Graph-based Bilingual Word Embedding for Statistical Machine Translation

RUI WANG, Shanghai Jiao Tong University & National Institute of Information and Communications Technology
HAI ZHAO, Shanghai Jiao Tong University
SABINE PLOUX, National Center for Scienti□c Research
BAO-LIANG LU, Shanghai Jiao Tong University
MASAO UTIYAMA, National Institute of Information and Communications Technology
EIICHIRO SUMITA, National Institute of Information and Communications Technology

Bilingual word embedding has been shown to be helpful for Statistical Machine Translation (SMT). However, most existing methods su□er from two obvious drawbacks. First, they only focus on simple contexts such as an entire document or a □xed sized sliding window to build word embedding and ignore latent useful information from the selected context. Second, the word sense but not the word should be the minimal semantic unit; however, most existing methods still use word representation.

To overcome these drawbacks, this paper presents a novel Graph-based Bilingual Word Embedding (GBWE) method that projects bilingual word senses into a multi-dimensional semantic space. First, a bilingual word co-occurrence graph is constructed using the co-occurrence and pointwise mutual information between the words. Then, maximum complete sub-graphs (cliques), which play the role of a minimal unit for bilingual sense representation, are dynamically extracted according to the contextual information. Consequently, correspondence analysis, principal component analyses, and neural networks are used to summarize the clique-word matrix into lower dimensions to build the embedding model.

Without contextual information, the proposed GBWE can be applied to lexical translation. In addition, given the contextual information, GBWE is able to give a dynamic solution for bilingual word representations, which can be applied to phrase translation and generation. Empirical results show that GBWE can enhance the performance of lexical translation, as well as Chinese/French-to-English and Chinese-to-Japanese phrase-based SMT tasks (IWSLT, NTCIR, NIST, and WAT).

CCS Concepts: • **Computing methodologies** → **Machine translation**; *Natural language processing*; Arti□cial intelligence;

Author's addresses: I Rui Wang (Part of this work was done when the □rst author was in SJTU, while he currently works at NICT), Hai Zhao (corresponding author) and Bao-Liang Lu: Shanghai Jiao Tong University, Shanghai, China; emails : wangrui@nict.go.jp, zhaohai@cs.sjtu.edu.cn, and bllu@sjtu.edu.cn. II Sabine Ploux (corresponding author): National Center for Scienti□c Research, Lyon, France. sploux@isc.cnrs.fr. III Masao Utiyama and Eiichiro Sumita: National Institute of Information and Communications Technology, Kyoto, Japan; emails : mutiyama@nict.go.jp and eiichiro.sumita@nict.go.jp.
Part of this study has been published as "A Bilingual Graph-based Semantic Model for Statistical Machine Translation" [80] in IJCAI-2016. This paper extends the previous bilingual semantic model, which was speci□cally designed for SMT, into a universal bilingual word embedding model. The detailed additional materials include lexical translation tasks, several dimension reduction techniques and additional comparisons with related studies. The codes have been released at https://github.com/wangruinlp/gbwe.

2375-4699/2017/10-ART0 $15.00
https://doi.org/0000001.0000001

Additional Key Words and Phrases: Statistical Machine Translation, Bilingual Word Embedding, Lexical Translation

ACM Reference format:
Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2017. Graph-based Bilingual Word Embedding for Statistical Machine Translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 0, 0, Article 0 (October 2017), 24 pages.
https://doi.org/0000001.0000001

1 INTRODUCTION

Bilingual word embedding can enhance many cross-lingual natural language processing tasks, such as lexical translation, cross-lingual document classi□cation, and Statistical Machine Translation (SMT) [23, 24, 40, 42, 51, 54, 69, 78, 89]. Bilingual word embedding can be considered to be a *cross-lingual* projection [76, 87] of monolingual word embedding [52, 59]. According to the *cross-lingual* projection object, there are three primary types of bilingual embedding methods.

1) Each language is embedded separately, and then the transformation of projecting one embedding onto the other are learned using word translation pairs. Mikolov et al. proposed a linear projection method [51], which was further extended with a normalized objective method [81] and a canonical correlation analysis [19, 46]. Zhang et al. propose a series of methods, such as earth mover and adversarial training, for cross-lingual projection and word embedding transformation by using Non-Parallel Data [86–88].

2) Parallel sentence/document-aligned corpora are used for learning word or phrase representation [18]. Recently, Neural Network (NN)-based projection methods have been widely used for this type of embedding [21, 26, 27, 42, 85]. One typical method is to use the aligned phrase pair from phrase-table to train the neural network translation model and estimate the phrase translation probabilities [21, 67].

3) Monolingual embedding and bilingual projection objectives are optimized jointly [1, 36, 48, 68, 77, 90]. Typically, a large monolingual corpus for monolingual embedding and a small parallel sentence-aligned corpus for bilingual projection are needed [16, 23].

Most of these methods use bag-of-word, n-grams, skip-grams, or other local co-occurrence to exploit monolingual word embedding and then use various cross-lingual projection methods to summarize the bilingual relationship. The question that arises is: Can we construct the cross-lingual relationship and monolingual word embedding together?

It is known that sense gives a more exact meaning formalization than the word itself and a graph can obtain a more global relationship than a contextual relationship. For a better and more exact semantic representation, we propose ***Bilingual Contexonym Cliques (BCCs)***, which are extracted from a bilingual Pointwise Mutual Information (PMI) based word co-occurrence graph. BCCs play the role of a minimal unit for bilingual sense representation. Several dimension reduction methods are used to summarize the BCC-word matrix into lower dimensional vectors for word representation. This study extends a previous monolingual word embedding method [62] that requires bilingual lexicons or synonyms for bilingual mapping[1] and has never been applied to SMT.

The remainder of this paper is organized as follows. We discuss related bilingual word embedding methods in Section 2. The proposed ***Graph-based Bilingual Word Embedding (GBWE)*** method is introduced in Section 3. The GBWE is applied to lexical translation as preliminary experiment in

[1] http://dico.isc.cnrs.fr/en/index.html

Section 4. With contextual information, GBWE can estimate the phrase translation probability (Section 5) and generate additional phrase pairs (Section 6) to enhance the SMT. The SMT experiments and analyses are given in Section 7. We conclude our work in Section 8.

2 RELATED WORK

Word embedding for vector representation is usually built in two-steps [4, 66, 79]. The first step concerns selecting the detailed contexts related to a given word. The second step is to summarize the relationship between the word and its contexts into lower dimensions. For bilingual word embedding, it is also necessary to project from one language space to another.

2.1 Context Selection

For context selection, four categories can be identified.

1) The first category extracts the word or word relationship information from the full text, which is usually regarded as document level processing and includes bag-of-words, Vector Space Models, Latent Semantic Analysis [41], and Latent Dirichlet Allocation [8].

2) The second category uses a sliding window, such as n-grams, skip-grams or other local co-occurrence relationships [44, 50, 52, 59, 90].

3) The third category used sub-word unit such as characters or the dictionary as the "context" information [61, 74, 83]. In addition to these methods, there are several other methods that attempt to solve the memory and space problem in word embedding [45, 63], cross-domain word embedding [82].

4) The fourth category, which has seldom been considered, uses a much more sophisticated graph style context. Ploux and Ji [62] described a graph based semantic matching model using bilingual lexicons and monolingual synonyms[2]. They then represented words using individual monolingual co-occurrences [34]. Saluja et al. [64] proposed a graph based method to generate translation candidates using monolingual co-occurrences. Oshikiri et al. proposed spectral graph based cross-lingual word embeddings [58]. This paper would extend the graph-based monolingual word embedding method [62] to bilingual word embedding with parallel sentences for SMT.

2.2 Relationship Summarizing

For relationship summarizing (dimension reduction), NN have recently become very popular for word/phrase embedding and SMT [14, 17, 21, 21, 29, 42, 50, 52, 71, 73, 85, 90]. In addition, there are also several studies that use matrix factorization [59, 68] for word embedding, such as Singular-Value Decomposition (SVD) [65], Correspondence Analysis (CA) [34, 62], Principal Component Analysis (PCA) [43, 45] and canonical correlation analysis [19, 46].

Most of the above existing methods only apply one dimension reduction method. This paper would introduce PCA, CA and NN methods to summarize the BCC-word relationship.

