Bilingual Autoencoders with Global Descriptors for Modeling Parallel Sentences

Biao Zhang[1], **Deyi Xiong**[2*], **Jinsong Su**[1], **Hong Duan**[1] **and Min Zhang**[2]
Xiamen University, Xiamen, China 361005[1]
Soochow University, Suzhou, China 215006[2]
zb@stu.xmu.edu.cn, {jssu,hduan}@xmu.edu.cn
{dyxiong, minzhang}@suda.edu.cn

Abstract

Parallel sentence representations are important for bilingual and cross-lingual tasks in natural language processing. In this paper, we explore a bilingual autoencoder approach to model parallel sentences. We extract sentence-level global descriptors (e.g. *min*, *max*) from word embeddings, and construct two monolingual autoencoders over these descriptors on the source and target language. In order to tightly connect the two autoencoders with bilingual correspondences, we force them to share the same decoding parameters and minimize a corpus-level semantic distance between the two languages. Being optimized towards a joint objective function of reconstruction and semantic errors, our bilingual antoencoder is able to learn continuous-valued latent representations for parallel sentences. Experiments on both intrinsic and extrinsic evaluations on statistical machine translation tasks show that our autoencoder achieves substantial improvements over the baselines.

1 Introduction

Neural sentence modeling that learns continuous-valued vector representations for sentences in a low-dimensional latent semantic space, has recently attracted considerable interests in the field of nature language processing (NLP). A variety of models have been proposed (Collobert and Weston, 2008; Socher et al., 2011; Mikolov et al., 2011; Hermann and Blunsom, 2013; Kalchbrenner and Blunsom, 2013; Kalchbrenner et al., 2014; Kim, 2014; Ma et al., 2015; Tai et al., 2015; Zhang et al., 2015a). Most of these models, however, focus on monolingual cases where sentences from a single language are modeled. They do not explore semantic correspondences among parallel sentences. On account of this, monolingual neural sentence models are not naturally fit for bilingual or cross-lingual NLP tasks, such as machine translation, cross-lingual classification and information retrieval.

In order to induce sentence representations in a bilingual rather than monolingual semantic space, researchers have proposed a few approaches, following efforts of bilingual word embeddings (Klementiev et al., 2012; Zou et al., 2013). These studies either explore *recursive/recurrent* neural networks (Zhang et al., 2014; Su et al., 2015) or *bag-of-words* based neural networks (Yih et al., 2011; Chandar et al., 2014; Lauly et al., 2014; Hermann and Blunsom, 2014; Zhou et al., 2015) to learn bilingual sentence representations. The former recursively compose sentences from the bottom up, taking into account bilingual constraints from word alignments. Due to the complexity, they are not easy to be scalable. Additionally, they also suffer from errors and noises of word alignments. In contrast, the latter are relatively simple and scalable. However, they often heavily rely on only one descriptor of the bag-of-words embeddings (e.g. the *avg* representation) and hence are weak in capturing fine-grained complex linguistic phenomena. Therefore we believe that modeling parallel sentences still remains a serious challenge.

Inspired by works on multimodal autoencoders (Ngiam et al., 2011; Wang et al., 2014; Feng et al., 2014; Feng et al., 2015), we explore a novel bilingual autoencoder to model parallel sentences, which is able to incorporate global semantic information into sentence representations. The overall architecture

[*] Corresponding author.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers,
pages 2548–2558, Osaka, Japan, December 11-17 2016.

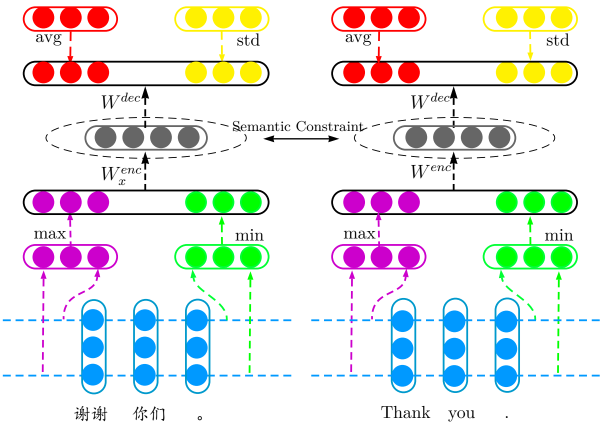

Figure 1: The architecture of our bilingual autoencoder visualized with a parallel sentence pair. Word embeddings are represented in blue color, while the *avg*, *std*, *min* and *max* descriptors are indicated in red, yellow, green and purple colors, respectively. Specifically, we use gray color to denote the hidden layer. The subscripts x and y indicate the source and target language respectively, and the superscripts enc and dec indicate the encoding and decoding process. Each dash line with its corresponding color depicts the information related with that specific descriptor, and the black ellipse around the hidden layer means that the bilingual semantic constraint is built at the corpus rather than sentence level.

