ELSEVIER

CrossMark

Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 56 (2019) 52–72

www.elsevier.com/locate/csl

Learning English–Chinese bilingual word representations from sentence-aligned parallel corpus☆

An-Zi Yen[a], Hen-Hsen Huang[a,*], Hsin-Hsi Chen[a,b,**]

[a] *Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan*
[b] *MOST Joint Research Center for AI Technology and All Vista Healthcare, National Taiwan University, Taipei, Taiwan*

Received 1 September 2017; Accepted 12 January 2019
Available online 24 January 2019

Abstract

Representation of words in different languages is fundamental for various cross-lingual applications. In the past researches, there was an argument in using or not using word alignment in learning bilingual word representations. This paper presents a comprehensive empirical study on the uses of parallel corpus to learn the word representations in the embedding space. Various non-alignment and alignment approaches are explored to formulate the contexts for Skip-gram modeling. In the approaches without word alignment, concatenating A and B, concatenating B and A, interleaving A with B, shuffling A and B, and using A and B separately are considered, where A and B denote parallel sentences in two languages. In the approaches with word alignment, three word alignment tools, including GIZA++, TsinghuaAligner, and fast_align, are employed to align words in sentences A and B. The effects of alignment direction from A to B or from B to A are also discussed. To deal with the unaligned words in the word alignment approach, two alternatives, using the words aligned with their immediate neighbors and using the words in the interleaving approach, are explored. We evaluate the performance of the adopted approaches in four tasks, including bilingual dictionary induction, cross-lingual information retrieval, cross-lingual analogy reasoning, and cross-lingual word semantic relatedness. These tasks cover the issues of translation, reasoning, and information access. Experimental results show the word alignment approach with conditional interleaving achieves the best performance in most of the tasks.

Keywords: Cross-lingual applications; Distributed word representation; Word alignment

1. Introduction

In distributed word representation, words are mapped to a low dimensional vector space, in which both syntax and semantics can be captured. Using monolingual distributional information has become a common technique in various tasks, such as tagging, chunking (Collobert et al., 2011), parsing (Chen et al., 2014) and sentiment analysis (Kim, 2014). One of popular methods to learn word representations is the Skip-gram model (Mikolov et al., 2013a),

☆ This paper has been recommended for acceptance by Roger Moore.

* Corresponding author.

** Corresponding author at: Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.
E-mail addresses: hhhuang@nlg.csie.ntu.edu.tw (H.-H. Huang), hhchen@ntu.edu.tw (H.-H. Chen).

https://doi.org/10.1016/j.csl.2019.01.002

which is trained to predict the context of a target word in the same sentence.

Distributed word representations can also be learned over different language pairs so that semantically close words in two languages are embedded close together in the same vector space. The bilingual representations can be applied to various applications, e.g., cross-lingual word similarity (Faruqui and Dyer, 2014), unsupervised cross-language part-of-speech tagging (Gouws and Søgaard, 2015), word translation (Mikolov et al., 2013b), and bilingual named entity recognition (Wang et al., 2013).

Recently, the approaches of learning bilingual word representations from parallel corpus with word alignment (Luong et al., 2015; Klementiev et al., 2012) or without word alignment (Coulmance et al., 2015; Gouws et al., 2015; Hermann and Blunsom, 2014; Sarath Chandar et al., 2014) have been proposed. In this paper, we empirically compare the impact of word alignment on learning bilingual word representations from sentence-aligned parallel corpus. In the approaches without word alignment, we experiment four different types of methods to formulate the contexts for Skip-gram modeling, including concatenating sentences in parallel corpus, interleaving words in parallel sentence, shuffling the words in bilingual sentence (Vulić and Moens, 2015), and learning two monolingual word representations separately (Gouws et al., 2015). In the approaches with word alignment, we first apply different word alignment tools, including GIZA++ (Och and Ney, 2003), TsinghuaAligner (Liu and Sun, 2015) and fast_align (Dyer et al., 2013), to align words in sentence-aligned parallel corpus, and then induce word alignment links to learn bilingual semantics. For those unaligned words, Luong et al. (2015) use their immediate aligned neighbors. Alternatively, we introduce a conditional interleaving mechanism to predict a source language's pivot word in the target language context when the pivot word is unaligned. The target language context is determined by the position of the pivot word in the source language corresponding to the position in the target language.

The main contributions of this paper are five-fold. (1) We investigate the approaches of learning bilingual word representations with and without word alignment comprehensively. (2) We propose an approach which learns bilingual word representations from sentence-aligned parallel corpus with word alignment and a conditional interleaving mechanism. (3) We investigate the impact of different word alignment tools and alignment directions in learning bilingual word representations. (4) We evaluate the bilingual representations on four different tasks, including bilingual dictionary induction, cross-lingual information retrieval, cross-lingual analogy reasoning, and cross-lingual word semantic relatedness, which cover the issues of translation, reasoning and information access. (5) We release the datasets for bilingual dictionary induction, cross-lingual analogy reasoning and cross-lingual word semantic relatedness measurement.

The organization of this paper is as follows. Section 2 surveys the related work. Section 3 presents our methodology. Section 4 shows the experimental setups and discusses the results of each of the four tasks. Section 5 concludes the remarks.

2. Related work

Previous works have found that using translational context results has better representations and plays an important role in a lot of NLP tasks, including word alignment (Zhao and Xing, 2006), machine translation (Tam et al., 2007), cross-lingual sentiment analysis (Boyd-Graber and Resnik, 2010) and bilingual lexicon extraction (Vulić and Moens, 2013).

Skip-gram and continuous bag-of-words (CBOW) models proposed by Mikolov et al. (2013a) are simple single-layered architectures that learn useful words representations from raw text. Skip-gram and CBOW encode geometrical distributional properties of words in the embedding space. Other models about learning the distributed representation of words are also proposed (Pennington et al., 2014; Levy and Goldberg, 2014; Lebret and Collobert, 2014).

In recent years, several approaches have been explored to train and align bilingual word embeddings. We category these approaches into three kinds: the approaches based on monolingual corpus, those based on sentence-aligned parallel corpus, and those based on word-aligned parallel corpus. We describe the related work of these approaches in Sections 2.1, 2.2, and 2.3, respectively. Section 2.4 gives an overview the tasks used for evaluating the bilingual word representations.

2.1. Approaches based on monolingual corpus

Zou et al. (2013) utilize the idea of transferring linguistic knowledge into resource-poor languages and learn the bilingual word representations utilizing word alignments extracted by the Berkeley Aligner to constrain bilingual equivalence in the objective function.

Mikolov et al. (2013b) propose an approach that trains the distributed word representation on monolingual corpus separately and then maps word representations across languages by using a small bilingual dictionary. They employ a stochastic gradient descent version of linear projection to transform the source language word vectors to the target language space. Similar to Mikolov et al. (2013b), Bhattacharya et al. (2016) obtain word translation from distributed word representation by using linear regression to learn a projection from the source language space to the target one.

The model proposed by Xiao and Guo (2014) learns bilingual word representations by using Wikitionary to build the connections between bilingual word pairs. Gouws and Søgaard (2015) replace the text in the monolingual corpus with a random translation using a small bilingual dictionary. They produce mixed context−target pairs and use this corpus for training task-specific bilingual word representations. However, this method does not handle polysemy because only a few translations are valid in the context. Duong et al. (2016) utilize the CBOW model to learn bilingual word representations based on a monolingual corpus and a bilingual dictionary.

Mogadala and Rettinger (2016) use the distributed memory model as the monolingual objective and jointly optimize the bilingual regularization function based on the availability of any types of corpus.

2.2. Approaches based on parallel corpus with sentence-alignment

Some approaches rely on sentence-aligned parallel corpus without word alignment. Faruqui and Dyer (2014) extend the method proposed by Mikolov et al. (2013b) and use canonical correlation analysis to project the source and the target language embeddings into a joint space simultaneously, where the bilingual dictionary is obtained from a parallel corpus. Lu et al. (2015) learn the monolingual word vectors same as Faruqui and Dyer (2014). For the bilingual word vectors, instead of linear mappings, they learn a nonlinear transformation by utilizing deep canonical correlation analysis (DCCA).

Hermann and Blunsom (2014) learn the word and document representations by two neural network architectures. The source sentence is read and encoded into a fixed-length vector, and then the translation is resulted from the encoded vector by the decoder. Given a source sentence, the neural network is jointly trained to maximize the probability of a correct translation. Chandar et al. (2014) also use auto-encoders to learn a mapping from a bag-of-words representation of an input phrase to that of an output phrase.

Following the Skip-Gram model (2013a), Gouws et al. (2015) propose the BilBOWA (Bilingual Bag-of-Words without Alignments) model that predicts a word in its context but minimizes the L2-loss between the word vectors of parallel sentences. Different from the work of Gouws et al. (2015), Coulmance et al. (2015) propose the Trans-gram model which uses two cross-lingual objectives and aligns the word vector of the target language to the context vector of the source language. Pham et al. (2015) extend the non-compositional paragraph vector model of Le and Mikolov (2014) to force bilingual sentences sharing the same sentence vector.

The simplest model proposed by Vulić and Moens (2015) is based on document-aligned comparable data called Bilingual Word Embedding Skip-gram model (BWESG) where bilingual semantic space is created by combining and shuffling the comparable sentences. Their model outperforms latent Dirichlet allocation model and unigram language model in both monolingual information retrieval and cross-lingual information retrieval tasks.

