Communications in Information and Systems
Volume 18, Number 1, 53–72, 2018

Topically-informed bilingually-constrained recursive autoencoders for statistical machine translation

Zhiwei Ruan and Rongrong Ji

Learning high-quality phrase vector representations is one of important research topics in statistical machine translation (SMT). Towards phrase embeddings, most existing works mainly explore syntactic and semantic clues among internal words within phrases, which are however insufficient for representation learning due to the lack of context information. In this paper, we propose topically-informed bilingually-constrained recursive autoencoders for SMT, which substantially extends the conventional bilingually-constrained recursive autoencoders by exploiting latent topics in two ways. First, we introduce topical contexts to induce topical phrase embeddings. Second, word topic assignments from a latent topic model are leveraged to constrain the learning of word and topic embeddings, both of which form the base of the contextual phrase embedding learning in the proposed model. Experiment results on Chinese-English translation show that the proposed model significantly improves the translation quality on NIST test sets.

1. Introduction

In the past decade, statistical machine translation (SMT) has made great progress and consequently attracted extensive research focus. Generally, translation models [3, 16] are trained with bilingual data, which suffer from the issue of data sparsity and the difficulty to exploit semantic information. Therefore, learning accurate semantic representations of translation units is crucial for SMT. With the rapid development of deep learning, it has become a trend to convert translation units into vector representations. To this end, most prior work mainly focus on learning bilingual word embeddings to improve individual components of SMT systems [1, 6, 9, 14, 18, 30, 31, 37, 42, 46]. However, SMT systems translate sentences with sequences of synchronous rules or phrases, rather than translating words separately. Hence, phrase-level

compact vector representation becomes a key issue in deep neural network based SMT.

Inspired by the success of monolingual phrase embeddings [5, 22, 25–27], many approaches have been presented to implement bilingual phrase embeddings for SMT [4, 8, 21, 28, 32, 44]. Similar to that in multi-modal cases [24, 33, 34, 36, 43], which are dedicated to learn a common space for data of different modalities, the intuition behind them is that a phrase and its correct translation should share the same semantic meaning, and thus, they should be embedded closely to each other in the shared embedding space. However, only semantic compositions of internal words within phrases are considered in the existing work, while the semantic information beyond phrases are ignored, which limits the potential of the learned phrase vector representations.

In this paper, we propose Topically-informed Bilingually-constrained Recursive Auto-encoders (TBRAE), which address the aforementioned issue in learning contextual bilingual phrase embeddings. Incorporating contextual clues into phrase vector representations, our model substantially extends the Bilingually-constrained Recursive Auto-encoders (BRAE) [44] by exploiting latent topics in two ways. Our first inspiration comes from the observation that the meanings of phrases are often context-dependent. Hence, we represent the document-level context of each phrase with its document-topic distribution, which can be incorporated with the RAE to produce the topical phrase embedding. Our second inspiration derives from the observation that the word topic assignments outputted by the latent topic model reflect the semantic correlations between words and topics, which can be used to constrain the learning of their embeddings. To this end, we design word-topic semantic constraints to encourage words with similar topic assignments to be placed closely in the embedding space. Comparing with BRAE, the TBRAE model not only considers the document-level context beyond the phrases, but also directly models the interactions between word and topic embeddings, both of which are the bases of topical phrase embeddings. To summarize, the main contributions of our work are the following:

- We enhance the representation capability of phrase embeddings by introducing the topic-based document-level context.
- We exploit the word topic assignments outputted by topic model to constrain the learning of word and topic embedding, both of which directly affect the final phrase semantic representations. To the best of our knowledge, this has not been investigated before.

- We integrate two phrase-level similarity features based on the TBRAE model to enhance a state-of-the-art SMT system, which achieves significant improvements on Chinese-English translation.

2. Background

We give brief introductions to the Latent Dirichlet Allocation (LDA) [2] and BRAE, both of which are related to the proposed TBRAE model.

2.1. LDA

In the last decade, topic models have drawn much attention and been applied successfully in various NLP tasks. Among these topic models, LDA is the most commonly used one at present, and therefore we use it to mine topics in our work. Based on the "bag-of-words" assumption, LDA views each document as a mixture of underlying topics, and generates each word according to the multinomial distribution conditioned on a topic. After training, LDA learns two types of parameters. The first one is the document-topic distribution recording the topic distribution of each document, which is often used to capture the document-level context in topic-based SMT [7, 11, 13, 39, 41, 45]. The second one relates to the topic-word distribution that represents each topic as a distribution over words. Based on this distribution, LDA samples a topic to generate each word in a document. The word topic assignments reflect the semantic correlations between words and topics, which can be used as the semantic constraints when learning phrase embeddings.

2.2. BRAE

The BRAE model is the bilingual variant of RAE, which jointly learns two RAEs for the source and target phrase embeddings. Figure 1 shows the framework of the RAE and BRAE models. Its basic idea is that translation equivalents share the same semantic meaning, and thus, they can supervise each other to learn their semantic vector representations.

