A Bilingual Graph-Based Semantic Model for Statistical Machine Translation

Rui Wang,[1] **Hai Zhao,**[1,2*] **Sabine Ploux,**[3 *] **Bao-Liang Lu,**[1,2] **and Masao Utiyama**[4]
[1]Department of Computer Science and Eng.
[2]Key Lab of Shanghai Education Commission for Intelligent Interaction and Cognitive Eng.
Shanghai Jiao Tong University, Shanghai, China
[3]Centre National de la Recherche Scientifique, CNRS-L2C2, France
[4]National Institute of Information and Communications Technology, Kyoto, Japan
wangrui.nlp@gmail.com, {zhaohai, blu}@cs.sjtu.edu.cn, sploux@isc.cnrs.fr, mutiyama@nict.go.jp

Abstract

Most existing bilingual embedding methods for Statistical Machine Translation (SMT) suffer from two obvious drawbacks. First, they only focus on simple context such as word count and co-occurrence in document or sliding window to build word embedding, ignoring latent useful information from selected context. Second, word sense but not word form is supposed to be the minimal semantic unit while most existing works are still for word representation. This paper presents Bilingual Graph-based Semantic Model (BGSM) to alleviate such shortcomings. By means of maximum complete sub-graph (clique) for context selection, BGSM is capable of effectively modeling word sense representation instead of the word form itself. The proposed model is applied to phrase pair translation probability estimation and generation for SMT. The empirical results show that BGSM can enhance SMT both in performance (up to +1.3 BLEU) and efficiency in comparison against existing methods.

1 Introduction

Continuous representations of words onto multi-dimensional vectors enhance traditional natural language processing [Zhao *et al.*, 2010; Zhao and Kit, 2011; Zhao *et al.*, 2013; Zhang and Zhao, 2013], especially Statistical Machine Translation (SMT), by measuring similarities of words using distances of corresponding vectors [Bengio *et al.*, 2003; Mikolov *et al.*, 2013b; Wang *et al.*, 2016a]. Most of early works are derived from cognitive processing such as WordNet [Miller *et al.*, 1990], in which the lexicon is organized conceptually as a set of terms associated with a partition into synsets[1], though words organized in this way are not conveniently represented as vectors.

Word embedding for vector representation is usually built in two-steps. The first is to determine the detailed context related to a given word. The second is to summary the relationship between word and its context into lower dimensions.

For context determination: 1) The first category is to extract the word or word relations from the entire text, which is usually regarded as document level processing, such as bag-of-word, LSA, and LDA. 2) The second category is to use sliding window, such as n-grams, skip-grams and other local co-occurrence relation [Mikolov *et al.*, 2013a; Zou *et al.*, 2013; Levy and Goldberg, 2014; Pennington *et al.*, 2014; Vulić and Moens, 2015]. 3) The third category, which has been seldom considered, uses much more sophisticated graph style context. Ploux and Ji [2003] describe a graph based semantic matching model using bilingual lexicons and monolingual synonyms[2]. They later represent words using individual monolingual co-occurrences [Ji and Ploux, 2003]. Saluja *et al.* [2014] propose a graph method to generate translation candidates using monolingual co-occurrences.

For relationship summarising, neural networks are very popular for bilingual word embeddings and SMT [Mikolov *et al.*, 2013a; Wang *et al.*, 2013; Zou *et al.*, 2013; Zhang *et al.*, 2014; Gao *et al.*, 2014; Wang *et al.*, 2014; Lauly *et al.*, 2014; Wang *et al.*, 2015; 2016b]. There are also some works which use matrix factorization [Ploux and Ji, 2003; Pennington *et al.*, 2014; Shi *et al.*, 2015] and canonical correlation analysis [Faruqui and Dyer, 2014; Lu *et al.*, 2015] for word embedding.

