Information Sciences 424 (2018) 287–300

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A neural generative autoencoder for bilingual word embeddings

Jinsong Su[a], Shan Wu[a], Biao Zhang[a], Changxing Wu[b], Yue Qin[a], Deyi Xiong[c,*]

[a] Xiamen University, Xiamen 361005, China
[b] Virtual Reality and Interactive Techniques Institute, East China Jiaotong University, Nanchang 330013, China
[c] Soochow University, Suzhou 215006, China

ARTICLE INFO

Article history:
Received 18 July 2017
Revised 19 September 2017
Accepted 29 September 2017
Available online 3 October 2017

Keywords:
Bilingual word embeddings
Neural generative autoencoder
Cross-lingual document classification
Translation probability modeling

ABSTRACT

Bilingual word embeddings (BWEs) have been shown to be useful in various cross-lingual natural language processing tasks. To accurately learn BWEs, previous studies often resort to discriminative approaches which explore semantic proximities between translation equivalents of different languages. Instead, in this paper, we propose a neural generative bilingual autoencoder (NGBAE) which introduces a latent variable to explicitly induce the underlying semantics of bilingual text. In this way, NGBAE is able to obtain better BWEs from more robust bilingual semantics by modeling the semantic distributions of bilingual text. In order to facilitate scalable inference and learning, we utilize deep neural networks to perform the recognition and generation procedures, and then employ stochastic gradient variational Bayes algorithm to optimize them jointly. We validate the proposed model via both extrinsic (cross-lingual document classification and translation probability modeling) and intrinsic (word embedding analysis) evaluations. Experimental results demonstrate the effectiveness of NGBAE on learning BWEs.

1. Introduction

The studies of word embeddings mainly focus on how to exploit the context of each word to learn its semantic meaning, which is formally represented as a dense, low-dimensional and real-valued vector. Due to the potential advantages of alleviating data sparsity and encoding syntactic and semantic information among words, word embeddings have been widely used in many natural language processing (NLP) tasks, such as language modeling [1], sentiment analysis [42], word sense discrimination [20], machine translation [54], opinion mining [14] and so on. Currently, word embedding has become an indispensable component in many NLP models.

When cross-lingual NLP tasks are in focus, an extension of interest from monolingual to bilingual word embeddings (BWEs) has naturally occurred. BWEs aim to embed the words of different languages into the same semantic space, allowing the cross-lingual semantic computation at the word level. In this aspect, previous studies often resort to discriminative approaches, which explore BWEs based semantic proximities between translation equivalents in different languages. However, these methods are prone to overfitting because they essentially memorize the training data as isolated points in the semantic space. As a result, the potential benefits of the learned BWEs for the subsequent NLP applications are limited.

* Corresponding author.
E-mail addresses: jssu@xmu.edu.cn (J. Su), wushan@stu.xmu.edu.cn (S. Wu), zb@stu.xmu.edu.cn (B. Zhang), wuchangxing@ecjtu.edu.cn (C. Wu), qinyue@stu.xmu.edu.cn (Y. Qin), dyxiong@suda.edu.cn (D. Xiong).

https://doi.org/10.1016/j.ins.2017.09.070
0020-0255/

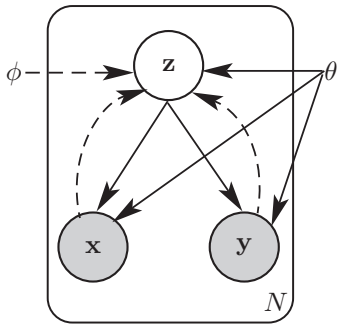

Fig. 1. Graphical illustration of NGBAE. We use solid and dashed lines to indicate the generation and recognition step respectively. Here, x and y are the bag-of-words representation of source and target sentence, and θ and ϕ are the corresponding model parameters, respectively. We use gray and white to represent observed and latent variables respectively.

In this paper, we propose a neural generative bilingual autoencoder (NGBAE) for learning BWEs. Similar to the conventional bilingual antoencoder [4], our model is able to learn the hidden encoder representations of paired bag-of-words sentences without word alignments, where word representations can then easily be extracted and used in the context of a supervised NLP task. The difference is that we introduce a latent variable to explicitly induce the underlying semantics of bilingual bag-of-words sentences. Specifically, given a parallel sentence with its bag-of-words representations in source and target language, denoted by x and y, instead of distinguishing the correct x/y from the corrupted ones [17] or modeling the joint probability $p(x, y)$ directly [4], we assume that there is a latent variable z from the underlying bilingual semantic space which generates both x and y[1]:

$$p(x, y) = \int_z p(x, y|z)p(z)dz \tag{1}$$

Learning through this generative framework, our model is able to not only capture the data generation distribution $p(x, y)$, but also produce the explicit bilingual semantic representation $z \in \mathbb{R}^{d_{sem}}$, where d_{sem} is a fixed dimension size. Compared with the previous discriminative methods, NGBAE has the following advantages: 1) It is able to explicitly model holistic properties of bilingual semantics, and 2) It learns bilingual semantics as soft ellipsoidal regions rather than individual points in latent space, forcing the learned bilingual representations to fill the space rather than memorizing the training data as isolated points. Hence, the learned bilingual semantic representations are more robust and therefore lead to better BWEs.

We further illustrate the graphical model of NGBAE in Fig. 1. There are two major components in NGBAE: 1) a *generation module* (solid lines in Fig. 1) that estimates the generation probability of observed bilingual text conditioned on the latent variable, denoted by $p(x, y|z)$, and 2) a *recognition module* (dashed lines in Fig. 1) that infers the latent variable based on the observed bilingual sentence, denoted as $p(z|x, y)$. Since the posterior distribution over z is usually intractable, we apply deep neural networks to approximate this module, as implemented in the conventional variational autoencoder [24,40]. Particularly, to enable efficient inference and learning, we adopt the *reparameterization method* [24,40] to bridge these two modules. In this way, our NGBAE becomes an end-to-end neural network endowed with the stochastic optimization ability to enhance its generality.

The major contributions of this paper lie in the following threefold:

- We thoroughly analyze and point out the drawbacks of previous studies for BWEs. Specifically, previous studies often resort to discriminative approaches which essentially only memorize the training data as isolated points in the semantic space, and therefore, these methods are prone to overfitting.
- We introduce a variational bilingual autoencoder to capture semantic proximities between bilingual translation equivalents, thus leading to more robust BWEs.
- We conduct a series of experiments on Chinese-English, English-French, English-German, English-Spanish language pairs, and quantitatively and qualitatively evaluate our model on both extrinsic and intrinsic tasks, including cross-lingual document classification, translation probability modeling, and bilingual word embedding analysis. Experimental results strongly demonstrate the generality and effectiveness of our model.

The rest of this paper is organized as follows. We summarize and compare the related work in Section 2. Then, we describe the proposed NGBAE model in detail in Section 3. Experimental results are reported and analyzed in Section 4. Finally, Section 5 presents conclusions and directions for future research.

[1] All the variables (including the latent variable) are multivariate.

2. Related work

Recently, learning appropriate semantic representations for different data modalities has gradually become hotspot in many research domains, such as image-based 3-D human pose recovery [18], video-based human pose recovery [19], image retrieval [51], image privacy protection [52], cross-language text classification [11] and so on.

