has attracted much attention (Mikolov et al. 2013a; Pennington et al. 2014; Zheng et al. 2017; Feng and Zheng 2018). These distributed representations have proven useful for many natural language processing tasks, such as language modelling (Bengio et al. 2003), sentiment analysis (Socher et al. 2013b) and syntactic parsing (Socher et al. 2013a). It is also possible to learn such word vector representations across different languages (Klementiev et al. 2012; Mikolov et al. 2013b; Hermann and Blunsom 2014; Gouws et al. 2015; Wei and Deng 2017; Søgaard et al. 2018), where similar words from multiple languages are clustered in the shared vector space. Multilingual word embeddings have been considered as an important building block for many cross-lingual tasks, including machine translation (Zou et al. 2013), parsing (Guo et al. 2015), and information retrieval (Vulić and Moens 2015).

Word alignment is often considered as a necessary pre-processing step for learning bilingual word embeddings, in which the words from two languages are first aligned automatically or semi-automatically and then the bilingual word embeddings are learned from the dataset with their words aligned explicitly. However, automatic word alignment is still a challenging task and results are unreliable for the subsequent learning of bilingual word embeddings. Some researchers have chosen to use the results of word alignment produced by an external tool like GIZA++ (Och and Ney 2003) from parallel data, but such word alignments are usually not good enough, and the alignment errors will propagate to the following word-embedding learning process. Others drop the step of automatic word alignment for this reason, but it is then impossible to fully leverage the implicit word-level information contained in a parallel corpus. In this study, we try to make better use of parallel data at both (explicit) sentence level and (implicit) word level, but the word alignment is not considered as a separate and predetermined step.

We propose a novel method to learn bilingual word alignments and word embeddings jointly, in which both tasks are reinforced mutually and gradually and can benefit from each other. The learned word alignment can be viewed as a distribution learned for each word in a sentence from a (source) language over all the words of its aligned translational equivalent from another (target) language. Under- or over-alignment problems might occur if no constraints are imposed because some words may not be aligned at all or aligned to too many words. Therefore, two criteria are proposed and enforced during the word alignment to deal with these problems: *coverage* and *sparsity*. That is, each word in a sentence should have at least one semantic equivalent in the parallel translation, but the number of such corresponding words is limited (note that a word may be aligned to a phrase in other language). We carried out two sets of experiments to evaluate our BWESA method. The first is to evaluate the quality of the learned bilingual word embeddings on two tasks: cross-lingual document classification (CLDC) and word translation. The second one is to assess the results of the word alignment using alignment error rate (AER). Our proposed BWESA approach achieved

state-of-the-art or comparable results on all these tasks. The effect of word alignment information was also confirmed by the experiments.

The main contributions of this paper are: (i) we propose a novel method to learn bilingual word embeddings and alignments from parallel corpus in a joint fashion; (ii) we recommend applying the two criteria of "coverage" and "sparsity" during word alignment to deal with under- and over-alignment problems; and (iii) our model achieved state-of-the-art or comparable results on cross-lingual document classification and word translation tasks.

2 Related work

Bilingual word embedding learning methods aim to embed the words from different languages into a shared continuous vector space, where the learned word embeddings yield a useful characteristic that similar words from multiple languages are close to each other in the space. Those methods can be roughly divided into three categories with respect to their training objectives: mapping offline, mapping online, and joint training.

2.1 Mapping offline

Mapping Offline methods first learn to obtain two monolingual word embeddings separately, and then to compute the mapping between the two different vector spaces by using extra resources (such as a dictionary). Learning a mapping matrix is arguably the most common way to obtain bilingual word embeddings by constructing a dictionary from Google Translate (Mikolov et al. 2013b), leveraging a seed dictionary (Artetxe et al. 2017) or employing the singular value decomposition (Smith et al. 2017). Although word embeddings can be learned at less computational cost by these methods, they might be incapable of capturing the phenomena of homonymy and polysemy that widely exist within and across languages because these methods usually consider only one translation per word. Although not requiring a parallel corpus is an advantage, offline mapping methods rely on the assumption that underlying embeddings should have a similar structure, which is known as the *isometry* assumption. However, this assumption can not be taken for granted and some researchers have shown that this assumption does not hold generally (Søgaard et al. 2018; Nakashole and Flauger 2018), and can severely degrade the performance of these methods.

2.2 Mapping online

Mapping Online methods try to learn sentence-level representations (often derived from their word embeddings) for different languages by making the learned representations of each pair of parallel sentences stay close to each other in a shared vector space. Those word embeddings are learned in an indirect way, and the word-level alignment is often not forced directly. Hermann and Blunsom (2014) proposed to

learn bilingual word embeddings by aligning the representations of parallel sentences while keeping sufficient distances between those of dissimilar ones. Chandar et al. (2014) used an autoencoder-based framework to produce the representation of a sentence, which can both reconstruct the bag-of-words for that sentence and those for the aligned translation. Kočiskỳ et al. (2014) proposed to learn both bilingual word embeddings and alignments based on FASTALIGN (Dyer et al. 2013), a variation of IBM model 2 (Brown et al. 1993), but they do not leverage the same or similar semantics conveyed in the parallel sentences directly when learning word alignment. Wei and Deng (2017) presented a variational autoencoder-based method, where a continuous latent variable is used to model the underlying semantics of each pair of parallel sentences and guide the reconstruction of these sentence pairs. Bilingual word embeddings are obtained indirectly in those methods by making the sentence pair well-aligned, and these methods might fail to fully capture the intrinsic semantics and syntactic characteristics at the level of their words.