Sense gives more exact meaning formulization than the word form itself. However, most of existing methods embed words as vectors, instead of sense information. Motivated by these inconveniences, we propose ***Bilingual Contexonym Cliques (BCCs)***, which are extracted from bilingual Point-wise Mutual Information (PMI) based word co-occurrence graph. BCC plays a role of minimal unit for bilingual sense representation. Correspondence Analysis (CA) is

*Corresponding authors. R. Wang, H. Zhao and B. L. Lu were partially supported by Cai Yuanpei Program (CSC No. 201304490199 and No. 201304490171), National Natural Science Foundation of China (No. 61170114 and No. 61272248), National Basic Research Program of China (No. 2013CB329401), Major Basic Research Program of Shanghai Science and Technology Committee (No. 15JC1400103), Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04), and Key Project of National Society Science Foundation of China (No. 15-ZDA041).

[1]synset is a small group of synonyms labeled as a concept.
[2]http://dico.isc.cnrs.fr

then used for summarizing BCC-word matrix into lower dimension vectors for word representation. This work extends previous monolingual method [Ploux and Ji, 2003], which needs bilingual lexicons or synonyms for bilingual mapping and has been never used for SMT.

The remaining of this paper is organized as follows: the proposed ***Bilingual Graph-based Semantic Model (BGSM)*** is introduced in Section 2, and then it is applied to phrase translation probability estimation in Section 3 and phrase pair generation in Section 4 to enhance SMT. The experiments and analysis are given in Section 5. Section 6 summarizes this work.

2 Bilingual Graph-based Semantic Model

2.1 Graph Constructing

Formally, words are considered as nodes (vertices) and co-occurrence relationships of words are considered as edges of graph. An edge-weighted graph derived from a bilingual corpus is formalized as, $G = \{W, E\}$, where W is node set and E is edge set which is weighted by co-occurrence relationship introduced 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 together 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 word) w_i and w_j, if they are in the same BS, they are called co-occurrences for each other and marked as n_i and n_j in the graph G. **As node n_i in graph is always referred to word w_i, we will not distinguish them throughout this paper.** The *Edge Weight* (EW) connecting nodes n_i and n_j is defined by a PMI score,

$$EW = \frac{Co(n_i, n_j)}{fr(n_i) \times fr(n_j)}, \quad (1)$$

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

For nearly all languages, stop words such as *of, a, the* in English or *de, une, la* in French have a wide distribution, and this results in that most nodes in the graph are unnecessarily connected with these stop word nodes. A filter with an EW threshold γ is thus set to prune these poorly connected edges. γ will be tuned using development data, to let the resulted graph keep the more useful edges based on the empirical result of each individual task.

2.2 Context-Dependent Clique Extraction

A clique defined in graph theory means a maximum, complete sub-graph [Luce and Perry, 1949]. For a subset of nodes with edges in the whole graph, if every two nodes in the subset are connected to each other, this subset of nodes form a clique. Suppose that both N_1 and N_2 are subsets of N in G. If $N_1 \subset N_2$, then N_1 cannot be a clique (maximality).

Graph problems, such as finding all cliques from a graph, are mostly associated with high computational complexity (*Clique Problem*). The *Clique Problem* related to our model has been shown *NP-complete* [Karp, 1972], and it is time consuming or even impossible to find the cliques from the whole graph built from a very large corpus (such as millions of sentences) without any pruning. In addition, not all of the nodes are useful for word representations, because some nodes do not have any connection with input contextual words (sparsity). These nodes actually have not a direct impact over clique extraction. For a word n and its contextual words $\{n_1, n_2, ..., n_i, ...n_t\}$ as input[3], only the co-occurrence nodes n_{ij} of each n_i (including n itself) are indeed useful and then actually extracted. The set of nodes $\{n_{ij}\}$ with their weighted edges form an extracted graph $G_{extracted}$ for further cliques extraction. So the number of nodes in the extracted graph $G_{extracted}$, $|N_{extracted}|$, is given by,

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

In practice, the $|N_{extracted}|$ is much smaller than $|V|$ (vocabulary size of bilingual corpus). For a typical corpus (IWSLT in Section 5.1), $|N_{extracted}|$ is around 371.2 on average and $|V|$ is 162.3K. Thus the clique extraction in practice is quite efficient as it works over a quite small sized graph.

Clique extraction may follow a standard routine in [Luce and Perry, 1949]. As the clique in this paper is to represent a fine grained bilingual sense of a word, it is called Bilingual Contexonym Clique (BCC). Similar but more fine grained than synset (a small group of synonyms labeled as concept) defined in WordNet [Miller *et al.*, 1990], the BCC now is the minimal unit for bilingual meaning representation.