is illustrated in Figure 1.[1] It is fast and scalable as we do not use complex recursiveness or recurrence. Comparing with conventional bag-of-words based autoencoders, our model explores several complementary sentence-level statistical descriptors, i.e., *min* (minimum), *max* (maximum), and *avg* (average), *std* (standard deviation) computed over all word vectors, to capture high-level global features. These descriptors alleviate the weakness of conventional bag-of-words representations in feature extraction and their insensitiveness to semantic constraints between sentences. The proposed autoencoder further encodes global descriptors (e.g., *min*, *max*) into hidden representations, and then decodes them into the other descriptors (e.g., *avg*, *std*). In order to capture bilingual correspondences, we force our autoencoder to share the decoding parameters across two languages (i.e., the same W^{dec} in Figure 1), and further optimize the parameters with respect to a corpus-level semantic constraints between the source and target language. This architecture tightly connects two monolingual autoencoders and bridges the gaps between the source and target semantic spaces.

To evaluate the effectiveness of our proposed bilingual autoencoder, we conduct both intrinsic and extrinsic evaluations on statistical machine translation (SMT) tasks. The intrinsic evaluation measures the capability of our model in semantic similarity calculation, while the extrinsic evaluation examines whether our model can be used to improve machine translation quality. Results on the NIST 2006 and 2008 datasets show that our autoencoder can significantly outperform the baseline methods. The main contributions of our work lie in the following three aspects:

- The proposed autoencoder learns bilingual representations for parallel sentences using global descriptors. To the best of our knowledge, this architecture has never been investigated before.
- We incorporate the corpus- rather than sentence-level semantic constraint into sentence embeddings, and share the decoding parameters to further bridge the semantic spaces of two different languages.
- Our model does not rely on word alignments. It is scalable, simple yet effective.

2 Related Work

Previous studies that are related to our work can be roughly divided into three groups: neural sentence modeling, multimodal autoencoders and bilingual autoencoders. We will briefly describe them in this

[1] For illustration, we set the dimensionality of word embedding to 3. Besides, we also omit the dash lines from word embeddings to *avg* and *std* descriptors. Without loss of generality, throughout this paper we assume the input and target to be *min*, *max* and *avg*, *std* respectively.

section and highlight the differences of our work from them.

neural sentence modeling (NSM): NSM is able to provide distributed sentence representations for many NLP tasks. One natural approach to NSM is to produce sentence representations from word embeddings with recursive or recurrent neural networks via composition (Socher et al., 2011; Mikolov et al., 2011; Hermann and Blunsom, 2013). To preserve sequence information over time, Tai et al. (2015) further incorporate LSTMs into tree-structured recursive networks. Unlike these composition-based methods, convolutional neural models utilize convolution layers to explicitly capture task-specific n-grams (Collobert and Weston, 2008; Kalchbrenner et al., 2014; Kim, 2014; Ma et al., 2015; Zhang et al., 2015a). Kalchbrenner and Blunsom (2013) combine convolutional and recurrent neural networks to model discourse representations. Different from these monolingual studies, our model aims at generating bilingual sentence representations. Furthermore we use autoencoders instead of recursive or recurrent neural networks. The descriptors in our model can be considered as a shallow convolution layer over an entire sentence (Zhang et al., 2015a).

multimodal autoencoders: Sentence modeling is performed only on text modality. However, information from different modalities might be complementary to each other. Motivated by this, Ngiam et al. (2011) develop deep multimodal autoencoders based on restricted Boltzmann machines (RBM) to jointly learn features over audio and video modalities. Instead of RBM, Wang et al. (2014) exploit stacked autoencoders for multimodal retrieval, and Feng et al. (2014) propose correspondence autoencoders for cross-modal retrieval. Feng et al. (2015) further investigate the utilization of deep correspondence RBM for cross-modal retrieval.

bilingual autoencoders: As a special case of multimodal autoencoders where each language is viewed as a modality, bilingual autoencoders have drawn attention in recent years. Based on recursive autoencoders, Zhang et al. (2014) incorporate phrase-level bilingual constraints into phrase representation learning. Along this line, Su et al. (2015) further exploit bilingual subphrase correspondences via word alignments for learning bilingual phrase structures and representations. Dissimilarly, Zhou et al. (2015) employ denoising autoencoders to learn bilingual embeddings. Lauly et al. (2014) directly perform encoding and decoding between source and target sentences based on bag-of-words representations. Chandar et al. (2014) exploit joint reconstruction and cross-lingual correlations on bilingual autoencoders. Unlike these models, our model builds autoencoders on global descriptors extracted from word embeddings, and incorporates bilingual semantics via corpus-level bilingual constraints into the hidden layer.