2.3 Joint training

Joint Training methods learn bilingual word representations by taking both monolingual and bilingual objectives into account. The word embeddings for each language are first separately trained from the monolingual corpus, and the obtained embeddings are then further tuned to satisfy the bilingual constraints defined either from pre-computed word alignments (Zou et al. 2013), or via coarse alignments under a uniform distribution assumption (Gouws et al. 2015). Luong et al. (2015) proposed a variant of skip-gram to learn BWEs by improving on the prediction of contextual words from both the monolingual and cross-lingual sentences. These methods make it possible to leverage both relatively small but valuable amounts of parallel data as well as large unlabelled monolingual texts. However, the performance of the learned BWEs is strongly sensitive to the quality of the predetermined word alignments, and good word alignments have generally been hard to achieve up to now.

In this study, we follow the line of the joint training strategy, but our model is different from others in that it is capable of learning bilingual word embeddings and word alignments jointly. We show that these two tasks can benefit each other in such a joint learning manner. Word alignments do not need to be predetermined before the training starts and are given opportunities to be improved gradually as the learning progresses, which leads to better bilingual word embeddings.

3 Models

We here describe our BWESA (Bilingual Word Embeddingswith Soft Alignment) method that can learn bilingual word embeddings and alignments automatically and simultaneously. Our objective function can be factorized into three parts. The first is designed for monolingual purposes, the second is for bilingual use-cases, and the last one is for word alignments, respectively denoted as $loss_{mono}$, $loss_{bi}$, and $loss_{align}$. The loss function can be formalized as in (1):

$$L_{bwe} = \alpha \cdot loss_{mono} + \beta \cdot loss_{bi} + \gamma \cdot loss_{align} \quad (1)$$

where α, β, and γ govern the relative importance of the three different parts. The first term can be further decomposed into two components, each for one language. Considering that the under-alignment or over-alignment problems will be harmful to the results of word alignment, we advocate to fulfil the two criteria of "coverage" and "sparsity" on word alignments during the learning process. Specifically, each word of a sentence should be aligned to at least one equivalent in the parallel sentence (i.e. fulfilling coverage) and at the same time, the cardinality of those semantic equivalences should be limited to a small number (i.e. fulfilling sparsity).

3.1 Monolingual objective

We chose to apply the skip-gram with negative sampling strategy (Mikolov et al. 2013a) to train the word embeddings from monolingual data since it has been widely used and can be performed at low computational cost. The philosophy behind skip-gram is that a word tends to have similar meanings to its neighbouring ones, and thus its feature representation can be trained by using the current word to predict its context (or neighbouring) words.

Specifically, for each word w in a vocabulary, $Con(w)$ consists of all contexts in which the word w occurs in a corpus. The loss function for the word w can be formalized as in (2):

$$\begin{aligned} loss(w) = & -\sum_{c \in Con(w)} [\log \sigma(r_w \cdot r_c) \\ & -\sum_{n \in Neg(w)} \log \sigma(r_w \cdot r_n)] \end{aligned} \quad (2)$$

where r_w is the distributed vector representation of w, and σ is denoted as the sigmoid function. The negative sampling method has been widely used to learn word embeddings (Mikolov et al. 2013a; Zheng et al. 2017; Feng and Zheng 2018), where the word embeddings are trained by maximizing the conditional likelihood of the current words given their contexts by the gradient ascent algorithm, which can be factorized with respect to the current word (positive) and its negative samples using logistic regression as in Eq. (2). For each positive word, a set of k negative words, denoted as $Neg(w)$, is randomly sampled from the vocabulary according to their frequencies. We need to train monolingual word embeddings twice, each for a language, and thus the loss for monolingual purpose, denoted as $loss_{mono}$, is defined as the sum of two parts, as in (3):

$$loss_{mono} = \sum_{w^e \in V^e} loss(w^e) + \sum_{w^f \in V^f} loss(w^f), \quad (3)$$

where w^e denotes a word in the vocabulary V^e, extracted from the corpus of a source language, and w^f a word in the vocabulary V^f from a target language. The "source" and "target" are just used to name different languages, and can be used interchangeably without affecting the results.

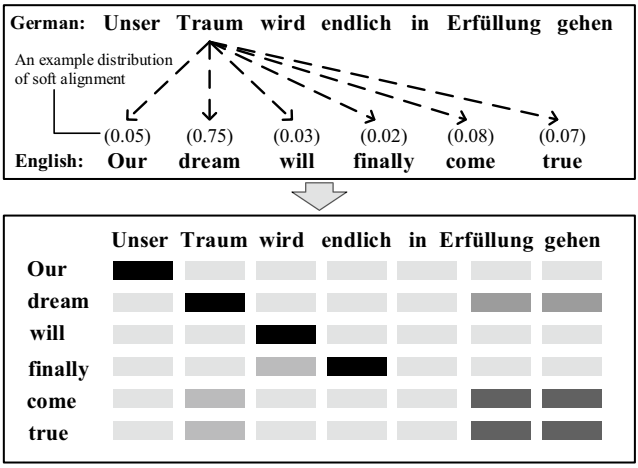

Fig. 1 The upper half shows an example distribution of soft alignment for the word "Traum" in a German sentence over all the words in the parallel English sentence, where most of the weights are given to the word "dream" which carries a similar meaning as the German word "Traum." The lower half illustrates a similarity matrix for a pair of sentences, in which the colour of each element shows the degree of similarity between the two corresponding words. The darker the colour is, the more similar they are semantically

3.2 Bilingual objective

In a parallel corpus, two sentences in a pair convey the same meaning, and each word (as a smaller semantic unit) in a sentence should have its own correspondence in the parallel sentence. We assume that there exists a soft alignment distribution for each word in a source sentence over the words from the target equivalent. Figure 1 illustrates a simple example of the soft alignment distribution for the German word "Traum", where most of the weight of the distribution is given to the English word "dream" which carries a similar meaning to the German word "Traum."