2.3 Bilingual Projection

One straightforward way is to use some seed translation pairs as gold standard to learn the transformation matrix between two language spaces [51]. Canonical correlation analysis [19, 46] was then be applied to learn this transformation matrix. Several neural network-based methods were proposed to learn bilingual word/phrase embedding using parallel sentences [21, 26, 27, 42, 85].

Recently, unsupervised methods were investigated, Cao et al. proposed a distribution matching method to learn bilingual word embeddings from monolingual data [11] . Artetxe et al. exploited the structural similarity of embedding spaces, and worked with as little bilingual evidence as

[2] http://dico.isc.cnrs.fr

a 25 word dictionary or even an automatically generated list of numerals [2]. Earth mover and adversarial training using Non-Parallel Data were also proposed for cross-lingual projection and word embedding transformation [86–88].

In this paper, we use the bilingual contexonym cliques to represent the bilingual word relationship.

2.4 Multi-sense Representations

It is known that a word may belong to various senses (i.e. be polysemous). There are several studies that focus on sense-speci□c word embedding. Guo et al. [25] propose an NN based recurrent neural network based word embedding method that makes use of previous contextual word information. Jauhar et al. [32] used semantic vector space models for multi-sense representation learning. Šuster et al. [72] learned multi-sense embedding with discrete autoencoders using both monolingual and bilingual information. Iacobacci et al. [30, 31] proposed SensEmbed, which can be applied to both word and relational similarities.

The key to distinguishing word sense is the contextual information, which is applied to nearly all of the above methods. For lexicon translation, we do not make use of contextual information; however, for SMT, we need to consider contextual information. In this paper, we evaluate the proposed method in both lexical translation and SMT.

2.5 Neural Machine Translation

Recently, Neural Machine Translation (NMT) has set new state-of-the-art benchmarks on many translation tasks [3, 14, 33, 49, 75, 84]. In NMT, word embedding, alignment (attention) and translation prediction are jointly learned by a neural network; therefore, independent word embedding is not so necessary in NMT as that in SMT. However, SMT still outperforms NMT in some low-resource language pair and domain-speci□c tasks [39].

3 GRAPH-BASED BILINGUAL WORD EMBEDDING

We □rst illustrate the whole pipeline of the graph-based bilingual word embedding in Fig. 1.

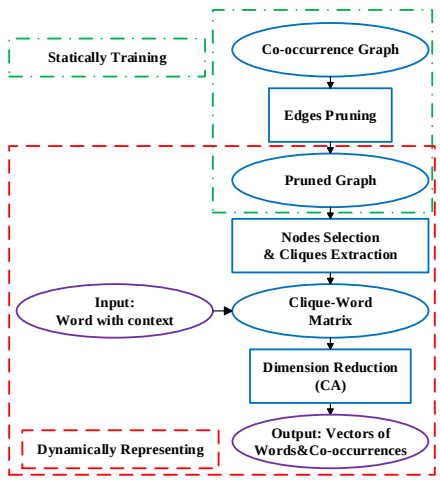


Fig. 1. The whole pipeline of the graph-based bilingual word embedding (CA as dimension reduction for example). The statically training includes Section 3.1, where the bilingual co-occurrence graph is constructed using the entire corpus. The dynamically representing includes Sections 3.2 and 3.3, where the clique-word matrix and dimension reduction are conducted according to the input word and its context.

3.1 Bilingual Co-occurrence Graph Construction

An edge-weighted graph can be derived from a bilingual corpus through formally regarding words by nodes (vertices) and their co-occurrence relationships as edges,

$$G = \{W, E\}, \tag{1}$$

where W is the node set and E is the set of edges weighted by a co-occurrence relationship defined as follows. For a given bilingual parallel corpus, each source sentence $S_F = (w_{f_1}, w_{f_2}, ..., w_{f_k})$ and its corresponding target sentence $S_E = (w_{e_1}, w_{e_2}, ..., w_{e_l})$ are combined to construct a *Bilingual Sentence* $(BS) = (w_{f_1}, w_{f_2}, ..., w_{f_k}, w_{e_1}, w_{e_2}, ..., w_{e_l})$. For words (either source or target words) w_i and w_j, if they are in the same BS, they are called co-occurrences and are marked n_i and n_j on the graph G. **Because the node n_i in G is always referred as word w_i, we will not distinguish between them in this paper.** The *Edge Weight* (EW) connecting nodes n_i and n_j is defined by a modified Pointwise Mutual Information (PMI) measure,

$$EW = \frac{\text{Co}(n_i, n_j)}{\text{fr}(n_i) \times \text{fr}(n_j)}, \tag{2}$$

where $\text{Co}(n_i, n_j)$ is the co-occurrence counting of n_i and n_j and $\text{fr}(n)$ stands for how many times n occurs in the corpus.

For nearly all languages, stop words such as *of, a, the* in English or *de, une, la* in French that have a wide distribution result in most nodes in the graph being unnecessarily connected. Therefore, a filter is set to prune these non-informative connecting edges [9] with an EW less than a threshold[3] of γ, which is tuned using the development data to allow the resulting graph to retain the more useful edges based on the empirical results of a given task.

3.2 Bilingual Contexonym Clique Extraction

In this paper, **A maximum clique defines a maximum complete sub-graph** [47]. If every two nodes in the subset of nodes with edges in the graph are connected to each other by an edge, this subset of nodes forms a clique. Suppose that both N_1 and N_2 are complete graphs in G. If $N_1 \subset N_2$, N_1 cannot be a maximum clique. In the remainder of this paper, the "clique" indicates the maximum clique.

Figure 2 illustrates an example of how to define cliques in an undirected graph. Figure 2 shows that $\{n_1, n_2, n_3, n_4\}$, $\{n_2, n_5\}$, and $\{n_5, n_6, n_7\}$ form three cliques. However, $\{n_1, n_3, n_4\}$ is not a clique, because it is a subset of $\{n_1, n_2, n_3, n_4\}$.

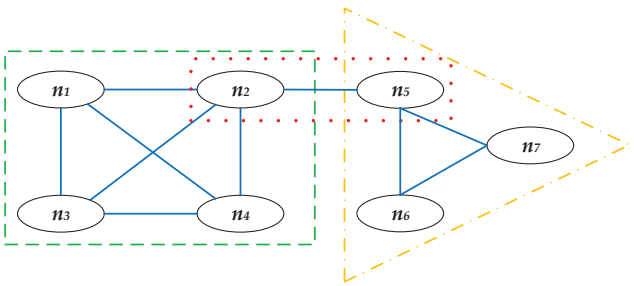


Fig. 2. Three cliques are formed with $\{n_1, n_2, n_3, n_4\}$ (in green), $\{n_2, n_5\}$ (in red) and $\{n_5, n_6, n_7\}$ (in yellow).

[3]The tuning experiments of parameter γ will be further shown in Figure 5 of Section 7.3.1.

Clique extraction is a non-trivial task. There are two reasons why some nodes in the graph may need to be pruned prior to clique extraction. I) Sparsity of the graph. There are nodes that do not connect to any input words (the words to be translated). Take Figure 2 as example, if the input word is n_2, the n_6 and n_7 do not connect to n_2. Therefore, these nodes actually have no direct impact over clique extraction or further word representations. II) Computational complexity. Graph problems are primarily associated with high computational complexity such as finding all cliques in a graph (the *Clique Problem*). The *Clique Problem* has been shown to be *NP-complete* [35]. Without any pruning, it is time consuming or even impossible to find all the cliques in a graph built from a very large corpus (such as millions of sentences).

For an input phrase (a word n and its contextual words $\{n_1, n_2, ..., n_i, ...n_t\}$)[4], only the co-occurrence nodes (co-occurrence words) n_{ij} of each n_i (including n itself) are defined as useful[5]. The set of nodes $\{n_{ij}\}$ with their weighted edges form an extracted graph $G_{\text{extracted}}$ for further clique extraction.