Taking the word *work_e* and *readers_e* as an example (without context), two groups of BCCs (in alphabetical order) are given in Table 1. It shows that different word senses can be distinguished by BCCs. The BCCs containing *employees_e*, *travail_f* (work) and *unemployed_e* may indicate the meaning of *job*, while the BCCs containing *readers_e* may indicate the meaning of *literature*.

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*}...

Table 1: BCC examples. Suffixes '*_e*' and '*_f*' are used to indicate English or French, respectively. The English words in parentheses are corresponding translations.

[3]For SMT task, the words in aligned phrases are used as context, please refer to Section 3 for details.

Note that edges pruning is static, and nodes selection is dynamic depending on input word sequence. The proposed node selection and clique extraction follow [Ploux and Ji, 2003], except that we use bilingual co-occurrence graph rather than monolingual synonym or hypo(hypero)nym graphs. BCCs can be regarded as loose synsets, as only strongly related words can be nodes in a clique that possess full connections, and different senses will naturally result in roughly different cliques from our empirical observations, though noise or improper connections also exist at the same time. To obtain concise semantic vector representations, a dimension reduction will be performed.

2.3 Semantic Spatial Representation

Correspondence Analysis (CA) [Benzcri, 1973] assesses the extent of matching between two variables. It 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.

To project words onto lower dimensional semantic space, CA is conducted over the clique-word matrix constructed from the relation between BCCs and words. An initial correspondence matrix $M = \{m_{ij}\}$ is built, where $m_{ij} = 1$ if the BCC in row i contains the word in column j, and 0 if not. Normalized correspondence matrix $\mathcal{P} = \{p_{ij}\}$ is directly derived from M, where $p_{ij} = m_{ij}/N_M$, and N_M is grand total of all the elements in M. Let the row and column marginal totals of $\mathcal{P}$ be r and c which are the vectors of row and column masses, respectively, and D_r and D_c be the diagonal matrices of row and column masses. Coordinates of the row and column profiles with respect to principal axes are computed by using the Singular Value Decomposition (SVD) as follows.

Principal coordinates of rows $\mathcal{F}$ and columns $\mathcal{G}$:

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

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

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

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

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

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

Following standard setting of CA [Benzcri, 1973], top principal dimensions (axes) of vectors are chosen for word and clique representation. Bilingual Graph-based Semantic Model (BGSM) is consequently constructed from these principal dimensions. In short, a word with its context are used as input of BGSM, and vectors of the word[4] and its bilingual co-occurrences are output.

[4]In fact both BCCs and words can be represented as vectors.

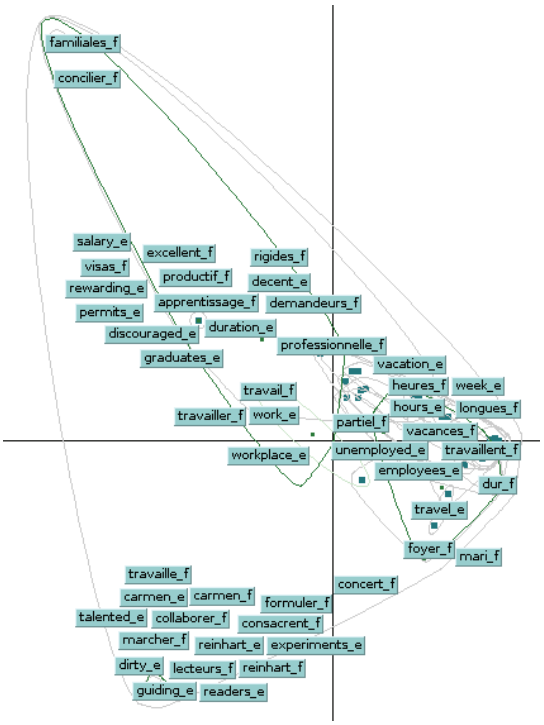

Figure 1: Spatial map of *work_e* (without context) as input (The green lines indicate the borderline of clusters).