In this work, we mainly focus on the BWEs that are the foundation of many cross-lingual NLP tasks. With the rapid development of deep learning, many approaches have been proposed to learn word embeddings with deep neural networks. In this respect, word embeddings are originally the by-products of training neural language models [1]. Then, Mikolov et al. [35] presented *Skip-Gram* and *Continuous Bag-of-Words* models, both of which are simple yet effective for learning word representations. Later, neural networks based word embedding modeling becomes a hot topic in the field of NLP [3,7,21,28–30,39,44,50]. Meanwhile, because of the abilities of encoding both syntactic and semantic information of words and dealing with the data sparsity, word embeddings have been widely used in many NLP tasks.

Moreover, along with the rise of cross-lingual NLP tasks, the researches on word representation learning are extended to learning BWEs. Originally, Klementiev et al. [25] used word embeddings learned by a multitask neural network language model with a regularization term that encourages pairs of frequently aligned words to have similar word embeddings. Mikolov et al. [36] first learned monolingual word embeddings of different languages, and then established a linear mapping between different semantic spaces from a small bilingual corpus. Zou et al. [54] initially trained word embeddings in the source language, and then initialized and refined word embeddings in the target language according to word alignments. Because these two methods do not require further training of word representations and consequently gain an edge of fast training speed. Recently, some models are proposed to jointly learn word representations under various cross-lingual semantic constraints. For instance, Hermann and Blunsom [16,17] proposed to learn multilingual word embeddings using parallel data in conjunction with a multilingual objective function for compositional vector models. Furthermore, Kočiský et al. [27] presented a probabilistic model to concurrently learn alignments and BWEs by marginalizing over word alignments. Chandar et al. [4] investigated bilingual autoencoder models to learn hidden encoder representations of paired bag-of-words sentences. Zhou et al. [53] incorporated sentiment information into bilingual word embeddings. Gouws et al. [12] applied a novel bag-of-words cross-lingual sampled objective to regularize two noise contrastive language models. Vulić and Moens [48,49] designed an adapted bilingual Skip-Gram model to learn BWEs from document-aligned corpus.

From different point of view, Luong et al. [32] improved BWEs by preserving clustering structures of words in each language. Lu et al. [31] explored how to incorporate bilingual information into word embeddings via deep canonical correlation analysis. Shi et al. [41] put forward a framework with matrix factorization to learn BWEs. Soyer et al. [43] presented a novel neural network based architecture for BWEs according to the observation that phrases have obvious closer semantic relation with their sub-phrases than with other randomly sampled phrases. Coulmance et al. [8] introduced a new method called Trans-gram, in which sentence alignments rather than word alignments are applied to learn word embeddings aligned across different languages. Oshikiri et al. [38] extended Eigenwords, the spectral monolingual word embeddings based on canonical correlation analysis (CCA), to cross-lingual tasks with sentence-alignment. Guo et al. [15] designed a representation learning framework and devoted it into cross-lingual settings. Upadhyay et al. [47] extensively compared the performances of different methods on generating cross-lingual word embeddings. Duong et al. [10] modeled polysemy to learn bilingual correspondences from monolingual data by the mean of EM translation selection.

Other studies related to our work include autoencoders [2] and variational neural models [24,40], both of which have drawn much attention. Autoencoders are neural networks which learn a reduced dimensional representation of fixed-size inputs such as image patches or bag-of-words representations of text documents. They can be used to efficiently learn useful feature encodings for subsequent tasks [18,19,42,51]. Variational neural networks are presented to perform efficient inference and learning in directed probabilistic models on large-scale dataset [24,40]. Different from the conventional mean-field approaches, they employed deep neural networks to approximate the intractable posterior distribution in continuous space. Because of its flexibility and scalability, variational neural network has been widely applied into different tasks, e.g. semi-supervised learning [23], RNN modeling [5], question answering [34] and image generation [13], etc. In this work, we extend the variational monolingual autoencoder [24] into the filed of bilingual semantic modeling, which, to the best of our knowledge, has never been investigated before. Finally, it should be noted that our model essentially introduces noise to refine BWEs. Therefore, our work is also related to [22] which also introduces Gaussian noise into document modeling but not via the variational approach.

3. Neural generative bilingual autoencoder

In this section, we describe the proposed NGBAE model in detail. Notations of our model are presented in Table 1. We first give an illustration of the overall model architecture in Section 3.1. Then, we describe how to apply neural networks to implement recognition, reparameterization and generation module in Sections 3.2–Section 3.4, respectively. Finally, we depict the objective function and training procedure in Section 3.5.

Table 1
Notations in the NGBAE model.

Symbol	Meaning
x	the bag-of-words representation of source sentence
y	the bag-of-words representation of target sentence
z	the vector representation of latent variable
r_x	the basic recognition vector of source sentence
r_y	the basic recognition vector of target sentence
V_x	the size of source vocabulary
V_y	the size of target vocabulary
$\mathcal{N}_z$	the diagonal Gaussian distribution used to model $q_\phi(z\|x, y)$
μ	the mean of $\mathcal{N}_z$
σ^2	the variance of $\mathcal{N}_z$
ϵ	the noise introduced in reparameterization trick
g_x	the vector representation of source sentence in the latent variable semantic space
g_y	the vector representation of target sentence in the latent variable semantic space
x'	the predicted probability vector of words in source sentence
y'	the predicted probability vector of words in target sentence
d_x	the vector dimension of source sentence in the source-side semantic space
d_y	the vector dimension of target sentence in the target-side semantic space
d_{sem}	the vector dimension in the latent semantic space
L	sample number

3.1. Overall architecture

Assuming we have a training set of such (x, y) pairs, where x and y are the bag-of-words representation of the source and target sentence, respectively, we expect to learn vectorial representations of words from the training set. For this, we propose a bilingual autoencoder model, which encodes x and y respectively as the sum of the representations of their words, followed by a non-linear transformation. Particularly, we introduce a latent variable to explicitly induce the underlying semantics of a bilingual sentence.

According to the variational theory, NGBAE formulates the generative model with a latent variable z, as shown in Eq. (1). To optimize the objective function in Eq. (1), NGBAE introduces an alternative variational lower bound with the following two properties: 1) it should be flexible to optimize, and 2) maximizing this lower bound can pull up the real log-likelihoods. More specifically, we apply the *Jenson's inequation* to derive the lower bound $\mathcal{L}$ of the defined joint distribution in Eq. (1) as follows

$$\log p_\theta(x, y) = \log \int_z \frac{q(z)}{q(z)} p_\theta(x, y|z) p(z) dz \geqslant \mathbb{E}_{q(z)}[\log p_\theta(x, y|z)] \text{-} KL[q(z)||p(z)] = \mathcal{L} \tag{2}$$

where θ parameterizes the generation distribution $p_\theta(x, y|z)$, $p(z)$ denotes the prior distribution typically modeled using a standard Gaussian distribution, i.e. $z \sim \mathcal{N}(0, \mathbf{I})$, and $q(z)$ denotes the variational distribution, and $KL(q(\cdot)||p(\cdot))$ is the *Kullback-Leibler divergence* between the distributions $q(\cdot)$ and $p(\cdot)$. Theoretically, the inequality of the above lower bound becomes equality when $q(z)$ is identical to the true posterior distribution $p(z|x, y)$. Therefore, we can consider $q(z)$ as an approximation of $p(z|x, y)$. To have a tight lower bound, we often condition $q(z)$ as $q_\phi(z|x, y)$, of which inference is the key to the success of NGBAE.