Besides neural network methods, Shi et al. (2015) learn the bilingual word representations by utilizing matrix co-factorization framework. They define monolingual objectives in the form of matrix decomposition. They jointly optimize two monolingual objectives with the bilingual objective acting as a bilingual regularizer during factorizing monolingual co-occurrence matrices.

2.3. Approaches based on parallel corpus with word-alignment

In the approaches of using sentence-aligned parallel corpus with word alignment, Klementiev et al. (2012) jointly optimize the mono-lingual and cross-lingual objectives simultaneously by minimizing the summation of monolingual loss functions for each language and the cross-lingual loss function. Wu et al. (2014) propose an approach to bilingual word representation by using the information of the phrase in source sentence and context of the phrase. Their model learns the low dimension of phrases by contextual information and aligned phrases. Kočiský et al.

(2014) proposed the distributed word alignment (DWA) model, which is a probabilistic model that jointly learns word alignments and bilingual distributed word representations. Their model is an extension of the fast_align model (Dyer et al., 2013). Instead of using the standard multinomial translation probability, DWA model uses a similarity measurement over distributed word representations by modifying the log-bilinear language model. Luong et al. (2015) propose Bilingual Skip-gram (BiSkip) whose learning objective is to predict words across languages where the context of words is expanded to include bilingual links obtained from word alignments.

Upadhyay et al. (2016) compare cross-lingual embedding models that require different forms of cross-lingual supervision. They suggest that the model with richer cross-lingual supervision can perform well in cross-lingual semantic tasks. In addition, as the experimental results in Upadhyay et al. (2016), the approaches using parallel corpus generally outperform those without the information of parallel corpus. Furthermore, word-level alignment corpus is rare and the amount of data is less than that of sentence-aligned parallel corpus. In contrast, sentence-aligned parallel corpus is easy to obtain and practical.

Word alignment introduces useful clues, but error alignment or unaligned words may have side effects in learning word representations. Different from the above approaches, this paper focuses on the selection and the uses of the bilingual contexts from a sentence-aligned parallel corpus without and with adopting word alignments. We examine the effects of different word alignment tools and alignment directions, and deal with those unaligned cases. Moreover, we also exhaustively discuss which approaches are preferred in which tasks in the experiments.

2.4. *Evaluating on different tasks*

We summarize the experiments for evaluating the quality of cross-lingual embeddings. In previous works, the cross-lingual embeddings have been evaluated on three different tasks, including the cross-lingual document classification (CLDC) introduced by Klementiev et al. (2012), the word translation task used by Mikolov et al. (2013b), and monolingual word semantic relatedness measurement. We show the results in Tables 1, 2 and 3, respectively. The best performances are highlighted in bold.

Table 1 shows the results of previous cross-lingual embeddings evaluated on the CLDC task. The goal of the CLDC task is to classify documents in the target language using only labeled documents in the source language. Klementiev et al. (2012) induce cross-lingual embeddings for English−German pairs, classify a subset of the English and German sections of the Reuters RCV1/RCV2 multilingual corpus to four categories: CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets). In this task, Pham et al. (2015) achieve accuracy higher than 90% on the CLDC task in both directions.

Table 2 shows the results of the word translation task used by Mikolov et al. (2013b). The words are extracted from the publicly available WMT11 English−Spanish. They extract the top 5 K frequent word pairs to train a translation matrix, and evaluate their method on the remaining 1 K. And then they use the online Google Translate service to derive two sets of translation: English to Spanish and Spanish to English. Table 2 shows the evaluation of the precision P@k as the fraction of target translations that are within the top-k words returned by their methods.

Table 1
Results on Reuters cross-lingual document classification task.

Method	En → De	De → En
Majority class	46.8%	46.8%
Klementiev et al. (2012)	77.6%	71.1%
Hermann and Blunsom (2014)	83.7%	71.4%
Sarath Chandar et al. (2014)	91.8%	72.8%
Kočiský et al. (2014)	83.1%	75.4%
Gouws et al. (2015)	86.5%	75%
Luong et al. (2015)	87.6%	77.8%
Coulmance et al. (2015)	87.8%	78.7%
Pham et al. (2015)	**92.7%**	**91.5%**
Duong et al. (2016)	86.3%	76.8%
Mogadala and Rettinger (2016)	88.1%	78.9%

Table 2
Results on the translation task.

Method	En → Es P@1	En → Es P@5	Es → En P@1	Es → En P@5
Edit distance	13%	18%	24%	27%
Word co-occurrence	19%	30%	20%	30%
Mikolov et al. (2013b)	33%	35%	**51%**	52%
Gouws et al. (2015)	39%	44%	**51%**	55%
Coulmance et al. (2015)	**45%**	**61%**	47%	**62%**

In the word translation task, Mikolov et al. (2013b) report the scores of two baseline strategies based on edit-distance and word co-occurrence, respectively. Coulmance et al. (2015) achieve a precision@1 of 45% for translation from English to Spanish. Both Mikolov et al. (2013b) and Gouws et al. (2015) achieve a precision@1 of 51% for translation from Spanish to English.

Table 3 shows the results of monolingual word semantic relatedness measurement on 3 datasets: WS-de (353 pairs), WS-en (353 pairs) and RW-en (2034 pairs) (Finkelstein et al., 2001; Luong et al., 2013, 2015). These datasets contain 353, 353, and 2034 word pairs, respectively, and the semantic similarity between each word pair is labeled by human annotators. Good word embeddings should produce the similarity score correlated with human judgment. The approach proposed by Duong et al. (2016) outperforms other methods.

Most previous studies investigated bilingual word representations between English and one of European languages. The bilingual word representations between English and Chinese are few explored. In this work, we focus on learning bilingual word representations between English and Chinese. A variety of strategies for word-level alignment are proposed. We compare our approaches with previous methods by evaluating on four different tasks, including bilingual dictionary induction, cross-lingual information retrieval, cross-lingual analogy reasoning, and cross-lingual word semantic relatedness.

3. Methodology

The goal of bilingual word embedding is to learn word representations for all words in both source language and target language in such a way that similar representations must be semantically close and similar representations must be assigned to similar words across languages.

In the previous studies, several strategies for learning bilingual word representations have been proposed. However, the comparison between the non-alignment and alignment approaches are unexplored. In this paper, two families of approaches – say, sentence-aligned only and sentence-aligned with word alignment, are compared. In the former family, we investigate various methods to learn bilingual representations from sentence-aligned parallel corpus without word alignment. In the latter family, we learn the bilingual representations from the bilingual signals aligned by various word alignment tools.

Section 3.1 introduces the parallel corpus we use to train the bilingual word representations. In Section 3.2, three word-alignment tools used in experiments are described and compared. Section 3.3 introduces Skip-gram, the word representation learning algorithm. Section 3.4 shows a number of methods that train the bilingual word representation on the parallel corpus without word-alignment. In Section 3.5, we propose the strategies that train the bilingual word representation by using the word-alignment information.

3.1. Corpus

For training the bilingual word representations, our material is an English–Chinese parallel corpus, named UM-Corpus[1] (Tian et al., 2014), which consists of 2.2 million parallel sentences from eight domains, including News, Spoken, Laws, Thesis, Education, Science, Subtitle, and Microblog. All the sources are collected from the online journals (national and international), official websites, online language learning resources (e.g. online dictionary and translation portals), TED, and Microblogs. The crawled HTML documents are parsed, and the contents are extracted. Tian et al. (2014) utilize existing mature algorithms to accelerate the building process, such as document alignment, sentence boundary detection and statistical sentence alignment approach. For quality concern, the final result is manually verified.

[1] http://nlp2ct.cis.umac.mo/um-corpus/index.html.

Table 3
Spearman’s rank correlation for monolingual word semantic relatedness measurement.

Method	WS-de	WS-en	RW-en
Klementiev et al. (2012)	0.238	0.132	0.073
Sarath Chandar et al. (2014)	0.346	0.398	0.205
Hermann and Blunsom (2014)	0.283	0.198	0.136
Gouws and Søgaard (2015)	0.674	0.718	0.310
Luong et al. (2015)	0.474	0.493	0.253
Lu et al. (2015)	–	0.708	–
Duong et al. (2016)	**0.711**	**0.762**	**0.440**

3.2. *Word alignment tools*

Word alignment is aimed at identifying the correspondence between words in two languages. In this paper, we utilize three different bilingual word alignment tools, including GIZA++ (Och and Ney, 2003), TsinghuaAligner (Liu and Sun, 2015), and fast_align (Dyer et al., 2013).

GIZA++ trains the IBM Models (Brown et al., 1993) and the Hidden Markov Model (HMM) (Vogel et al., 1996) using the EM algorithm, and applies these models to compute Viterbi alignments for statistical machine translation. TsinghuaAligner takes the translation probabilities derived from GIZA++ as the central feature and introduces log-linear models into word alignment. Fast_align is a variation of the lexical translation models. The model of word alignment configurations is a log-linear reparameterization of Model 2.

Word alignment tool aligns words in a source language sentence to the corresponding words in a target language sentence. For a sentence-aligned parallel corpus, either side can be considered as the source or the target. Different alignment direction may produce different word-alignment results, and thus affect the learning of bilingual word representations. In particular, the two languages come from different language families, e.g., English and Chinese in the UM-Corpus, which is our experimental dataset.