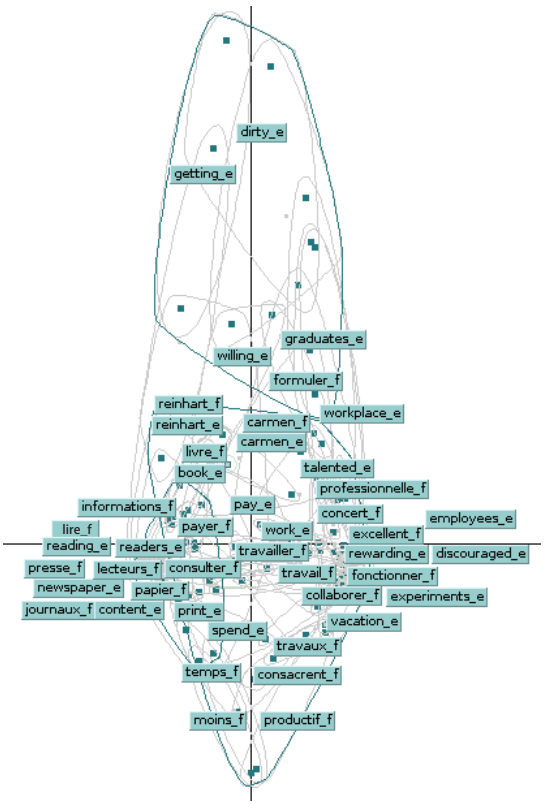

Figure 2: Spatial map of *work*

To visualize the results, top two dimensions are chosen and illustrated into spatial map. We illustrate the spatial relationship between BCCs and words in the same map, instead of the spatial map of words only. BCCs are represented by points and words by regions. Label of word is approximately (to avoid overlapping) placed at the barycentre of region (gray line) delineated by a set of BCCs that contain the word. BCCs are clustered [Ward Jr, 1963] into three groups (green line). We only present several typical words, and the words too far from most of other words are discarded.

Figures 1 and 2 illustrate the spatial representation of all the co-occurrence words when *work_e* is input without context and with *readers_e* as context using BGSM, respectively. The following observations can be obtained from them:

1) Several good translation pairs are shown, such as (*work_e, travailler_f*), (*readers_e*, *lecteurs_f*), (*reading_e, lire_f*) and (*book_e, livre_f*).

2) For *work_e* as input, the word *work_e* itself is placed at the center and the other words are mainly clustered into three semantic groups: *employment* (the right), *evaluation of job* (the upper left), and *literary works*[5] (the bottom). However, we cannot determine which sense of *work_e* belongs to.

3) For *work_e*+*readers_e* as input, we can identify that the sense of *work_e* by the words close to both of *work_e* and *readers_e*, such as *book_e*, *print_e* and *paper_f*.

3 Phrase Translation Probability Estimation

BGSM represents words as vectors dynamically on various geometric coordinates according to contextual words. For each word in a source phrase of SMT, its contextual words are fixed, so all the translation candidate target words can be represented as vectors in the same geometric coordinates. This makes it possible to apply BGSM into phrase-based SMT for selecting translated phrase candidates.

3.1 Bilingual Phrase Semantic Representation

The phrase table of phrase-based SMT model can be simply formalized as[6],

$$(P_F,\ P_E,\ scores,\ word\text{-}alignment), \tag{2}$$

where $P_F(w_{f_1}, w_{f_2}, ..., w_{f_k})$ and $P_E(w_{e_1}, w_{e_2}, ..., w_{e_l})$ are source and its aligned target phrases, respectively, and $scores$ indicate various feature scores including direct translation probability, lexical weighting and phrase penalty. The phrase length is limited to 7, which is the default setting for phrase-based SMT.

BGSM represents words in phrase table as six-dimension vectors. It should be noted that the clique extraction depends on contextual words, and CA then projects clique-word matrix onto corresponding semantic geometric coordinates accordingly. So the same contextual words should be used for all the words in P_F and all its aligned P_E, in order to represent them on the same geometric coordinates. For each word w_{f_i} or w_{e_j} (where $1 \leq i \leq k$, $1 \leq j \leq l$), in 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 target words in all the aligned P_E are used as its 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 as vector V_{wf_i} (or V_{we_j}) (The co-occurrence word w_{co} can also be represented as vector $V_{w_{co}}$). Note that all the source and target words for the same source phrase P_F are represented as vectors in the same geometric coordinates. Some words may not belong to any BCC (partially because the graph is pruned). These *unknown* words would be represented as a default vector.