We take the result of the ReLU nonlinear transformation over the dot product of two words' embeddings as their similarity score, and such scores are further normalized to approximate the aligned distribution. The similarity score, a_{ij}, can be computed and then normalized to $\hat{a}_{ij}$, as in (4):

$$a_{ij} = ReLU(r_{w_i^e} \cdot r_{w_j^f}), \quad \hat{a}_{ij} = \frac{a_{ij}}{\sum_k a_{ik}} \tag{4}$$

where w_i^e denotes the i-th word in a sentence from the source language e, and w_j^f denotes the j-th word in the parallel one from the target language f. The similarity score defined by the dot product gradually forces the two word vectors closer to one another during the training process, and finally causes all the words from the two different languages to be embedded in the same vector space. The ReLU function is used to produce the scores so that two dissimilar words can have a zero score, and

more importantly, the sparsity property can be guaranteed to some extent since many word pairs will receive a scored of zero or close to zero.

$$dist(w_i^e) = ||r_{w_i^e} - \sum_j \hat{a}_{ij} r_{w_j^f}||^2 \tag{5}$$

We define a distance we want to reduce in the vector space in Eq. (5) to reflect how well the meaning of a source word is represented in its parallel equivalent. Specifically, such distance is formalized as the Euclidean metric between the embedding of the i-th word in a sentence and the weighted average of those of all the words in the parallel sentence. The estimated distribution $\hat{a}_{ij}$ for the i-word from the source language is used as the weights.

Likewise, $dist(w_j^f)$ can also be calculated for the j-th target word in the same way. Therefore, the loss function for the bilingual purpose can be defined as in (6):

$$loss_{bi} = \frac{1}{N} \sum_{(S_e,S_f)\in D} \left(\sum_{i=1}^{|S_e|} dist\left(w_i^e\right) + \sum_{j=1}^{|S_f|} dist\left(w_j^f\right) \right) \tag{6}$$

where $\mathcal{D} = \{(s_e, s_f)_n\}_{n=1}^N$ is a dataset consisting of parallel sentence pairs (s_e, s_f), and N is the number of those pairs.

In our model, the 'soft' word alignment is derived from the similarity scores estimated between the word embeddings by taking the semantic equivalences at their sentence level as a guide. The derived alignments are used to learn the bilingual word embeddings, and in turn, the learned embeddings are further used to improve the quality of the word alignments. In this way, the word alignments and embeddings can be learned jointly and reinforced mutually.

3.3 Coverage and sparsity

We introduce the two criteria of "coverage" and "sparsity" for the word alignment process, and the alignment loss $loss_{align}$ has two parts designed to fulfill these two criteria: $loss_{cov}$ (for coverage) and $loss_{spa}$ (for sparsity). The coverage criterion means that each word of a sentence should be aligned to at least one equivalent in the parallel sentence, and this criterion is proposed to treat the under-alignment problem. The loss to enforce the coverage criterion for a source language is defined as in (7):

$$loss_{cov_e} = \frac{1}{N} \sum_{(s_e,s_f)\in\mathcal{D}} \sum_{j=1}^{|s_f|} \left(1 - \sum_{i=1}^{|s_e|} \hat{a}_{ij}\right)^2 \tag{7}$$

For a target language, the $loss_{cov_f}$ can be defined in a similar way, and we take the sum of the two losses as the training objective for the coverage criterion. To meet the "sparsity" criterion, the cardinality of semantic equivalences of each word in any sentence should be limited to a reasonably small number. As discussed above, this criterion is implicitly guaranteed via ReLU nonlinear transformation defined in Eq. (4). The loss function for the word alignment can be written as in (8):

$$loss_{align} = loss_{cov_e} + loss_{cov_f} \tag{8}$$

In summary, to jointly obtain better vector representations of words for the source and target languages, the word embeddings are first trained with the monolingual loss of Eq. (3) and then trained by using the bilingual loss of Eq. (6) as well as the word alignment loss of Eq. (8). A similar negative sampling strategy like skip-gram (Mikolov et al. 2013a) was used to fulfill the monolingual training objective, and a good (soft) word alignment distribution is learned by leveraging a sentence-level parallel corpus to meet the proposed two criteria of "coverage" and "sparsity" for the word-level alignment.

4 Experiments

We conducted three sets of experiments to evaluate our BWESA method by comparing it to other representative methods: word translation and cross-lingual document classification for the learned bilingual word embeddings, and AER for the obtained word alignments.

4.1 Training details

4.1.1 Training datasets

English-German (en-de) and English-French (en-fr) branches of the Europarl v7 corpus (Koehn 2005) were used to train the models for comparison, which contain 1.9M en-de parallel sentences with 49.7M English and 52.0M German words, and 2.0M en-fr parallel sentences with 55.7M English and 61.9M French words. The models were also evaluated on an English-Chinese (en-zh) dataset. Those two languages belong to different language families, and they are much more different from each other than en-de or en-fr pairs. The dataset for en-zh was extracted from LDC,[1] which consists of 2.5M parallel sentences with 80.8M English and 72.0M Chinese words.

4.1.2 Hyperparameters and initialization

We tuned the hyperparameters by trying only a few different settings on the validation set. In our experiments, we set the number of negative samples to 64, window size to 5, subsampling rate to 0.0001, and initialized learning rate to 0.1. The dimensionality of word embeddings was set to 40 for both en-de and en-fr pairs, and 100 for en-zh. We first let $\alpha + \beta = 1$, and tuned γ within $\{0.5, 1.0, 2.0, 4.0, 8.0\}$. In this way, we can observe the model's behaviour without the constraint on word alignment, and see how much we can improve performance by introducing this constraint

[1] https://www.ldc.upenn.edu/.

later in an incremental manner. The experimental results show that the performance is relatively insensitive to the values of γ, but we chose to set $\gamma = 0.5$ as this yielded a slightly better performance than other values on the validation set.