Given the BRAE model with parameters θ, there are two kinds of errors involved for the phrase pair (f, e): (1) **reconstruction error** $E_{rec}(f, e; \theta)$ used to measure how well the learned vector representations $\vec{f}$ and $\vec{e}$ represent the phrases f and e, respectively. Similar to RAE, the reconstruction error of each phrase is defined as the sum of the reconstruction error at each node in its optimal binary tree, which is obtained in a greedy fashion [27]; (2) **semantic error** $E_{sem}(f, e; \theta)$ that evaluates the semantic distances between $\vec{f}$ and $\vec{e}$.

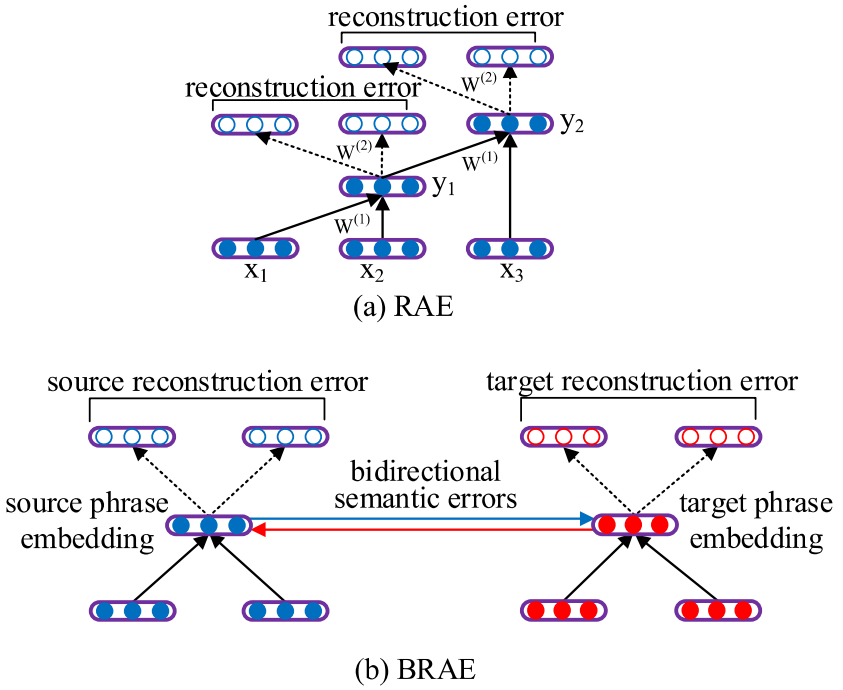

Figure 1: The illustration of the RAE and BRAE architectures.

Assuming there exist bidirectional transformations between the source- and target-side semantic embedding spaces, Zhang et al. first separately learn phrase embeddings for two languages, and then fine-tune them according to the semantic distance between translation equivalents [44]. Furthermore, to enhance the semantic error with positive and negative examples, Zhang et al. propose the *max-semantic-margin error* to minimize the semantic distance between translation equivalents and maximize that between non-translation pairs simultaneously [44].

3. The TBRAE model

In this section, we first give an overview of the TBRAE model, followed by the methods of modeling words, phrases and contexts, word-topic semantic constraints, respectively. Afterwards, we describe the model objective and the strategy for model training. Figure 2 provides the architecture of the TBRAE model, which is an extension of the BRAE [44]. It consists of four components: (1) ***two recursive auto encoders*** that separately summarize the semantic

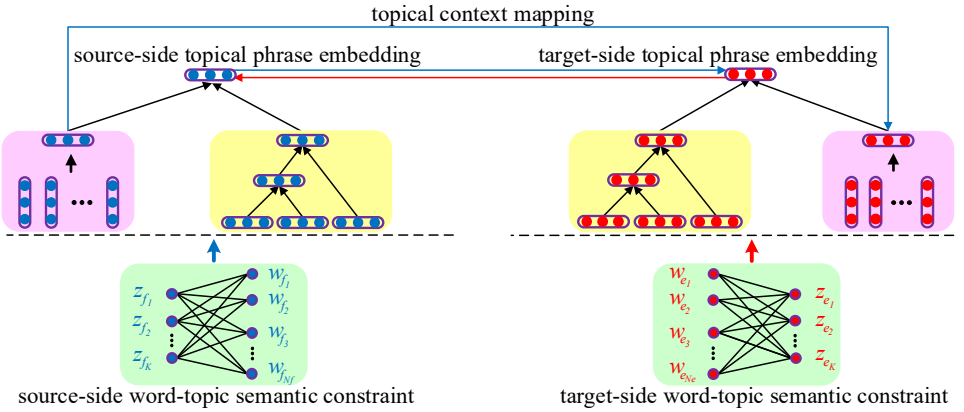

Figure 2: The illustration of the TBRAE architecture. w_* and z_* represent the words and topics for two languages, respectively. Rectangles in mauve stand for the context representation models while canary yellow and light green rectangles represent the RAE-based phrase embeddings and the word-topic semantic constraints separately.

meanings of the source and target phrases; (2) ***two context representation models*** which are respectively used to model the source-side and the target-side topical contexts; (3) ***bidirectional semantic constraints for topical phrase embeddings*** that minimize the bidirectional semantic distances between phrases and their translations in topical contexts; (4) ***two word-topic semantic constraints*** which exploit the word topic assignments to constrain word and topic embeddings in two languages, respectively.