Our work is similar to the work of Zhang et al. (2015b) in that we both aim at modeling parallel sentences under the semantic constraints. The difference lies on the following two aspects: 1) they employ a chunk-based convolutional neural network to model sentences, while we use a much simpler autoencoder network; and 2) they impose the bilingual constraint strictly on the sentence level. Instead, we explore a much more relax constraints on the corpus level. As Chandar et al. (2014)'s bilingual autoencoder is most closely related to ours, we use their autoencoder as our baseline and describe it with more details in the next section.

3 Bilingual Bag-of-Words Autoencoder

The Bilingual Bag-of-Words Autoencoder (BBoWAE) (Chandar et al., 2014) builds two separate feed-forward autoencoders based on bag-of-words representations. The two autoencoders are jointly reconstructed with cross-lingual correlations. Figure 2 shows the overall architecture. Given a parallel sentence pair $(\mathbf{x}, \mathbf{y})$, BBoWAE model first generates corresponding bag-of-words based sentence representations, and then encodes them into the hidden layer:

$$a(\mathbf{x}) = c + W^{\mathcal{X}} v(\mathbf{x}),\ \phi(\mathbf{x}) = h(a(\mathbf{x})); \quad a(\mathbf{y}) = c + W^{\mathcal{Y}} v(\mathbf{y}),\ \phi(\mathbf{y}) = h(a(\mathbf{y})) \tag{1}$$

where $v(\cdot)$ converts sentence into a fixed-size but sparse binary vector, $W^{\mathcal{X}}$ and $W^{\mathcal{Y}}$ implicitly represent language specific word embeddings, c is the bias term shared by both autoencoders and $h(\cdot)$ is an element-wise non-linear function.

Upon these hidden layers, BBowAE model further performs a reconstruction of the original sentence

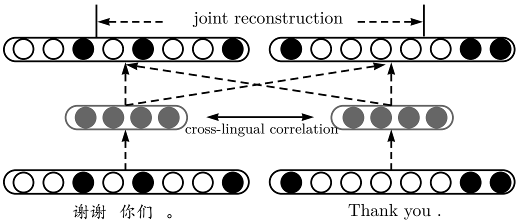

Figure 2: The architecture of BBoWAE model. We use black and white colors to indicate 1 and 0 respectively, which form the bag-of-words representation of a sentence. Notice that the gray nodes are real-valued.

in both languages:

$$\hat{v}(\mathbf{x})_{\mathbf{x}} = h(W^{\mathcal{X}^T}\phi(\mathbf{x}) + b^{\mathcal{X}}), \hat{v}(\mathbf{y})_{\mathbf{x}} = h(W^{\mathcal{Y}^T}\phi(\mathbf{x}) + b^{\mathcal{Y}}) \tag{2}$$

$$\hat{v}(\mathbf{x})_{\mathbf{y}} = h(W^{\mathcal{X}^T}\phi(\mathbf{y}) + b^{\mathcal{X}}), \hat{v}(\mathbf{y})_{\mathbf{y}} = h(W^{\mathcal{Y}^T}\phi(\mathbf{y}) + b^{\mathcal{Y}}) \tag{3}$$

where $b^{\mathcal{X}}$ and $b^{\mathcal{Y}}$ are bias terms, $h(\cdot)$ here is the sigmoid non-linearity and $\hat{v}(\mathbf{x})_{\mathbf{y}}$ denotes the reconstructed representation of $v(\mathbf{x})$ from $v(\mathbf{y})$. The other three reconstructed vectors are similarly defined.

To favour more meaningful bilingual representations, BBoWAE incorporates a cross-lingual correlation error and different reconstruction errors into the objective function:

$$\ell(\mathbf{x}, \mathbf{y}) + \ell(\mathbf{y}, \mathbf{x}) + \ell(\mathbf{x}, \mathbf{x}) + \ell(\mathbf{y}, \mathbf{y}) + \beta\ell([\mathbf{x}, \mathbf{y}], [\mathbf{x}, \mathbf{y}]) - \lambda cor(a(\mathbf{x}), a(\mathbf{y})) \tag{4}$$

where β, λ are hyperparameters, and $[\mathbf{x}, \mathbf{y}]$ represents the concatenation of the two sentences. Each loss function $\ell(\cdot)$ is the cross-entropy error between the original bag-of-words representation $v(\cdot)$ and reconstructed representation $\hat{v}(\cdot)$. (e.g. $v(\mathbf{y})$ vs. $\hat{v}(\mathbf{y})_{\mathbf{x}}$).