We take a bilingual sentence as an example to demonstrate the problems in applying the word alignment tools. Figs. 1–3 show the word alignments from English sentence to Chinese sentence by the above three tools. Figs. 4–6 show the word-aligned results in the reverse direction, i.e., from Chinese sentence to English sentence. The red links denote the wrong word alignments. Figs. 1,$^{-3}$ show the alignment of TsinghuaAligner is better. In this case, “第一輪 (diyilun)” should be aligned with “first” and “round”, however, fast_align aligns “floored” with “第一輪 (diyilun)”; GIZA++ aligns “floored” with “第一輪 (diyilun)”, “比賽 (bisai)”, and “猛擊 (mengji)”.

Figs. 2 and 5 show the word alignments of TsinghuaAligner from Chinese sentence to English sentence and from English sentence to Chinese sentence are almost same. However, there are still some unaligned words. Interestingly, GIZA++ and fast_align “第一輪 (diyilun)” to “first” and “round’ correctly in Figs. 4 and 6, although fast_align also aligns “第一輪 (diyilun)” to “floored”. The result of word alignment will influence the bilingual word representations in the embedding space. That is, the wrong alignments propagate the error to the step of word representation learning. Specifically, the unaligned words may learn incorrect semantic relations between Chinese and English. In this work, we propose methods to deal with the error alignments and unaligned problems resulting from word alignment tools.

3.3. *Skip-gram*

The idea of representing words as vectors is proposed by Rumelhart et al. (1986). Before introducing the methods we propose, we briefly introduce the Skip-gram model (Mikolov et al., 2013a), which is a backbone of many

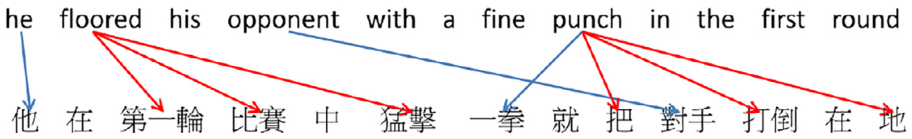
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地

Fig. 1. Applying GIZA++ from English to Chinese.

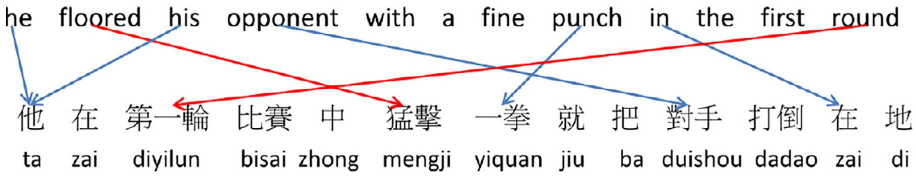
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地
ta zai diyilun bisai zhong mengji yiquan jiu ba duishou dadao zai di

Fig. 2. Applying TsinghuaAligner from English to Chinese.

approaches to learning word representations. Skip-gram learns word representations using simple neural network architecture for statistical language modeling. Its training objective is to learn the representation of a word by predicting the word's context in the same sentence. Given a pivot word *w* and its context words *cw*, the probability of *cw* and *w* is defined by the softmax function, where $\overrightarrow{w}$ and $\overrightarrow{cw}$ are word representations of *w* and *cw*, respectively:

$$p(cw|w) = \frac{1}{1 + \exp\left(-\overrightarrow{w} \cdot \overrightarrow{cw}\right)}$$

The objective of the Skip-gram model is to estimate the log probability of *cw* to be in the context of *w*, and the learning goal is to maximize the average log probability:

$$J = \frac{1}{M} \sum_{S \in T} \sum_{w \in S} \sum_{cw \in S[i-l:i+l]} \log p(cw|w)$$

where *T* is a training corpus, *S* is a sentence in *T*, $S[i-l : i+l]$ is a word window in *S* centered on *w*, and *M* is total number of words in *T*.

An efficient approach of deriving word embeddings presented by Mikolov et al. (2013a) is negative-sampling, which can improve both the training speed and the quality of word embeddings. Given (*w, cw*), a pair of word *w* and the context *cw* in the training corpus, the model samples a set of "negative pivot-context" pairs (*w, cw'*) where $cw' \neq w$.

After training, closely related words that predict similar context words should have similar vector representations. In Sections 3.4 and 3.5, we will describe how to extend this idea to cross-lingual contexts.

3.4. Approaches without word alignment

In addition to the past researches (Gouws et al., 2015; Vulić and Moens, 2015), we employ the following methods to formulate the context for learning bilingual word representations from English−Chinese parallel corpus without word alignment.

Concat(EC) and Concat(CE): These approaches are simple. We just concatenate each pair of aligned sentences from two languages directly and train word embeddings with Skip-gram model. Concat(EC)/Concat(CE) means putting English sentence before/after the corresponding Chinese Sentence. Consider an English-Chinese parallel sentence "I saw her duck with a telescope"-"我 用 望遠鏡 看 她的 鴨子". In the pseudo sentence produced by Concat (EC), the context of "telescope" will be "duck", "with", "a", "我 (I)", "用 (with)", and "望遠鏡 (telescope)" when

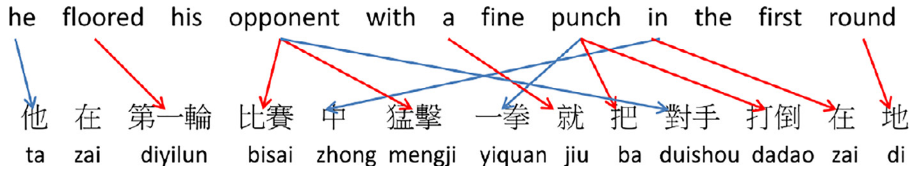
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地
ta zai diyilun bisai zhong mengji yiquan jiu ba duishou dadao zai di

Fig. 3. Applying fast_align from English to Chinese.

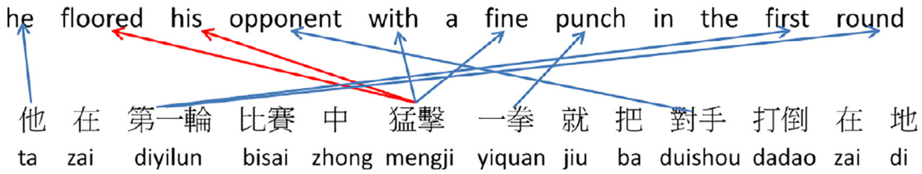
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地
ta zai diyilun bisai zhong mengji yiquan jiu ba duishou dadao zai di

Fig. 4. Applying GIZA++ from Chinese to English. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

the window size is 7. It shows an interesting linguistic phenomenon - preposition phrase tends to succeed English verb and precede Chinese verb.

Interleaving: The main idea of the Interleaving approach is to extract bilingual signals from contexts in a parallel sentence. It tries to estimate the probabilities of a target word w_t to appear in the context of target language cw_t and in the context of source language cw_s, respectively. Let each English sentence S_e be aligned with Chinese sentence S_c in the parallel corpus $T_{e,c}$. Assume the lengths of S_e and S_c are m and n, respectively, and M is total number of English words. The context picked for a target word w_e in S_e is a word window $cw_e \in S_e[i_{e-l} : i_{e+l}]$ where i_e is the position of w_e in S_e. On the other hand, the context picked for the target word w_e in S_c is the set of words cw_c from the position $\frac{i_e}{k}-l$ to $\frac{i_e}{k}+l$ in S_c where k depends on m and n. The following equation is the objective function Ω_e of aligning English target word with English and Chinese contextual words, where k is $\frac{\max(m,n)}{\min(m,n)}$.

$$\Omega_e = \frac{1}{M}\sum_{(S_e,S_c\in T_{e,c})}\sum_{w_e\in S_e}\left[\sum_{cw_e\in S_e[i_e-l:i_e+l]}\log p(cw_e|w_e)+\sum_{cw_c\in S_c\left[\frac{i_e}{k}-l:\frac{i_e}{k}+l\right]}\log p(cw_c|w_e)\right]$$

Fig. 7 shows a parallel sentence “he floored his opponent with a fine punch in the first round” – “他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地” to demonstrate the computation of Ω_e, where “fine” is the pivot word and l is 2.

Similarly, we define Ω_c from the Chinese side in the parallel corpus. We sum up the objective functions Ω_e and Ω_c as follows, and let it be the objective function of the Interleaving approach.

$$J_{Interleaving} = \Omega_e + \Omega_c$$

3.5. Approaches with word alignment

Unaligned words are one of the major problems in using word alignment tools. This paper proposes an approach that integrates word alignment and the interleaving approach. We call this approach **BiCIn** (**bi**lingual word representations with word alignment and **c**ondition **in**terleaving). Here, interleaving is applied on condition only for those unaligned words. The algorithm is shown as follows.