3.2 Semantic Similarity Measurement

Because the numbers of word alignments in phrase pairs are different, *Normalized Euclidean Distance* (NED) is adopted to measure the distance between source and target phrases incorporated with IBM word-alignment model:

$$NED(P_F, P_E) = \sqrt{\frac{\sum_{align(i,j)} ED^2(V_{wf_i}, V_{we_j})}{\sum_{i,j} align(w_{f_i}, w_{e_j})}}, \tag{3}$$

where $ED(V_{wf_i}, V_{we_j})$ stands for *Euclidean Distance* between word vectors V_{wf_i} and V_{we_j}, $align(i, j)$ is from the word-alignment model in Eq. (2), and $\sum_{i,j} align(w_{f_i}, w_{e_j})$ is total number of word alignments between P_F and P_E.

As there are usually multiple P_E that are aligned to P_F, N_{P_E} is noted as the number of aligned P_E. To let the similarity score $Sim(P_E|P_F)$ be a probability distribution, $Sim(P_E|P_F)$ =1 if $N_{P_E} = 1$; otherwise, $Sim(P_{E_i}|P_F)$ is given by,

$$Sim(P_{E_i}|P_F) = \frac{\sum_j NED^2(P_F, P_{E_j}) - NED^2(P_F, P_{E_i})}{(N_{P_E} - 1) \times \sum_j NED^2(P_F, P_{E_j})}. \tag{4}$$

Using the same pipeline, $Sim(P_F|P_E)$ can also be calculated. Both $Sim(P_E|P_F)$ and $Sim(P_F|P_E)$ can be added as features for SMT decoding.

4 Bilingual Phrase Generation (BPG)

A few phrases are outside corpus but share the similar meaning as those inside the corpus. Take the source French phrase *la bonne réponse* as example, the corresponding aligned target English phrase *the right answer* is in the corpus and 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 candidates for translation.

Since the BGSM can be used to represent words as vectors and measure their similarities by computing vector distance, it is possible to generate new (maybe better) phrases with similar meaning as the original one for phrase table.

4.1 Phrase Pair Generation

As mentioned in Section 3, for each word w (source or target), both of the source words in P_F and target words are

[5]Part of borderline of this cluster is overlapped by some words.

[6]The alignment is from standard IBM alignment model [Berger *et al.*, 1994].

used as its contextual words (Strategy-B). 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 source phrase P_F or target phrase P_E, each word inside will be tentatively replaced by the nearest word in its corresponding co-occurrence according to word vector distance (here, Euclidean distance is actually 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[7].

$Sim(P'_E|P_F)$ and $Sim(P'_F|P_E)$ can be calculated using Eqs. (3) and (4). They are also being the phrase transition probabilities for the generated (P_F, P'_E) and (P'_F, P_E), respectively, as no such probabilities exist in the original phrase table. The updated lexical weighting $lex(P'_E|P_F)$ and inverse lexical weighting $lex(P'_F|P_E)$ are computed by IBM model [Berger *et al.*, 1994].

The generated phrases are filled-up [Bisazza *et al.*, 2011] into original phrase table. That is, a penalty score is added as feature: for original phrase pairs, the penalty is set as one; for the generated ones the penalty is set as *natural logarithm base* e (= 2.71828...). All scores weights in phrase table will be further tuned using MERT [Och, 2003].

4.2 Prase-table Size Tuning

Using the phrase generation approach, a lot of new phrase pairs can be generated. We need to select the most reasonable ones inside them. The *Distance Ratio* (DR) of normalized distance in Eq. (4) between the generated phrase pair (P_F, P'_E) and the original phrase pair (P_F, P_E),

$$DR(P'_E, P_E) = \frac{NED(P_F, P'_E)}{NED(P_F, P_E)}, \tag{5}$$

is used to measure the usefulness of generated phrase pairs.