Before training the model, we employ LDA to obtain the topic distributions of documents and the word topic assignments. The former is available for the topical phrase embeddings, while the latter is used to constrain the learning of word and topic embeddings.

3.1. RAE-based phrase modeling

In our model, each word in the vocabulary V corresponds to an n-dimensional real-valued vector, and all the vectors are stacked into a word embedding matrix $L_w \in \mathbb{R}^{n \times |V|}$. Regarding the phrase p as the meaningful composition of its internal words, we apply an RAE to learn its vector representation $\vec{p}$, as shown in Figure 1(a). The vector representations (x_1, x_2, x_3) of the ordered words in the phrase serve as the input to the RAE. For two children vectors

$\vec{x}_1$ and $\vec{x}_2$, the parent vector $\vec{y}_1$ is

$$\vec{y}_1 = f(W^{(1)}[\vec{x}_1; \vec{x}_2] + b^{(1)}) \tag{1}$$

where $W^{(1)} \in \mathbb{R}^{n \times 2n}$ is a parameter matrix, $b^{(1)} \in \mathbb{R}^{n \times 1}$ is a bias term, and f is an element-wise activation function such as $tanh(\cdot)$, which is used in our experiments. In this way, $\vec{y}_1$ is also a n-dimensional vector. To evaluate how well $\vec{y}_1$ represents its children, we reconstruct the vector representations $\vec{x}'_1$ and $\vec{x}'_2$ of the original children nodes in the following way:

$$[\vec{x}'_1; \vec{x}'_2] = f(W^{(2)}\vec{y}_1 + b^{(2)}) \tag{2}$$

where $W^{(2)} \in \mathbb{R}^{2n \times n}$ and $b^{(2)} \in \mathbb{R}^{2n \times 1}$. Considering two children vectors $\vec{y}_1$ and $\vec{x}_3$, we then further apply Eq. (1) to compute the parent vector $\vec{y}_2$. This combination and reconstruction process of auto-encoder repeats at each node until the entire phrase vector is generated. To obtain the optimal binary tree and phrase representation for p, we employ a greedy algorithm [27] to minimize the sum of the *reconstruction error* at each node in the binary tree $T(p)$:

$$E_{rec}(p; \theta) = \sum_{y \in T(p)} \frac{1}{2} \parallel [\vec{c}_1; \vec{c}_2]_y - [\vec{c}'_1; \vec{c}'_2]_y \parallel^2 \tag{3}$$

where y represents a non-leaf node in $T(p)$. It has two children vectors $\vec{c}_1$ and $\vec{c}_2$ which are reconstructed as the vectors $\vec{c}'_1$ and $\vec{c}'_2$.

3.2. Context modeling

Inspired by topic-based SMT, we use the document-topic distribution to represent the document-level context of phrases in each document. To make topical context computable, we regard each topic z in the topic set Z outputted by LDA as a pseudo word, of which semantic representation is also an n-dimensional real-valued vector. Thus, all topic embeddings can also be stacked to form a matrix $L_z \in \mathbb{R}^{n \times |Z|}$.

For the phrase p with the vector representation $\vec{p}$ in the document d, we define the semantic representation $\vec{dc}$ of its document-level context as the weighted sum of topic embeddings $\vec{z}$:

$$\vec{dc} = \sum_z p(z|d) \cdot \vec{z} \tag{4}$$

During training, we apply the above-mentioned approach to obtain the topical contexts of the phrase pair (f, e) in two languages, which are represented as $\vec{dc}_f$ and $\vec{dc}_e$, respectively. It has the advantage of exploiting extra monolingual corpora for better topic modeling. However, only source language documents are available during translation. To obtain the target-side topical context, we conduct topical context mapping in the source-to-target direction, and minimize the semantic distance $E_{tcm}(f|e;\theta)$ between the topical context embeddings in two languages as below:

$$E_{tcm}(f|e;\theta) = \frac{1}{2}\|\vec{dc}_e - f(W^{(3)}_{f2e} \cdot \vec{dc}_f + b^{(3)}_{f2e})\|^2 \tag{5}$$

where $W^{(3)}_{f2e} \in \mathbb{R}^{n\times n}$ is the mapping matrix and $b^{(3)}_{f2e} \in \mathbb{R}^{n\times 1}$ is the corresponding bias term.

3.3. Bilingually-constrained topical phrase embeddings

Similar to the BRAE model, our TBRAE model exploits cross-lingual semantic equivalence to learn the topical phrase embeddings. The intuition behind TBRAE is that the source- and target-side parts of each phrase pair share the same semantic meanings under topical contexts. Thus, they can consider each other as the gold vector representation to learn the phrasal semantic representations under the topical contexts in two languages.

To model this intuition, we first introduce a standard neural network layer to produce the topical phrase embedding $\vec{p}_{dc}$

$$\vec{p}_{dc} = g(W^{(4)}[\vec{p}; \vec{dc}] + b^{(3)}) \tag{6}$$

where $W^{(4)} \in \mathbb{R}^{n\times 2n}$ and $b^{(4)} \in \mathbb{R}^{n\times 1}$.