First, we utilize word alignment tool GIZA++ (GA), TsinghuaAligner (TA), and fast_align (FA) separately to align words between English sentence and Chinese sentence, and train the Skip-gram model with word alignment information. Then, we use the Interleaving approach to deal with the words which are not aligned. Formally, given an alignment link between a word w_e in S_e and a word w_c in S_c. The BiCIn model uses w_e to predict context words in

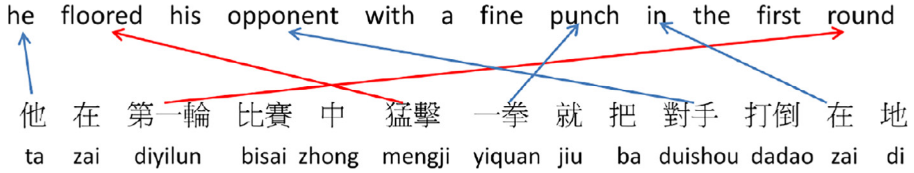
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地
ta zai diyilun bisai zhong mengji yiquan jiu ba duishou dadao zai di

Fig. 5. Applying TsinghuaAligner from Chinese to English. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

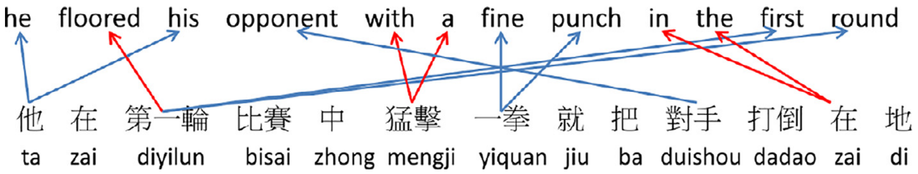
he floored his opponent with a fine punch in the first round
他 在 第一輪 比賽 中 猛擊 一拳 就 把 對手 打倒 在 地
ta zai diyilun bisai zhong mengji yiquan jiu ba duishou dadao zai di

Fig. 6. Applying fast_align from Chinese to English. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

S_e centered around $cw_e \in S_e[a_e - l{:}a_e + l]$ and uses w_e to predict w_c in S_c from position $a_c - l$ to $a_c + l$ where a_e and a_c are positions of w_e and w_c in S_e and S_c, respectively. The following equation is the objective function Φ_e of aligning English target word vectors with English context vectors and Chinese context vectors where A_e is aligned words in S_e. On the other hand, U_e is unaligned words in S_e.

$$\Phi_e = \frac{1}{M} \sum_{(S_e, S_c \in T_{e,c})} \left\{ \sum_{w_e \in A_e} \left[\sum_{cw_e \in S_e[a_e - l : a_e + l]} \log p(cw_e | w_e) + \sum_{cw_c \in S_c[a_c - l : a_c + l]} \log p(cw_c | w_e) \right] \right.$$

$$\left. + \sum_{w_e \in U_e} \left[\sum_{cw_e \in S_e[i_e - l : i_e + l]} \log p(cw_e | w_e) + \sum_{cw_c \in S_c\left[\frac{i_e}{k} - l : \frac{i_e}{k} + l\right]} \log p(cw_c | w_e) \right] \right\}$$

Φ_c is defined in the similar way from the Chinese side. We sum up the objective function Φ_e and Φ_c as follows and let it be the objective function of BiCIn.

$$J_{BiCIn} = \Phi_e + \Phi_c$$

Fig. 8 shows the idea of the BiCIn method for learning bilingual word representations where the word alignment link is based on the example in Fig. 2. Here “punch” is aligned with “一拳 (yiquan)”. For the unaligned word “fine”, the computation of Φ_e in BiCIn is the same as the interleaving approach shown in Fig. 7.

4. Experiments

We evaluate the qualities of the bilingual word representations with the following four tasks:

1. Bilingual dictionary induction.
2. Cross-lingual information retrieval.
3. Cross-lingual analogy reasoning.
4. Cross-lingual word semantic relatedness.

We compare the approaches proposed in Sections 3.4 and 3.5 with the methods proposed in previous work including BWESG (Shuffle) (Vulić and Moens, 2015), the Bilingual Skip-gram Model (BiSkip) (Luong et al., 2015), and the Bilingual Bag-of-Words without Alignments (BilBOWA) (Gouws et al., 2015). All models are trained by the Skip-gram model with dimension 300, window size tuned over {10, 20, 30, 40}, and 15 negative samples. The word

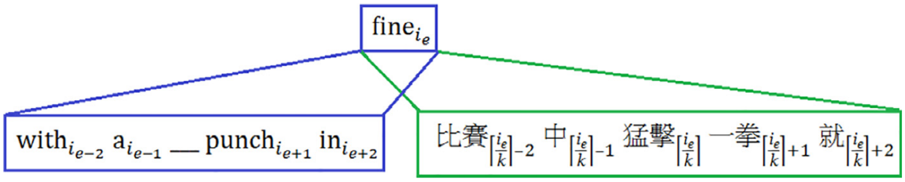
fine_{i_e}
with_{i_e-2} a_{i_e-1} — punch_{i_e+1} in_{i_e+2}
$比賽_{\left[\frac{i_e}{k}\right]-2}$ $中_{\left[\frac{i_e}{k}\right]-1}$ $猛擊_{\left[\frac{i_e}{k}\right]}$ $一拳_{\left[\frac{i_e}{k}\right]+1}$ $就_{\left[\frac{i_e}{k}\right]+2}$

Fig. 7. An example for the interleaving approach.

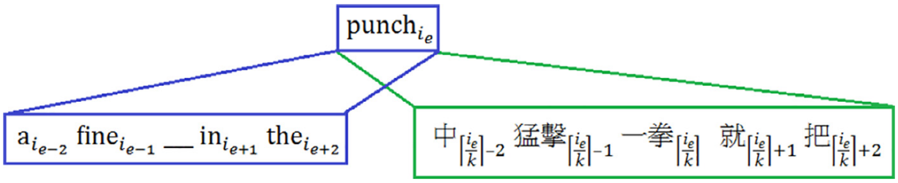
punch_{i_e}
$\text{a}_{i_{e-2}}$ $\text{fine}_{i_{e-1}}$ __ $\text{in}_{i_{e+1}}$ $\text{the}_{i_{e+2}}$
中$_{\lceil \frac{i_e}{k} \rceil -2}$ 猛擊$_{\lceil \frac{i_e}{k} \rceil -1}$ 一拳$_{\lceil \frac{i_e}{k} \rceil}$ 就$_{\lceil \frac{i_e}{k} \rceil +1}$ 把$_{\lceil \frac{i_e}{k} \rceil +2}$

Fig. 8. An example for the word alignment with conditional interleaving.

Table 4
Bilingual dictionary induction (EC).

Alignment	Method	MRR	
Without word alignment	Shuffle	0.4652	
	BilBOWA	0.3538	
	Concat(EC)	0.4569	
	Concat(CE)	0.4570	
	Interleaving	0.4768	
With word alignment	Method	MRR(E)	MRR(C)
	GA+BiSkip	0.4451	0.3018
	TA+BiSkip	0.4462	0.3002
	FA+BiSkip	0.4501	0.3791
	GA+BiCIn	0.4755	0.4757
	TA+BiCIn	0.4774	0.4824
	FA+BiCIn	0.4802	**0.4867**

alignments for training the BiSkip model are generated using the systems specified in Section 3.1 and the other parameters follow (Upadhyay et al., 2016).

A word alignment tool aligns a source language sentence to a target language sentence. In our experimental English-Chinese parallel corpus, the directions can be either English to Chinese or Chinese to English. Both directions are experimented in this study. We report the results of English and Chinese as source language in bilingual dictionary induction task, and the better one is reported in the rest tasks due to the length limits. In all experiments, McNemar's test is adopted for significance test ($p<0.001$). The numbers with underline denote the results are better than Shuffle, BilBOWA and BiSkip with significance level ($p<0.001$).

The datasets for bilingual dictionary induction, cross-lingual analogy reasoning and cross-lingual word semantic relatedness measurement are released in our website.[2]

4.1. Bilingual dictionary induction

This task aims at judging the quality of bilingual word representations based on a bilingual dictionary. We adopt a dictionary which contains 306,519 bilingual pairs. There are two subtasks: (1) English to Chinese (EC), and (2) Chinese to English (CE). For the EC task, there are 45,104 English words and each word has at least one Chinese translation. For the CE task, there are 173,790 Chinese words along with its English translation words.

For each bilingual entry (w_1, w_2) in the dictionary, we report the top 10 neighbors of w_1 based on the learned bilingual word representations, and compute the RR (reciprocal rank) of w_2. Tables 4 and 5 show the MRR (mean reciprocal rank) of the methods in EC task and CE task, respectively, where MRR(E) and MRR(C) mean the source language is English and Chinese, respectively.

For the approaches without word alignment, Interleaving is the best one, Shuffle is the next and BilBOWA is the worst. The reasons are as follows. The cross-lingual loss function of BilBOWA is counting distance between bag-of-words representations of two aligned sentences, while the loss function of the other approaches is extracting linguistic features of two aligned sentences.

For the approaches with word alignment, BiCIn outperforms BiSkip no matter which alignment tools are used. Because word alignment errors will generate wrong links and BiSkip uses the nearest neighbor alignment or an aver-

[2] http://nlg18.csie.ntu.edu.tw/BiCin/Dataset.zip.

Table 5
Bilingual dictionary induction (CE).

Alignment	Method	MRR	
Without word alignment	Shuffle	0.3836	
	BilBOWA	0.2921	
	Concat(EC)	0.3791	
	Concat(CE)	0.3685	
	Interleaving	0.3951	
With word alignment	Method	MRR(E)	MRR(C)
	GA+BiSkip	0.3741	0.2140
	TA+BiSkip	0.3748	0.2132
	FA+BiSkip	0.3779	0.2853
	GA+BiCIn	0.4074	0.3997
	TA+BiCIn	0.4100	0.4002
	FA+BiCIn	**0.4145**	0.4028

Table 6
Results in Single Language IR (SLIR).

