# BattRAE: Bidimensional Attention-Based Recursive Autoencoders for Learning Bilingual Phrase Embeddings

**Biao Zhang,**[1] **Deyi Xiong,**[2] **Jinsong Su**[1*]
Xiamen University, Xiamen, China 361005[1]
Soochow University, Suzhou, China 215006[2]
zb@stu.xmu.edu.cn, dyxiong@suda.edu.cn, jssu@xmu.edu.cn

## Abstract

In this paper, we propose a bidimensional attention based recursive autoencoder (BattRAE) to integrate clues and source-target interactions at multiple levels of granularity into bilingual phrase representations. We employ recursive autoencoders to generate tree structures of phrases with embeddings at different levels of granularity (e.g., words, sub-phrases and phrases). Over these embeddings on the source and target side, we introduce a *bidimensional attention network* to learn their interactions encoded in a *bidimensional attention matrix*, from which we extract two soft attention weight distributions simultaneously. These weight distributions enable BattRAE to generate compositive phrase representations via convolution. Based on the learned phrase representations, we further use a bilinear neural model, trained via a max-margin method, to measure bilingual semantic similarity. To evaluate the effectiveness of BattRAE, we incorporate this semantic similarity as an additional feature into a state-of-the-art SMT system. Extensive experiments on NIST Chinese-English test sets show that our model achieves a substantial improvement of up to *1.63* BLEU points on average over the baseline.

## 1 Introduction

As one of the most important components in statistical machine translation (SMT), translation model measures the translation faithfulness of a hypothesis to a source fragment (Och and Ney 2003; Koehn, Och, and Marcu 2003; Chiang 2007). Conventional translation models extract a huge number of *bilingual phrases* with conditional translation probabilities and lexical weights (Koehn, Och, and Marcu 2003). Due to the heavy reliance of the calculation of these probabilities and weights on surface forms of bilingual phrases, traditional translation models often suffer from the problem of data sparsity. This leads researchers to investigate methods that learn underlying semantic representations of phrases using neural networks (Gao et al. 2014; Zhang et al. 2014; Cho et al. 2014; Su et al. 2015).

Typically, these neural models learn bilingual phrase embeddings in a way that embeddings of source and corresponding target phrases are optimized to be close as much as possible in a continuous space. In spite of their success, they either explore **clues** (linguistic items from contexts) only at a single level of granularity or capture **interactions** (alignments between source and target items) only at the same level of granularity to learn bilingual phrase embeddings.

| Src (Ref) | Tgt |
|---|---|
| ***shijie*** *geda chengshi* *(major cities in the world)* | *in other major cities* |
| *dui* ***jingji xuezhe*** *(to the economists)* | *to* ***economists*** |

Table 1: Examples of bilingual phrases ((romanized) Chinese-English) from our translation model. The important words or phrases are highlighted in bold. **Src** = source, **Tgt** = target, **Ref** = literal translation of source phrase.

We believe that clues and interactions from a single level of granularity are not adequate to measure underlying semantic similarity of bilingual phrases due to the high language divergence. Take the Chinese-English translation pairs in Table 1 as examples. At the word level of granularity, we can easily recognize that the translation of the first instance is not faithful as Chinese word "*shijie*" (*world*) is not translated at all. While in the second instance, semantic judgment at the word level is not sufficient as there is no translation for single Chinese word "*jingji*" (*economy*) or "*xuezhe*" (*scholar*). We have to elevate the calculation of semantic similarity to a higher sub-phrase level: "*jingji xuezhe*" vs. "*economists*". This suggests that clues and interactions between the source and target side at multiple levels of granularity should be explored to measure semantic similarity of bilingual phrases.

In order to capture multi-level clues and interactions, we propose a bidimensional attention based recursive autoencoder (BattRAE). It learns bilingual phrase embeddings according to the strengths of interactions between linguistic items at different levels of granularity (i.e., words, sub-phrases and entire phrases) on the source side and those on the target side. The philosophy behind BattRAE is twofold: 1) Phrase embeddings are learned from weighted clues at different levels of granularity; 2) The weights of clues are calculated according to the alignments of linguistic items at different levels of granularity between the source and target side. We introduce a *bidimensional attention network* to

*Corresponding author.