Luong et al. (2015) and Gouws et al. (2015) found that for both en-de and en-fr, setting the dimensionality of word embeddings to 40 sufficed as these two languages are quite similar to each other. For a fair comparison, we followed their setting in our experiments. However, English and Chinese (en-zh) belong to different language families, and they are much more different from each other than en-de or en-fr pairs. Accordingly, we chose to enlarge the model capacity by increasing the size of word embeddings to 100. Our preliminary experiments showed that the size of word embedding generally has a limited impact on the performance if it is large enough. We tuned the values of k within {8, 16, 32, 64, 128} and set the number of negative samples to 64, which yields the best performance on the validation set.

The problem of learning bilingual word embeddings is that it has a very large search space, which makes it extremely difficult to learn good word embeddings starting with a random initialization. Accordingly, we "warm-up" the model by using information from cross-lingual word co-occurrence statistics to speed up the training process. In the first several iterations, the similarity scores between words are calculated based on the word co-occurrence, as in (9):

$$s_{ij} = \frac{count(w_i^e, w_j^f)}{count(w_i^e)}, \tag{9}$$

where $count(w_i^e, w_j^f)$ is the frequency of co-occurrence of w_i^e and w_j^f word pairs in the training corpus, and $count(w_i^e)$ is the total occurrence of the i-th word for a source language. The scores of stop words or other high-frequency words need to be scaled properly, and the Inverse Document Frequency (IDF) that reflects how important a word is to a sentence in a corpus is combined to calculate such scores, as in (10):

$$s'_{ij} = s_{ij} * idf(w_j^f), \tag{10}$$

where $idf(w_j^f)$ is the calculated IDF of word w_j^f. The score s'_{ij} will be normalized to obtain $\hat{s}_{ij}$ in the same way defined as Eq. (4). In the first few iterations, $\hat{s}_{ij}$ was used instead of $\hat{a}_{ij}$ defined in Eqs. (4) and (5).

4.2 Bilingual word embedding evaluation

To evaluate the learned bilingual word embeddings experimentally, BWESA was compared to the following representative models:

(1) DistribReps (Klementiev et al. 2012): They formulate the word embedding learning for a pair of languages as a multitask learning problem where each task corresponds to a single word, and task relatedness is derived from co-occurrence statistics in bilingual parallel data.

(2) BICVM (Hermann and Blunsom 2014): they leverage parallel data and learn to align the embeddings of semantically equivalent sentences, while maintaining sufficient distance between those of dissimilar sentences. The idea behind their method is that, given enough parallel data, a shared representation of two parallel sentences would be forced to capture the common elements and words between these two sentences.
(3) BAE (Chandar et al. 2014): they use autoencoder-based methods for cross-language learning of vector word representations that are coherent between two languages by learning to reconstruct the bag-of-words representations of aligned sentences, within and between languages.
(4) BilBOWA (Gouws et al. 2015): they train bilingual word embeddings on monolingual data and extract a bilingual signal from a set of sentence-aligned data with a sampled bag-of-words cross-lingual objective, which is used to regularize two noise-contrastive language models for cross-lingual feature learning.
(5) BiSkip (Luong et al. 2015): they extended the skip-gram model to learn bilingual representations by using the co-occurrence context information within a language and meaning-equivalent signals across languages.
(6) CLSim (Shi et al. 2015): they proposed a matrix co-factorization framework for learning cross-lingual word embeddings, in which the monolingual training objective is defined in the form of matrix decomposition, and cross-lingual constraints are forced by information derived from parallel corpora.
(7) BRAVE-S (Mogadala and Rettinger 2016): they proposed a model to learn bilingual word embeddings of words from sentence-aligned parallel corpora with the elastic net regularization proposed by Zou and Hastie (2005).
(8) BiVAE (Wei and Deng 2017): they presented a variational autoencoding approach for training bilingual word embeddings where a continuous latent variable is introduced to explicitly model the underlying semantics of the parallel sentence pairs and to guide the generation of the sentence pairs.
(9) Adv-Refine-CSLS (Conneau et al. 2017): they explored building a bilingual dictionary between two languages without using any parallel corpora by aligning monolingual word-embedding spaces in an unsupervised way with adversarial training and a refinement procedure.
(10) DP (Li et al. 2019): they proposed a method to induce a word alignment by estimating the relevance between a pair of words (x, y) from a source language and a target one. The relevance score is estimated by removing a word x from a source sentence, and calculating the difference in the probabilities of generating a word y in the target (translated) sentence before and after the word x being removed with the help of a machine translation model.
(11) E-SGNS (Ormazabal et al. 2020): The core idea of their method is to fix target language embeddings and learn from scratch a set of embeddings for a source language that is aligned with the target one. They use an extension of skip-gram (Mikolov et al. 2013a) that leverages translated context words as anchor points and apply self-learning and iterative restarts to reduce the dependency on the initial dictionary. They proposed three methods to build an initial dictionary, and we chose to compare the version with unsupervised mapping initialization because this version achieved the best result in word alignment on average.