Alignment	Method	E-T	E-D	C-T	C-D
–	NTCIR-4	0.3576	0.3469	0.3146	0.3255
–	SG (E)	0.1411	0.1351	–	–
	SG (C)	–	–	0.0817	0.0874
Without word alignment	Shuffle	0.1698	**0.1643**	0.1143	0.0862
	BilBOWA	0.0694	0.0641	0.0439	0.0452
	Concat (EC)	0.1566	0.1458	0.1073	**0.0913**
	Concat (CE)	0.1627	0.1572	0.0965	0.0800
	Interleaving	0.1625	0.1598	0.1080	0.0867
With word alignment	GA+BiSkip	**0.1808**	**0.1637**	**0.1165**	0.0817
	TA+BiSkip	**0.1820**	0.1599	0.1155	0.0801
	FA+BiSkip	0.1683	0.1525	0.1079	0.0871
	GA+BiCIn	0.1690	**0.1696**	**0.1162**	**0.1021**
	TA+BiCIn	**0.1735**	0.1593	0.1136	0.0909
	FA+BiCIn	0.1697	0.1575	**0.1163**	**0.1039**

age of the two neighbor alignments, the wrong links result in the incorrect word cross-lingual semantics when a word is unaligned. Take word alignment links in Fig. 3 as an example. The word “first” is unaligned, so BiSkip will use the alignment of its neighbor “round”. However, the neighbor’s alignment is wrong.

In addition, comparing the performance of BiSkip and BiCIn with different word alignment tools and different directions shown in Tables 4 and 5, BiSkip prefers applying fast_align to align words and using English as the source language. The performance is worst when the source language is Chinese. In contrast, BiCIn is more robust. The performance of BiCIn is similar no matter which word alignment tool and which alignment direction are adopted. Consequently, using BiCIn to learn bilingual word representations is more appropriate than BiSkip in this task.

Comparing the results of the best approaches without word alignment (i.e., Interleaving) and the best approach with word alignment (i.e., BiCIn), the word level alignment information improves the performance.

4.2. Cross-lingual information retrieval

We use the dataset in the NTCIR-4[3] Cross-Lingual Information Retrieval (CLIR) task to evaluate different approaches. NTCIR-4 provides two subtasks: Single Language Information Retrieval (SLIR) and Bilingual Information Retrieval (BLIR). In SLIR, the topic set and the document set are written in the same language. In BLIR task, a topic in one language is used to access the documents in another language. The NTCIR-4 dataset contains 347,376 English documents and 381,375 Chinese documents. Total 60 topics are given. Each topic consists of four fields, i.e., Title (T), Description (D), Narrative (N), and Concepts (C).

[3] http://research.nii.ac.jp/ntcir/ntcir-ws4/ws-en.html.

In SLIR, we execute T-run and D-run in which Title and Description fields are used, respectively. In BLIR, we execute D-run only. Topic and document embeddings are constructed by summing up word vectors by equal weight, and then documents are ranked by computing the cosine similarity between topic embedding and each document embedding. Mean Average Precision (MAP) score is considered as the main evaluation metric.

Table 6 shows the result of SLIR runs on Chinese documents and English documents. In the notations “E-T”, “E-D”, “C-T” and “C-D”, “E” and “C” denote English and Chinese, and “T” and “D” denote T-run and D-run. For example, “E-T” means T-run on English document set. The top three MAP scores are shown in bold. In SLIR, we also compare bilingual embeddings with monolingual embeddings trained on the Skip-gram model with only English sentences or Chinese sentences in UM-Corpus, e.g., SG(E) is the Skip-gram model trained on English monolingual sentences only.

The best runs in NTCIR-4 are reported in the first line of Table 6. The performances of the methods explored in this paper are all below the best scores in NTCIR-4. The major reason is that the out-of-vocabulary (OOV) words appear in 6 English topics and 16 Chinese topics. Take the English topic, “Find articles containing Taiwan laborers' appeal in the "Chiutou" (Autumn Struggle) protest and the laborer policies proposed by Government in 1998″, as an example. “Chiutou”, which is not in the UM-corpus, is an OOV word. The best NTCIR-4 run builds knowledge ontology for some topic terms by using search engine on the Internet with manual verification to solve the OOV problem. In this task, we do not deal with the OOV problem with human intervention, so that the performances cannot be compared directly.

Vulić and Moens (2015) propose the BWESG (Shuffle) model, which performs better than the omnipresent standard query likelihood model and the latent Dirichlet allocation (LDA) model in both Monolingual Retrieval and Cross-Lingual Retrieval. It is the major target to be compared in the experiments.

Table 6 shows that Shuffle is better than the other approaches without word alignment in most runs. Comparing the results of the approaches without word alignment and the approaches with word alignment, the times of the latter approaches ranking in top three (in bold) are more than those of the former approaches. In the approaches with word alignment, the times of BiCIn ranking in top three are more than those of BiSkip. Moreover, BiCIn performs better than Monolingual Skip-gram model in all the runs.

Table 7 shows the results of the BLIR runs on English topics accessing to Chinese documents (E→C) and the BLIR runs on Chinese topics accessing to English documents (C→E). The top three MAP scores are shown in bold and the best runs in NTCIR-4 are reported in the first line.

Intuitively, cross-lingual information retrieval is more challenging than the monolingual one because it needs to cross the language boundaries by translating query and comprehending word semantics. In general, machine translation systems or bilingual dictionaries cannot cover all the words included in queries. The best run in NTCIR-4 solves the out-of-vocabulary problem in the translation process by expanding bilingual dictionaries with the web resources and collecting translation information of unknown words from the web manually.

For the approaches without word alignment, Shuffle performs better in C→E, and Interleaving performs better in E→C. Especially, the performances of Interleaving are better than those of BiSkip in both E→C and C→E tasks. That is consistent with the experiments in the bilingual dictionary induction.

Table 7
Results in Bilingual Information Retrieval (BLIR).

Alignment	Method	E→C	C→E
–	NTCIR-4	0.0663	0.2238
Without word alignment	Shuffle	0.0268	**0.0781**
	BilBOWA	0.0036	0.0177
	Concat(EC)	0.0235	0.0646
	Concat(CE)	0.0160	0.0678
	Interleaving	0.0283	0.0707
With word alignment	GA+BiSkip	0.0278	0.0664
	TA+BiSkip	0.0268	0.0642
	FA+BiSkip	0.0261	0.0612
	GA+BiCIn	**0.0286**	**0.0725**
	TA+BiCIn	**0.0294**	**0.0811**
	FA+BiCIn	**0.0294**	0.0715

In the approaches with word alignment, BiCIn performs better than BiSkip does in bilingual information retrieval no matter which alignment tools are adopted. The results show that BiCIn captures monolingual and bilingual word representations more precisely than the other bilingual word representation methods do.

4.3. *Cross-lingual analogy reasoning*

The word analogy task is first introduced by Mikolov et al. (2013a). The analogy reasoning questions are in the form of “a is to a* as b is to b*”. We are required to infer b* given the known identities of a, a*, and b. For instance, in the question “king: queen = man: __”, woman is returned as the answer. Similarly, France is returned for the question “Madrid: Spain = Paris: __”.

The goal of the cross-lingual analogy reasoning task is to evaluate the performance of the bilingual embeddings in capturing cross lingual syntactic and semantic regularities. We develop a cross-lingual analogy reasoning dataset based on the analogy reasoning dataset (Mikolov et al., 2013a), which consists of 8869 semantic questions and 10,675 syntactic questions in English. In the original dataset, there are five types of semantic questions, including capital-common-countries, capital-world, currency, city-in-state, and family. Moreover, there are nine types of syntactic questions in the dataset, including plural verbs, past tense, adjective to adverb, etc.

We translate English words in an analogy reasoning question into Chinese by Google Translate. We consider the following bilingual analogy reasoning questions in which the issues of translation and reasoning are covered. Note EECC means a and a* are in English, and b and b* are in Chinese. The other notations have the similar interpretations.

(a) EECC and CCEE: the questions relate to cross-lingual analogy reasoning only.
(b) ECEC and CECE: the questions relate to both cross-lingual analogy reasoning and translation.
(c) EEEC and CCCE: the questions relate to cross-lingual asymmetric reasoning and translation.

Some linguistic properties in English do not appear in Chinese. For example, Chinese does not have plural verbs, thus “eat” and “eats” have the same Chinese translation. In the experimental setup, we remove the analogy reasoning questions which are not suitable for cross-language evaluation. Table 8 shows the types and the statistics of the analogy reasoning questions in this study.

We deal with the analogy reasoning by the 3CosMul method (Levy et al., 2014) shown as follows, where ε=0.001. The performance is measured by accuracy at top 1 (acc@1) and accuracy at top 5 (acc@5).

$$\underset{\mathrm{b}^* \in V}{\operatorname{argmax}} \left(\frac{\cos(b^*, b)\cos(b^*,\ a^*)}{\cos(b^*, \mathrm{a}) + \varepsilon} \right)$$

4.3.1. *Results of monolingual analogy reasoning*

Figs. 9 and 10 show the results of English and Chinese analogy reasoning measured by acc@1 and acc@5, respectively. The method SG(E)/SG(C), which is the Skip-gram model trained on English/Chinese monolingual sentences only, is regarded as the baseline. All the methods adopted in the experiments are listed in the x-axis, acc@1 and acc@5 are shown in the *y*-axis with different colors, and the exact numbers are specified in the top of the bars.

Table 8
Types and statistics of analogy reasoning questions.