The term $\ell([\mathbf{x}, \mathbf{y}], [\mathbf{x}, \mathbf{y}])$ is the joint reconstruction term, where the two sentences are simultaneously presented as input and reconstructed; and the term $cor(a(\mathbf{x}), a(\mathbf{y}))$ is the sum of scalar correlations between $a(\mathbf{x})$ and $a(\mathbf{y})$ across all dimensions. To obtain a stochastic estimate of the correlation, small mini-batches are used during training.

Notice that the sentence representation $a(\mathbf{x})$/$a(\mathbf{y})$ is actually the *sum* of bag-of-words embeddings. Although BBoWAE is effective in learning bilingual word embeddings, it is weak in modeling parallel sentences due to the well known insufficiency of the sum representations. We instead resort to multiple global descriptors over word embeddings and further explore more suitable bilingual constraints.

4 Bilingual Autoencoder with Global Descriptors

We describe our model in two phases: autoencoder with global descriptors and bilingual semantic constraints. The former constitutes the basic structure for modeling monolingual sentences, while the latter explores bilingual constraints to connect the two autoencoders. After the description of the architecture, we present details for parameter inference.

4.1 Autoencoder with Global Descriptors

Our autoencoder is built upon distributed word embeddings, where each word in vocabulary V corresponds to a d-dimensional dense, real-valued vector, and all word vectors are stacked into a word embedding matrix $L \in \mathbb{R}^{d\times|V|}$, where $|V|$ is the vocabulary size.

Given an ordered list of n words in a sentence, we retrieve the i-th word representation from L with its corresponding vocabulary index I_i: $e(x_i) = L_{:,I_i} \in \mathbb{R}^d$. All word vectors in the sentence $\mathbf{x}$ produce the following output matrix: $M = (e(x_1), e(x_2), \ldots, e(x_n)) \in \mathbb{R}^{d\times n}$. From this matrix M, we investigate four sentence-level statistical descriptors: *min*, *max*, *avg* and *std* in each row r as follows:

$$g_r^{min} = \min{(M_{r,1:n})}, g_r^{max} = \max{(M_{r,1:n})}, g_r^{avg} = \frac{1}{n}\sum_{i=1}^{n} M_{r,i}, g_r^{std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_{r,i} - g_r^{avg})^2} \tag{5}$$

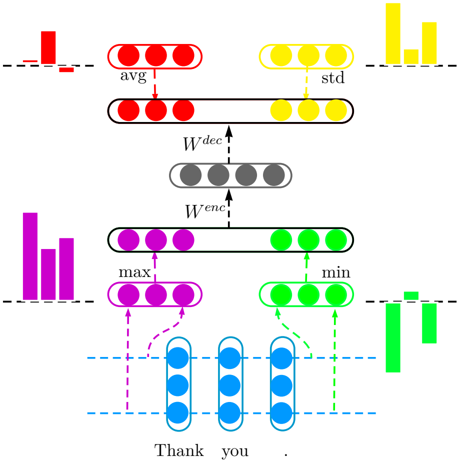

Figure 3: Monolingual autoencoder with global descriptors. The colored histogram on each corner indicates the corresponding descriptor for a toy example. Notice that the black dash line in each histogram represents the value 0.

Figure 3 illustrates the distribution of each descriptor for a toy example[2], from which we can find that these descriptors represent different aspects of the same sentence and are complementary to each other.

After obtaining these global descriptors, we select some of them as our encoding input (e.g. *min*, *max*), and leave the others as our decoding target (e.g. *std*, *avg*). Note that the neural network built on these descriptors is an autoencoder since the input and output layer in the network represent the same sentence.[3] Additionally, this autoencoder can naturally handle variable-length sentences (Notice that $g \in \mathbb{R}^d$). In particular, we concatenate the input descriptors into one vector, and encode it into the hidden layer shown as gray nodes in Figure 3:

$$hid = f(W^{enc}[g^{max}; g^{min}] + b^{enc}) \tag{6}$$

where $W^{enc} \in \mathbb{R}^{m \times 2d}$ and $b^{enc} \in \mathbb{R}^m$ are the encoding parameters. $hid \in \mathbb{R}^m$ is the sentence representation, and $f(\cdot)$ is an element-wise activation function such as $tanh(\cdot)$, which is used for all activation functions in our model. To prevent the hidden layer from being very small, we normalize all output vectors of the hidden layer to have a unit length, $hid = \frac{hid}{\|hid\|}$.