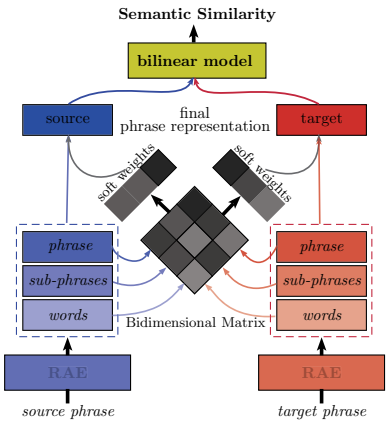


Figure 1: Overall architecture for the proposed BattRAE model. We use blue and red color to indicate the source- and target-related representations or structures respectively. The gray colors indicate real values in the bidimensional mechanism.

learn the strengths of these alignments. Figure 1 illustrates the overall architecture of BattRAE model. Specifically,

- First, we adopt recursive autoencoders to generate hierarchical structures of source and target phrases separately. At the same time, we can also obtain embeddings of multiple levels of granularity, i.e., words, sub-phrases and entire phrases from the generated structures. (see Section 2)
- Second, BattRAE projects the representations of linguistic items at different levels of granularity onto an attention space, upon which the alignment strengths of linguistic items from the source and target side are calculated by estimating how well they semantically match. These alignment scores are stored in a bidimensional attention matrix. Over this matrix, we perform row (column)-wise summation and softmax operations to generate attention weights on the source (target) side. The final phrase representations are computed as the weighted sum of their initial embeddings using these attention weights. (see Section3.1)
- Finally, BattRAE projects the bilingual phrase representations onto a common semantic space, and uses a bilinear model to measure their semantic similarity. (see Section 3.2)

We train the BattRAE model with a max-margin method, which maximizes the semantic similarity of translation equivalents and minimizes that of non-translation pairs (see Section 3.3).

In order to verify the effectiveness of BattRAE in learning bilingual phrase representations, we incorporate the learned semantic similarity of bilingual phrases as a new feature into SMT for translation selection. We conduct experiments with a state-of-the-art SMT system on large-scale training data. Results on the NIST 2006 and 2008 datasets show that BattRAE achieves significant improvements over baseline methods. Further analysis on the bidimensional attention matrix reveals that BattRAE is able to detect semantically related parts of bilingual phrases and assign higher weights to these parts for constructing final bilingual phrase embeddings than those not semantically related.

## 2 Learning Embeddings at Different Levels of Granularity

We use recursive autoencoders (RAE) to learn initial embeddings at different levels of granularity for our model. Combining two children vectors from the bottom up recursively, RAE is able to generate low-dimensional vector representations for variable-sized sequences. The recursion procedure usually consists of two neural operations: *composition* and *reconstruction*.

*Composition:* Typically, the input to RAE is a list of ordered words in a phrase $(x_1, x_2, x_3)$, each of which is embedded into a $d$-dimensional continuous vector.[1] In each recursion, RAE selects two neighboring children (e.g. $c_1 = x_1$ and $c_2 = x_2$) via some selection criterion, and then compose them into a parent embedding $y_1$, which can be computed as follows:

$$y_1 = f(W^{(1)}[c_1; c_2] + b^{(1)}) \quad (1)$$

where $[c_1; c_2] \in \mathbb{R}^{2d}$ is the concatenation of $c_1$ and $c_2$, $W^{(1)} \in \mathbb{R}^{d \times 2d}$ is a parameter matrix, $b^{(1)} \in \mathbb{R}^d$ is a bias term, and $f$ is an element-wise activation function such as $tanh(\cdot)$, which is used in our experiments.