Table 1 The accuracy (%) for word translation task on the open multi-lingual WordNet datasest. P@1, P@5 and P@10 denote top-1, top-5 and top-10 accuracy, respectively

Models	en-de			en-fr		
	P@1	P@5	P@10	P@1	P@5	P@10
DistribReps	46.1	55.4	61.5	53.2	62.0	65.8
BICVM	47.1	60.3	69.0	52.3	64.1	67.8
BAE	66.4	78.6	81.9	54.5	63.0	69.0
BilBOWA	65.0	76.3	80.6	64.1	65.4	77.5
BiSkip	67.6	77.3	80.9	64.3	75.4	76.6
CLSim	66.6	77.0	78.0	64.5	76.1	78.2
BRAVE-S	53.9	75.0	77.3	64.1	75.9	80.2
BiVAE	68.1	77.9	81.0	62.1	75.2	80.3
Adv-Refine-CSLS	62.3	75.7	81.3	61.2	74.9	79.8
DP	66.9	77.4	81.6	61.4	75.2	78.5
E-SGNS	68.5	78.6	82.3	65.0	76.4	80.4
BWECLCO	58.1	64.5	71.2	55.2	65.8	76.5
BWESA	70.3	79.0	83.0	65.5	76.8	81.7

(12) We also developed a new strong baseline, denoted as BWECLCO, which only uses the cross-lingual word co-occurrence to estimate the alignment distribution without the following word alignment learning step as BWESA.

4.2.1 Word translation

Word translation aims to select the most similar word from a target language for a given word from a source language (Mikolov et al. 2013b; Gouws et al. 2015). This task is often used to evaluate how well the similar words from different languages are aligned with each other in the learned vector space with the cosine distance. Following Upadhyay et al. (2016), the gold word pairs were extracted from the Open Multilingual WordNet (OMW) datasest released by Bond and Foster (2013), consisting of 19,675 en-de, 20,449 en-fr and 42,300 en-zh word pairs.

We report the Top-1, top-5 and top-10 accuracy (denoted by P@1, P@5 and P@10) achieved by different models in Tables 1 and 2. As we can see, BWESA produced state-of-the-art results on the OMW dataset for all three language pairs. Although E-SGNS achieved the best averaged top-1 accuracy of 57.4% among the previous models, it was surpassed by BWESA by a fairly significant margin (about 1% on average). The results reported in Table 2 also show that BiVAE outperformed the other competitors for the word translation task, even when the difference between the two languages is large. In addition, we note that the "fully fledged" BWESA model is superior to BWECLCO, with an average increment of 9.63% when word alignment learning is turned off, indicating that the joint solution for learning bilingual word embedding and alignment is preferable and both tasks can mutually benefit and reinforce each other during joint learning. The experimental results show that BWESA is capable of learning finer-grained (word-level) semantic equivalences from (sentence-level) parallel corpora, due to the fact that the alignment learning

Table 2 The accuracy (%) of word translation on open multi-lingual WordNet datasest and the accuracy (%) of cross-lingual document classification (CLDC) on Reuters RCV1/RCV2 multilingual corpora for English-Chinese pair

Models	Word translation			CLDC	
	P@1	P@5	P@10	en-zh	zh-en
BICVM	36.4	38.5	49.0	83.1	66.6
BilBOWA	38.0	43.9	55.3	76.6	72.8
BiSkip	36.1	49.0	56.7	85.5	75.7
CLSim	32.8	47.2	53.6	71.1	72.6
BRAVE-S	37.8	44.7	51.3	78.1	73.3
BiVAE	37.9	48.3	55.7	78.3	72.5
DP	38.2	48.7	56.1	79.8	74.1
E-SGNS	38.7	49.8	57.3	84.7	76.5
BWECLCO	32.7	42.6	52.3	74.8	62.1
BWESA	39.1	51.3	58.4	86.8	78.4

P@1, P@5 and P@10 denote top-1, top-5 and top-10 accuracy, respectively

strategy causes similar words, properly chosen by the continuously improving word alignment, to become closer in the shared vector space as training progresses.

4.2.2 Cross-lingual document classification

Cross-lingual document classification (CLDC) can be used to assess the quality of the learned BWEs by training a classifier on one language and testing it on another. Following the settings of Klementiev et al. (2012), English, German, French and Chinese subsections of Reuters RCV1/RCV2 multilingual corpora were used for evaluation, and the documents labeled with one of CCAT, ECAT, GCAT, or MCAT topics are considered for this task. In this experiment, 15,000 documents were extracted from RCV1/2, in which 5000 documents were randomly selected as the test set, and the rest was taken as the training set.

Following Klementiev et al. (2012), three additional baseline systems (Majority Class, Glossed, and MT systems) are also listed in Table 3 for comparison. The Majority Class simply labels all the documents to be classified with the category having the most samples in the training set. The glossed system works as follows: a classifier is first trained over the documents from a source language; for a document written in another language, every word in the document is replaced with its most frequently aligned word from the source language; finally, the document with its words replaced is labeled by the classifier trained in the first step. The MT system is different from the Glossed system in that the documents to be classified are translated into the source language not by using word-level replacement, but by applying a phrase-based statistical MT tool.

The results reported in Tables 2 and 3 show that BWESA achieved consistently higher performance over the competitors on almost all the CLDC datasets considered. Although CLSim Shi et al. (2015) achieved the best result on the en-de subtask, BWESA outperformed CLSim on the other six language pairs by a significant

Table 3 The accuracy (%) of cross-lingual document classification (CLDC) task on Reuters RCV1/RCV2 multilingual corpora

Models	Accuracy			
	en-de	de-en	en-fr	fr-en
Majority class	46.8	46.8	22.5	25.0
Glossed system	65.1	68.6	74.2	70.2
MT system	68.1	67.4	76.3	71.1
DistribReps	77.6	71.1	74.5	61.9
BICVM	86.4	74.7	83.3	63.0
BAE	91.8	74.2	84.6	74.2
BilBOWA	86.5	75.0	88.5	79.8
BiSkip	90.7	80.3	90.2	77.7
CLSim	**92.7**	80.1	86.7	79.9
BRAVE-S	89.7	80.1	82.5	79.5
BiVAE	91.0	80.4	87.6	79.8
DP	90.5	77.9	88.6	80.3
E-SGNS	91.5	80.6	90.8	81.4
BWECLCO	85.3	78.0	86.7	76.7
BWESA	91.7	**80.9**	**91.6**	**83.3**

The bold fonts are used to highlight the best results

margin (4.93% on average), highlighting the potential of BWESA for practical CLDC, an important downstream task for BWEs. Another noteworthy result of these experiments is the success of the joint learning strategy, which boosts classification accuracy by about 8.18% on average.