A threshold ε is set up to keep the most useful generated phrase pair only. Namely, for a source phrase P_F, only the P'_E whose $DR(P'_E, P_E)$ smaller than ε are selected as the generated phrase pair (P_F, P'_E). Using the same pipeline, the size of generated source candidate phrases is also tuned. For SMT task, the threshold is tuned according to SMT performance on development data.

5 Experiments

5.1 Setting up

To evaluate BGSM in various language and domain SMT systems, Corpora of IWSLT-2014 French to English (EN) [Cettolo *et al.*, 2012], NTCIR-9 Chinese to English [Goto *et al.*, 2011] and NISTOpenMT08[8] are chosen.

[7]Two or more words can be replaced, but it may lead to serious sense bias, and the experiments also show that replacing more than two words does not perform well.

[8]Zou et al. [2013] only released their word vectors rather than their code (http://ai.stanford.edu/~wzou/mt/), so we have to conduct experiments on NIST08 Chinese-English translation task as they did for comparison. The training data consists of part of NIST OpenMT06, United Nations Parallel Text (1993-2007) and corpora of [Galley *et al.*, 2008] that were used by [Zou *et al.*, 2013]. NIST Eval 2006 is used as development data and NIST Eval 2008 as test data.

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

Table 2: Sentence statistics on parallel corpora.

5.2 Baseline Systems

The same basic settings for the IWSLT-2014, NTCIR-9 and NIST08 translation baseline systems are complied. The standard Moses phrase-based SMT system is applied [Koehn *et al.*, 2007] together with GIZA++ [Och and Ney, 2004] for alignment, SRILM [Stolcke, 2002] for language modeling and MERT for tuning (we run MERT three times and record the average BLEU score on test data). The paired bootstrap re-sampling test[9] is performed. Significant tests are done for each round of test. Marks at the right of BLEU scores indicate whether our proposed methods are significantly better/worse than the corresponding baseline ('++/−−': significantly better/worse at *significance level* $\alpha = 0.01$; '+/−': $\alpha = 0.05$). All the experiments in this paper are conducted on the same machine with 2.70GHz CPU.

As the proposed BGSM is a bilingual word embedding method and applied to SMT as new features, we only compare with most related bilingual word embedding or generation methods for SMT. For phrase pair translation probability estimation task, two typical neural network based bilingual embedding methods, Continuous Space Translation Model (CSTM[10]) [Schwenk, 2012] and [Zou *et al.*, 2013], are select as baselines. The embedding of each method is added as features to the phrase-based SMT baseline, with all the other setting the same. For bilingual phrase generation methods, CSTM is used as the same way as [Schwenk, 2012]'s generation method. We also compare with [Saluja *et al.*, 2014], which uses graph method for translation candidate generation[11].

5.3 Results and Analysis

Only the best performed results (for both the baselines and proposed methods) on development data are chosen to be evaluated on test data and shown. The parameters for BGSM are set as follows: 1) Vector dimensions are 6; 2) Threshold γ for edge weight pruning EW in Eq. (1) is 3×10^{-4}; 3) Threshold ε for phrase table tuning DR in Eq. (5) is 1.31.

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

[10]The recommended settings of CSTM [Schwenk, 2012] are followed. That is, phrase length limit is set as 7, shared 320-dimension projection layer for each word (that is 2240 for 7 words), 768-dimension projection layer, 512-dimension hidden layer. The dimensions of input/output layers are the same as the size of vocabularies of source/target words.

[11]The implementation and settings in [Saluja *et al.*, 2014] are followed except morphological generation.

Phrase Pair Translation Probability Estimation Results

Table 3 indicates that BGSM can improve SMT performance up to +0.85 BLEU, and outperforms CSTM or Zou's methods up to +0.67 BLEU. Besides, the Strategy-B performs better than Strategy-A, which attributes to more contextual (both target and source) information used for the former while only source for the latter.

	IWSLT	NTCIR	NIST
Baseline	31.80	32.19	30.12
+Zou	N / A	N / A	30.36
+CSTM	32.19	32.37	30.25
+BGSM-A	32.32+	32.56	30.38
+BGSM-B	**32.61++**	**33.04++**	**30.44+**

Table 3: Phrase pair translation probability estimation results (BLEU).