Notice that all the embedding parameters for two languages are learned separately, and therefore the produced source- and target-side topical phrase embeddings $\vec{f}_{dc}$ and $\vec{e}_{dc}$ are located in different vector spaces. For this, we apply a semantic transformation matrix to map a topical phrase embedding to the semantic space in the other language, and then minimize the semantic distance between its transformed vector and the embedding vector of its translation. Formally, we calculate the semantic distance between $\vec{f}_{dc}$ and $\vec{e}_{dc}$ in the target-side embedding space as follows:

$$E_{sem}(f|e;\theta) = \frac{1}{2}\|\vec{e}_{dc} - f(W^{(5)}_{f2e}\vec{f}_{dc} + b^{(5)}_{f2e})\|^2 \tag{7}$$

where $W_{f2e}^{(5)} \in \mathbb{R}^{n \times n}$ is the semantic mapping matrix in the source-to-target direction, and $b_f^{(5)} \in \mathbb{R}^{n \times 1}$ is the corresponding bias term.

Then, we follow Zhang et al. to introduce a *max-semantic-margin error* $E_{sem}^*(f|e;\theta)$ to exploit the positive example (f, e) and the negative example (f, e') simultaneously [44]:

$$E_{sem}^*(f|e;\theta) = max\{0, E_{sem}(f|e;\theta) - E_{sem}(f|e',\theta) + 1\} \tag{8}$$

where e' is another candidate translation of f or a bad translation that replaces the words in e with randomly chosen target language words.

Note that the above-mentioned semantic transformation can also be performed in the target-to-source direction. Therefore, the semantic distance is bidirectional. Due to the limitation of space, we do not describe it.

3.4. Word-topic semantic constraint modeling

As stated previously, word and topic embeddings constitute the basis of the contextual phrase semantic representations in TBRAE. Inspired by Tang et al., who study heterogeneous text network embeddings [29] , we exploit the semantic correlation between words and topics for better topical phrase embeddings.

The basic idea of our approach is as follows. In LDA model training, one topic is sampled to generate each word in the document. Thus, the topic assignments of each word reflect its semantic information in the topic space. According to the principle of maximum likelihood estimation, we first define the empirical conditional probability $\hat{p}(z|w)$ of topic z given word w based on the topic assignments of words

$$\hat{p}(z|w) = \frac{count(w, z)}{\sum_{z'} count(w, z')} \tag{9}$$

where $count(w, z)$ denotes the number of times z sampled to generate w. Thus, the semantic correlation between words w and w' in the topic space can actually be determined by their conditional distributions $\hat{p}(z|w)$ and $\hat{p}(z|w')$. If they are semantically related, then w and w' should be represented closely in the embedding space.

To preserve the semantic correlations between words when learning topical phrase embeddings, we also define the conditional probability $p(z|w)$ based

on the word and topic embeddings as follows:

$$p(z|w) = \frac{exp(\vec{w} \cdot \vec{z})}{\sum_{z'} exp(\vec{w} \cdot \vec{z'})} \tag{10}$$

where $\vec{w}$ denotes the embedding vector of word w. Then, we choose the Kullback-Leibler divergence $E_{wt}(w;\theta)$ to encourage $p(*|w)$ to be close to $\hat{p}(*|w)$:

$$E_{wt}(w;\theta) = \lambda_w \cdot d(p(*|w), \hat{p}(*|w)) \tag{11}$$

Here we introduce the weight λ_w that is defined as the frequency of w to distinguish the effects of different words. Omitting some constants, the objective function imposed on w becomes

$$E_{wt}(w;\theta) = \sum_{z} count(w,z) log p(z|w) \tag{12}$$

Similarly, the word-topic semantic constraints mentioned here can apply to the embedding learning in two languages.

3.5. The model objective

The above objective functions either act on each phrase pair or are used as global word-topic semantic constraints to learn the topical phrase embeddings.

For a phrase pair (f, e), there are three kinds of errors involved: (1) ***reconstruction error*** $E_{rec}(f, e; \theta)$: how well the learned vector representations $\vec{f}$ and $\vec{e}$ represent the phrases f and e respectively? (2) ***topical context mapping error*** $E_{tcm}(f|\ e; \theta)$: what is the semantic distance between the learned vector representations of topical contexts $\vec{dc}_f$ and $\vec{dc}_e$? (3) ***semantic error*** $E_{sem}(f, e; \theta)$: what is the semantic distance between the learned topical phrase embeddings $\vec{f}_{dc}$ and $\vec{e}_{dc}$? Formally, the joint error of (f, e) is defined as below:

$$\begin{aligned} E(f, e; \theta) = {} & \alpha \cdot E_{rec}(f, e; \theta) + \beta \cdot E_{tcm}(f|e; \theta) \\ & + (1 - \alpha - \beta) \cdot E_{sem}(f, e; \theta) \end{aligned} \tag{13}$$

where the hyper-parameters α and β are used to weight different error functions, $E_{rec}(f, e; \theta)$ is the sum of $E_{rec}(f; \theta)$ and $E_{rec}(e; \theta)$, while $E_{sem}(f, e; \theta)$ equals $E^*_{sem}(f|e; \theta)$ plus $E^*_{sem}(e|f; \theta)$.