4.3 Word alignment evaluation

Both the word translation and cross-lingual document classification tasks were used to evaluate the bilingual word embeddings learned by our model and by other approaches. In this experimental setting, we would like to see how well the words from different languages are aligned, and whether the word alignment has indeed been improved by BWESA.

4.3.1 Alignment error rate

AER is often used to measure how well words are aligned by comparing the model's proposed alignments with the gold ones annotated by humans. We chose to use the inverse of alignment error rate (i.e. 1−AER) suggested by Koehn (2009) as the evaluation metric. The higher the inverse rate, the better the word alignment will be. Like Levy et al. (2017), we first leveraged the Edinburgh Bible Corpus and a subset of the Europarl corpus (180K sentences) to train cross-lingual word embeddings,

Table 4 The results of word alignment reported in the inverse of alignment error rate (1− AER)

Dataset	Language	Models						
		IBM-model1	IBM-model3	Dice	BilBOWA	BiSkip	BWE-CLCO	BWESA
GRACA	en-fr	0.4192	0.5593	0.4357	0.4404	0.4333	0.4720	**0.5632**
	fr-en	0.4381	0.5743	0.4275	0.4643	0.4642	0.4484	**0.5852**
	en-es	0.4182	0.5668	0.4920	0.4803	0.4968	0.5178	**0.5894**
	es-en	0.4572	0.5918	0.4683	0.5112	0.4376	0.5081	**0.6171**
	en-pt	0.3795	0.4742	0.1763	0.4755	0.4752	0.4827	**0.5149**
	pt-en	0.4060	0.5172	0.1776	0.4741	0.4307	0.4787	**0.5733**
HAN-SARDS	en-fr	0.4836	**0.6493**	0.5285	0.3628	0.3520	0.3687	0.4774
	fr-en	0.5190	**0.6773**	0.5214	0.3673	0.3701	0.3744	0.4802
LAMVERT	en-es	0.3802	**0.5273**	0.3585	0.3422	0.3386	0.3439	0.4563
	es-en	0.3561	**0.5748**	0.3544	0.3234	0.2902	0.3203	0.4301
MIHAL-CEA	en-ro	0.0969	0.1375	0.0994	0.0978	0.0785	0.1105	**0.1399**
	ro-en	0.0989	0.1420	0.0982	0.0910	0.0863	0.1012	**0.1585**
HOL-MQVIST	en-sv	0.3341	0.5154	0.3579	0.3661	0.3939	0.2484	**0.5287**
	sv-en	0.3223	0.4600	0.3571	0.3721	0.3625	0.2618	**0.4764**
CAKMAK	en-tr	0.2998	**0.3127**	0.3027	0.1392	0.1495	0.1712	0.2751
	tr-en	0.2998	**0.3127**	0.3081	0.1568	0.1140	0.1788	0.2829

The higher the inverse rate, the better the word alignment will be

The bold fonts are used to highlight the best results

and then sixteen manually annotated word alignment datasets were used to evaluate the word alignments produced by different models.

We evaluate the proposed BWESA and BWECLCO for the word alignment in the inverse of AER by comparing to BilBOWA (Gouws et al. 2015) and BiSkip (Luong et al. 2015) that have been tested on the word-alignment task. We also listed IBM-model1 and IBM-model3 (Brown et al. 1993) as two strong baselines for comparison because they were particularly designed for word alignment by using the bilingual word co-occurrence statistics. In contrast, the Dice system (Och and Ney 2003) was selected for comparison, in which the Dice coefficient was introduced to measures the similarity between cross-lingual words based on the number of aligned (parallel) sentences in which they co-occur. Similar the "coverage" and "sparsity" criteria were applied in IBM-model3. In this study, we redefined those two criteria to fit the case where distributed representations are used.

As shown in Table 4, BWESA achieved the highest performance for ten different language pairs on GRACA, MIHALCEA, and HOLMQVIST datasets. Although BWESA did not outperform IBM-Model3 on the other three pairs, it still performs competitively. Note that IBM-Model3 was tailored for word alignment using many features based on linguistic knowledge. The experimental

results show that our BWESA model can effectively learn high-quality bilingual word embeddings and relatively reliable word alignments in a joint manner.

5 Qualitative analysis

In this section, qualitative analyses were performed evaluating the effectiveness of BWESA on two aspects: neighbouring word discovery and word embedding visualization. The language pair English-German was used.

5.1 Nearest neighbour words

We randomly sampled five words from English and German to retrieve their top-5 nearest neighbour words within and across languages based on the Cosine similarity. The results listed in Table 5 demonstrate that the nearest neighbour words discovered by our model are generally semantically coherent. For example, English and German words describing "time" concepts (such as "moment" and "Zeit") are well clustered, indicating that the desired word clustering is well formed for these two languages.

5.2 Visualization

To illustrate how well the bilingual word representations were learned by BWESA, we plotted a two-dimensional projection of word representations produced by BWESA in Fig. 2. We used the *t*-SNE algorithm (van der Maaten and Hinton 2008) to perform the projection for the illustration. English and German word pairs were randomly extracted from the Open Multilingual WordNet data (Bond and Foster 2013). The distances between any two words are calculated using the Cosine similarity. All English words are indicated using green, and for each English word, its associated German word is shown in blue if their similarity score is greater than a given threshold (say 0.8). If not, the corresponding German word is shown in yellow. We can see that there are many more blue words than yellow ones. This shows that BWESA can produce better word representations, which helps to improve the accuracy of word translation, cross-lingual document classification, and word alignment.