The number of nodes $|N_{\text{extracted}}|$, in the extracted graph $G_{\text{extracted}}$, is computed by

$$|N_{\text{extracted}}| = \left| \bigcup_{\forall i,j} \{n_{ij}\} \right|. \tag{3}$$

Take Figure 2 as example again, if the input word is n_2, the $G_{\text{extracted}}$ is $\{n_1, n_2, n_3, n_4, n_5\}$. In practice, $|N_{\text{extracted}}|$ is much smaller than $|V|$ (the vocabulary size of a bilingual corpus). For a typical corpus (IWSLT in Section 7.1), $|N_{\text{extracted}}|$ is approximately 371.2 on average and $|V|$ is 162.3K. Therefore, clique extraction in practice is quite efficient because it works over a very small graph[6].

Table 1. Several example of BCCS.

Words	BCCs
work_e	{*employees_e*, *travail_f* (work), *unemployed_e*, *work_e* } {*heures_f* (hours), *travaillent_f* (to work, third-person plural form), *travailler_f* (work), *week_e*, *work_e* } {*readers_e*, *work_e* }...
readers_e	{*informations_f* (information), *journaux_f* (newspapers), *online_e*, *readers_e*} {*journaux_f* (newspapers), *lire_f* (read), *newspaper_e*, *presse_f* (press), *readers_e*, *reading_e*} {*readers_e*, *work_e*}...

Note: The suffixes "*_e*" and "*_f*" are used to indicate English or French, respectively. The English words in parentheses are the corresponding translations.

Clique extraction may follow a standard routine [47]. Because the clique in this paper represents the fine-gained bilingual sense of a word given a set of its contextual words, it is called a ***Bilingual Contexonym Clique (BCC)***. Similar but more fine grained than a synset (a small group of synonyms

[4]For SMT tasks, the words in aligned phrases are used as contextual words, please refer to Section 5 for details.

[5]We can also use the immediate co-occurrence nodes as seed words and select more connected nodes (co-occurrence words) with these seed words. However, empirical results show that the computational cost increases exponentially when using these two-step co-occurrence nodes as input and that the performance does not improve. Therefore, we adopted the immediate co-occurrence strategy in this paper.

[6]Please refer to Section 7.3.5 to see efficiency comparison in details.

labeled as a concept) defined in WordNet [53], the BCC plays the role of a minimal unit for bilingual meaning representation. Therefore, *BCC-Word* relationship can obtain a more exact semantic relationship between a word and its senses, compared to using simple bag-of-words or sliding window contexts.

Taking the word *work_e* (The suffixes "_e" and "_f" are used to indicate English or French, respectively) and *readers_e* as an example (without context), some of the BCCs (in alphabetical order) are listed in Table 1. This shows that BCCs can distinguish multiply word meaning. The BCC containing *employees_e*, *travail_f* (work) and *unemployed_e* may indicate the sense of *job*. The BCC containing *readers_e* may indicate the sense of *literature*.

Note that edge pruning is static and that node selection is dynamic depending on the input word sequences. The proposed node selection and clique extraction follows Ploux & Ji's study [62], except that we use a bilingual co-occurrence graph rather than monolingual synonym or hypo(hypero)nym graphs. BCCs can be regarded as loose synsets because only strongly related words can be nodes in a clique that possesses full connections and different senses will naturally result in roughly different cliques from our empirical observations, even though noise or improper connections also exist simultaneously.

3.3 Dimension Reduction & Semantic Spatial Representation

To obtain concise semantic vector representation, three dimension reduction methods are introduced. Both Principal Component Analysis (PCA) [60] and Correspondence Analysis (CA) [28] can summarize a set of possibly correlated variables into a smaller number of variables which is also called principal components in PCA. All these variables are usually in a vector presentation, therefore the processing is performed as a series of matrix transformations. The importance of every output components may be measured by a predefined variable, which is variance in PCA and is called inertia in CA. The most difference between them is that CA treats rows and columns equivalently. For either method, we can select top ranked components according to their importance measure so that dimension reduction can be achieved. Besides PCA and CA, we also apply a neural network-based method to dimension reduction and obtain the principal dimensions.

3.3.1 Principal Component Analysis. In this paper, PCA was conducted over the clique-word matrix constructed from the relationship between the BCCs and the words. An initial correspondence matrix $X = \{x_{ij}\}$ is built, where x_{ij} = 1 if $word_i \in BCC_j$ and 0 if not. Take the example in Table 1 again, part of BCC-word initial matrix is shown in Table 2.

Table 2. A BCC-word initial matrix example.

BCC-Word	$word_1$ (*work_e*)	$word_2$ (*travail_f*)	$word_3$ (*employees_e*)	...
BCC_1	1	0	1	
BCC_2	1	1	0	
BCC_3	1	0	0	
...				

We want to linearly transform this matrix X (whose vectors are normalized to zero mean), into another matrix $Y = PX$, whose covariance matrix C_Y maximizes the diagonal entries and minimizes the off-diagonal entries (the diagonal matrix):

$$C_Y = \frac{(PX)(PX)^T}{n-1} = \frac{P(XX^T)P^T}{n-1} = \frac{PSP^T}{n-1}, \tag{4}$$

where $S = XX^T = EDE^T$. E is an orthonormal matrix whose columns are the orthonormal eigenvectors of S, and D is a diagonal matrix that has the eigenvalues of S as its (diagonal) entries. By choosing the rows of P to be the eigenvectors of S, we ensure that $P = E^T$ and vice-versa. The principal components (the rows of P) are the eigenvectors of S, and in order of *importance*.

3.3.2 Correspondence Analysis. Similar to PCA, CA also determines the first n factors of a system of orthogonal axes that capture the greatest amount of variance in the matrix. CA is primarily applied to categorical rather than continuous data [5, 28]. It assesses the extent of matching between two variables and determines the first n factors of a system of orthogonal axes that capture the greatest amount of variance in the matrix. The first axis (or factor) captures the largest variations, the second axis captures the second largest, and so on. CA has been applied to several related semantic tasks [34, 62].

In this paper, PCA and CA use the same initial BCC-word matrix as original BCC-word matrix X. A normalized correspondence matrix $\mathcal{P} = \{p_{ij}\}$ is directly derived from X, where $p_{ij} = x_{ij}/N_X$, and N_X is $\sum_{i,j} x_{ij}$ (the grand total of all the elements in X). Let the row and column marginal totals of $\mathcal{P}$ be r and c, which are the vectors of the row and column masses, respectively, and let D_r and D_c be the diagonal matrices of the row and column masses, respectively. The coordinates of the row and column profiles with respect to the principal axes are computed by using the Singular Value Decomposition (SVD).

The principal coordinates of rows $\mathcal{F}$ and columns $\mathcal{G}$ are:

$$\mathcal{F} = {D_r}^{-\frac{1}{2}} U\Sigma, \quad \mathcal{G} = {D_c}^{-\frac{1}{2}} V\Sigma, \tag{5}$$

where U, V and Σ (the diagonal matrix of the singular values in descending order) are derived from the matrix of the standardized residuals S and the SVD,

$$S = U\Sigma V^* = {D_r}^{-\frac{1}{2}} (\mathcal{P} - rc^*) {D_c}^{-\frac{1}{2}}, \tag{6}$$

where $*$ denotes the conjugate transpose and $U^*U = V^*V = I$.

According to the above processes, CA projects the BCCs ($\mathcal{F}$) and words ($\mathcal{G}$) into the semantic geometric coordinates as vectors. The inertia χ^2/N_M is used to measure the semantic variations of the principal axes for $\mathcal{F}$ and $\mathcal{G}$:

$$\chi^2/N_M = \sum_i \sum_j \frac{(x_{ij} - r_i c_j)^2}{r_i c_j}. \tag{7}$$

Following the standard setting of CA [5], the top principal dimensions (axes) (please refer to Figure 4 for dimension tuning experiments) of the vectors are chosen for the word and clique representation.

3.3.3 Neural Network (NN). We applied the NN-based Continuous Bag-of-Words (CBOW) Model structure in word2vec [50], and BCC is considered to be bilingual Bag-of-Words. The difference between CBOW and CA/PCA is that the model of CA/PCA is fixed and the NN parameters of CBOW model should be trained before being using. Therefore, we need some samples to train the CBOW model. The original monolingual CBOW use the monolingual words in the n-grams as the input bag-of-words. In comparison, we use the bilingual words in BCCs as the input bag-of-words.