As illustrated in Fig. 2, there are mainly three steps when generating a bilingual sentence: *Recognition, Reparameterization* and *Generation*. Following previous work [24,40], we adopt deep neural networks to model these three modules:

(a) **Bilingual Neural Recognition Network.** It models $q_\phi(z|x, y)$ as a diagonal Gaussian distribution $\mathcal{N}_z(\mu, \text{diag}(\sigma^2))$ (Fig. 2(a));
(b) **Reparameterization.** It samples the latent representation $z \in \mathbb{R}^{d_{sem}}$ from $q_\phi(z|x, y)$ (Fig. 2(b));
(c) **Bilingual Neural Generation Network.** It models $p_\theta(x, y|z)$ to generate the words of bilingual sentence independently from z according to Bernoulli distribution (Fig. 2(c)).

In contrast to the conventional bilingual autoencoder (BAE) [4], NGBAE differs significantly in the (a) and (b) modules, where the integration of latent variable z endows it with both the enriched probabilistic interpretation and the representation ability of bilingual semantics. We will give a detailed description in the following subsections.

3.2. Bilingual neural recognition network

As shown in Fig. 2 (a), the bilingual neural recognition network is built upon the binary bag-of-words representation of bilingual text, where the input bilingual text is converted into a fixed-size but sparse binary vector (x, y) such that x_i/y_j is 1 if word i/j occurs and otherwise 0. Using this network, we first employ a simple multilayer perceptron to perform non-linear

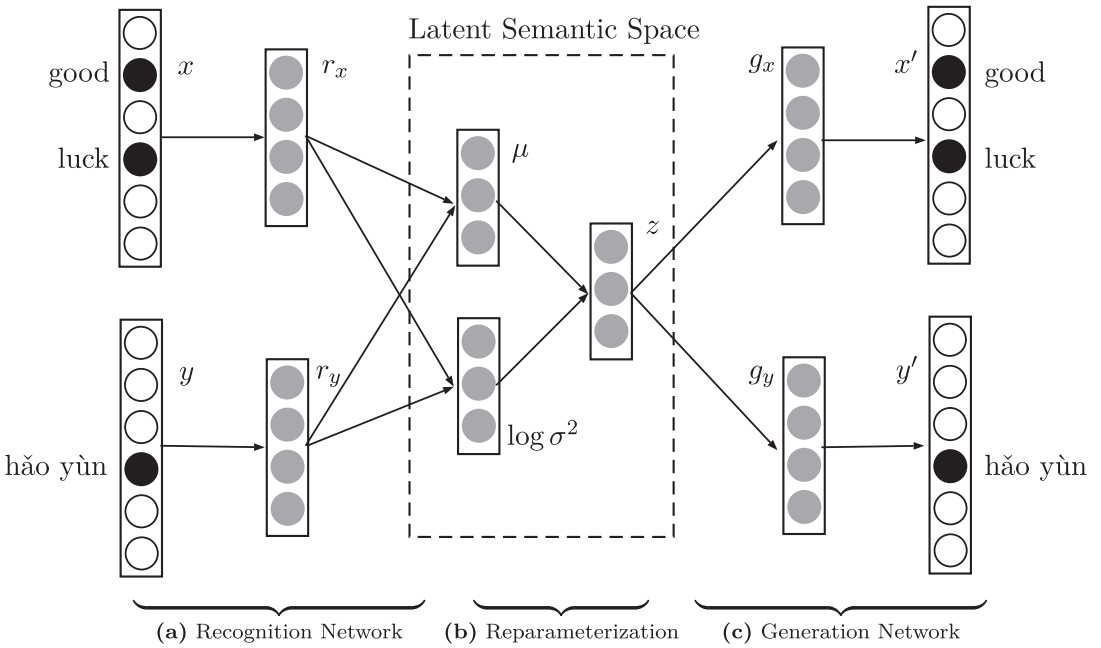

Fig. 2. Overview of NGBAE with a bilingual sentence "(good luck, hǎoyùn)". We use white, gray, black color to indicate 0s, real values and 1s respectively.

transformations converting the binary input to the basic recognition vectors in two languages, which are denoted as r_x and r_y, respectively:

$$r_x = f(W_x^{(r)} x + b_x^{(r)}) \tag{3}$$

$$r_y = f(W_y^{(r)} y + b_y^{(r)}) \tag{4}$$

where $W_x^{(r)} \in \mathbb{R}^{d_x \times |V_x|}$, $W_y^{(r)} \in \mathbb{R}^{d_y \times |V_y|}$, $b_x^{(r)} \in \mathbb{R}^{d_x}$, $b_y^{(r)} \in \mathbb{R}^{d_y}$ are the model parameters, d_x and d_y are the dimensions of source and target representation, $|V_x|$ and $|V_y|$ are the sizes of source vocabulary V_x and target vocabulary V_y, respectively, and $f(\cdot)$ is an element-wise nonlinear function such as the $tanh(\cdot)$ which is used in our experiments. Note that the equality between d_x and d_y is not required in our work.

Then, we obtain the parameters of $\mathcal{N}_z$ through the following linear regression functions:

$$\mu = W_x^{(\mu)} r_x + W_y^{(\mu)} r_y + b^{(\mu)} \tag{5}$$

$$\log \sigma^2 = W_x^{(\sigma)} r_x + W_y^{(\sigma)} r_y + b^{(\sigma)} \tag{6}$$

where $b^{(\mu)}, b^{(\sigma)} \in \mathbb{R}^{d_{sem}}$, $W_x^{(\mu)}, W_x^{(\sigma)} \in \mathbb{R}^{d_{sem} \times d_x}$, $W_y^{(\mu)}, W_y^{(\sigma)} \in \mathbb{R}^{d_{sem} \times d_y}$ are the model parameters, and both μ and $\log \sigma^2$ are d_{sem}-dimension vectors.

Using the nonlinear function in Eq. (3), we ensure that the inferred variational distribution $q_\phi(z|x, y)$ fits well with the relatively complicated posterior. Therefore, it would be meaningful to re-generate the bilingual text from the learned latent semantic representation z.

3.3. Reparameterization

Given the recognition module, we then focus on how to obtain the latent representation z from the induced Gaussian distribution $\mathcal{N}_z(\mu, \text{diag}(\sigma^2))$. A natural solution is to perform sampling from $\mathcal{N}_z$ since $\mathcal{N}_z$ is completely known. However, this may break the connection between the recognition and generation model, leading to the difficulty in the model inference and learning. Here we adopt the *reparameterization trick* [24,40] leveraging the "*location-scale*" property of Gaussian distribution to deal with this problem:

$$z = \mu + \sigma \odot \epsilon \tag{7}$$

where $\odot$ denotes an element-wise product operation, and ϵ acts as a noise term obeying the standard Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. In doing so, NGBAE becomes an end-to-end neural network, and thus we are able to perform stochastic back-propagation to optimize the whole model via *stochastic gradient ascent* algorithm.

Intuitively, the introduction of ϵ makes the learned latent representation fit well with the semantic space rather than memorize the training data as isolated representations in the space. Therefore, NGBAE can be regarded as a regularized version of the standard bilingual autoencoder.

3.4. Bilingual neural generation network

As illustrated in Fig. 2 (c), the bilingual neural generation model aims at reconstructing the bilingual text (x', y') from the common latent semantic representation z. As implemented in [24], we establish a symmetric generation model using two neural networks, both of which are similar to that of the above-mentioned recognition model. Specifically, we first project z into the semantic space of each language separately:

$$g_x = f(W_x^{(g)} z + b_x^{(g)}) \tag{8}$$

$$g_y = f(W_y^{(g)} z + b_y^{(g)}) \tag{9}$$

where $W_x^{(g)} \in \mathbb{R}^{d_x \times d_{sem}}$, $W_y^{(g)} \in \mathbb{R}^{d_y \times d_{sem}}$, $b_x^{(g)} \in \mathbb{R}^{d_x}$ and $b_y^{(g)} \in \mathbb{R}^{d_y}$ are the model parameters.