6 Conclusion

We have presented BWESA for learning bilingual word embeddings and alignments in a joint way, in which both tasks can mutually benefit and reinforce each other during the learning process. BWESA is able to learn bilingual word representations from parallel corpora without explicit word-level alignment information in a weakly supervised manner. Two criteria of "coverage" and "sparsity" were reintroduced for

Table 5 Examples of nearest neighbours

Words	German word neighbours	English word neighbours
Januar (January)	April (April), März (March), Oktober (October), Juli (July), Dezember (December)	February, april, august, May, march
Recht (law)	Grundrecht (fundamental right), Freiheit (freedom), Anrecht (right), Anspruch (right), Liberalisierung(liberalization)	Right, dignity, rights, Principle, freedom
Monat (month)	Woche (week), Zeit (time), Tag (day), Jahr (year), Dekade (decade)	Month, week, time, day, year
Bürger (citizen)	Unionsbürger (EU citizens), Mitbürger (fellow citizens), Völker (people), Wähler(voter), Bevölkerungen (populations)	Citizens, voters, consumers, Citizen, people
Sagen (say)	bemerken (notice), erklären (explain), klarstellen (clarify), erinnern (recall), mitteilen (tell)	Say, recall, tell, Understand, remember
People	Menschen (people), Eltern (parents), Kinder (children), Reisenden (travelers), Leute (people)	Inhabitants, citizens, children, Populations, voters
My	meinem (my), mein (my), meine (my), Ihre (your), seine (his)	Your, our, his, her, mine
Union	Union (union), Allianz (alliance), Gemeinschaft (community), EU (EU), Parlament (parliament)	Parliament, community, institutions, EU, alliance
Time	Zeit (time), jetzt (now), nunmehr (now), nun (now), Dauer (duration)	Moment, decade, weeks, timescale, now
Need	müssen (must), brauchen (need), sollten (should), fordern (demand), benötigen (need)	Require, necessary, demand, Must, should

The words in parentheses are the corresponding translations in English

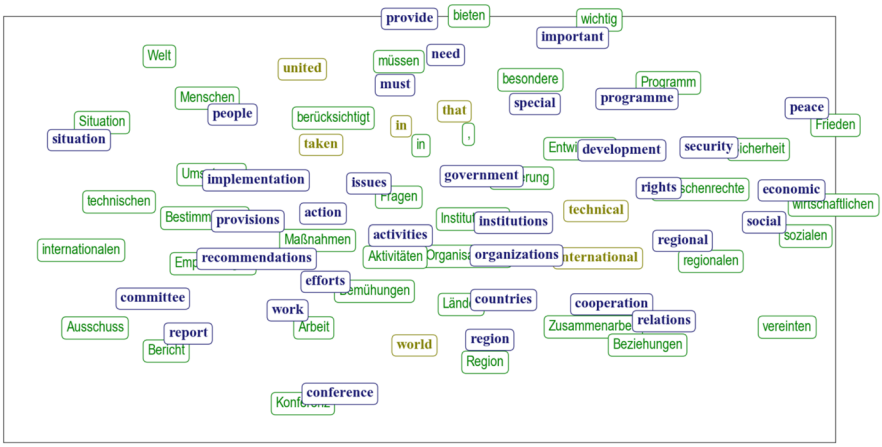

Fig. 2 Two-dimensional projection of the mappings among German and English word embeddings by the *t*-SNE algorithm

learning better word alignments in the case of distributed representations to deal with the under- and over-alignment problems. Extensive experimental results show that BWESA achieved state-of-the-art or comparable results on various cross-lingual tasks, including document classification, word translation, and word alignment.

For future work, it would be interesting to see whether bilingual word embeddings can be learned in a completely unsupervised way. Besides this avenue, we are aware that recently proposed mT5 (Xue et al. 2021), XLM (Conneau and Lample 2019), mBART (Liu et al. 2020) and Unicoder (Huang et al. 2019) could benefit from the idea of soft word alignment, which helps to learn bilingual word representations from the parallel corpora without requiring explicit word-level alignment. We leave this as future work because very large architectures are required to train such contextualized representations at the cost of great computational power, time, and resources.

References

Artetxe M, Labaka G, Agirre E (2017) Learning bilingual word embeddings with (almost) no bilingual data. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol 1. Vancouver, Canada, pp 451–462

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155

Bond F, Foster R (2013) Linking and extending an open multilingual wordnet. In: Proceedings of the 51st annual meeting of the association for computational linguistics (vol 1: Long Papers). Sofia, Bulgaria, pp 1352–1362

Brown PF, Della Pietra VJ, Della Pietra SA, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19(2):263–311

Chandar SA, Lauly S, Larochelle H, Khapra M, Ravindran B, Raykar VC, Saha A (2014) An autoencoder approach to learning bilingual word representations. In: Proceedings of the conference on neural information processing systems. Montreal, Canada, pp 1853–1861

Conneau A, Lample G (2019) Cross-lingual language model pretraining. In: Proceedings of the conference on neural information processing systems. Vancouver, Canada, p 10

Conneau A, Lample G, Ranzato M, Denoyer L, Jégou H (2017) Word translation without parallel data. In: Proceedings of the international conference on learning representations. Vancouver, Canada, p 14

Dyer C, Chahuneau V, Smith NA (2013) A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics. Human Language Technologies, Atlanta, USA, pp 644–648