Bilingual Phrase Pair Generation Results

'Baseline + BPG' indicates adding the generated phrase pairs to the original phrase table. 'BPG + BGSM' indicates adding the generated phrase pairs, as well as replacing the translation probabilities in original phrase table with the $Sim(P_E|P_F)$ and $Sim(P_F|P_E)$ calculated by BGSM (Section 4.1).

Corpora	Methods	Phrase Pairs	BLEU
IWSLT	Baseline	9.8M	31.80
	+CSTM	23.1M	32.19
	+Saluja	31.5M	32.35
	+BPG	25.6M	32.37
	+BPG+BGSM	25.6M	**33.13++**
NTCIR	Baseline	71.8M	32.19
	+CSTM	297.8M	32.42
	+Saluja	341.3M	32.68
	+BPG	312.6M	32.54+
	+BPG+BGSM	312.6M	**33.47++**

Table 4: Bilingual Phrase Generation (BPG) results.

The results in Table 4 indicate that the proposed BPG and BGSM methods can work well together and enhance SMT performance significantly up to +1.33 BLEU. They also outperform state-of-the-art method up to +0.79 BLEU.

Efficiency Comparison

We compare the efficiencies for model training and computing the probability scores of phrases pairs using CSTM and BGSM. Two thousand phrase pairs are randomly selected from the whole IWSLT-2014 FR-EN corpus. The CPU time of training models (the whole corpus) and calculating their probability scores (2,000 sentences) is shown in Table 5.

The results in Table 5 demonstrate that BGSM is much more efficient than CSTM, especially for training, the former can be more than 50 times as fast as the later.

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

Table 5: CPU time on IWSLT-2014.

6 Conclusion

Existing word embedding methods usually only consider simple context such as document or sliding window for word relationship building and later word representation. Instead, this paper focuses on sense representation in terms of bilingual background. Using a graph constructed from a bilingual corpus, Bilingual Contexonym Clique (BCC) is proposed for better sense characterization. A BCC-word matrix is then built from dynamic sense-sensitive context in the graph and correspondence analysis is to summarize the matrix into lower dimensions as Bilingual Graph Semantic Model (BGSM).

BGSM word embedding is applied to phrase pair translation probability estimation and generation. The experimental results show that the proposed model can enhance phrase-based SMT decoding and achieve a significant improvement with high computational efficiency. It also outperforms the existing related word embedding methods for SMT.

References

[Bengio *et al.*, 2003] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 2003.

[Benzcri, 1973] J.-P. Benzcri. L'analyse des correspondances. In *L'Analyse des Donnes*, 1973.

[Berger *et al.*, 1994] A. L. Berger, P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, and L. Ureš. The Candide system for machine translation. In *HLT*, 1994.

[Bisazza *et al.*, 2011] A. Bisazza, N. Ruiz, M. Federico, and B. Kessler. Fill-up versus interpolation methods for phrase-based smt adaptation. In *IWSLT*, 2011.

[Cettolo *et al.*, 2012] M. Cettolo, C. Girardi, and M. Federico. Wit^3: Web inventory of transcribed and translated talks. In *EAMT*, 2012.

[Faruqui and Dyer, 2014] M. Faruqui and C. Dyer. Improving vector space word representations using multilingual correlation. In *EACL*, 2014.

[Galley *et al.*, 2008] M Galley, P. Chang, D. Cer, J. Finkel, and C. Manning. NIST open machine translation 2008 evaluation: Stanford university's system description. In *NISTOpenMT*, 2008.

[Gao *et al.*, 2014] J. Gao, X. He, W.-T. Yih, and L. Deng. Learning continuous phrase representations for translation modeling. In *ACL*, 2014.

[Goto *et al.*, 2011] I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou. Overview of the patent machine translation

task at the NTCIR-9 workshop. In *NTCIR-9 Workshop*, 2011.

[Ji and Ploux, 2003] H. Ji and S. Ploux. Lexical knowledge representation with contexonyms. In *MT Summit*, 2003.

[Karp, 1972] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*. Springer, 1972.

[Koehn *et al.*, 2007] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *ACL*, June 2007.