To implement reconstruction, we then further stack a Sigmoid layer to predict the probabilities of each word in x and y:

$$x' = h(W_{x'} g_x + b_{x'}) \tag{10}$$

$$y' = h(W_{y'} g_y + b_{y'}) \tag{11}$$

where $h(\cdot)$ is the Sigmoid function, $W_{x'} \in \mathbb{R}^{V_x \times d_x}$, $W_{y'} \in \mathbb{R}^{V_y \times d_y}$, $b_{x'} \in \mathbb{R}^{V_x}$ and $b_{y'} \in \mathbb{R}^{V_y}$.

We assume that $p_\theta(x, y|z)$ is a multivariate Bernoulli distribution. In this way, the logarithm of $p_\theta(x, y|z)$ can be computed by comparing the reconstructed representation to the original binary bag-of-words representation as follows:

$$\begin{aligned} \log p_\theta(x, y|z) = &\sum_i x_i \log x'_i + (1 - x_i) \log(1 - x'_i) \\ &+ \sum_i y_i \log y'_i + (1 - y_i) \log(1 - y'_i) \end{aligned} \tag{12}$$

3.5. Model objective and learning

The objective function of NGBAE on the training instance (x, y) is defined as follows:

$$\begin{aligned} \mathcal{L}(x, y; \theta, \phi) \approx\ & \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x, y|z)] - KL(q_\phi(z|x, y)||p_\theta(z)) \\ =\ & \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x, y|z^{(l)}) + \frac{1}{2} \sum_{k=1}^{d_{sem}} [1 + \log({\sigma_k}^2) - {\mu_k}^2 - {\sigma_k}^2] \end{aligned} \tag{13}$$

where $z^{(l)} = \mu + \sigma \odot \epsilon^{(l)}$ and $\epsilon^{(l)} \sim \mathcal{N}(0, \mathbf{I})$. The first term is the expectation of $\log p_\theta(x, y|z)$ over $q_\phi(z|x, y)$. Following previous work [24,40], we employ the *Monte Carlo* method to estimate this expectation: $\mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x, y|z)] \approx \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(x, y|z^{(l)})$ where L is the number of samples. The second term is the analytical form of the *Kullback–Leibler divergence*. Under the assumption that $q_\phi(z|x, y)$ follows $\mathcal{N}_z(\mu, \mathrm{diag}(\sigma^2))$ and $p_\theta(z)$ obeys $\mathcal{N}(0, \mathbf{I})$, we can directly unfold $-KL(q_\phi(z|x, y)||p_\theta(z))$ as $\frac{1}{2} \sum_{k=1}^{d_{sem}} [1 + \log({\sigma_k}^2) - {\mu_k}^2 - {\sigma_k}^2]$[2].Both terms are differentiable so that the gradient of $\mathcal{L}$ can be calculated through conventional back-propagation algorithm.

Finally, it should be noticed that words in different languages locate in different semantic spaces in the above procedure. For this, we first employed the transformation matrix $W_x^{(\mu)}$ and $W_y^{(\mu)}$ to project them into the common latent semantic space, and then investigate the effectiveness of the transformed BWEs in various NLP tasks.

4. Experiments

We conducted experiments on English-Chinese, English-German, English-French and English-Spanish language pairs to testify our model in terms of both extrinsic and intrinsic evaluations. We first carried out two groups of experiments to investigate the effectiveness of the proposed model: cross-lingual document classification and translation probability modeling, and then deeply analyzed the learned BWEs.

[2] We refer the readers to [24] for detail if you are not familiar to this.

Table 2
Datasets for training BWEs. #BS = the number of bilingual sentence, #SW = the size of source (English) vocabulary, and #TW = the size of target (Chinese/German/French/Spanish) vocabulary.

Language Pair	#BS	#SW	#TW
English-Chinese	2,500,000	30,000	30,000
English-German	1,920,209	43,614	50,110
English-French	2,007,723	43,614	35,891
English-Spanish	1,965,734	43,614	38,519

Table 3
Document numbers in the English-Chinese document classification experiment. Note that we extract monolingual documents from English-Chinese parallel corpus.

Language	Data Set	Laws	News	Subtitles	Thesis
English	Train Set	1,769	3,786	2,525	1,920
	Validation Set	175	382	255	188
	Test Set	874	1,878	1,282	966
Chinese	Train Set	1,769	3,786	2,525	1,920
	Validation Set	175	382	255	188
	Test Set	874	1,878	1,282	966

4.1. Setup

Data Set. To learn English-Chinese BWEs, we used LDC corpus[3] consisting of 2.5 million sentence pairs. As for the training data of English-German, English-French, English-Spanish, we used sections of the Europarl corpus[4] containing roughly 2 million parallel sentences for each language pair mentioned above. We followed [25] to tokenize the sentences using NLTK,[5] remove punctuations and lowercase all words. We did not remove stopwords since they are important contextual words for learning BWEs. We filtered out words which occur fewer than five times and induced representations of $d_{sem} = 40$ for the vocabulary words in our data sets. The datasets for training BWEs are illustrated in Table 2.

Network Setting. For the hyper-parameters of NGBAE, we set $d_x = d_y = 400$, $d_{sem} = 40$, $L = 1$ and other hyper-parameters empirically according to our preliminary experiments and other previous studies [24]. With respect to parameter initialization, we randomly initialized all model parameters according to a normal distribution ($\mu = 0$, $\sigma = 0.01$). When training our model, we employed the *Adagrad* algorithm [9] to set the learning rate as 0.01 and the maximum iteration number as 25. At each batch, we randomly selected 100 sentence pairs to update model parameters. The model was trained for approximately four days. Since the training can be performed off-line, we believe that the training time is not critical to the real world usage as the learned BWEs are often fixed in the subsequent NLP tasks. We implemented our model using Theano.[6][45]

4.2. Cross-lingual document classification

In the first group of experiments, we evaluated our model on cross-lingual document classification task. As implemented in [25], we represent each document as an average of representations of all of its tokens weighted by their *idf* scores. Then, we trained a multi-class document classifier for 10 epochs using the averaged perceptron algorithm [6]. To investigate the generality of our model, we carried out experiments in both directions for the same language pair. In particular, we trained a classifier with English documents that is then used to classify Chinese documents, and vice versa.

For English-Chinese experiments, we used the documents selected from UM-Corpus [46] as experimental data, where each document is labeled with only one single topic among the four labels: *Laws, News, Subtitles* and *Thesis*. For English-German, English-French, and English-Spanish experiments, we extracted the documents from sections of the Reuters RCV1/RCV2 corpora. Following [25], we only used documents which were assigned exactly one of the 4 top level categories in the topic hierarchy (CCAT (Corporate/Industrial), ECAT (Economics), GCAT (Government/Social), and MCAT (Markets)). We chose 10,000/5,000 document in each language as the training/test set, and kept another 1000 as the validation set for hyper-parameter tuning.

Tables 3 and 4 gives the statistics of the various data sets. In addition to the conventional BAE, we also compared our model with the following approaches:

[3] The corpora include LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and the Hansards, Laws and News parts of LDC2004T08.
[4] http://www.statmt.org/europarl/
[5] http://www.nltk.org/
[6] http://www.deeplearning.net/software/theano/

Table 4
Document numbers in the English-German, English-French, English-Spanish document classification experiments.