*Reconstruction:* After the composition, we obtain the representation for the parent $y_1$ which is also a $d$-dimensional vector. In order to measure how well the parent $y_1$ represents its children, we reconstruct the original child nodes via a reconstruction layer:

$$[c_1'; c_2'] = f(W^{(2)} y_1 + b^{(2)}) \quad (2)$$

here $c_1'$ and $c_2'$ are the reconstructed children, $W^{(2)} \in \mathbb{R}^{2d \times d}$ and $b^{(2)} \in \mathbb{R}^{2d}$. The minimum Euclidean distance between $[c_1'; c_2']$ and $[c_1; c_2]$ is usually used as the selection criterion during composition.

These two standard processes form the basic procedure of RAE, which repeat until the embedding of the entire phrase is generated. In addition to phrase embeddings, RAE also constructs a binary tree. The structure of the tree is determined by the used selection criterion in composition. As generating the optimal binary tree for a phrase is usually intractable, we employ a greedy algorithm (Socher et al. 2011b) based on the following reconstruction error:

$$E_{rec}(\mathbf{x}) = \sum_{y \in T(\mathbf{x})} \frac{1}{2} \parallel [c_1; c_2]_y - [c_1'; c_2']_y \parallel^2 \quad (3)$$

Parameters $W^{(1)}$ and $W^{(2)}$ are thereby learned to minimize the sum of reconstruction errors at each intermediate node $y$ in the binary tree $T(\mathbf{x})$. For more details, we refer the readers to (Socher et al. 2011b).

Given an binary tree learned by RAE, we regard each level of the tree as **a level of granularity**. In this way, we can use

[1] Generally, all these word vectors are stacked into a word embedding matrix $L \in \mathbb{R}^{d \times |V|}$, where $|V|$ is the size of the vocabulary.

RAE to produce embeddings of linguistic expressions at different levels of granularity. Unfortunately, RAE is unable to synthesize embeddings across different levels of granularity, which will be discussed in the next section.

Additionally, as illustrated in Figure 1, RAEs for the source and target language are learned separately. In our model, we assume that phrase embeddings for different languages are from different semantic spaces. To make this clear, we denote dimensions of source and target phrase embeddings as $d_s$ and $d_t$ respectively.

## 3 Bidimensional Attention-Based Recursive Autoencoders

In this section, we present the proposed BattRAE model. We first elaborate the bidimensional attention network, and then the semantic similarity model built on phrase embeddings learned with the attention network. Finally, we introduce the objective function and training procedure.

### 3.1 Bidimensional Attention Network

As mentioned in Section 1, we would like to incorporate clues and interactions at multiple levels of granularity into phrase embeddings and further into the semantic similarity model of bilingual phrases. The clues are encoded in multi-level embeddings learned by RAEs. The interactions between linguistic items on the source and target side can be measured by how well they semantically match. In order to jointly model clues and interactions at multiple levels of granularity, we propose the *bidimensional attention network*, which is illustrated in Figure 2.

We take the bilingual phrase ("*dui jingji xuezhe*", "*to economists*") in Table 1 as an example. Let's suppose that their phrase structures learned by RAE are "(*dui*, (*jingji*, *xuezhe*))" and "(*to*, *economists*)" respectively. We perform a postorder traversal on these structures to extract the embeddings of words, sub-phrases and the entire source/target phrase. We treat each embedding as a column and put them together to form a matrix $M_s \in \mathbb{R}^{d_s \times n_s}$ on the source side and $M_t \in \mathbb{R}^{d_t \times n_t}$ on the target side. Here, $n_s$=5 and $M_s$ contains embeddings from linguistic items ("*dui*", "*jingji*", "*xuezhe*", "*jingji xuezhe*", "*dui jingji xuezhe*") at different levels of granularity. Similarly, $n_t$=3 and $M_t$ contains embeddings of ("*to*", "*economists*", "*to economists*"). $M_s$ and $M_t$ form the input layer for the bidimensional attention network.