A	a*	b	b*	Abbr.	#Questions
English	English	Chinese	Chinese	EECC	9266
English	Chinese	English	Chinese	ECEC	4633
English	English	English	Chinese	EEEC	4633
Chinese	Chinese	English	English	CCEE	9266
Chinese	English	Chinese	English	CECE	4633
Chinese	Chinese	Chinese	English	CCCE	4633
English	English	English	English	EEEE	19,544
Chinese	Chinese	Chinese	Chinese	CCCC	4633

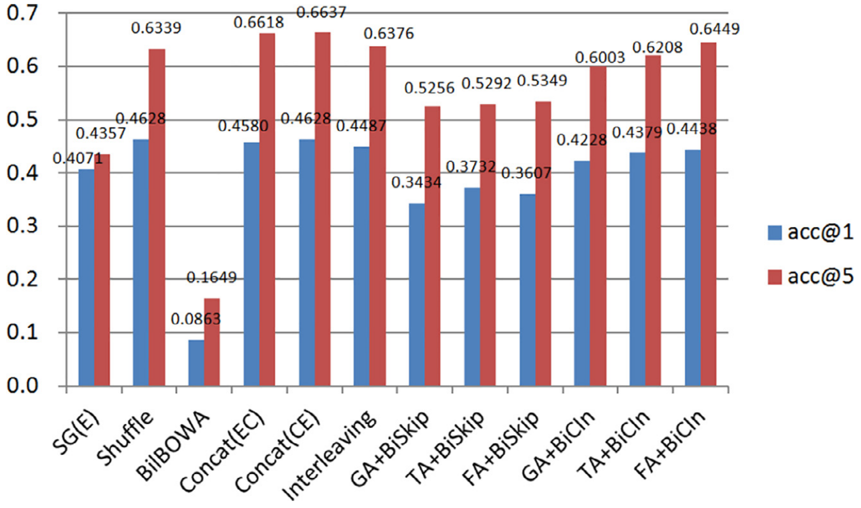

Fig. 9. Results of English analogy reasoning. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Interestingly, the performances of the proposed bilingual embedding models in Sections 3.4 and 3.5, i.e., Concat (EC), Concat(CE), Interleaving, and BiCIn, are better than those of monolingual embeddings, i.e., SG(E) and SG (C), in monolingual analogy reasoning questions. This is because similar representations are assigned to similar words across languages in bilingual vector space. That can help infer word relationship in one language if the bilingual embedding space captures correct semantic regularities for the words in another language.

For illustration, the t-SNE algorithm is used to project the word vectors to two-dimensional figures. Figs. 11 and 12 show the directions from capital to the country in SG(E) and SG(C), where green lines denote the direction are different from the others. The different directions from the capital to the country will result in wrong analogy reasoning. On the other hand, Fig. 13 shows the directions from the capital to the country in bilingual embeddings learned by TA+BiCIn. Note that the directions from the capital to the country either in English or Chinese shown in Fig. 13 are almost similar. These examples demonstrate that bilingual information is helpful to monolingual analogy reasoning task.

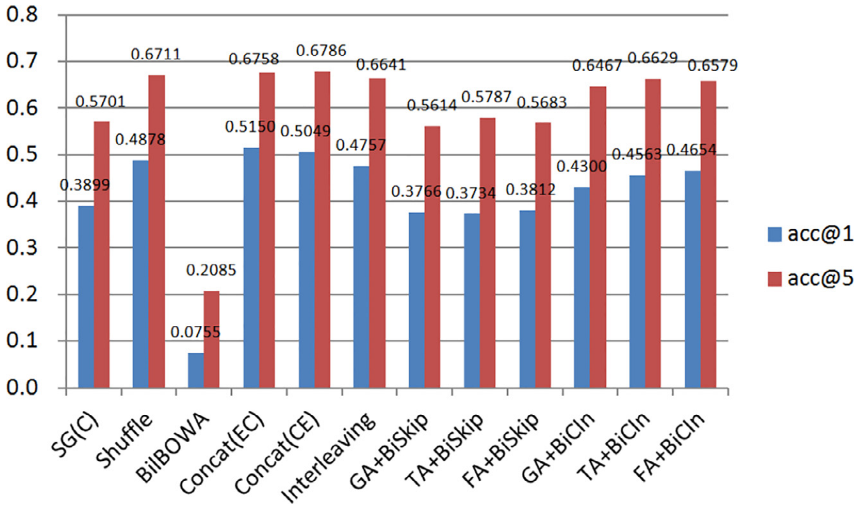

Fig. 10. Results of Chinese analogy reasoning. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

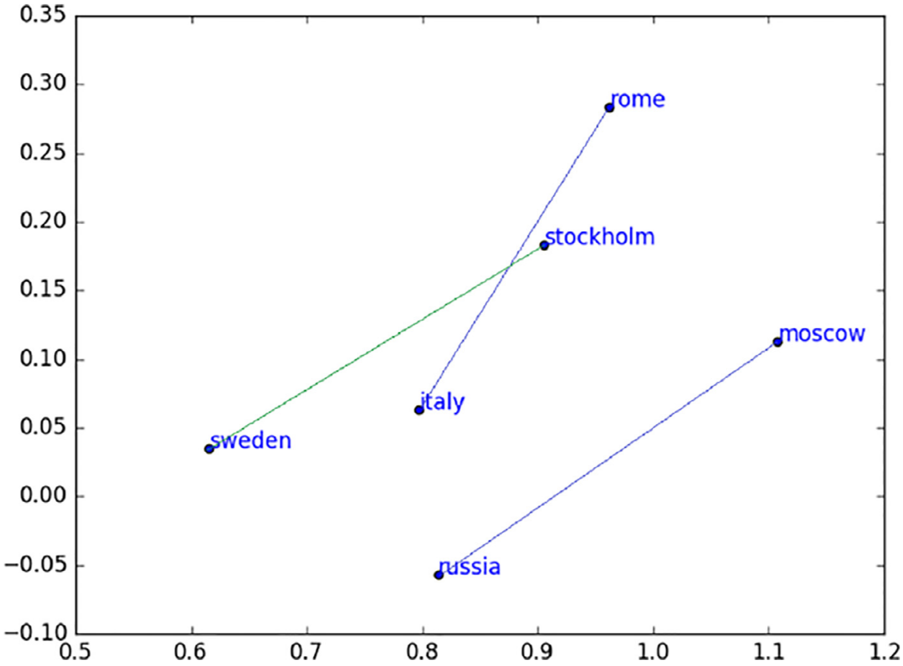

Fig. 11. Visualization of English words about capital and country in monolingual analogy reasoning by SG(E). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

4.3.2. Results of cross-lingual analogy reasoning

Fig. 14 shows the results of the approaches without word alignment, i.e., Shuffle, BilBOWA, Concat(EC), Concat (CE) and Interleaving, and the approaches with word alignment, i.e., BiSkip and BiCIn, for EECC, ECEC, and

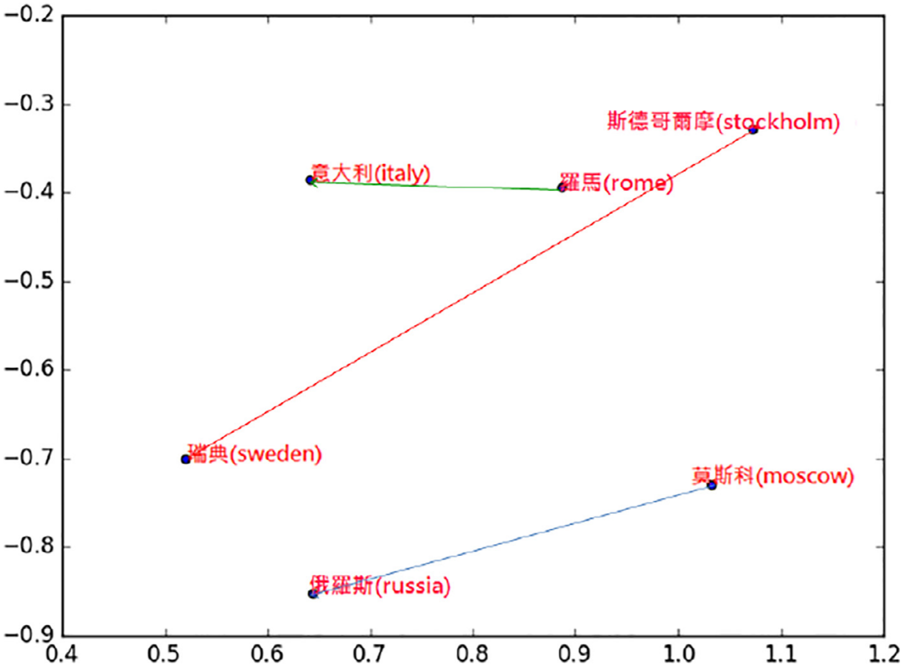

Fig. 12. Visualization of Chinese words about capital and country in monolingual analogy reasoning by SG(C). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