Besides, we impose the word-topic semantic constraint, mentioned in Section 3.4, on words in two languages. Thus, the final objective over the training set D becomes:

$$J_{TBRAE} = \frac{1}{N} \sum_{(f,e)\in D} E(f,e;\theta) + \gamma \cdot \left(\frac{1}{|V_f|} \sum_{f\in V_f} E_{wt}(f;\theta) + \frac{1}{|V_e|} \sum_{e\in V_e} E_{wt}(e;\theta) \right) + R(\theta) \tag{14}$$

where γ is the hyper-parameter used to reflect the effect of the word-topic semantic constraints, $E_{wt}(*;\theta)$ denotes the error functions of the word-topic semantic constraints for two languages, and $R(\theta)$ is the regularization term involving the following parameter sets:[1] (1) θ_{L_w}: the word embedding matrices; (2) θ_{L_z}: the topic embedding matrices; (3) θ_{rec}: the RAE parameter matrices $W^{(1)}$, $W^{(2)}$ and bias terms $b^{(1)}$, $b^{(2)}$; (4) θ_{tcm}: the topical context mapping matrix $W^{(3)}_{f2e}$ and bias term $b^{(3)}_{f2e}$; (5) θ_{tp}: the topical phrase embedding parameter matrices $W^{(4)}$ and bias terms $b^{(4)}$; (6) θ_{sem}: the phrase semantic transformation matrices $W^{(5)}$ and bias terms $b^{(5)}$. Here we assign parameter sets different weights for regularization:

$$R(\theta) = \frac{\lambda_{L_w}}{2}||\theta_{L_w}||^2 + \frac{\lambda_{L_z}}{2}||\theta_{L_z}||^2 + \frac{\lambda_{rec}}{2}||\theta_{rec}||^2 + \frac{\lambda_{tcm}}{2}||\theta_{tcm}||^2 + \frac{\lambda_{tp}}{2}||\theta_{tp}||^2 + \frac{\lambda_{sem}}{2}||\theta_{sem}||^2 \tag{15}$$

We apply a similar co-training style algorithm as [44] to train the model parameters. Specifically, for each phrase pair, we fix its target-side contextual phrase representation to update its source-side parameters, and vice versa. In this process, we apply mini-batch to tune parameters based on gradients over the joint error. In each batch, we only use the forced decoding phrase pairs of a document for model training. Note that the word-topic semantic constraints are imposed on words rather than phrases. For this, we just consider the topic assignments of words occurring in the document and calculate their semantic constraint errors. This procedure repeats until either the joint error (shown in Eq. (14)) reaches a local minimum or the number of iterations is larger than the pre-defined one (25 is used in experiments).

[1]Note that the source and target languages have different sets of parameters.

4. Experiments

To validate the effectiveness of our TBRAE model, we conducted experiments on NIST Chinese-English translation task. Given a bilingual phrase (f, e) used to translate the source document, we successively implemented RAE-based phrase embeddings, topical context modelings and mapping, topical phrase embeddings, and semantic transformations. Finally, we calculated the cosine similarities between f and e in two directions, which are used as two additional features in the log-linear framework of SMT system.

4.1. Setup

Following Zhang et al. we chose a phrase-based translation system with a maximum entropy based reordering model (MEBTG) [35, 40] as our experiment system [44]. Our training data consists of the FBIS corpus and the Hansards part of LDC2004T07 corpus, with 1.0M parallel sentences (25.2M Chinese words and 29M English words). Following Zhang et al., we performed forced decoding with Leaving-One-Out [38] on the parallel sentences to collect 3.65M phrase pairs of high quality for the model training [44]. We trained a 5-gram language model on the Xinhua portion of Gigaword corpus using *SRILM* Toolkits.[2] Besides, we used NIST MT05 and MT06/MT08 data set as the development and test set, respectively. For translation results, we chose the case-insensitive BLEU-4 [23] as the evaluation metric, and performed paired bootstrap sampling [15] to calculate the statistical significance in BLEU score differences.

For the topic model, we used the GibbsLDA++[3] for estimation and inference. Following [10, 39], we set the parameters as follows: the topic number N_z=30, the hyper-parameters α_z=50/N_z and β_z=0.1, and the iteration number N_{iter}=1000.

In the experiments, we set the vector dimension as 50 and the learning rate as 0.01, as implemented in [44]. To tune the hyper-parameters, we randomly selected 250,000 forced decoding bilingual phrases as training set, 5000 as development set, and another 5000 as test set. We first incrementally drew α from 0.05 to 0.3, β from 0.05 to 0.25 with step 0.05, γ from 0.01 to 0.1 with step 0.01, and λ_* exponentially from 10^{-5} to 10^{-2}, and then determined the optimal hyper-parameters according to the overall error of the proposed model on the test set. Noting that too many hyper-parameters will lead to difficulties

[2]http://www.speech.sri.com/projects/srilm/download.html
[3]http://gibbslda.sourceforge.net/

in model training, we first learned word embeddings, RAE parameters, and some hyper-parameters such as λ_{L_w}, λ_{rec}, λ_{sem} using the BRAE model. Then, we fixed these parameters, hyper-parameters and tuned the others during the TBARE model training. Finally, we set α=0.1, β=0.1, γ=0.08, λ_{L_w}=10^{-5}, λ_{L_t}=10^{-5}, λ_{rec}=10^{-2}, λ_{tcm}=10^{-5}, λ_{tp}=10^{-3} and λ_{sem}=10^{-5}.