Language	Data Set	CCAT	ECAT	GCAT	MCAT
English	Train Set	4,395	869	2,287	2,449
	Validation Set	449	96	228	227
	Test Set	2,236	386	1,150	1,228
German	Train Set	4,099	617	3,029	2,255
	Validation Set	312	62	397	229
	Test Set	1,495	331	2,045	1,129
French	Train Set	2,078	662	6,071	1,189
	Validation Set	219	65	574	142
	Test Set	1,122	290	3,001	587
Spanish	Train Set	1,461	915	905	6,719
	Validation Set	155	84	118	643
	Test Set	784	455	430	3,331

- **Majority Class**: It directly assigns test documents the most frequent class in the training set. Note that the training and test documents belong to different languages, and the topic distributions of data sets differ greatly in different languages.
- **MT**: It first translates test documents to the language of the training documents using MOSES,[7] and then classifies the translated documents as the documents in the source language. MOSES is an open source phrase-based translation system with default configurations and a 5-gram language model trained on the target portion of training data (same as the one used for inducing our bilingual embeddings).
- **BWEs-Multitask** [25]: It learns BWEs using a multitask neural network language model with a regularization term that encourages pairs of frequently aligned words to have similar word embeddings. We followed closely the setup used by [25] and compared with their method, for which word representations are publicly available.[8]
- **BiCVM** [17]: It implements multilingual word embeddings using parallel data in conjunction with a multilingual objective function for compositional vector models. We directly used their open source code.[9]
- **BAE** [4]. It models bag-of-words representations of aligned sentences and involves four-direction reconstruction terms (source to source, source to target, target to source, target to target) as well as a correlation term. Different from our model, no latent variable is explored in BAE. We directly implemented their model using their open source code.[10]
- **BilBOWA** [12]. It learns BWEs using a novel sampled bag-of-words cross-lingual objective, which is used to regularize two noise-contrastive language models for efficient cross-lingual feature learning. We directly implemented their model using their open source code.[11]

Table 5 summarizes the classification results in two directions obtained by different models, respectively. In exception to English → German, NGBAE always outperforms the other models by large margins. For these results, we speculate that our model is fit to large-scale corpus and can model the observed variability from data with noises, and that it does not significantly outperform the baseline on sparse data. In the English-German experiments, we followed [25] to construct vocabularies using the words occurring no fewer than five times. In this way, the size of German vocabulary is larger than that of English vocabulary (50,110 vs 43,614). However, in the English-German corpus, the number of unique German tokens is about three times of that of unique English tokens (376,295 vs 113,247). Obviously, there are more low-frequency German tokens, leading to more serious data sparsity in the German corpus. Therefore, our model does not perform well enough in the English → German experiment.

4.3. Translation probability modeling

In the second group of experiments, we compared the BWEs learned by different models from a different perspective. For this, we first induced two types of lexical translation probability:

1) **Translation Probabilities Based on BWEs**. As implemented in Maaten and Hinton [33], we converted the cross-lingual similarity based on BWEs to the word-level translation probability. Formally, the source-to-target word-level translation probability between the source word x_i and the target word y_j is defined as follows:

$$p_{bwe}(y_j|x_i) = \frac{exp(-||\vec{x}_i - \vec{y}_j||^2)}{\sum_{y_k \in tran(x_i)} exp(-||\vec{x}_i - \vec{y}_k||^2)} \tag{14}$$

[7] http://www.statmt.org/moses/
[8] http://people.mmci.uni-saarland.de/aklement/data/distrib/
[9] https://github.com/karlmoritz/bicvm
[10] http://www.sarathchandar.in/crl.html
[11] https://github.com/gouwsmeister/bilbowa

Table 5
Experimental results of the cross-lingual document classification. Here, we directly used the publicly available BWEs of [25] for the BWEs-Multitask model, and ran the released open-source tools to obtain BWEs for the BiCVM, BAE, and BilBOWA models.

Language Pair($\mathbf{L_1}$ – $\mathbf{L_2}$)	Model	$\mathbf{L_1} \rightarrow \mathbf{L_2}$	$\mathbf{L_2} \rightarrow \mathbf{L_1}$
English-Chinese	Majority Class	37.6	37.6
	MT	63.3	60.4
	BWEs-Multitask	–	–
	BiCVM	70.0	70.4
	BAE	75.6	73.3
	BilBOWA	76.5	73.0
	NGBAE	**78.2**	**74.2**
English-German	Majority Class	40.9	44.7
	MT	68.1	67.4
	BWEs-Multitask	77.6	71.1
	BiCVM	86.9	74.3
	BAE	**91.8**	74.2
	BilBOWA	86.5	75.0
	NGBAE	91.3	**77.8**
English-French	Majority Class	22.4	23.0
	MT	76.3	71.1
	BWEs-Multitask	74.5	61.9
	BiCVM	78.4	68.8
	BAE	84.6	74.2
	BilBOWA	82.4	68.4
	NGBAE	**88.4**	**79.1**
English-Spanish	Majority Class	15.7	24.6
	MT	52.0	58.4
	BWEs-Multitask	51.3	63.0
	BiCVM	62.8	65.6
	BAE	59.4	64.4
	BilBOWA	61.7	65.1
	NGBAE	**65.9**	**71.0**

where $tran(x_i)$ denotes the target-side translation set that can be easily extracted from the word-aligned training corpus. To better depict our experimental results, we refer to translation probability distributions using BWEs-Multitask, BiCVM, BAE, BilBOWA and NGBAE as $p^{bm}_{bwe}(*|x_i)$, $p^{bicvm}_{bwe}(*|x_i)$, $p^{bae}_{bwe}(*|x_i)$, $p^{bilbowa}_{bwe}(*|x_i)$, and $p^{ngbae}_{bwe}(*|x_i)$, respectively .

2) **Translation Probabilities Based on Word Alignments**. We ran GIZA++ [37] toolkit to word-align the training corpus with the option "*grow-diag-final-and*", and followed Koehn et al. [26] to calculate the lexical translation probability as follows:

$$p_{mle}(y_j|x_i) = \frac{Count(x_i, y_j)}{\sum_{y_k \in tran(x_i)} Count(x_i, y_k)} \tag{15}$$

where $Count(x_i, y_j)$ is the alignment count of x_i and y_j.

Then, we computed the KL distances between $p^{*}_{bwe}(*|x_i)$ and $p_{mle}(^{*}|x_i)$, respectively. Intuitively, word alignments reflect the word-level bilingual semantic correspondence to some extent. If the translation probability distribution derived by our model is closer to that of word alignments, we have reasons to believe that NGBAE is able to learn better BWEs than the other models. Likewise, we defined the target-to-source translation probability distributions $p^{bm}_{bwe}(*|y_j)$, $p^{bicvm}_{bwe}(*|y_j)$, $p^{bae}_{bwe}(*|y_j)$, $p^{bilbowa}_{bwe}(*|y_j)$, $p^{ngbae}_{bwe}(*|y_j)$ and $p_{mle}(^{*}|y_j)$, and carried out the same experiments. In the specific implementation, we investigated the translation probability distributions for words with different word frequency thresholds: 0, 10, 50, and 100.

The final results of translation probability modeling in two directions are provided in Table 6. Compared with the other approaches, the proposed NGBAE enables the derived translation probability distributions to be closer to those of word alignments in most cases. These results robustly demonstrate the effectiveness of our model from another angle. When setting word frequency threshold as 0, we observe that the KL distances are smaller than the other cases. The underlying reason is that the translation probability distributions of the once-appearing words remain unchanged using different approaches, and therefore, the whole translation probability distributions are relatively similar.