Upon the hidden layer, we stack a decoding layer to *reconstruct* the other descriptors:

$$out = f(W^{dec}hid + b^{dec}) \tag{7}$$

where $W^{dec} \in \mathbb{R}^{m \times 2d}$ and $b^{dec} \in \mathbb{R}^{2d}$ are the decoding parameters.

To integrate the information contained in different descriptors into sentence embedding learning, we train our autoencoder to minimize the following Euclidean distance error with respect to the target descriptors $[g^{avg}; g^{std}]$:

$$E_{ae}(\mathbf{x}) = \frac{1}{2}\|[g^{avg}; g^{std}] - out\|^2 \tag{8}$$

4.2 Bilingual Semantic Constraints

In order to incorporate bilingual correspondences between the source and target sentences into the above-mentioned autoencoder, we employ two kinds of bilingual constraints.

The first bilingual constraint is to share decoding parameters across the two autoencoders, that is, using the same W^{dec} and b^{dec} for them. This ensures that the relations between sentence representations (in the hidden layer) and target descriptors (in the output layer) are consistent across the source and target language. In this way, bilingual information is propagated from one language to the other language

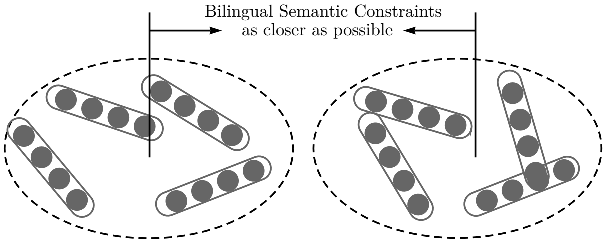

Figure 4: Corpus-level bilingual semantic constraints. The left and right dashed ellipses indicate bilingual semantic spaces.

through these parameters.

The second bilingual constraint is that the center of bilingual spaces should be as close as possible. Therefore we minimize a corpus-level semantic distance defined as follows:

$$E_{sem}(\mathcal{C}) = \sum_{k=1}^{\lceil N/B \rceil} \frac{1}{2} \| \frac{1}{B} \sum_{i=1}^{B} \mathbf{x}_{k,i} - \frac{1}{B} \sum_{i=1}^{B} \mathbf{y}_{k,i} \|^2 \tag{9}$$

where $\mathcal{C}$ represents the entire parallel corpus, which has N parallel sentences with a batch size of B.

Figure 4 shows this constraint, where we try to shorten the distance between the centers of the source and target language spaces. Different from previous works (Zhang et al., 2014; Chandar et al., 2014; Su et al., 2015), we define this distance at the corpus rather than sentence level. The reasons for this are twofold: 1) The potential equivalent translations for a source sentence could be many, and 2) the number of non-equivalent translations for one sentence must be enormous. Therefore, defining the semantic distance at the sentence level would be too strict, and may bring in noises to our model. In contrast, our corpus-level constraint is much more relax and general. Notice that if we set $B = 1$, this constraint goes back to sentence level; If we set $B \to \infty$, it becomes a constraint upon the whole training corpus. Intuitively, B controls the strictness of the semantic constraints.

4.3 Parameter Inference

There are two kinds of errors involved in the overall objective function: the *autoencoder error* in Eq. (8) and the *semantic error* in Eq. (9). Given the training corpus $\mathcal{C}$, the joint training objective is:

$$J(\mathcal{C}) = \frac{\alpha}{N} \sum_{i=1}^{N} (E_{ae}(\mathbf{x}_i) + E_{ae}(\mathbf{y}_i)) + (1-\alpha) E_{sem}(\mathcal{C}) + R(\theta) \tag{10}$$

α is a balance factor, $\theta = \{L_{\mathbf{x}}, L_{\mathbf{y}}, W_{\mathbf{x}}^{enc}, W_{\mathbf{y}}^{enc}, W^{dec}\}$, and $R(\theta)$ is the regularization term: $R(\theta) = \frac{\lambda_L}{2}\|\theta_L\|^2 + \frac{\lambda_W}{2}\|\theta_W\|^2$. For regularization, we divide θ into two sets: θ_L for $(L_{\mathbf{x}}, L_{\mathbf{y}})$ and θ_W for $(W_{\mathbf{x}}^{enc}, W_{\mathbf{y}}^{enc}, W^{dec})$. Accordingly, λ_L and λ_W are the corresponding coefficients.