[Lauly *et al.*, 2014] S. Lauly, H. Larochelle, M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha. An autoencoder approach to learning bilingual word representations. In *NIPS*, 2014.

[Levy and Goldberg, 2014] O. Levy and Y. Goldberg. Linguistic regularities in sparse and explicit word representations. In *CONLL*, 2014.

[Lu *et al.*, 2015] A. Lu, W. Wang, M. Bansal, K. Gimpel, and K. Livescu. Deep multilingual correlation for improved word embeddings. In *NAACL*, 2015.

[Luce and Perry, 1949] R. D. Luce and A. D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 1949.

[Mikolov *et al.*, 2013a] T. Mikolov, Q. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv*, 2013.

[Mikolov *et al.*, 2013b] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Miller *et al.*, 1990] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 1990.

[Och and Ney, 2004] F. J. Och and H. Ney. The alignment template approach to statistical machine translation. *Computational linguistics*, 2004.

[Och, 2003] F. J. Och. Minimum error rate training in statistical machine translation. In *ACL*, 2003.

[Pennington *et al.*, 2014] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[Ploux and Ji, 2003] S. Ploux and H. Ji. A model for matching semantic maps between languages (french/english, english/french). *Computational Linguistics*, 2003.

[Saluja *et al.*, 2014] A. Saluja, H. Hassan, K. Toutanova, and C. Quirk. Graph-based semi-supervised learning of translation models from monolingual data. In *ACL*, 2014.

[Schwenk, 2012] H. Schwenk. Continuous space translation models for phrase-based statistical machine translation. In *COLING*, 2012.

[Shi *et al.*, 2015] T. Shi, Z. Liu, Y. Liu, and M. Sun. Learning cross-lingual word embeddings via matrix co-factorization. In *ACL*, 2015.

[Stolcke, 2002] A. Stolcke. SRILM-an extensible language modeling toolkit. In *ICSLP*, 2002.

[Vulić and Moens, 2015] I. Vulić and M.-F. Moens. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *ACL*, 2015.

[Wang *et al.*, 2013] R. Wang, M. Utiyama, I. Goto, E. Sumita, H. Zhao, and B.-L. Lu. Converting continuous-space language models into n-gram language models for statistical machine translation. In *EMNLP*, 2013.

[Wang *et al.*, 2014] R. Wang, H. Zhao, B.-L. Lu, M. Utiyama, and E. Sumita. Neural network based bilingual language model growing for statistical machine translation. In *EMNLP*, 2014.

[Wang *et al.*, 2015] R. Wang, H. Zhao, B.-L. Lu, M. Utiyama, and E. Sumita. Bilingual continuous-space language model growing for statistical machine translation. *IEEE/ACM TASLP*, 2015.

[Wang *et al.*, 2016a] P.-L. Wang, Y. Qian, H. Zhao, and F. Soong. Learning distributed word representations for bidirectional LSTM recurrent neural network. In *NAACL*, 2016.

[Wang *et al.*, 2016b] R. Wang, M. Utiyama, I. Goto, E. Sumita, H. Zhao, and B.-L. Lu. Converting continuous-space language models into n-gram language models with efficient bilingual pruning for statistical machine translation. *ACM TALLIP*, 2016.

[Ward Jr, 1963] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 1963.

[Zhang and Zhao, 2013] J. Zhang and H. Zhao. Improving function word alignment with frequency and syntactic information. In *IJCAI*, 2013.

[Zhang *et al.*, 2014] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong. Bilingually-constrained phrase embeddings for machine translation. In *ACL*, 2014.

[Zhao and Kit, 2011] H. Zhao and C. Kit. Integrating unsupervised and supervised word segmentation: The role of goodness measures. *Information Sciences*, 2011.

[Zhao *et al.*, 2010] H. Zhao, C.-N. Huang, M. Li, and B.-L. Lu. A unified character-based tagging framework for Chinese word segmentation. *ACM TALLIP*, 2010.

[Zhao *et al.*, 2013] H. Zhao, X. Zhang, and C. Kit. Integrative semantic dependency parsing via efficient large-scale feature selection. *Journal of Artificial Intelligence Research*, 2013.

[Zou *et al.*, 2013] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, 2013.