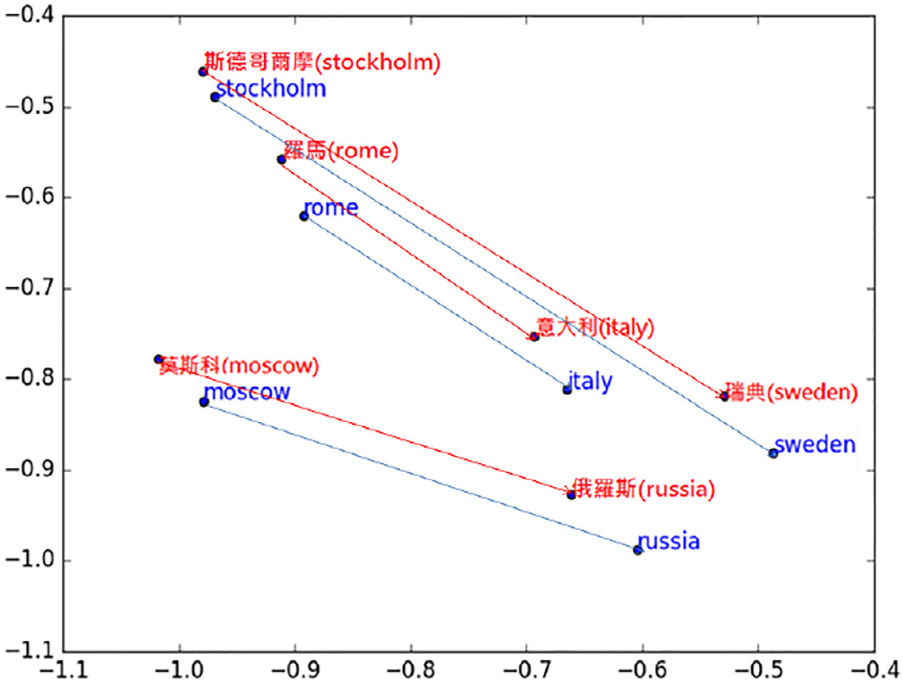

Fig. 13. Visualization of English and Chinese words about capital and country in monolingual analogy reasoning by TA+BiCIn.

EEEC questions, where b* is in Chinese. Fig. 15 shows the results of cross-lingual analogy reasoning for CCEE, CECE, and CCCE questions, where b* is in English. The two metrics acc@1 and acc@5 are shown in the same bar with different colors for the same method and the same question type. Moreover, the exact numbers of the top 3 methods for each question type are also listed for reference.

The best three methods measured by acc@5 are the approaches without word alignment. In particular, the performances of concatenating methods are the best and significant with $p<0.001$ using the McNemar’s test comparing with Shuffle, BilBOWA and BiSkip. The performances of the BiCIn methods are not as good as expected when

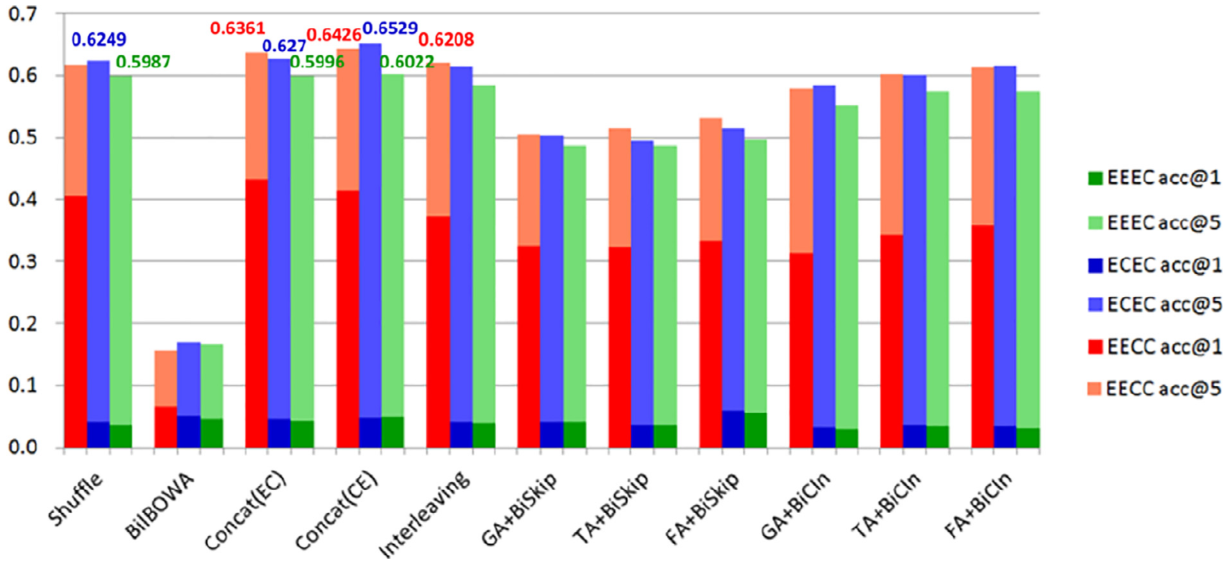

Fig. 14. Results of cross-lingual analogy reasoning (b* is in Chinese).

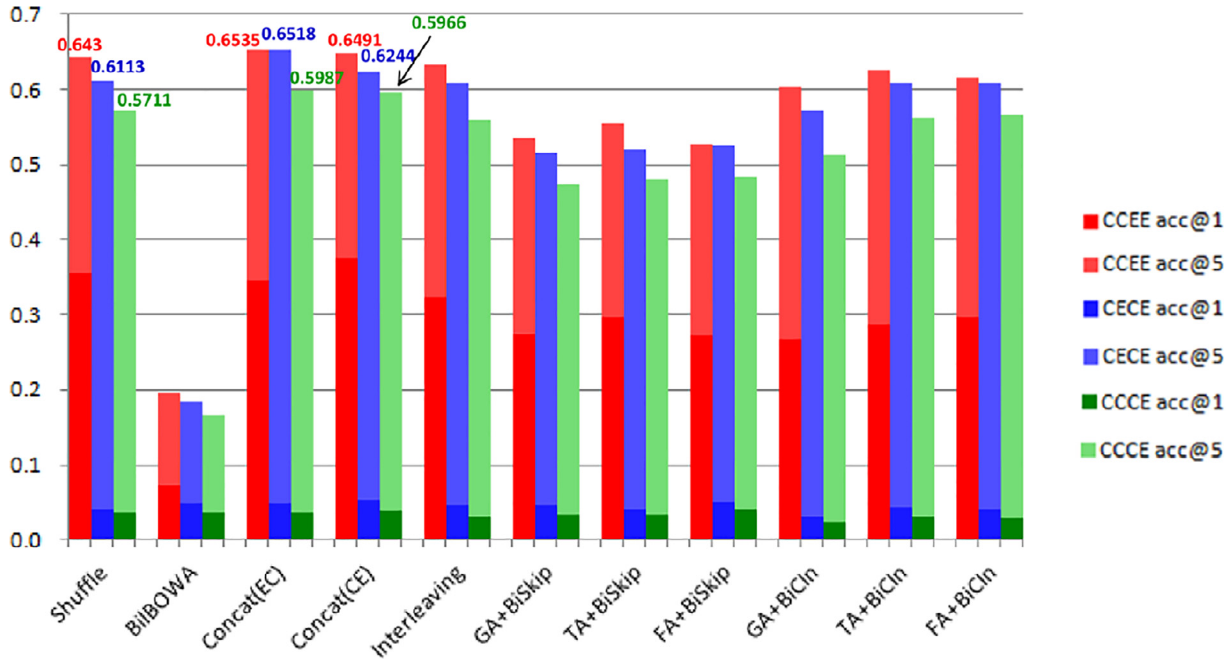

Fig. 15. Results of cross-lingual analogy reasoning (b* is in English). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

measured by acc@1, however, their performances rise greatly when measured by acc@5. This is because BiCIn tends to rank the translation of word *b* at top 1, but it can return the correct answers at top 5. Table 9 shows the TA +BiCIn method returns the wrong answer at top 1 for some questions. Actually, it returns the translation of word *b* in the question at top 1, and returns the correct answer at top 2. Although BilBOWA and BiSkip perform well when measured by acc@1, the number of cross-lingual analogy reasoning questions which can be solved is less than our methods (i.e., Interleaving or BiCIn) from the metric acc@5.

4.4. Cross-lingual word semantic relatedness

The task of word semantic relatedness is to measure the association degree between a pair of words. WordSim353 (Finkelstein et al., 2001) contains 353 English word pairs which are assigned similarity ratings by an average of 13 to 16 human judgments. Spearman’s rank correlation coefficient (Myers et al., 1995) between rankings measured by the cosine similarity of two words’ vectors learned by a model and word similarity ratings assigned by humans is the metric to evaluate the model’s performance.

To develop a dataset for cross-lingual semantic relatedness measurement, we use Google Translate to obtain the Chinese translation of 353 English word pairs and remove duplicate ones. For example, the English word pair is (law, lawyer), and the bilingual word pairs are (law, 律師 (lawyer)) and (法律 (law), lawyer). Finally, we obtain 697 bilingual word pairs and 353 monolingual word pairs. Table 10 shows the Spearman correlation of the methods in the experiments.

Table 9
Some Top-1 answers returned by TA+BiCIn.

Analogy reasoning questions	Top-1 answer
Helsinki: 芬蘭 (Finland) = Athens: 希臘 (Greece)	雅典 (Athens)
Father: 母親 (Mother) = Uncle: 阿姨 (Aunt)	叔叔 (Uncle)
Beijing: China = Hanoi 越南 (Vietnam)	河內 (Hanoi)
Father: Mother = Son: 女兒 (Daughter)	兒子 (Son)

Table 10
Monolingual and bilingual word semantic relatedness experiments.