We employ the toolkit Word2Vec (Mikolov et al., 2013) to pretrain word embeddings for each language on the corresponding part of $\mathcal{C}$, and randomly initialize the other parameters using normal distribution ($\mu = 0, \sigma = 0.01$). To optimize these parameters, we apply the L-BFGS algorithm to the gradient of Eq. (10).

5 Sentence Representation and Semantic Similarity

Given a parallel sentence pair $(\mathbf{x}, \mathbf{y})$, our model will generate sentence representations $(hid_{\mathbf{x}}, hid_{\mathbf{y}})$ with the trained parameters θ^*. Different from Eq. (6) that only uses the input descriptors to calculate the hidden layer, we use both the input and target descriptors to compute the final hidden representation

[2] We sampled the word embeddings uniformly from -1 to 1. Finally, we obtained *Thank*:$(0.95, 0.64, -0.23)$, *you*:$(-0.82, 0.54, 0.72)$ and .:$(-0.12, 0.10, -0.52)$ respectively for illustration.

[3] Therefore, the *autoencoder* in this paper is an extension of the conventional autoencoder.

of $\mathbf{x}$ or $\mathbf{y}$ as these descriptors provide different and complementary views of a sentence. The final hidden representations of the sentence $\mathbf{x}$ and $\mathbf{y}$ are defined as follows:

$$hid_{\mathbf{x}} = f(W_{\mathbf{x}}^{enc,*}[g_{\mathbf{x}}^{min}; g_{\mathbf{x}}^{max}] + {W^{dec,*}}^{T}[g_{\mathbf{x}}^{avg}; g_{\mathbf{x}}^{std}] + b^{enc}) \tag{11}$$

$$hid_{\mathbf{y}} = f(W_{\mathbf{y}}^{enc,*}[g_{\mathbf{y}}^{min}; g_{\mathbf{y}}^{max}] + {W^{dec,*}}^{T}[g_{\mathbf{y}}^{avg}; g_{\mathbf{y}}^{std}] + b^{enc}) \tag{12}$$

The corresponding semantic similarity is measured by the Euclidean distance between the two hidden representations:

$$Sim(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|hid_{\mathbf{x}} - hid_{\mathbf{y}}\|^2 \tag{13}$$

The smaller $Sim(\cdot, \cdot)$ is, the more semantically similar the parallel sentence pair is. All the following intrinsic and extrinsic evaluations are based on this similarity measurement.

6 Experiments

In this section, we carried out a series of experiments to validate the effectiveness of our proposed bilingual autoencoders on NIST Chinese-English translation tasks using large-scale bilingual training data. In particular, we investigated 1) whether our model is able to distinguish parallel sentence pairs from nonparallel sentences, and 2) whether our model can improve machine translation quality.

6.1 Setup

Our translation decoder is a state-of-the-art hierarchical phrased-based SMT system (Chiang, 2007). The bilingual training data is the combination of LDC2003E14, LDC2004T07, LDC2005T06, LDC2005T10 and LDC2004T08 (Hong Kong Hansards/Laws/News), which contains 2.9M sentence pairs with 80.9M Chinese words and 86.4M English words. We used a 4-gram language model which was trained on the Xinhua section of the English Gigaword corpus (306M words) using the SRILM[4] toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing.

In addition to the baseline decoder, we also compared our bilingual autoencoder against the abovementioned BBoWAE model (Chandar et al., 2014). To train this model, we used the same bilingual corpus and their open source code[5] with the same hyperparameters as they used. We employed the objective in Eq. (4) except for the correlation term as an entropy-based semantic similarity measure.[6]

Since our model is general in terms of input and target descriptors, we investigated two variants of the proposed bilingual autoencoder: 1) *min* and *max* as the input descriptors, while *avg* and *std* as the target descriptors (MM2AS); and 2) the opposite direction (AS2MM). Throughout all experiments, we set $B = 100$, $\alpha = 0.100$, $\lambda_L = 10^{-6}$, $\lambda_W = 10^{-3}$ and $m = 2d$. With respect to the dimensionality of word embeddings, we tried four different dimensions from 25 to 100 with an increment of 25 each time.

We used the NIST evaluation set of MT05 as our development set, and sets of MT06/MT08 as the test sets. Case-insensitive NIST BLEU (Papineni et al., 2002) was used to measure translation performance. We used minimum error rate training (MERT) (Och, 2003) to optimize the feature weights. In order to alleviate the instability of MERT, we followed Clark et al. (2011) to run MERT three times and report average BLEU scores over the three runs for all our MT experiments.