4.4. Word embedding analysis

To illustrate the effectiveness of the proposed model at learning semantic BWEs, we show some examples in Table 7. Using the BWEs of the NGBAE model, we searched and listed the most semantically similar words of the example words.

Table 6
Experimental results of the translation probability modeling. We highlight the smallest KL distance in bold for each experiment.

Direction	Model	Word Frequency Threshold			
		0	10	50	100
English → Chinese	BWEs-Multitask	–	–	–	–
	BiCVM	0.843	1.538	1.586	1.683
	BAE	0.914	1.782	1.925	1.889
	BilBOWA	0.783	1.409	1.612	1.717
	NGBAE	**0.732**	**1.141**	**1.216**	**1.300**
Chinese → English	BWEs-Multitask	–	–	–	–
	BiCVM	0.800	1.658	1.748	1.761
	BAE	0.928	1.986	2.070	2.065
	BilBOWA	0.796	1.616	1.882	1.820
	NGBAE	**0.717**	**1.439**	**1.326**	**1.412**
English → German	BWEs-Multitask	0.665	1.320	1.555	1.664
	BiCVM	0.677	1.245	1.504	**1.603**
	BAE	0.704	1.309	1.611	1.660
	BilBOWA	0.582	1.184	1.614	1.784
	NGBAE	**0.480**	**1.076**	**1.482**	1.681
German → English	BWEs-Multitask	0.663	1.518	2.033	2.258
	BiCVM	0.610	1.307	1.613	1.826
	BAE	0.975	2.012	2.074	2.000
	BilBOWA	0.455	1.245	1.702	2.031
	NGBAE	**0.349**	**1.049**	**1.558**	**1.794**
English → French	BWEs-Multitask	0.801	1.669	2.001	2.152
	BiCVM	0.583	1.084	1.314	1.425
	BAE	0.765	1.447	1.635	1.672
	BilBOWA	0.540	1.036	1.379	1.487
	NGBAE	**0.509**	**1.006**	**1.178**	**1.234**
French → English	BWEs-Multitask	0.833	1.652	2.017	2.187
	BiCVM	0.557	1.085	1.502	1.681
	BAE	0.877	1.654	1.770	1.778
	BilBOWA	0.555	1.083	1.398	1.507
	NGBAE	**0.531**	**1.060**	**1.199**	**1.238**
English → Spanish	BWEs-Multitask	0.528	1.271	1.705	1.903
	BiCVM	0.740	1.067	1.230	1.340
	BAE	0.715	1.384	1.560	1.594
	BilBOWA	0.650	1.187	1.367	1.431
	NGBAE	**0.506**	**1.021**	**1.184**	**1.236**
Spanish → English	BWEs-Multitask	0.667	1.428	1.885	2.102
	BiCVM	0.810	1.125	1.316	1.440
	BAE	0.908	1.723	1.795	1.772
	BilBOWA	0.618	1.183	1.478	1.473
	NGBAE	**0.531**	**1.099**	**1.250**	**1.291**

We observe 1) that bilingual words that form a translation pair are close to each other and 2) that close words within the same language are also syntactically/semantically related.

Finally, we visualized the learned BWEs to check whether NGBAE is capable of learning some bilingual semantic correspondence. To do this, we selected the most frequent 40 Chinese words and their English translation words with the highest translation probability based on word alignments, and then used the *t-SNE* toolkit[12] [33] to show these BWEs. As shown in Fig. 3, we observe that although we do not exploit any word alignment information, semantically equivalent bilingual words locate closely to each other in the learned semantic space.

4.5. Additional experiments

To investigate the robustness of our model under various conditions, we conducted additional Chinese-English experiments.

4.5.1. Effect of varying the amount of supervised training data

We first evaluated the effect of varying the amount of supervised training data for training the classifiers: 100, 200, 500, 1,000, 5,000 and 10,000. The experimental results are summarized in Fig. 4. We find that NGBAE outperforms the other models at almost all data sizes.

[12] https://lvdmaaten.github.io/tsne/

Table 7
Examples of English words along with the closest words both in Chinese and English, selected according to the Euclidean distance between the embeddings learned by NGBAE.

Example Word	Language	Most Semantically Similar Words
january	English	*january, march, july*
	Chinese	yīyuè, sānyuè, yīrì
oil	English	*oil, petroleum, crude*
	Chinese	shíyóu, yuányóu, yóu
president	English	*president, presidential, inaugural*
	Chinese	zǒngtǒng, xièrèn, zhǔxí
microsoft	English	*microsoft, intel, pc*
	Chinese	wēiruǎn, ibm, pc
said	English	*said, saying, pointed*
	Chinese	shuō, tándào, bìngchēng
market	English	*market, markets, marketplace*
	Chinese	shìchǎng, hángyè, shìchǎnghuà

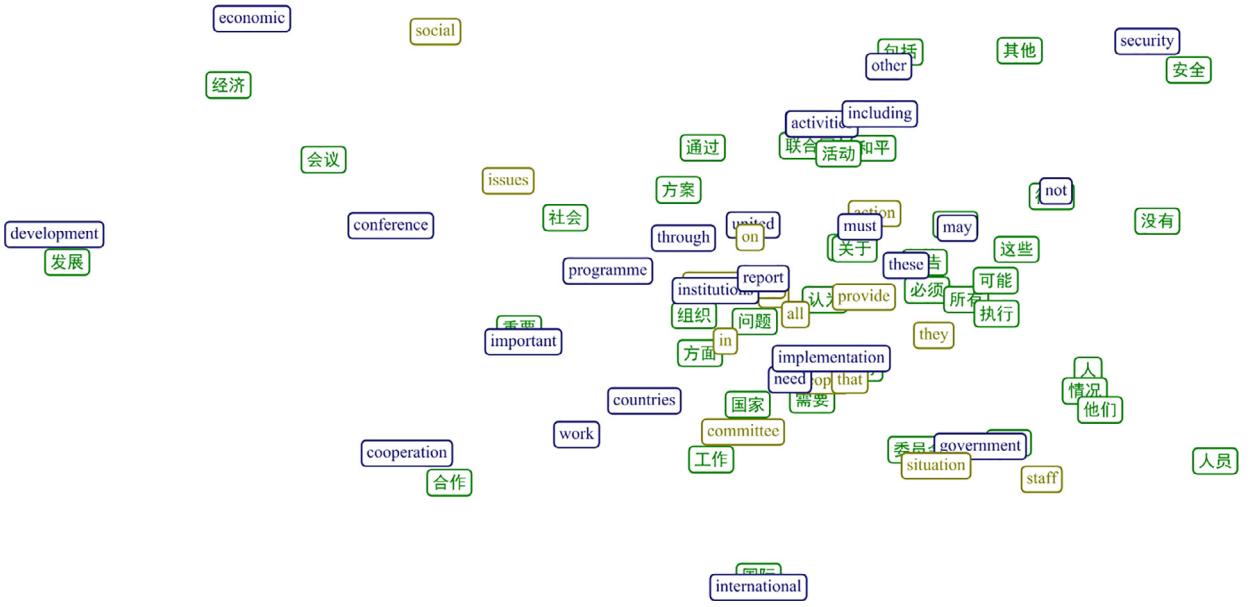

Fig. 3. 2-D visualization of the learned bilingual word embeddings: Chinese words are plotted in green boxes, while English words in blue/light yellow if their distances to the corresponding Chinese words with the highest translation probabilities are less/greater than the threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5.2. Effect of word embedding dimensions

We also conducted additional English-Chinese experiments to study the effects of word embedding dimensions, which are very important hyper-parameters in BWEs. We tried four different dimensions from 20, 40, 80, to 160. The experimental results are displayed in Table 8. We can observe that the performance of our model is always superior to those of the other models. These results demonstrate the effectiveness of our model once again.