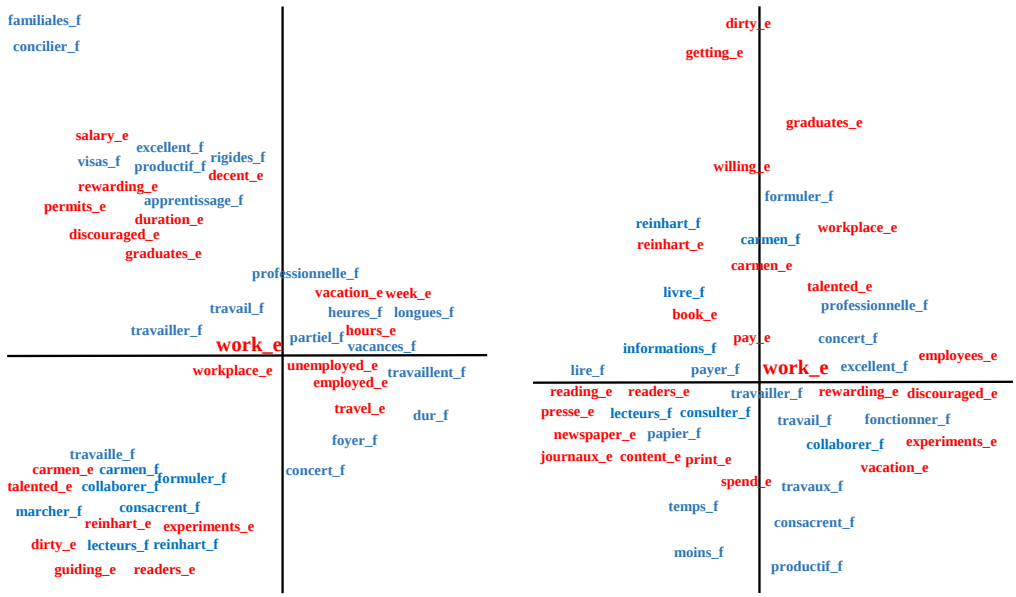
For CA and PCA, only the BCCs related with one word itself or it context are used to construct the BCC-word matrix. However, the NN-based method cannot be directly applied to this size of BCC-word matrix summarization as can PCA/CA, because only using these BCCs is too sparse (there is only several hundreds of BCCs for each word on average) to train a robust CBOW model

for each word. Therefore, we use the BCCs for all of the words to train a single CBOW model for all of the words. That is, all the BCCs of each word in the corpus are pre-computed, and they are then used as a whole as the input of the CBOW Model. The window size of CBOW is set to eight. We discard the BCCs containing more than eight words and set the projections of the missing words to zero for BCCs containing less than eight words.

Graph-based Bilingual Word Embedding (GBWE) is consequently constructed from one of these principal dimensions (PCA, CA, or NN). In short, a word (itself or with its context) is used as input for the GBWE and vectors of the word and its (it self's or together with its context's) bilingual co-occurrence words are output.

3.4 Visualization

To visualize the results, the top two dimensions were chosen and illustrated via a spatial map (CA is adopted for example). We only present a few typical words due to the limited space.



work_e (without context) as input. Both *work_e* and its context *readers_e* as input.

Fig. 3. Word representation illustration (CA is adopted for example). The English words are in red and the French words are in blue.

Figure 3 illustrates the spatial representation of all the co-occurrence words when we input *work_e* without context and with *readers_e* as context using GBWE (CA), respectively. We can obtain the following observations from these graphs:

1) There are several good translation pairs, where the French and English words are close in distance. For example, (*work_e, travailler_f*) and (*readers_e*, *lecteurs_f*) in both Figures 3 (left) and 3 (right), (*hours_e*, *heures_f*) in Figure 3 (left), and (*reading_e, lire_f*) and (*book_e, livre_f*) in Figure 3 (right).

2) For *work_e* as input in Figure 3 (left), the word *work_e* itself is placed at the center and the other words around it are about *employment* (such as *hours_e* at right), *evaluation of job* (such as

salary_e at upper left), and *publication* (such as *experiments_e* at bottom). The senses of the words around *work_e* are quite different and it's hard to determine which sense *work_e* belongs to.

3) For *work_e*+*readers_e* as input in Figure 3 (right), we can determine the sense of *work_e* by the words close to both *work_e* and *readers_e*, such as *book_e*, *print_e* and *paper_f*.

4 PRELIMINARY EXPERIMENT–LEXICAL TRANSLATION

GBWE was firstly evaluated on lexical translation task as a preliminary experiment. Lexical translation can be viewed as one-word phrase translation case, where contextual information is not necessary.

Following the previous lexical translation settings of Mikolov et al. [51], the 6000 most frequent words from the WMT11 Spanish-English (Sp-En) data[7] were translated into the target languages using online Google Translation (individually for English and Spanish). Because the Mikolov method requires translation-pairs for training, they used the first 5000 most frequent words to learn the "*translation matrix*" and the remaining 1000 words were used as a test set. The proposed method only uses parallel sentences for training; therefore, we used the first 5000 most frequent words for dimension tuning and remaining 1000 test-pairs for evaluation. To translate a source word, we find its k nearest target words using the Euclidean distance and then evaluate the translation precision $P@k$ as the fraction of the target translations that are within the top-k returned words. We also evaluated these methods on the IWSLT-2014 French-English (Fr-En) task[8] with the same settings as the WMT11 task.

Three methods reported in Mikolov et al. [51] were used as baselines, the *Edit Distance*, *Word Co-occurrence*, and *Translation Matrix* methods, together with two state-of-the-art bilingual word embeddings: BilBOWA [23] and Oshikiri et al.'s method[58][9]. Their default settings were followed.

4.1 Dimension Tuning

Figure 4 shows the dimension tuning (using the average score of four sub-tasks) experiments on the development data of the IWSLT-2014 task. Because the Mikolov method requires this data for transformation matrix training, their default dimension of 300 was applied. As mentioned in Section 3.3.3, the BCC-word matrix of the PCA/CA-based method is an extracted graph; therefore, the original dimension is much smaller than that of the NN based method, where nearly all the BCCs in the entire graph are used as input for the NN models. The best performing dimension on the development data was evaluated on the test data.

4.2 Evaluation Results

Similar to Mikolov method [51], we also discarded word pairs whose Google translations were out-of-vocabulary. The evaluation results on the test data are shown in Table 3. For WMT11, the baseline results are from the reports of the corresponding papers. For IWSLT14, the baseline results are from our re-implementations of the corresponding methods, because neither BilBOWA nor Mikolov [51] implemented their method in IWSLT14 task. The results in bold indicates the best outperforming system for each task. The numbers in parentheses show how much the best results outperformed the best baseline results.

As shown in Table 3:

1) The GBWE-based methods achieved the best performances in seven out of eight sub-tasks.

[7] http://www.statmt.org/wmt11/

[8] https://wit3.fbk.eu

[9] The settings of their method in the WMT task are different from ours; therefore we only compared their performance in the IWSLT task.

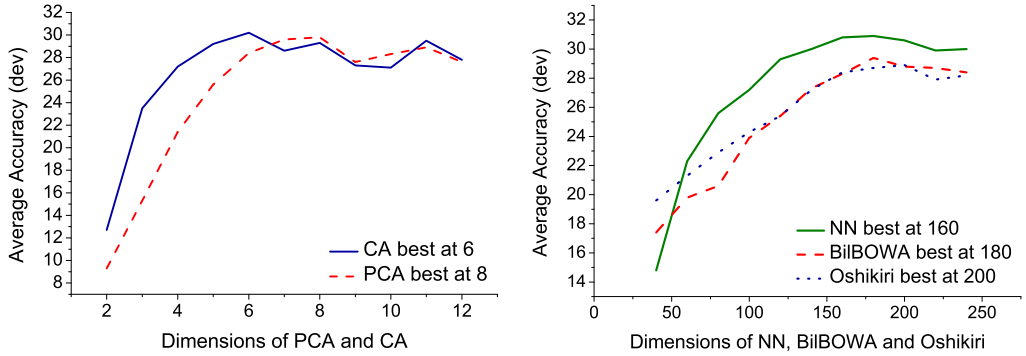


Fig. 4. Dimension tuning for the IWSLT lexical translation task

Table 3. Lexicon Translation Evaluation Results.