6.2 Intrinsic Evaluation: Semantic Analysis

The first group of experiments aims at analyzing the ability of our model in distinguishing parallel sequences from nonparallel sequences, at both the phrase and sentence level. For convenience, we used MM2AS model and set $d = 25$ in all the following experiments.

[4] http://www.speech.sri.com/projects/srilm/download.html

[5] http://www.sarathchandar.in/crl.html

[6] We make this choice due to the following two reasons: 1) the correlation term is meaningless for a single sentence pair; 2) the combined reconstruction errors in Eq. (4) provide a balanced and comprehensive similarity measure as they calculate cross entropies of $\mathbf{x}$ to $\mathbf{x}$, $\mathbf{y}$ to $\mathbf{y}$, $\mathbf{x}$ to $\mathbf{y}$, $\mathbf{y}$ to $\mathbf{x}$ and $[\mathbf{x}, \mathbf{y}]$ to $[\mathbf{x}, \mathbf{y}]$.

Source Phrase	Target Candidates
充满 激情	imbued with passions full of passion full of excitement
宾至如归 的 感觉	feeling of being at home feel at home feel welcomed
促进 可 持续 发展	promotion of sustainable development promote sustainable development promote sustained growth

Table 1: Source phrases in our phrase table with their top-3 target translations selected by our model.

System	MT06	MT08	Avg
BBoWAE	77.73	74.57	76.15
Our Model	87.18	85.85	**86.52**

Table 2: Accuracy on recognizing parallel sentence pairs.

6.2.1 Analysis on Phrasal Semantic Similarities

To have a deep look into what our model measures similarities of bilingual phrases, we show some examples from our scored phrase table in Table 1. For each variable-length source phrase, we extract the top-3 target candidates according to their semantic similarity scores calculated in Eq. (13).

Take the source phrase " 宾至如归 的 感觉" as an example, our model succeeds in distinguishing the most semantically equivalent translation "*feeling of being at home*" from the less equivalent translation "*feel welcomed*". Although the candidate "*feel at home*" has almost the same meaning, this candidate is a verb phrase which is not consistent with the noun phrase at the source side from the perspective of syntax. This indicates that the proposed bilingual model is able to capture some semantic and syntactic properties of bilingual sequences.

6.2.2 Evaluation on Sentential Semantic Similarities

To testify the ability of our model at the sentence level, we further collected source sentences and their reference translations from the test sets MT06/MT08 as our parallel sentence pairs. We randomly sampled l target sentences from the target vocabulary based on reference translations and combined sampled target sentences and their corresponding source sentences as nonparallel sentence pairs.[7]

For both our model and BBoWAE model, we test whether the model can assign a higher similarity score to a parallel sentence pair than a sampled nonparallel sentence pair. We employ accuracy as our evaluation metric, and conducted 5 evaluations, for each of which we sampled 3 different target sentences for each source sentence. The final average results are shown in Table 2. We find that our model significantly outperforms BBoWAE model by an absolute improvement of 10%. This demonstrates that our model can learn better semantic representations for parallel sentences than BBoWAE does.

6.3 Extrinsic Evaluation: Machine Translation

The second group of experiments were carried out to study the effectiveness of our model in calculating semantic similarities for SMT task: *decoding with phrasal semantic similarities*. This needs to measure the semantic similarity between a source sequence and its translation candidate according to Eq. (13).

In addition to the conventional four translation probabilities (phrase translation probabilities and lexical weights in both directions), we incorporate the phrasal semantic similarities as an additional feature into our baseline SMT system.

Table 3 summarizes the results, from which we observe that:

1) Our model can significantly improve translation quality on all testsets for any dimensions. Particularly, AS2MM model obtains the best result 28.08 when $d = 100$, which outperforms Baseline and BBoWAE model by up to 1.46 and 1.1 BLEU points respectively.

[7] The length of a sampled sentence is uniformly sampled from 1 to the length of the corresponding reference sentence.

System	d	MT06	MT08	Avg
Baseline	-	30.04	23.19	26.62
BBoWAE	40	30.52	23.43	26.98
MM2AS	25	31.11↑	23.90↑	27.51
	50	31.42⇑	24.19⇑	27.81
	75	31.18↑	23.98↑	27.58
	100	31.11↑	23.91↑	27.51
AS2MM	25	31.29⇑	24.34⇑	27.82
	50	31.10↑	24.06⇑	27.58
	75	31.17↑	23.92↑	27.55
	100	31.64⇑	24.51⇑	28.08

Table 3: Experiment results for different dimensions with phrasal semantic similarities on the test sets. **Avg** = average BLEU scores on the test sets. ⇑/↑:significantly better than BBoWAE ($p < 0.01/0.05$, respectively).