Alignment level	Method	Bilingual	English	Chinese
–	SG(E)	–	0.6824	–
	SG(C)	–	–	0.6476
Without word alignment	Shuffle	0.6681	0.6781	0.6694
	BilBOWA	0.5389	0.5079	0.5577
	Concat(EC)	0.6637	0.6535	0.6429
	Concat(CE)	0.6613	0.6656	0.6381
	Interleaving	0.6884	0.6898	0.6712
With word alignment	GA+BiSkip	0.6354	0.6469	0.6097
	TA+BiSkip	0.6385	0.6592	0.6270
	FA+BiSkip	0.6524	0.6818	0.6475
	GA+BiCIn	0.7007	**0.7212**	**0.6919**
	TA+BiCIn	**0.7058**	0.7041	0.6863
	FA+BiCIn	0.6853	0.7040	0.6873

In the monolingual word semantic relatedness experiments, Interleaving is the best one in the approaches without word alignment and is also better than the Skip-gram models trained on monolingual corpus, i.e., SG(E) and SG(C). Shuffle also performs better than SG does in the Chinese word semantic relatedness experiment.

In the bilingual word semantic relatedness experiments, Interleaving is the best one in the approaches without word alignment and is also better than BiSkip, which requires word alignment. The reason is that the representations of the unaligned words might be affected by BiSkip. In contrast, the unaligned words in BiCIn are dealt with by the interleaving approach, which achieves a better performance than BiSkip does. BiCIn outperforms BiSkip no matter which word alignment tools are applied. In addition, BiCIn improves the performance of Interleaving in both monolingual and cross-lingual word semantic relatedness. We can conclude that word alignment is required in word semantic relatedness task.

In Table 3, Duong et al. (2016) achieve the 0.762 Spearman's rank correlation coefficient on monolingual word semantic relatedness task. However, the learned embedding is combined dictionary from both Panlex and Wiktionary. Without dictionary information, their performance is 0.686, which is inferior to Interleaving.

4.5. Summary and discussion

Table 11 summarizes the experimental results of the four tasks. We address the major issues denoted by the solid circle (•) behind each task. All these tasks have the translation issue. CLIR and analogy reasoning tasks have one additional issue. Moreover, the solid diamond (◆), the up-pointing triangle (▲), and the down-pointing triangle (▼) denote the best, the next best, and the worst methods, respectively. Overall, BiCIn performs the best in three tasks; Interleaving performs the next best in three tasks; Shuffle performs the next best in two tasks; and BilBOWA performs the worst in all the tasks. The dictionary induction and the semantic relatedness face the same issue, so that the preferred and dispreferred methods are consistent. The analogy reasoning task prefers the approaches without word alignment except BilBOWA.

Table 11
Result summary.

Task	Dictionary induction	CLIR	Analogy reasoning	Semantic relatedness
Translation	●	●	●	●
Reasoning			●	
Access		●		
Shuffle		▲ (C→E)	▲	
BilBOWA	▼	▼	▼	▼
Concat			◆	
Interleaving	▲	▲ (E→C)		▲
BiSkip				
BiCIn	◆	◆		◆

BiCIn is less sensitive to the alignment direction and alignment tools. Comparatively, BiSkip is more sensitive to the alignment errors and unaligned words. In conclusion, we recommend using the BiCIn method if word alignment tool is available. If there are no such tools, the Interleaving method is recommended.

5. Conclusions

In this paper, we investigate the issues of using or not using word alignment in learning bilingual word representations and explore various non-alignment and alignment approaches to generate contexts for training the Skip-gram model. In the approaches without word alignment, we explore concatenating, interleaving, shuffling parallel sentence, and using parallel sentence separately. In the approaches with word alignment, three word alignment tools are applied.

For the case of unaligned words, BiSkip utilizes immediate neighbors' alignments. In contrast, we introduce BiCIn to learn bilingual word representations of the unaligned words by conditional interleaving. We evaluate these approaches in four tasks, including (1) bilingual dictionary induction, (2) cross-lingual information retrieval, (3) cross-lingual analogy reasoning, and (4) cross-lingual word semantic relatedness.

From the experimental results of the four tasks, we observe the word alignment approach with conditional interleaving (BiCIn) performs the best in most of the tasks. In addition, we also propose Interleaving which is a state-of-the-art approach of bilingual word representation without word alignment. It performs well in bilingual dictionary induction task and cross-lingual word semantic relatedness task. In the cross-lingual analogy reasoning task, the approaches without word alignment are better than the approaches with word alignment.

In the future, we will examine the uses of BiCIn and Interleaving in the tasks which concern other issues such as classification. Besides English-Chinese, other language pairs can be explored.

Acknowledgments

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-106-2923-E-002-012-MY3, MOST-107-2634-F-002-011-, MOST-108-2634-F-002-008-, MOST-107-2218-E-009-050-, and AS-TP-107-M05.

References

Bhattacharya, P., Goyal, P., Sarkar, S., 2016. Using Word Embeddings for Query Translation for Hindi to English Cross Language Information Retrieval. *Computación y Sistemas* 20 (3), 435–447.

Boyd-Graber, J., Resnik, P., 2010. Holistic sentiment analysis across languages: multilingual supervised latent dirichlet allocation. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2010). ACL, pp. 45–55.

Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L., 1993. The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. 19 (2), 263–311.

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

Coulmance, J., Marty, J.M., Wenzek, G., Benhalloum, A., 2015. Trans-gram, fast cross-lingual word-embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2015). ACL, pp. 1109–1113.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.P., 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537.

Duong, L., Kanayama, H., Ma, T., Bird, S., Cohn, T., 2016. Learning crosslingual word embeddings without bilingual corpora. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). ACL, pp. 1285–1295.

Dyer, C., Chahuneau, V., Smith, N.A., 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* (NAACL-HLT 2013). ACL, pp. 644–648.

Faruqui, M., Dyer, C., 2014. Improving vector space word representations using multilingual correlation. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (EMNLP 2014). ACL, pp. 462–471.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E., 2001. Placing search in context: the concept revisited. In: *Proceedings of the 10th International Conference on World Wide Web* (WWW 2001). New York, NY, USA. ACM, pp. 406–414. doi: 10.1145/371920.372094.

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

Gouws, S., Bengio, Y., Corrado, G., 2015. BilBOWA: fast bilingual distributed representations without word alignments. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, 37, pp. 748–756.

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

Kim, Y., 2014. Convolutional neural networks for sentence classification. 2014. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2014). ACL, pp. 1746–1751.

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

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

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

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

Levy, O., Goldberg, Y., Ramat-Gan, I., 2014. Linguistic regularities in sparse and explicit word representations. In: Proceedings of the Eighteenth Conference on Computational Language Learning (CoNLL 2014). ACL, pp. 171–180.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning* (ICML-14), pp. 1188–1196.

Liu, Y., Sun, M., 2015. Contrastive unsupervised word alignment with non-local features. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (AAAI 2015). AAAI, pp. 2295–2301.

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

Luong, T., Pham, H., Manning, C.D., 2015. Bilingual word representations with monolingual quality in mind. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2015). ACL, pp. 151–159.

Luong, T., Socher, R., Manning, C.D., 2013. Better word representations with recursive neural networks for morphology. In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning* (CoNLL 2013), pp. 104–113.

Mikolov, T., Chen, K., Corrado, G.S., Dean, J., 2013a. Efficient Estimation of Word Representations in Vector Space. In CoRR, abs/1301.3781.

Mikolov, T., Le, Q.V., Sutskever, I., 2013b. Exploiting Similarities Among Languages for Machine Translation. In CoRR, abs/1309.4168.

Mogadala, A., Rettinger, A., 2016. Bilingual word embeddings from parallel and no-parallel corpora for cross-language text classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (NAACL-HLT 2016). ACL, pp. 692–702.

Myers, J.L., Well, A., Lorch, R.F., 1995. Research Design & Statistical Analysis. Routledge.

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

Pham, H., Luong, T., Manning, C.D., 2015. Learning distributed representations for multilingual text sequences. In: Proceedings of NAACL-HLT 2015, pp. 8–94.

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

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. doi: 10.1038/323533a0.

Sarath Chandar, A.P., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V.C., Saha, A., 2014. An autoencoder approach to learning bilingual word representations. In: *Proceedings of* Advances in Neural Information Processing Systems (NIPS 2014). NIPSF, pp. 1853–1861.

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

Tam, Y.C., Lane, I., Schultz, T., 2007. Bilingual-LSA based LM adaptation for spoken language translation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. ACL, pp. 520–527.

Tian, L., Wong, D.F., Chao, L.S., Quaresma, P., Oliveira, F., Lu, Y., Li, S., Wang, Y., Wang, L., 2014. UM-corpus: a large english-chinese parallel corpus for statistical machine translation. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). ELRA, pp. 1837–1842.

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 (ACL 2016). ACL, pp. 1661–1670.

Vogel, S., Ney, H., Tillmann, C., 1996. Hmm-based word alignment in statistical translation. In: *Proceedings of the 16th Conference on Computational Linguistics (COLI*NG 1996). 2, ICCL, pp. 836–841.

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

Vulić, I., Moens, M.F., 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '15). New York, NY, USA. ACM, pp. 363–372. doi: 10.1145/2766462.2767752.

Wang, M., Che, W., Manning, C.D., 2013. Joint word alignment and bilingual named entity recognition using dual decomposition. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL 2013). ACL, pp. 1073–1082.

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

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

Zhao, B., Xing, E.P., 2006. BiTAM: bilingual topic AdMixture models for word alignment. In: *Proceedings of the COLING/ACL on Main conference poster sessions* (COLING-ACL), pp. 969–976.

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