WMT11	En-Sp P@1	Sp-En P@1	En-Sp P@5	Sp-En P@5	Time /hours
Edit Distance [51]	13	18	24	27	-
Word Co-occurrence [51]	30	19	20	30	-
Translation Matrix [51]	33	35	51	52	-
BilBOWA [23]	39	**44**	51	55	-
GBWE-PCA	34	40	55	53	2.7
GBWE-CA	36	42	**56(+5)**	**60(+5)**	3.2
GBWE-NN	**41(+2)**	37	53	54	28.7
IWSLT14	En-Fr P@1	Fr-En P@1	En-Fr P@5	Fr-En P@5	Time /hours
Translation Matrix [51]	23	27	37	32	2.1
BilBOWA [23]	26	25	38	31	1.4
Oshikiri [58]	24	26	32	33	1.6
GBWE-PCA	18	26	38	**41(+8)**	**0.9**
GBWE-CA	22	**29(+2)**	**41(+3)**	36	1.1
GBWE-NN	**27(+1)**	28	36	38	12.3

Note: For WMT11, the baseline results are from the reports of the corresponding papers. For IWSLT14, the baseline results are from our re-implementations of the corresponding methods, because neither BilBOWA nor Mikolov [51] implemented their method in IWSLT14 task.

2) The GBWE-CA achieved the best performances among the three dimension reduction methods in four out of eight sub-tasks.

3) The model training and calculating CPU time of GBWE-PCA/CA are slightly better than those of existing methods. GBWE-NN is more time consuming than GBWE-PCA/CA.

Since the proposed GBWE methods works in the preliminary lexical translation task, we would try to phrase translation task, where the contextual information is necessary.

5 PHRASE TRANSLATION PROBABILITY ESTIMATION

Contextual information is important and should be considered for phrase-based translation task. GBWE represents words as vectors dynamically in various geometric space according to the contextual words. For each word in a source phrase of the SMT, its contextual words are fixed, so that all the translation candidate target words can be represented as vectors in the same geometric space. This makes GBWE capable of selecting translated phrase candidates in phrase-based SMT.

5.1 Bilingual Phrase Semantic Representation

The phrase-table of phrase-based SMT can be simply formalized as[10]

$$(P_F,\ P_E,\ \text{scores},\ \text{word-alignment}), \tag{8}$$

where $P_F(w_{f_1}, w_{f_2}, ...w_{f_i}, ..., w_{f_k})$ and $P_E(w_{e_1}, w_{e_2}, ..., w_{e_j}, ..., w_{e_l})$ are the source and its aligned target phrase, respectively, and scores indicates the various feature scores, including directed translation probability, lexical weights, and the phrase penalty. The phrase length is limited to seven, which is the default setting for the phrase-based SMT. The word-alignment indicates the word alignment information between $w_{f_i} \in P_F$ and $w_{e_j} \in P_E$.

GBWE (CA is adopted here) is applied to represent words in the phrase-table as vectors. Note that the clique extraction depends on the contextual words and that CA then projects a clique-word matrix into the corresponding semantic geometric space accordingly. Therefore, the same contextual words should be used for all the words in P_F, and all its aligned P_E, to represent them in the same geometric space. For each word w_{fi} (or w_{ei}) in a phrase pair (P_F, P_E), we consider two strategies for selecting the context words,

Strategy-A: Only the source words in P_F are used as the contextual words $\{w_{f_1}, w_{f_2}, ..., w_{f_k}\}$.

Strategy-B: Both the source words in P_F and the target words in all the aligned P_{E_β} are used as its the contextual words $\{w_{f_1}, w_{f_2}, ..., w_{f_k}, w_{e_1}, w_{e_2}, ..., w_{e_l}\}$.

Word w_{fi} (or w_{ei}) is represented[11] as a vector V_{wf_i} (or V_{we_j}). The co-occurrence word w_{co} can also be represented as a vector $V_{w_{co}}$, which is described in Section 3.3. Note that all the source and target words for the same source phrase P_F are represented as vectors in the same geometric space.

5.2 Semantic Similarity Measurement

Because the lengths of the phrases are different, the *Phrase Distance* (PD) is adopted to measure the distance between the source and the target phrases incorporated by the word-alignment model:

$$\text{PD}(P_F, P_E) = \sqrt{\frac{\sum_{\text{align}(w_{f_i}, w_{e_j})} \text{ED}^2(V_{wf_i}, V_{we_j})}{\sum_{i,j} |\ \text{align}(w_{f_i}, w_{e_j})|}}, \tag{9}$$

where $\text{ED}(V_{wf_i}, V_{we_j})$ stands for the *Euclidean Distance* between the word vectors V_{wf_i} and V_{we_j}, $align(w_{f_i}, w_{e_j})$ is from the word-alignment model in Eq. (8), and $|\sum_{i,j} \text{align}(w_{f_i}, w_{e_j})|$ is the sum of alignments between w_{f_i} and w_{e_j}.

Because there are usually multiple P_E (P_{E_m}) that are aligned with P_F, the distance is normalized to ensure that the summary of $\text{PD}(P_F, P_{E_m})$ for each P_F is equal to one. Therefore the *Normalized Phrase Distance* (NPD) from P_F to P_{E_m} is adopted:

$$\text{NPD}(P_{E_m}|P_F) = \frac{\text{PD}(P_F, P_{E_m})}{\sum_m \text{PD}(P_F, P_{E_m})}. \tag{10}$$

[10]The alignment is from the standard IBM alignment model [6, 10].

[11]The *unknown* words are empirically represented as default vectors.

Using the same pipeline, NPD($P_F|P_E$) can also be calculated. Both NPD($P_E|P_F$) and NPD($P_F|P_E$) can be additional phrase-table features for phrase-based SMT decoding.

6 BILINGUAL PHRASE GENERATION (BPG)

A few phrases that are not in the corpus may share a similar meaning as those inside the corpus. Takes the source French phrase *la bonne réponse* as an example; the corresponding aligned target English phrase *the right answer* is in the corpus and the phrase-table. The other phrases, such as *the correct answer* or *the right response*, may not be in the corpus or phrase-table; however, they are also good translation candidates.

Because the GBWE can be used to represent words as vectors and to measure their similarities by measuring their distance, it is possible to find similar words to replace the original words in the phrase-table to generate a new phrase with a similar meaning as the original one.

6.1 Phrase Pair Generation

Section 3 focused on phrase pairs inside the phrase table; however, in this section, we focus on measuring the similarity between phrase pairs outside the phrase-table and selecting new phrase pairs to enhance the SMT.

As mentioned in Section 5, for each word w (source or target), both the source words in P_F and the target words in P_E are used as its contextual words (Strategy-B). The word w and its co-occurrence words $\{w_{co}\}$ are represented as vectors.

For an aligned word pair (w_{f_i}, w_{e_j}), we find a new translation replacement w'_{e_j} in $\{w_{co}\}$ to help generate new phrases. For either the source phrase P_F or the target phrase P_E, each word in the phrase is tentatively replaced by the nearest word in its corresponding co-occurrence according to the word vector distance (here, the Euclidean distance is adopted). However, only one word replacement with the minimal distance for either phrase will be chosen and implemented to generate two new phrases P'_E and P'_F, respectively[12].

NPD($P'_E|P_F$) and NPD($P'_F|P_E$) can be calculated using Eqs. (9) and (10). Because there are no original phrase translation probabilities $\psi(P'_E|P_F)$ or $\psi(P'_F|P_E)$ for the generated (P_F, P'_E) and (P_E, P'_F) in the original phrase-table, NPD($P'_E|P_F$) and NPD($P'_F|P_E$) are used as $\psi(P'_E|P_F)$ or $\psi(P'_F|P_E)$ instead. The updated lexical weighting lex($P'_E|P_F$) and inverse lexical weighting lex($P'_F|P_E$) are computed using IBM models [6].

The generated phrases are *filled-up* [7] into the original phrase-table. That is, a penalty score is added as a feature. For original phrase pairs, the penalty is set to one; for the generated pairs, the penalty is set to the *natural logarithm base* e (= 2.71828...). All the score weights in phrase-table are further tuned using MERT [56].