2) The BBoWAE model, designed for learning bilingual word embeddings, is not good at modeling high-level and long sequence representations. The translation result of this model is 26.98, only 0.36 points higher than the Baseline. Additionally, as one major difference between BBoWAE and our model is the corpus-level semantic constraints, this result further demonstrates the superiority of this constraint.

3) It's interesting that the overall result of AS2MM model is slightly better than that of MM2AS model (27.76 vs. 27.60 on average). The reason may be that when back-propagating autoencoder errors onto the hidden and input layer, AS2MM model is easier and more straightforward than MM2AS.

7 Conclusion and Future Work

In this paper, we have presented a simple yet effective and scalable bilingual autoencoder for parallel sentence modeling. We incorporate global descriptors and corpus-level semantic constraints into bilingual sentence representations. Experiment results show that our approach achieves substantial improvements against the baseline models.

For the future work, we would like to explore more variants of our bilingual autoencoder, e.g. taking the *avg*, *max* as inputs and the *std*, *min* as targets. Besides, we will further enhance our autoencoder with semantic and deeper descriptors and verify our model on other bilingual or cross-lingual tasks, such as cross-lingual sentiment classification.

Acknowledgements

The authors were supported by National Natural Science Foundation of China (Grant Nos. 61303082, 61403269, 61432013, 61525205 and 61672440), Natural Science Foundation of Fujian Province (Grant No. 2016J05161) and Natural Science Foundation of Jiangsu Province (Grant No. BK20140355). We also thank the anonymous reviewers for their insightful comments.

References

Sarath Chandar, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Proc. of NIPS*, pages 1853–1861.

David Chiang. 2007. Hierarchical phrase-based translation. *Comput. Linguist.*, pages 201–228, June.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of HLT*, pages 176–181.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proc. of ICML*, pages 160–167.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proc. of ACMMM*, pages 7–16.

Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2015. Deep correspondence restricted boltzmann machine for cross-modal retrieval. *Neurocomputing*, pages 50 – 60.

Karl Moritz Hermann and Phil Blunsom. 2013. The Role of Syntax in Vector Space Models of Compositional Semantics. In *Proc. of ACL*, August.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proc. of ACL*, pages 58–68, June.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *Proc. of CVSC*.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *Proc. of ACL*, June.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*, pages 1746–1751, October.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proc. of COLING*, pages 1459–1474, December.

Stanislas Lauly, Alex Boulanger, and Hugo Larochelle. 2014. Learning multilingual word representations using a bag-of-words autoencoder. *NIPS Workshop*.

Mingbo Ma, Liang Huang, Bowen Zhou, and Bing Xiang. 2015. Dependency-based convolutional neural networks for sentence embedding. In *Proc. of ACL-IJCNLP*, pages 174–179, July.

Tomas Mikolov, Stefan Kombrink, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2011. Extensions of recurrent neural network language model. In *Proc. of ICASSP*, pages 5528–5531.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, pages 3111–3119.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proc. of ICML*, pages 689–696.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. of EMNLP*, pages 151–161.

Andreas Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proc. of ICSLP*, volume 2, pages 901–904.

Jinsong Su, Deyi Xiong Xiong, Biao Zhang, Yang Liu, Junfeng Yao, and Min Zhang. 2015. Bilingual correspondence recursive autoencoder for statistical machine translation. In *Proc. of EMNLP*, September.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proc. of ACL-IJCNLP*, pages 1556–1566, July.

Wei Wang, Beng Chin Ooi, Xiaoyan Yang, Dongxiang Zhang, and Yueting Zhuang. 2014. Effective multi-modal retrieval based on stacked auto-encoders. *Proc. of VLDB Endow.*, pages 649–660, April.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proc. of CoNLL*, pages 247–256, June.

Jiajun Zhang, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proc. of ACL*, pages 111–121, June.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015a. Shallow convolutional neural network for implicit discourse relation recognition. In *Proc. of EMNLP*, September.

Jiajun Zhang, Dakun Zhang, and Jie Hao. 2015b. Local translation prediction with global sentence representation. In *Proc. of IJCAI*, IJCAI'15, pages 1398–1404.

HuiWei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proc. of ACL-IJCNLP*, pages 430–440, July.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Pro. of EMNLP*, pages 1393–1398, October.