4.5.3. Experimental results on large scale dataset

Finally, we studied the generality of our model on large scale dataset. To this end, we increased the English-Chinese dataset to contain 10 million sentence pairs. This corpus contains our original 2.5 million sentence pairs and 7.5 million sentence pairs from China Workshop on Machine Translation,[13] and repeated English-Chinese text classification and translation probability modeling experiments with the same hyper-parameters. Tables 9 and 10 provide the experimental results on cross-lingual document classification and translation probability modeling task, respectively. Obviously, no matter what the task is, our model significantly outperform the other models and is still effective on large scale dataset.

[13] It consists of casia2015 corpus, casict2011 corpus, casict2015 corpus, datum2015 corpus, neu2017 corpus from http://nlp.nju.edu.cn/cwmt-wmt/.

Table 8
Experimental results of the English-Chinese document classification with different dimensions. Please note that we did not reported the experimental results of BWEs-Multitask model, because there are no publicly available English-Chinese BWEs in [25].

Word Dimension	Model	**EN → CH**	**CH → EN**
20	Majority Class	37.6	37.6
	MT	62.1	59.8
	BWEs-Multitask	–	–
	BiCVM	69.1	69.2
	BAE	74.9	72.1
	BilBOWA	75.4	72.9
	NGBAE	**78.0**	**74.0**
40	Majority Class	37.6	37.6
	MT	63.3	60.4
	BWEs-Multitask	–	–
	BiCVM	70.0	70.4
	BAE	75.6	73.3
	BilBOWA	76.5	73.0
	NGBAE	**78.2**	**74.2**
80	Majority Class	37.6	37.6
	MT	63.9	60.7
	BWEs-Multitask	–	–
	BiCVM	70.2	70.8
	BAE	75.2	73.4
	BilBOWA	76.8	74.8
	NGBAE	**78.2**	**75.5**
160	Majority Class	37.6	37.6
	MT	63.2	60.1
	BWEs-Multitask	–	–
	BiCVM	70.0	69.4
	BAE	74.6	71.3
	BilBOWA	75.4	71.8
	NGBAE	**76.1**	**73.9**

Table 9
Experimental results of the English-Chinese document classification on large scale dateset.

Language Pair	Model	**EN → CH**	**EN → CH**	Training Time
English-Chinese	Majority Class	37.6	37.6	–
	MT	63.8	60.6	–
	BWEs-Multitask	–	–	–
	BiCVM	65.0	63.9	3.5 h
	BAE	76.1	73.6	64 h
	BilBOWA	76.4	74.7	6 h
	NGBAE	**79.4**	**76.3**	76 h

Table 10
Experimental results of the English-Chinese translation probability modeling on large scale dateset.

Direction	Model	Word Frequency Threshold			
		0	10	50	100
English → Chinese	BWEs-Multitask	–	–	–	–
	BiCVM	0.917	1.406	1.653	1.714
	BAE	0.923	1.799	1.865	1.887
	BilBOWA	0.864	1.405	1.684	1.816
	NGBAE	**0.808**	**1.329**	**1.404**	**1.445**
Chinese → English	BWEs-Multitask	–	–	–	–
	BiCVM	1.036	1.550	1.802	1.878
	BAE	1.102	1.811	1.967	2.087
	BilBOWA	0.969	1.461	1.743	1.884
	NGBAE	**0.883**	**1.353**	**1.472**	**1.544**

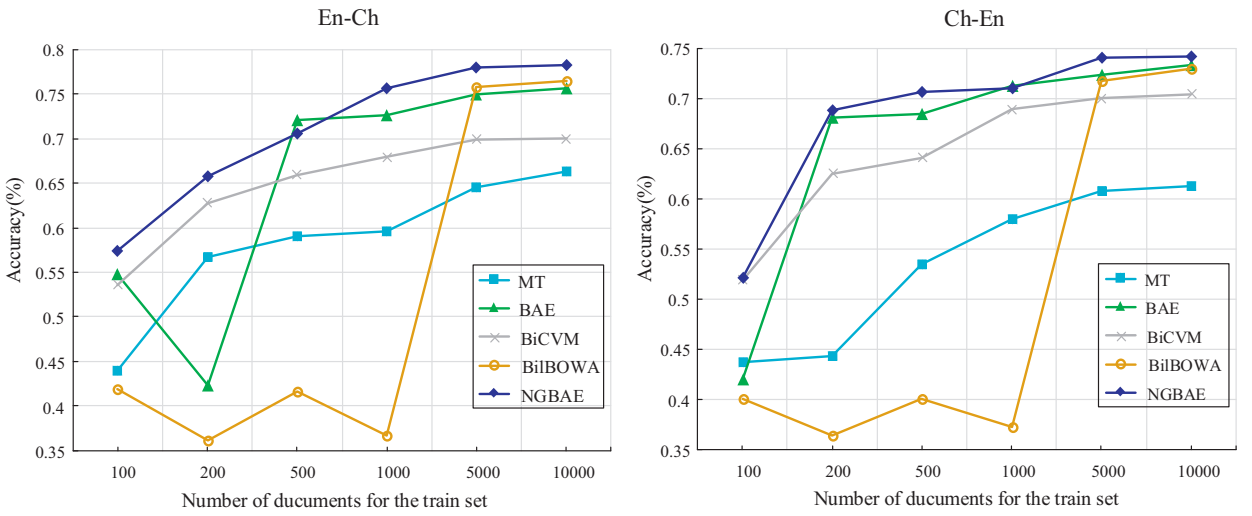

Fig. 4. English-Chinese classification accuracy results using different amounts of supervised training data.

5. Conclusion and future work

In this paper, we have presented a neural generative autoencoder for learning bilingual word embeddings, which incorporates a latent variable to explicitly model the underlying bilingual semantics. By virtue of the variational neural approach, we use a deep neural network to approximate the intractable posterior distribution. A reparameterization method is further introduced to bridge the gap between the recognition and generation model, enabling our model to be an end-to-end neural network which can be trained stochastically. Both extrinsic and intrinsic experiments on English-Chinese, English-German, English-French, and English-Spanish language pairs strongly demonstrate the effectiveness of our model.

In the future, we plan to exploit word alignments to improve our model. Additionally, the proposed model is general, and therefore we would like to check whether our model can be further improved using more complicated neural networks such as CNN and RNN.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (Nos. 61622209 and 61672440), Scientific Research Project of National Language Committee of China (Grant No. YB135-49), Natural Science Foundation of Fujian Province of China (No. 2016J05161).