6.2 Phrase-table Size Tuning

Using the phrase generation approach, numerous new phrase pairs can be generated. We need to select the most reasonable of these. The *Distance Ratio* (DR) of the normalized distance in Eq. (10) between the generated phrase pair (P_F, P'_E) and the original phrase pair (P_F, P_E),

$$\mathrm{DR}(P'_E, P_E) = \frac{\mathrm{NPD}(P_F, P'_E)}{\mathrm{NPD}(P_F, P_E)}, \tag{11}$$

is used to measure the usefulness of the generated word pairs.

[12]Two or more words can be replaced; However, this may lead to serious sense bias, and the experiments also show that replacing more than two words does not result in performance.

A threshold ε is set to retain only the most useful generated phrase pairs. Namely, for a source phrase P_F, only the P'_E whose DR(P'_E, P_E) is smaller than ε are selected as the generated word pair (P_F, P'_E). Using the same pipeline, the sizes of the generated source candidate phrases are also tuned. For the SMT task, the threshold is tuned using the SMT performance of the developmental data.

7 EXPERIMENTS

7.1 Set up

We evaluated the performance of GBWE in SMT using corpora with various language pairs, domains and sizes (from 186.8K to 2.4M sentences): 1) IWSLT-2014 French-to-English (EN) [13], with dev2010 and test2010/2011 as development (dev) and test data, respectively; 2) NTCIR-9 Chinese-to-English [22] and 3) NIST OpenMT08[13] Chinese to English, with NIST Eval 2006 and NIST Eval 2008 as development data and test data, respectively. The corpora statistics are shown in Table 4. **For the GBWE method, we only report the CA-based dimension reduction method, which performed the best in lexical translation task.**

Table 4. Sentences statistics on parallel corpora.

Corpus	IWSLT	NCTIR	NIST
training	186.8K	1.0M	2.4M
dev	0.9K	2.0K	1.6K
test	1.6K	2.0K	1.3K

7.2 Baseline Systems

The same basic settings for the IWSLT-2014, NTCIR-9, and NIST08 translation baseline systems were followed. The standard Moses phrase-based SMT system was applied [38] together with GIZA++ [57] for alignment, SRILM [70] for language modeling, and MERT for tuning. The tool, `mteval-v13a.pl`[14], was used to calculate the BLEU scores (we ran MERT three times and recorded the average BLEU score). The paired bootstrap resampling test[37][15] was then performed. Signi□-cance tests were done for each round of the test. The marks to the right of the BLEU scores indicate whether our proposed methods are signi□cantly better/worse than the corresponding baseline ("++/−−": signi□cantly better/worse at a *signi cance level* of $\alpha = 0.01$; "+/−": $\alpha = 0.05$). In addition, we evaluated the results using multeval[16] [15] and show them in Table 6. All the experiments in this paper were conducted on the same machine with 2.70GHz CPUs. **Note that all the bilingual embedding models were trained using the same corpus as the SMT systems.**

We are aware that there are several state-of-the-art end-to-end neural machine translation methods [3, 14, 33]. However, the proposed GBWE is a bilingual word embedding method and is applied to SMT as additional features; therefore, we only compare it with the most related bilingual word embedding and generation methods for SMT. For the phrase pair translation probability estimation task, three typical NN based bilingual embedding methods, the Continuous Space

[13] Zou et al. [90] (a typical bilingual embedding method for comparison) only released their word vectors rather than their code (http://ai.stanford.edu/~wzou/mt/); therefore, we have to conduct experiments on NIST08 Chinese-to-English translation task as they did for fair comparison. The training data consists of part of NIST OpenMT06, the United Nations Parallel Text (1993-2007), and the other corpora [12, 20] that were used by [90].

[14] Available at http://www.itl.nist.gov/iad/mig/tests/mt/2009/

[15] The implementation of our system follows http://www.ark.cs.cmu.edu/MT

[16] https://github.com/jhclark/multeval

Translation Model (CSTM) [67], BilBOWA [23], and Zou et al. [90]'s bilingual word embedding method, were selected as baselines. The embedding of each method was added as features to the phrase-based SMT baseline, with all the other setting the same. For the bilingual phrase generation methods, the CSTM was used to generate phrase pairs[17]. In addition, we compared our method with Saluja et al. [64], which also used a graph-based method to generate translation candidates[18].

7.3 Results and Analysis

7.3.1 Filter Tuning. A series of parameter-tuning experiments were conducted on the development dataset to select the proper threshold γ for the edge weight EW in Eq. (2) for the SMT. First, we roughly determined the right order of magnitude (such as 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}...), and then we finely tuned the values. Figure 5 illustrates filter tuning on the IWSLT-2014 corpus.

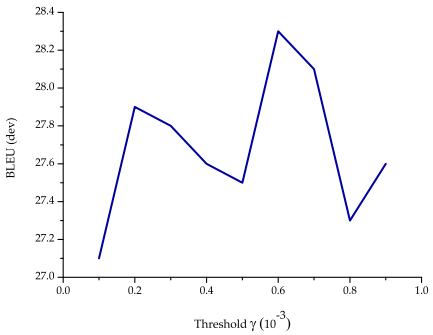


Fig. 5. Filter tuning on the development data in the IWSLT-2014 corpus.

7.3.2 Phrase Pair Translation Probability Estimation. Table 5 shows the performance of GBWE for the SMT. Zou *et al.* released their word vectors on NIST08 but not their codes; therefore, their method is only applied to NIST08 (Please refer to Footnote 13 for details).

Table 5. The phrase Pair Translation Probability Estimation Results (BLEU).

	IWSLT	NTCIR	NIST
Baseline	31.80	32.19	30.12
+Zou [90]	N / A	N / A	30.36
+BilBOWA [23]	31.42	31.87	29.63
+CSTM [67]	32.19	32.37	30.25
+GBWE-A	32.32+	32.56	30.38
+GBWE-B	**32.61++**	**33.04++**	**30.44+**

Note: In Tables 5 and 6, the marks to the right of the BLEU scores indicate whether our proposed methods are significantly better/worse than the corresponding baseline ("++/−−": significantly better/worse at a *significance level* of $\alpha = 0.01$; "+/−": $\alpha = 0.05$).

[17] They discuss and show phrase generation examples as experimental evidence in their paper, and we followed their basic idea.

[18] Their basic implements were followed except for the morphological generation.

As shown in Table 5:

1) GBWE can improve SMT performance up to +0.85 BLEU, which is better than the best performance of existing NN methods of up to +0.67 BLEU.

2) Strategy-B (described in Section 5.1) performs better than Strategy-A. The reason for this may be that both target and source contextual information is used for Strategy-B while only source contextual information is used for Strategy-A.

7.3.3 Bilingual Phrase Pair Generation. Figure 6 shows the generated phrase size tuning on the development data for the IWSLT/NIST corpus. The best performing systems on the development data would be evaluated on the test data. All of the BLEU were computed by mteval-v13a.pl. multeval is only used for measuring variances of test data selection (s_sel), optimizer instability (s_opt), and the p-value.

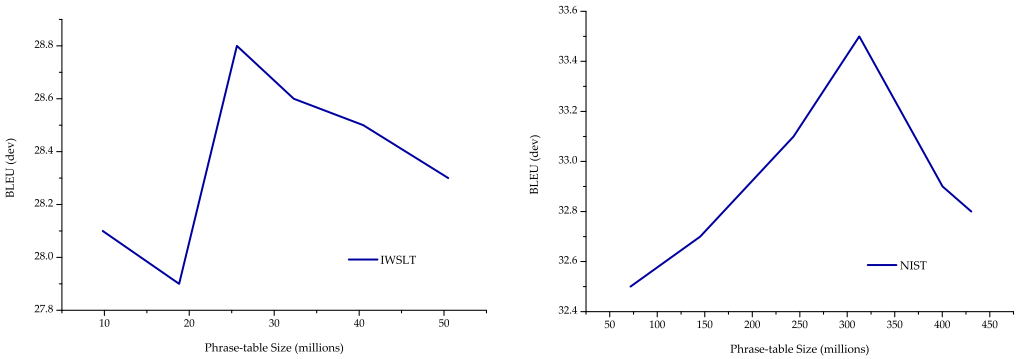


Fig. 6. Generated phrase size tuning on development data in the IWSLT/NIST corpus.