We further stack an attention layer upon the input matrices to project the embeddings from $M_s$ and $M_t$ onto a common attention space as follows (see the gray circles in Figure 2):

$$A_s = f(W^{(3)} M_s + b^A_{[:]}) \tag{4}$$

$$A_t = f(W^{(4)} M_t + b^A_{[:]}) \tag{5}$$

where $W^{(3)} \in \mathbb{R}^{d_a \times d_s}$, $W^{(4)} \in \mathbb{R}^{d_a \times d_t}$ are transformation matrices, $b^A \in \mathbb{R}^{d_a}$ is the bias term, and $d_a$ is the dimensionality of the attention space. The subscript $[:]$ indicates a *broadcasting* operation. Note that we use different transformation matrices but share the same bias term for $A_s$ and $A_t$. This will force our model to learn to encode attention semantics into the two transformation matrices, rather than the bias term.

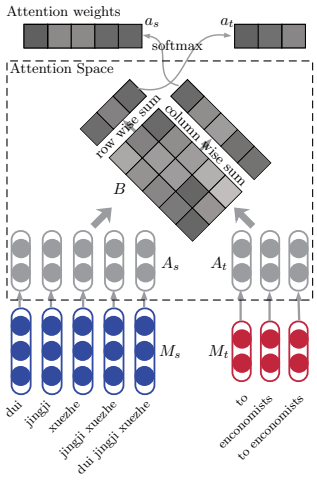


Figure 2: An illustration of the bidimensional attention network in BattRAE model. The gray circles represent the attention space. We use subscript $s$ and $t$ to indicate the source and target respectively.

On this attention space, each embedding from the source side is able to interact with all embeddings from the target side and vice versa. The strength of such an interaction can be measured by a semantically matching score, which is calculated via the following equation:

$$B_{i,j} = g(A^T_{s,i} A_{t,j}) \tag{6}$$

where $B_{i,j} \in \mathbb{R}$ is the score that measures how well the $i$-th column embedding in $A_s$ semantically matches the $j$-th column embedding in $A_t$, and $g(\cdot)$ is a non-linear function, e.g., the sigmoid function used in this paper. All matching scores form a matrix $B \in \mathbb{R}^{n_s \times n_t}$, which we call the *bidimensional attention matrix*. Intuitively, this matrix is a result of handshakes between the source and target phrase at multiple levels of granularity.

Given the bidimensional attention matrix, our next interest lies in how important an embedding at a specific level of granularity is to the semantic similarity between the corresponding source and target phrase. As each embedding interacts all embeddings on the other side, its importance can be measured as the summation of the strengths of all these interactions, i.e., matching scores computed in Eq. (6). This can be done via a row/column-wise summation operation over the bidimensional attention matrix as follows.

$$\tilde{a}_{s,i} = \sum_j B_{i,j}, \;\; \tilde{a}_{t,j} = \sum_i B_{i,j} \tag{7}$$

where $\tilde{a}_s \in \mathbb{R}^{n_s}$ and $\tilde{a}_t \in \mathbb{R}^{n_t}$ are the matching score vectors.

Because the length of a phrase is uncertain, we apply a Softmax operation on $\tilde{a}_s$ and $\tilde{a}_t$ to keep their values at the same magnitude: $a_s = Softmax(\tilde{a}_s), a_t = Softmax(\tilde{a}_t)$. This forces $a_s$ and $a_t$ to become real-valued distributions on the attention space. We call them *attention weights* (see

Figure 2). An important feature of this attention mechanism is that it naturally deals with variable-length bilingual inputs (as we do not impose any length constrains on $n_s$ and $n_t$ at all).

To obtain final bilingual phrase representations, we convolute the embeddings in phrase structures with the computed attention weights:

$$p_s = \sum_i a_{s,i} M_{s,i}, \;\; p_t = \sum_j a_{t,j} M_{t,j} \tag{8}$$

This ensures that the generated phrase representations encode weighted clues and interactions at multiple levels of granularity between the source and target phrase. Notice that $p_s \in \mathbb{R}^{d_s}$ and $p_t \in \mathbb{R}^{d_t}$ still locate in their language-specific vector space.