References

[1] Y. Bengio, R. Ducharme, P. Vincent, C. Janvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (2003) 1137–1155.
[2] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. Montreal, Greedy layer-wise training of deep networks, Proc. of NIPS2007, 2007.
[3] P. Bhatia, R. Guthrie, J. Eisenstein, Morphological priors for probabilistic neural word embeddings, in: Proc. of EMNLP2016, 2016, pp. 490–500.
[4] S. Chandar, S. Lauly, H. Larochelle, M.M. Khapra, B. Ravindran, V. Raykar, A. Saha, An autoencoder approach to learning bilingual word representations, in: Proc. of NIPS2014, 2014, pp. 1853–1861.
[5] J. Chung, K. Kastner, L. Dinh, K. Goel, A.C. Courville, Y. Bengio, A recurrent latent variable model for sequential data, in: Proc. of NIPS2015, 2015, pp. 2980–2988.
[6] M. Collins, Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms, in: Proceedings of EMNLP 2002, 2002, pp. 1–8.
[7] R. Cotterell, H. Schütze, J. Eisner, Morphological smoothing and extrapolation of word embeddings, in: Proc. of ACL2016, 2016, pp. 1651–1660.
[8] J. Coulmance, J.-M. Marty, G. Wenzek, A. Benhalloum, Trans-gram, fast cross-lingual word-embeddings, in: Proc. of EMNLP2015, 2015, pp. 1109–1113.
[9] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, Technical report, EECS Department, University of California, Berkeley, 2010.
[10] L. Duong, H. Kanayama, T. Ma, S. Bird, T. Cohn, Learning crosslingual word embeddings without bilingual corpora, in: Proc. of EMNLP2016, 2016, pp. 1285–1295.
[11] M.A.M. Garcia, R.P. Rodriguez, L.A. Rifon, Wikipedia-based cross-language text classification, Inf. Sci 406 (2017) 12–28.
[12] S. Gouws, Y. Bengio, G. Corrado, Bilbowa: fast bilingual distributed representations without word alignments, in: Proc. of ICML2015, 2015, pp. 748–756.
[13] K. Gregor, I. Danihelka, A. Graves, D. Wierstra, DRAW: a recurrent neural network for image generation, CoRR abs/1502.04623 (2015).
[14] X. Gu, Y. Gu, H. Wu, Cascaded convolutional neural networks for aspect-based opinion summary, Neural Process. Lett. (2017).
[15] J. Guo, W. Che, D. Yarowsky, H. Wang, T. Liu, A representation learning framework for multi-source transfer parsing, in: Proc. of AAAI2016, 2016, pp. 2734–2740.
[16] K.M. Hermann, P. Blunsom, Multilingual distributed representations without word alignment, in: Proc. of ICLR2014, 2014.
[17] K.M. Hermann, P. Blunsom, Multilingual models for compositional distributed semantics, in: Proc. of ACL2014, 2014, pp. 58–68.
[18] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Trans. Ind. Electron. 62 (2015) 3742–3751.
[19] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (2015) 5659–5670.

[20] E. Huang, R. Socher, C. Manning, A. Ng, Improving word representations via global context and multiple word prototypes, in: Proc. of ACL2012, 2012, pp. 873–882.
[21] S. Ji, H. Yun, P. Yanardag, S. Matsushima, S.V.N. Vishwanathan, Wordrank: learning word embeddings via robust ranking, in: Proc. of EMNLP2016, 2016, pp. 658–668.
[22] D. Kim, C. Park, J. Oh, H. Yu, Deep hybrid recommender systems via exploiting document context and statistics of items, Inf. Sci. 417 (2017) 72–87.
[23] D.P. Kingma, S. Mohamed, D.J. Rezende, M. Welling, Semi-supervised learning with deep generative models, in: Proc. of NIPS2014, 2014, pp. 3581–3589.
[24] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: Proc. of ICLR2014, 2014.
[25] A. Klementiev, I. Titov, B. Bhattarai, Inducing crosslingual distributed representations of words, in: Proc. of COLING2012, 2012, pp. 1459–1474.
[26] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of NAACL 2003, 2003, pp. 48–54.
[27] T. Kočiský, K.M. Hermann, P. Blunsom, Learning bilingual word representations by marginalizing alignments, in: Proc. of ACL2014, 2014, pp. 224–229.
[28] W. Ling, C. Dyer, A.W. Black, I. Trancoso, R. Fermandez, S. Amir, L. Marujo, T. Luis, Finding function in form: compositional character models for open vocabulary word representation, in: Proc. of EMNLP2015, 2015, pp. 1520–1530.
[29] P. Liu, X. Qiu, X. Huang, Learning context-sensitive word embeddings with neural tensor skip-gram mode, in: Proc. of IJCAI2015, 2015, pp. 1284–1290.
[30] Y. Liu, Z. Liu, T.-s. Chua, M. Sun, Topical word embeddings, in: Proc. of AAAI2015, 2015, pp. 2418–2424.
[31] A. Lu, W. Wang, M. Bansal, K. Gimpel, K. Livescu, Deep multilingual correlation for improved word embeddings, in: Proc. of NAACL2015, 2015, pp. 250–256.
[32] M.-T. Luong, H. Pham, C.D. Manning, Bilingual word representations with monolingual quality in mind, in: Proc. of NAACL2015, 2015, pp. 151–159.
[33] L.v.d. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[34] Y. Miao, L. Yu, P. Blunsom, Neural variational inference for text processing, Arxiv preprint. abs/1511.06038 (2015).
[35] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Arxiv preprint. abs/1301.3781 (2013).
[36] T. Mikolov, Q.V. Le, I. Sutskever, Exploiting similarities among languages for machine translation, Arxiv preprint. abs/1309.4168 (2013).
[37] F.J. Och, H. Ney, A systematic comparison of various statistical alignment models, Comput. Ling. (2003) 19–51.
[38] T. Oshikiri, K. Fukui, H. Shimodaira, Cross-lingual word representations via spectral graph embeddings, in: Proc. of ACL2016, 2016, pp. 493–498.
[39] P. Qian, X. Qiu, X. Huang, Investigating language universal and specific properties in word embeddings, in: Proc. of ACL2016, 2016, pp. 1478–1488.
[40] D.J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, in: Proc. of ICML2014, 2014.
[41] T. Shi, Z. Liu, Y. Liu, M. Sun, Learning cross-lingual word embeddings via matrix co-factorization, in: Proc. of ACL2015, 2015, pp. 567–572.
[42] R. Socher, J. Pennington, E.H. Huang, A.Y. Ng, C.D. Manning, Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proc. of EMNLP2011, 2011, pp. 151–161.
[43] H. Soyer, P. Stenetorp, A. Aizawa, Leveraging monolingual data for crosslingual compositional word representations, in: Proc. of ICLR2015, 2015.
[44] K. Stratos, M. Collins, D. Hsu, Model-based word embeddings from decompositions of count matrices, in: Proc. of ACL2015, 2015, pp. 1282–1291.
[45] Theano Development Team, Theano: a Python framework for fast computation of mathematical expressions, arXiv e-prints abs/1605.02688 (2016).
[46] L. Tian, D.F. Wong, L.S. Chao, P. Quaresma, F. Oliveira, L. Yi, Um-corpus: a large english-chinese parallel corpus for statistical machine translation, in: Proc. of LREC, 2014, pp. 1837–1842.
[47] S. Upadhyay, M. Faruqui, C. Dyer, D. Roth, Cross-lingual models of word embeddings: an empirical comparison, in: Proc. of ACL2016, 2016, pp. 1661–1670.
[48] I. Vulić, M.-F. Moens, Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction, in: Proc. of ACL2015, 2015, pp. 719–725.
[49] I. Vulić, M.-F. Moens, Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings, in: Proc. of SIGIR2015, 2015, pp. 363–372.
[50] W. Yin, H. Schütze, Learning word meta-embeddings, in: Proc. of ACL2016, 2016, pp. 1351–1360.
[51] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE Trans. Cybern. (2016) 1–11.
[52] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, Iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Trans. Inf. Forensics Secur. (2016) 1005–1016.
[53] H. Zhou, L. Chen, F. Shi, D. Huang, Learning bilingual sentiment word embeddings for cross-language sentiment classification, in: Proc. of ACL2015, 2015, pp. 430–440.
[54] W.Y. Zou, R. Rocher, D. Cer, C.D. Manning, Bilingual word embeddings for phrase-based machine translation, in: Proc. of EMNLP2013, 2013, pp. 1393–1398.