Table 6 shows the result of the Bilingual Phrase Pair Generation (BPG). “Baseline + BPG” indicates the addition of the generated phrase pairs to the original phrase-table. “GBWE + BPG” indicates the addition of the generated phrase pairs to the original phrase-table, as well as replacing the translation probabilities $\psi(P_E|P_F)$ and $\psi(P_F|P_E)$ in the original phrae-table with the NPD$(P_E|P_F)$ and NPD$(P_F|P_E)$ calculated by GBWE (see Section 6.1). For the baselines, we also compared with CSTM and Saluja et al. method [64]. As in Table 6 shown:

1) The proposed BPG method can slightly improved SMT performances, up to +0.57 BLEU.

2) The proposed BPG method and the GBWE method can work well together and enhance the SMT performance signi□cantly, up to +1.33 BLEU. They also outperform the best performing existing methods, which can enhance the SMT performance up to +0.79 BLEU. This indicates that the proposed BPG method worked synergistically with the GBWE translation probability estimation method.

7.3.4 Translation Examples. For the GBWE (CA) translation examples, we showed an translation example of in NIST Chinese-to-English task (to show the e□cient of proposed method in Asian languages) in Table 7.

The Chinese word “duoshao” originally has two primary meanings, one as how many (in most of the cases) and the other as “somewhat/rather”. As shown in Table 7, the baseline and Zou [90] did not fully consider the contextual information and translated it into “how many” or “many”. In comparison, the proposed GBWE method considered the context “gandao yihan” and translated into “somewhat”.

Table 6. The Bilingual Phrase Generation (BPG) Results.

Corpora	Methods	Phrase-table Size	BLEU	s_sel	s_opt	p-value
IWSLT	Baseline	9.8M	31.80	0.5	0.2	-
	+CSTM [67]	23.1M	32.19	0.5	0.2	0.2368
	+Saluja [64]	31.5M	32.35	0.5	0.2	0.0573
	+BPG	25.6M	32.37+	0.5	0.9	0.0389
	+BPG+GBWE-B	25.6M	**33.13++**	0.5	0.2	0.0001
NTCIR	Baseline	71.8M	32.19	0.3	0.1	-
	+CSTM [67]	297.8M	32.42	0.3	0.1	0.3024
	+Saluja [64]	341.3M	32.68	0.3	0.2	0.2231
	+BPG	312.6M	32.54	0.3	0.2	0.2617
	+BPG+GBWE-B	312.6M	**33.47++**	0.3	0.1	0.0001

Note: Phrase-table size indicates the number of phrase pairs in the phrase-table. "s_sel" indicates the variance due to the test set selection, which was calculated using bootstrap resampling for each optimizer run; this number reports the average variance over all the optimizer runs. "s_opt" indicates the variance due to the optimizer instability, which was calculated directly as the variance of the aggregate metric score over all the optimizer runs. "p-value" is the p-value calculated by approximate randomization [15].

Table 7. Translation examples of NIST Chinese-to-English task.

Methods	Translation
Source sentence	dan cong bisai jieguo laishuo , 2 bi 2 de bifen shi heli de , ye shi zhongguo nenggou jieshou de . guanjian shi bisai guocheng , duoshao lingren gandao yihan .
Reference	judging simply from the result of the match , 2-2 seems a reasonable score , and is also one that the chinese team can accept. the key problem is that the way the match went made people feel it was rather a pity .
Baseline	judging by the results of 2 2 , is reasonable , and is also the chinese team to accept . the key is in the process of competition , how many people feel regret .
+Zou [90]	from the results of the competition , 2 to 2 are reasonable , and it is also acceptable to china . the key is competition process , many regrettable .
+GBWE-B	simply judging by the match result , 2 2 score is reasonable and also acceptable to the chinese team. the key is that the match itself left people feeling somewhat disappointed .

Note: The words in red is the corresponding translation of the source word "duoshao".

In addition, we show some examples of phrase pairs generated using CSTM and GBWE of IWSLT French-to English task in Table 8. The NN-based CSTM tends to replace the articles, such as *the, a, an*. For GBWE, the stop words, such as *the, a, an*, are pruned prior to clique extraction (please refer to Section 3.1); therefore, more reasonable translation candidates are generated. These generated translation candidates enhanced translation diversity and thus help SMT performance.

Table 8. Examples of generated phrases of IWSLT French-to English task.

Source	Original Target	CSTM Generated	GBWE Generated
la bonne réponse	*the right answer*	*1. a right answer* *2. all right answer* *3. the right reply*	*1. the correct answer* *2. the right response* *3. the good answer*
nettoyer le jardin	*clean the garden*	*1. clean a garden* *2. clean the yard* *3. clean an garden*	*1. clean the yard* *2. clean the ground* *3. tidy the garden*

Note: We only show short target-side phrases for simpli□cation. The phrases are ranked by their Distance Ratio in Eq. (11). Some generated phrases overlap with existing phrases in the original phrase-table. The probabilities of the overlapping generated phrases are interpolated from the existing ones.

7.3.5 Efficiency Comparison. We compared the e□ciencies for model training and computed the probability scores of the phrases pairs using CSTM and GBWE. A total of 2000 phrase pairs were randomly selected from the entire IWSLT-2014 FR-EN corpus. The CPU time for the training models (the whole corpus) and calculating their probability scores (the 2000 phrase pairs) are shown in Table 9.

Table 9. CPU time for IWSLT-2014.

Methods	Training Time	Calculating Time
CSTM	59.5 Hours	17.1 Minutes
GBWE-A	1.1 Hours	8.9 Minutes
GBWE-B	1.1 Hours	15.6 Minutes

The results in Table 9 demonstrate that GBWE is much more e□cient than CSTM; especially for training GBWE can be more than 50 times as fast as CSTM. Because the translation probabilities are all pre-computed, the decoding time for each method is nearly the same.

7.3.6 WAT Chinese-to-Japanese Task. To verify the e□ectiveness of the proposed approaches on Asian languages, we also evaluated GBWE (CA) on the Chinese-to-Japanese Asian Scienti□c Paper Excerpt Corpus (ASPEC)[19] in the 4th Workshop on Asian Translation (WAT2017) [55]. The corpus statistic is showed in Table 10.

Table 10. Corpus statistic of WAT Chinese-to-Japanese task.

Corpora	Sentences	Tokens
Training	672.3K	22.2M
Development	2.1K	70.0K
Testing	2.1K	69.0K

The empirical results are shown in Table 11.

[19] http://lotus.kuee.kyoto-u.ac.jp/ASPEC/

Table 11. Results on the WAT Chinese-to-Japanese task.

Methods	BLEU
Baseline	33.13
+ GBWE-B	33.40
+ BPG	33.90+
+ BPG + GBWE-B	**34.30**++

As shown in Table 11. The empirical results indicate that the proposed GBWE also worked well on the WAT Chinese-to-Japanese task and signi□cantly improved the phrase-based SMT baseline.

8 CONCLUSION

In this paper, we proposed a novel cross-lingual sense unit BCC using a graph-based method. BCC can describe word senses better compared to simple bag-of-words or sliding window methods. A context-based dynamic bilingual BCC-word matrix was constructed, and then CA, PCA and NN were applied to summarize this matrix into lower dimensions. The GBWE was accordingly constructed for dynamical bilingual word embedding.

The usefulness of the proposed model was veri□ed via three bilingual processing tasks. 1) Lexical translation. The empirical results indicate that GBWE can predict several relevant translation candidates and enhance the lexical translation accuracy. 2) Phrase translation. We propose two strategies to select the useful contextual information for GBWE. The experimental results show that GBWE based features can improve the phrase-based SMT performance with high computational e□ciency. 3) Bilingual phrase generation. The experimental results show that the phrase pairs generated by GBWE can further improve the SMT performance and work well with GBWE-based features.

ACKNOWLEDGMENTS

Thanks for the helpful comments from the editors and reviewers. Rui Wang, Hai Zhao, and Bao-Liang Lu were partially supported by the National Key Research and Development Program of China (Grant No. 2017YFB1002501), National Key Research and Development Program of China (No. 2017YFB0304100), Cai Yuanpei Program (CSC No. 201304490199 and No. 201304490171) between China and France, Key Project of National Society Science Foundation of China (No. 15-ZDA041), National Natural Science Foundation of China (No. 61170114, No. 61673266, No. 61672343, and No. 61733011), Major Basic Research Program of Shanghai Science and Technology Committee (No. 15JC1400103), and The Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04). Rui Wang, Masao Utiyama, and Eiichiro Sumita are partially supported by the program “Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology” of MIC, Japan.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 2289–2294.