Model	**MT06**	**MT08**
MEGBT	29.66	21.52
BRAE	30.27	22.53
TBRAE(cont)	30.35^{**}	22.77^{**}
TBRAE(ss)	30.88^{**+}	23.34^{**++}
TBRAE(ts)	30.68^{**}	23.11^{**+}
TBRAE	31.16^{**++}	23.71^{**++}

Table 1: Experiment results on the test sets when setting dimension n=50. */**: significantly better than MEBTG (p<0.05/0.01), +/++: significantly better than BRAE (p<0.05/0.01)

4.2. Overall performance

First, we investigated the overall performance of the TBRAE model. Following Zhang et al., we set the dimensionality of the word and the topic embedding as 50 [44]. In addition to the conventional MEBTG system and the BRAE model, we also compared our model with its three variants: (1) TBRAE(cont) which explores only topical contexts while ignoring the word-topic semantic constraints; (2) TBRAE(ss) that uses only the cosine similarity feature in the source-side semantic space; (3) TBRAE(ts) that uses only the cosine similarity feature in the target-side semantic space.

Table 1 summarizes the comparison results of different models on the test sets. In all cases, TBRAE performs better than MEBTG, BRAE and TBRAE(cont), even if it use only one similarity feature. When using bidirectional similarity features together, TBRAE achieves the best performance, which is better than MEBTG, BRAE and TBRAE(cont) by 1.5/2.19, 0.89/1.18 and 0.81/0.94 BLEU points on the two sets, respectively. These experiments show that the exploitation of latent topics, especially word-topic constraints, contributes to outperforming MEBTG and BRAE both of which consider only internal semantic information of bilingual phrases.

4.3. Effect of embedding dimensionality

To investigate the generality of the TBRAE model, we tried four different dimensions from 25 to 100 with an increment of 25 each time.

TBRAE (Dimension)	**25**	**50**	**75**	**100**
MT06	30.95	31.16	30.87	30.70
MT08	23.54	23.71	23.32	23.43

Table 2: Experiment results for different dimensionalities.

The results are displayed in Table 2. We found that the TBRAE model is not consistently improved with the increment of dimensionality, and we can get satisfactory results when setting d= 50. These results are consistent with [28], which concludes that a larger dimension makes parameter tuning more difficult.

4.4. Result analysis

In order to know how our TBRAE model improves the SMT system more intuitively, we analyzed the experiment results from two angles.

Model	**BRAE**	**TBRAE(cont)**	**TBRAE**
ASG(ss)	0.0196	0.0267	0.1521
ASG(ts)	0.0147	-0.0043	0.1109

Table 3: The average cosine similarity gaps between MT05/06/08 phrase pairs and non-translation pairs. **ASG(ss)** and **ASG(ts)** denote average the similarity gaps in the source-side and the target-side semantic spaces, respectively.

Particularly, we first extracted phrase pairs from the word-aligned MT05, MT06, MT08 data sets and constructed a negative example for each phrase pair using the method described in Section 3.3. Then, we calculated the average cosine similarity gaps in two directions between phrase pairs and non-translation pairs using different models. Table 3 provides the calculation results. When using the TBRAE model, we observed that both of the average gaps in two directions are larger than other models. For these results, we

believe the TBRAE model is able to distinguish different translations more precisely by making better use of latent topics. These results echo the experiment results reported in Section 4.2.

SRC	... 反对党 大型 抗争 及 政局 [**混乱**] 。
Ref	... *massive opposition protests and political* [***chaos***] .
BRAE	... *large-scale protests of opposition parties and political* [***confusion***] .
TBRAE	... *large-scale protests of opposition parties and political* [***chaos***] .

Table 4: Translation result analysis.

To clearly understand the superiority of the TBRAE model on learning semantic phrase embeddings, we compared the best translations of the SMT system using the BRAE and TBRAE models. We found that our approach really improves translation quality by utilizing latent topics which are, on the contrary, ignored in BRAE. In the example shown in Table 4, BRAE fails to obtain the right translation for the word "混乱" due to the higher frequency of "*confusion*" than "*chaos*" in translation (203 vs 127). In fact, in the document related to *politic* topic, it is more likely that "混乱" is interpreted as "*chaos*" rather than "*confusion*", which is captured by our TBRAE model. As a result, the SMT system enhanced by TBRAE is able to correctly choose the translation "*chaos*" for "混乱".

5. Related work

Recently, learning bilingual text embeddings has attracted great attention, especially for SMT. Li et al. proposed an RAE-based ITG reordering classifier [17]. Kalchbrenner and Blunsom introduced recurrent continuous translation models that comprise a class of purely continuous sentence-level translation models [14]. Lu et al. applied the deep autoencoder to automatically learn new features for the phrase-based translation model [21]. Gao et al. presented a continuous-space phrase translation model to project bilingual phrases into the continuous-valued vector representations [8]. Zhang et al. proposed the BRAE model [44], which is the basis of our TBRAE model. Cho et al. proposed a novel Encoder-Decoder that consists of two RNNs for bilingual phrase embeddings [4] . Su et al. explored inner structures and semantic correspondence inside bilingual phrases for better phrase embeddings [28] . Hu et al.

proposed a context-dependent convolutional matching model to capture the semantic similarities between context-sensitive phrase pairs [12]. Significantly different from these studies, our model introduces latent topics to improve bilingual phrase embeddings.