Feng J, Zheng X (2018) Geometric relationship between word and context representations. In: Proceedings of the AAAI conference on artificial intelligence. New Orleans, Louisiana, USA, pp 5102–5109

Gouws S, Bengio Y, Corrado G (2015) BilBOWA: fast bilingual distributed representations without word alignments. In: Proceedings of the international conference on machine learning. Lille, France, pp 748–756

Guo J, Che W, Yarowsky D, Wang H, Liu T (2015) Cross-lingual dependency parsing based on distributed representations. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1: Long Papers). Beijing, China, pp 1234–1244

Hermann KM, Blunsom P (2014) Multilingual models for compositional distributed semantics. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (vol. 1: Long Papers). Baltimore, Maryland, USA, pp 58–68

Huang H, Liang Y, Duan N, Gong M, Shou L, Jiang D, Zhou M (2019) Unicoder: a universal language encoder by pre-training with multiple cross-lingual tasks. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). Hong Kong, China, pp 2485–2494

Klementiev A, Titov I, Bhattarai B (2012) Inducing crosslingual distributed representations of words. In: Proceedings of COLING 2012. Mumbai, India, pp 1459–1474

Kočiskỳ T, Hermann KM, Blunsom P (2014) Learning bilingual word representations by marginalizing alignments. In: Proceedings of the annual meeting of the association for computational linguistics, vol 2. Baltimore, Maryland, USA, pp 224–229

Koehn P (2005) Europarl: a parallel corpus for statistical machine translation. In: Proceedings of machine translation summit x: papers

Koehn P (2009) Statistical machine translation. Cambridge University Press, Cambridge

Levy O, Søgaard A, Goldberg Y (2017) A strong baseline for learning cross-lingual word embeddings from sentence alignments. In: Proceedings of the conference of the European chapter of the association for computational linguistics, vol 1. Valencia, Spain, pp 765–774

Li X, Li G, Liu L, Meng M, Shi S (2019) On the word alignment from neural machine translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics. Florence, Italy, pp 1293–1303

Liu Y, Gu J, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L (2020) Multilingual denoising pre-training for neural machine translation. Trans Assoc Comput Linguist 8:726–742

Luong T, Pham H, Manning CD (2015) Bilingual word representations with monolingual quality in mind. In: Proceedings of the conference of the North American chapter of the association for computational linguistics. Human Language Technologies, Denver, Colorado, pp 151–159

Mikolov T, Chen K, Corrado G, Dean J (2013a) Efficient estimation of word representations in vector space. arXiv:1301.3781

Mikolov T, Le QV, Sutskever I (2013b) Exploiting similarities among languages for machine translation. arXiv:1309.4168

Mogadala A, Rettinger A (2016) Bilingual word embeddings from parallel and non-parallel corpora for cross-language text classification. In: Proceedings of the conference of the North American chapter of the association for computational linguistics. Human Language Technologies, San Diego, California, USA, pp 692–702

Nakashole N, Flauger R (2018) Characterizing departures from linearity in word translation. In: Proceedings of the 56th annual meeting of the association for computational linguistics (vol 2: Short Papers). Melbourne, Australia, pp 221–227

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Ormazabal A, Artetxe M, Soroa A, Labaka G, Agirre E (2020) Beyond offline mapping: learning cross lingual word embeddings through context anchoring. arXiv:2012.15715

Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the conference on empirical methods in natural language processing. Doha, Qatar, pp 1532–1543

Shi T, Liu Z, Liu Y, Sun M (2015) Learning cross-lingual word embeddings via matrix co-factorization. In: Proceedings of the annual meeting of the association for computational linguistics and the international joint conference on natural language processing, vol 2. Beijing, China, pp 567–572

Smith SL, Turban DH, Hamblin S, Hammerla NY (2017) Offline bilingual word vectors, orthogonal transformations and the inverted softmax. arXiv:1702.03859

Socher R, Bauer J, Manning CD et al (2013a) Parsing with compositional vector grammars. In: Proceedings of the 51st annual meeting of the association for computational linguistics (vol 1: Long Papers), vol 1. Sofia, Bulgaria, pp 455–465

Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng A, Potts C (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the conference on empirical methods in natural language processing. Seattle, Washington, USA, pp 1631–1642

Søgaard A, Ruder S, Vulić I (2018) On the limitations of unsupervised bilingual dictionary induction. In: Proceedings of the 56th annual meeting of the association for computational linguistics (vol 1: Long Papers). Melbourne, Australia, pp 328–339

Upadhyay S, Faruqui M, Dyer C, Roth D (2016) Cross-lingual models of word embeddings: an empirical comparison. In: Proceedings of the 54th annual meeting of the association for computational linguistics (vol 1: Long Papers. Berlin, Germany

van der Maaten L, Hinton G (2008) Visualizing data using t-sne. J Mach Learn Res 9:2579–2605

Vulić I, Moens M-F (2015) Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the international ACM SIGIR conference on research and development in information retrieval. Santiago, Chile, pp 363–372

Wei L, Deng Z-H (2017) A variational autoencoding approach for inducing cross-lingual word embeddings. In: Proceedings of the international joint conference on artificial intelligence. AAAI Press, Melbourne, Australia, pp 4165–4171

Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C (2021) mT5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the conference of the North American chapter of the association for computational linguistics. Human Language Technologies

Zheng X, Feng J, Chen Y, Peng H, Zhang W (2017) Learning context-specific word/character embeddings. In: Proceedings of the AAAI conference on artificial intelligence. San Francisco, California USA

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67(5):768–768

Zou WY, Socher R, Cer D, Manning CD (2013) Bilingual word embeddings for phrase-based machine translation. In: Proceedings of the conference on empirical methods in natural language processing. Seattle, Washington, USA, pp 1393–1398