[2] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 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)*. Vancouver, Canada, 451–462.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.
[4] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, 238–247.
[5] Jean-Paul Benzécri. 1973. L'Analyse des Correspondances. In *L'Analyse des Données*, Vol. II. Dunod Paris.
[6] Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. La□erty, Robert L. Mercer, Harry Printz, and Luboš Ureš. 1994. The Candide System for Machine Translation. In *Proceedings of the Workshop on Human Language Technology (HLT '94)*. Stroudsburg, PA, USA, 157–162.
[7] Arianna Bisazza, Nick Ruiz, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2011. Fill-up versus interpolation methods for phrase-based SMT adaptation.. In *Proceedings of the International Workshop on Spoken Language Translation*. San Francisco, 136–143.
[8] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.
[9] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. 1976. *Graph theory with applications*. Vol. 290. Macmillan London.
[10] Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19, 2 (June 1993), 263–311.
[11] Hailong Cao, Tiejun Zhao, Shu ZHANG, and Yao Meng. 2016. A Distribution-based Model to Learn Bilingual Word Embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, 1818–1827.
[12] Daniel Cer, Michel Galley, Daniel Jurafsky, and Christopher D. Manning. 2010. Phrasal: A Statistical Machine Translation Toolkit for Exploring New Model Features. In *Proceedings of the NAACL HLT 2010 Demonstration Session*. Los Angeles, California, 9–12.
[13] Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*. Trento, Italy, 261–268.
[14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar, 1724–1734.
[15] Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA, 176–181.
[16] Jocelyn Coulmance, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, Fast Cross-lingual Word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, 1109–1113.
[17] Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, 1370–1380.
[18] Allyson Ettinger, Philip Resnik, and Marine Carpuat. 2016. Retro□tting Sense-Speci□c Word Vectors Using Parallel Text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California, 1378–1383.
[19] Manaal Faruqui and Chris Dyer. 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*. Gothenburg, Sweden, 462–471.
[20] Michel Galley, P Chang, Daniel Cer, Jenny R Finkel, and Christopher D Manning. 2008. NIST Open Machine Translation 2008 Evaluation: Stanford University's System Description. In *Unpublished working notes of the 2008 NIST Open Machine Translation Evaluation Workshop*. Citeseer.
[21] Jianfeng Gao, Xiaodong He, Wen-tau Yih, and Li Deng. 2014. Learning Continuous Phrase Representations for Translation Modeling. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, 699–709.
[22] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita, and Benjamin K. Tsou. 2011. Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop. In *Proceedings of NTCIR-9 Workshop Meeting*. Tokyo, Japan, 559–578.
[23] Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France.
[24] Stephan Gouws and Anders Søgaard. 2015. Simple task-speci□c bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies. Denver, Colorado, 1386–1390.
[25] Jiang Guo, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Sense-speci□c Word Embeddings By Exploiting Bilingual Resources. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland, 497–507.
[26] Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173* (2013).
[27] Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual Models for Compositional Distributed Semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, 58–68.
[28] Hermann O Hirschfeld. 1935. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 31. Cambridge Univ Press, 520–524.
[29] Eric Huang, Richard Socher, Christopher Manning, and Andrew Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea, 873–882.
[30] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 95–105.
[31] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, 897–907.
[32] Sujay Kumar Jauhar, Chris Dyer, and Eduard Hovy. 2015. Ontologically Grounded Multi-sense Representation Learning for Semantic Vector Space Models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, 683–693.
[33] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China, 1–10.
[34] Hyungsuk Ji and Sabine Ploux. 2003. Lexical Knowledge Representation with Contexonyms. In *Proceedings of the 9th Machine Translation*. 194–201.
[35] Richard M. Karp. 1972. Reducibility among combinatorial problems. *Complexity of Computer Computations* (1972), 85–103.
[36] Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. *Citeseer* (2012).
[37] Philipp Koehn. 2004. Statistical Signi□cance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, 388–395.
[38] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic, 177–180.
[39] Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. *CoRR* abs/1706.03872 (2017).
[40] Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning Bilingual Word Representations by Marginalizing Alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland, 224–229.
[41] Thomas K Landauer and Susan T. Dutnais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* (1997), 211–240.
[42] Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An Autoencoder Approach to Learning Bilingual Word Representations. In *Advances in Neural Information Processing Systems*. 1853–1861.
[43] Rémi Lebret and Ronan Collobert. 2014. Word Embeddings through Hellinger PCA. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden, 482–490.
[44] Omer Levy and Yoav Goldberg. 2014. Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan, 171–180.
[45] Shaoshi Ling, Yangqiu Song, and Dan Roth. 2016. Word Embeddings with Limited Memory. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany, 387–392.

[46] Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 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*. Denver, Colorado, 250–256.

[47] R.Duncan Luce and AlbertD. Perry. 1949. A method of matrix analysis of group structure. *Psychometrika* 14, 2 (1949), 95–116.

[48] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado, 151–159.

[49] Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. 2016. Vocabulary Manipulation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany, 124–129.

[50] Tomas Mikolov, Kai Chen, Greg Corrado, and Je□rey Dean. 2013. E□cient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[51]

[68] Tianze Shi, Zhiyuan Liu, Yang Liu, and Maosong Sun. 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 (Volume 2: Short Papers)*. Beijing, China, 567–572.
[69] Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. O□ine bilingual word vectors, orthogonal transformations and the inverted softmax. *ArXiv* (2017).
[70] Andreas Stolcke. 2002. SRILM-an extensible language modeling toolkit. In *Proceedings International Conference on Spoken Language Processing*. Seattle, 257–286.
[71] Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation Modeling with Bidirectional Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, 14–25.
[72] Simon Šuster, Ivan Titov, and Gertjan van Noord. 2016. Bilingual Learning of Multi-sense Embeddings with Discrete Autoencoders. *arXiv* (2016).
[73] Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
[74] Julien Tissier, Christopher Gravier, and Amaury Habrard. 2017. Dict2vec : Learning Word Embeddings using Lexical Dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 254–263.
[75] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling Coverage for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, 76–85.
[76] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. *arXiv* (2016).
[77] Ivan Vulić and Anna Korhonen. 2016. On the Role of Seed Lexicons in Learning Bilingual Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany, 247–257.
[78] Ivan Vulić and Marie-Francine Moens. 2015. Bilingual Word Embeddings from Non-Parallel Document-Aligned Data Applied to Bilingual Lexicon Induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Beijing, China, 719–725.
[79] Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2015. Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Di□erences for Lexical Relation Learning. *CoRR* (2015).
[80] Rui Wang, Hai Zhao, Sabine Ploux, Bao-Liang Lu, and Masao Utiyama. 2016. A Bilingual Graph-based Semantic Model for Statistical Machine Translation. In *Proceedings of the 25th International Joint Conference on Arti cial Intelligence*. New York, 2950–2956.
[81] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado, 1006–1011.
[82] Wei Yang, Wei Lu, and Vincent Zheng. 2017. A Simple Regularization-based Algorithm for Learning Cross-Domain Word Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 2888–2894.
[83] Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark, 286–291.
[84] Biao Zhang, Deyi Xiong, jinsong su, Hong Duan, and Min Zhang. 2016. Variational Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 521–530.
[85] Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained Phrase Embeddings for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland, 111–121.
[86] Meng Zhang, Yang Liu, Huanbo Luan, Yiqun Liu, and Maosong Sun. 2016. Inducing Bilingual Lexica From Non-Parallel Data With Earth Mover's Distance Regularization. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 3188–3198.
[87] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial Training for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, 1959–1970.
[88] Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth Mover's Distance Minimization for Unsupervised Bilingual Lexicon Induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Copenhagen, Denmark, 1924–1935.
[89] Meng Zhang, Yang Liu, Huan-Bo Luan, Maosong Sun, Tatsuya Izuha, and Jie Hao. 2016. Building Earth Mover's Distance on Bilingual Word Embeddings for Machine Translation.. In *The Thirty AAAI Conference on Arti cial Intelligence*. 2870–2876.
[90] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, 1393–1398.

Received October, 2017; revised January, 2018