Specifically, we exploited latent topics in two ways in TBRAE. Inspired by topic-based SMT [7, 11, 13, 39, 41, 45], we introduce topical contexts to implement the topical phrase embeddings. More importantly, we design two semantic constraints based on word topic assignments to refine the word and topic embeddings in our model. In this aspect, recently, Liu et al., introduced the latent topic model to globally cluster words into different topics according to their contexts [20] . Furthermore, Liu et al. used a tensor layer to capture more interactions between words and topics under different contexts [19]. Different from the methods mentioned here, our TBRAE model further exploits 2-step semantic correlations between words and topics for phrase embeddings, which, to the best of our knowledge, has never been investigated before.

6. Conclusions and future work

We have presented a topically-informed BRAE model which exploits latent topics to improve phrase embeddings for SMT. Topical contexts are introduced to enhance the determinativeness of phrase embeddings in different contexts. Word topic assignments are also used to constrain the learning of word and topic embeddings, both of which directly affect the learned contextual phrase embeddings. The experiment results on Chinese-English translation demonstrate the superiority of our model over a state-of-the-art baseline and BRAE [44].

There are some valuable research directions in the future. First, it is interesting to follow Liu et al. [19] to directly model the interaction between words and topics for bilingual phrase embeddings. Second, our model holds the potential to be extended to other phrase-based and even syntax-based systems.

7. Acknowledgments

This work is supported by the National Key R&D Program (No.2017YFC 0113000, and No.2016YFB1001503), Nature Science Foundation of China (No.U1705262, No.61772443, and No.61572410), Post Doctoral Innovative Talent Support Program under Grant BX201600094, China Post-Doctoral Science Foundation under Grant 2017M612134, Scientific Research Project

of National Language Committee of China (Grant No. YB135-49), and Nature Science Foundation of Fujian Province, China (No. 2017J01125 and No. 2018J01106).

References

[1] M. Auli, M. Galley, C. Quirk, and G. Zweig, *Joint language and translation modeling with recurrent neural networks*, in: Proc. of EMNLP 2013 (2013), 1044–1054.

[2] D. M. Blei, *Latent Dirichlet allocation*, Journal of Machine Learning (2003), 993–1022.

[3] D. Chiang, *Hierarchical phrase-based translation*, Computational Linguistics (2007), 201–228.

[4] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using RNN encoder–decoder for statistical machine translation*, in: Proc. of EMNLP 2014 (2014), 1724–1734.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, *Natural language processing (almost) from scratch*, Journal of Machine Learning Research **16** (2011), 2493–2537.

[6] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul, *Fast and robust neural network joint models for statistical machine translation*, in: Proc. of ACL 2014 (2014), 1370–1380.

[7] V. Eidelman, J. Boyd-Graber, and P. Resnik, *Topic models for dynamic translation model adaptation*, in: Proc. of ACL 2012 (2012), Short paper, 115–119.

[8] J. Gao, X. He, W.-t. Yih, and L. Deng, *Learning continuous phrase representations for translation modeling*, in: Proc. of ACL 2014 (2014), 699–709.

[9] E. Garmash and C. Monz, *Dependency-based bilingual language models for reordering in statistical machine translation*, in: Proc. of EMNLP 2014 (2014), 1689–1700.

[10] T. L. Griffiths and M. Steyvers, *Finding scientific topics*, in: Proc. of the National Academy of Sciences 2004 (2004).

[11] E. Hasler, P. Blunsom, P. Koehn, and B. Haddow, *Dynamic topic adaptation for phrase-based MT.*, in: Proc. of EACL 2014 (2014), 328–337.

[12] B. Hu, Z. Tu, Z. Lu, H. Li, and Q. Chen, *Context-dependent translation selection using convolutional neural network*, in: Proc. of ACL2015 Short Paper (2015), 536–541.

[13] Y. Hu, K. Zhai, V. Eidelman, and J. Boyd-Graber, *Polylingual tree-based topic models for translation domain adaptation*, in: Proc. of ACL2014 (2014), 1166–1176.

[14] N. Kalchbrenner and P. Blunsom, *Recurrent continuous translation models*, in: Proc. of EMNLP 2013 (2013), 1700–1709.

[15] P. Koehn, *Statistical significance tests for machine translation evaluation*, in: Proc. of EMNLP 2004 (2004), 388–395.

[16] P. Koehn, F. J. Och, and D. Marcu, *Statistical phrase-based translation*, in: Proc. of NAACL 2003 (2003), 48–54.

[17] P. Li, Y. Liu, and M. Sun, *Recursive autoencoders for ITG-based translation*, in: Proc. of EMNLP 2013, 567–577 (2013).

[18] L. Liu, T. Watanabe, E. Sumita, and T. Zhao, *Additive neural networks for statistical machine translation*, in: Proc. of ACL 2013 (2013), 791–801.

[19] P. Liu, X. Qiu, and X. Huang, *Learning context-sensitive word embeddings with neural tensor skip-gram model*, in: Proc. of IJCAI 2015 (2015), 1284–1290.

[20] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, *Topical word embeddings.*, in: Proc. of AAAI 2015 (2015), 2418–2424.

[21] S. Lu, Z. Chen, and B. Xu, *Learning new semi-supervised deep auto-encoder features for statistical machine translation*, in: Proc. of ACL 2014 (2014), 122–132.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in: Proc. NIPS 2013 (2013), 3111–3119.

[23] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *Bleu: A method for automatic evaluation of machine translation*, in: Proc. of ACL 2002 (2002), 311–318.

[24] Y. Peng, X. Huang, and J. Qi, *Cross-media shared representation by hierarchical learning with multiple deep networks*, in: IJCAI, 3846–3853 (2016).

[25] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, *Dynamic pooling and unfolding recursive autoencoders for paraphrase detection*, in: Proc. of NIPS 2011 (2011).

[26] R. Socher, C. D. Manning, and A. Y. Ng, *Learning continuous phrase representations and syntactic parsing with recursive neural networks*, in: Proc. of the NIPS 2010 Workshop (2010), 1–9.

[27] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, *Semi-supervised recursive autoencoders for predicting sentiment distributions*, in: Proc. of EMNLP 2011 (2011), 151–161.

[28] J. Su, D. Xiong, B. Zhang, Y. Liu, J. Yao, and M. Zhang, *Bilingual correspondence recursive autoencoder for statistical machine translation*, in: Proc. of EMNLP 2015 (2015), 1248–1258.

[29] J. Tang, M. Qu, and Q. Mei, *PTE: Predictive text embedding through large-scale heterogeneous text networks*, in: Proc. of KDD 2015 (2015), 1165–1174.

[30] K. M. Tran, A. Bisazza, and C. Monz, *Word translation prediction for morphologically rich languages with bilingual neural networks*, in: Proc. of EMNLP 2014 (2014), 1676–1688.

[31] A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang, *Decoding with large-scale neural language models improves translation*, in: Proc. of EMNLP 2013 (2013), 1387–1392.

[32] X. Wang, D. Xiong, and M. Zhang, *Learning semantic representations for nonterminals in hierarchical phrase-based translation*, in: Proc. of EMNLP 2015 (2015), 1391–1400.

[33] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, *Cross-modal retrieval with cnn visual features: A new baseline*, IEEE transactions on cybernetics **47** (2017), no. 2, 449–460.

[34] Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, and S. Yan, *Modality-dependent cross-media retrieval*, ACM Transactions on Intelligent Systems and Technology (TIST) **7** (2016), no. 4, 57.

[35] D. Wu, *Stochastic inversion transduction grammars and bilingual parsing of parallel corpora*, Computational Linguistics **23** (1997), no. 3, 377–403.

[36] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, *Sparse multi-modal hashing*, IEEE Transactions on Multimedia **16** (2014), no. 2, 427–439.

[37] H. Wu, D. Dong, X. Hu, D. Yu, W. He, H. Wu, H. Wang, and T. Liu, *Improve statistical machine translation with context-sensitive bilingual semantic embedding model*, in: Proc. of EMNLP 2014 (2014), 142–146.

[38] J. Wuebker, A. Mauser, and H. Ney, *Training phrase translation models with leaving-one-out*, in: Proc. of ACL 2010 (2010), 475–484.

[39] X. Xiao, D. Xiong, M. Zhang, Q. Liu, and S. Lin, *A topic similarity model for hierarchical phrase-based translation*, in: Proc. of ACL 2012 (2012), 750–758.

[40] D. Xiong, Q. Liu, and S. Lin, *Maximum entropy based phrase reordering model for statistical machine translation*, in: Proc. of ACL 2006 (2006), 521–528.

[41] D. Xiong and M. Zhang, *A topic-based coherence model for statistical machine translation*, in: Proc. of AAAI 2013 (2013).

[42] N. Yang, S. Liu, M. Li, M. Zhou, and N. Yu, *Word alignment modeling with context dependent deep neural network*, in: Proc. of ACL 2013 (2013), 166–175.

[43] Y. Yang, Y.-T. Zhuang, F. Wu, and Y.-H. Pan, *Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval*, IEEE Transactions on Multimedia **10** (2008), no. 3, 437–446.

[44] J. Zhang, S. Liu, M. Li, M. Zhou, and C. Zong, *Bilingually-constrained phrase embeddings for machine translation*, in: Proc. of ACL 2014 (2014), 111–121.

[45] B. Zhao and E. P. Xing, *BiTAM: Bilingual topic admixture models for word alignment*, in: Proc. of ACL 2006 (2006), 969–976.

[46] W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, *Bilingual word embeddings for phrase-based machine translation*, in: Proc. of EMNLP 2013 (2013), 1393–1398.

Fujian Key Laboratory of Sensing and Computing for Smart City
School of Information Science and Engineering
Xiamen University, Fujian 361005, China
E-mail address: zw_ruan@stu.xmu.edu.cn

Fujian Key Laboratory of Sensing and Computing for Smart City
School of Information Science and Engineering
Xiamen University, 361005, China
E-mail address: rrji@xmu.edu.cn