### 3.2 Semantic Similarity

To measure the semantic similarity for a bilingual phrase, we first transform the learned bilingual phrase representations $p_s$ and $p_t$ into a common semantic space through a non-linear projection as follows:

$$s_s = f(W^{(5)} p_s + b^s) \tag{9}$$

$$s_t = f(W^{(6)} p_t + b^s) \tag{10}$$

where $W^{(5)} \in \mathbb{R}^{d_{sem} \times d_s}$, $W^{(6)} \in \mathbb{R}^{d_{sem} \times d_t}$ and $b^s \in \mathbb{R}^{d_{sem}}$ are the parameters. Similar to the transformation in Eq. (4) and (5), we share the same bias term for both $s_s$ and $s_t$.

We then use a bilingual model to compute the semantic similarity score as follows:

$$s(f, e) = s_s^T S s_t \tag{11}$$

where $f$ and $e$ is the source and target phrase respectively, and $s(\cdot, \cdot)$ represents the semantic similarity function. $S \in \mathbb{R}^{d_{sem} \times d_{sem}}$ is a squared matrix of parameters to be learned. We choose this model because that the matrix $S$ actually represents an interaction between $s_s$ and $s_t$, which is desired for our purpose.

### 3.3 Objective and Training

There are two kinds of errors involved in our objective function: *reconstruction error* (see Eq. (3)) and *semantic error*. The latter error measures how well a source phrase semantically match its counterpart target phrase. We employ a max-margin method to estimate this semantic error. Given a training instance $(f, e)$ with negative samples $(f^-, e^-)$, we define the following ranking-based error:

$$\begin{aligned} E_{sem}(f, e) = & \, max(0, 1 + s(f, e^-) - s(f, e)) \\ & + max(0, 1 + s(f^-, e) - s(f, e)) \end{aligned} \tag{12}$$

Intuitively, minimizing this error will maximize the semantic similarity of the correct translation pair $(f, e)$ and minimize (up to a margin) the similarity of negative translation pairs $(f^-, e)$ and $(f, e^-)$. In order to generate the negative samples, we replace words in a correct translation pair with random words, which is similar to the sampling method used by Zhang et al. (2014).

For each training instance $(f, e)$, the joint objective of BattRAE is defined as follows:

$$J(\theta) = \alpha E_{rec}(f, e) + \beta E_{sem}(f, e) + R(\theta) \tag{13}$$

where $E_{rec}(f, e) = E_{rec}(f) + E_{rec}(e)$, parameters $\alpha$ and $\beta$ ($\alpha + \beta = 1$) are used to balance the preference between the two errors, and $R(\theta)$ is the regularization term. We divide the parameters $\theta$ into four different groups[2]:

1. $\theta_L$ : the word embedding matrices $L_s$ and $L_t$;
2. $\theta_{rec}$ : the parameters for RAE $W_s^{(1)}$, $W_t^{(1)}$, $W_s^{(2)}$, $W_t^{(2)}$ and $b_s^{(1)}$, $b_t^{(1)}$, $b_s^{(2)}$, $b_t^{(2)}$;
3. $\theta_{att}$ : the parameters for the projection of the input matrices onto the attention space $W^{(3)}, W^{(4)}$ and $b^A$;
4. $\theta_{sem}$ : the parameters for semantic similarity computation $W^{(5)}, W^{(6)}, S$ and $b^s$;

And each parameter group is regularized with a unique weight:

$$R(\theta) = \frac{\lambda_L}{2}\|\theta_L\|^2 + \frac{\lambda_{rec}}{2}\|\theta_{rec}\|^2 + \frac{\lambda_{att}}{2}\|\theta_{att}\|^2 + \frac{\lambda_{sem}}{2}\|\theta_{sem}\|^2 \tag{14}$$

where $\lambda_*$ are our hyperparameters.

To optimize these parameters, we apply the L-BFGS algorithm which requires two conditions: *parameter initialization* and *gradient calculation*.

*Parameter Initialization:* We randomly initialize $\theta_{rec}$, $\theta_{att}$ and $\theta_{sem}$ according to a normal distribution ($\mu$=0,$\sigma$=0.01). With respect to the word embeddings $\theta_L$, we use the toolkit Word2Vec[3] to pretrain them on a large scale unlabeled data. All these parameters will be further fine-tuned in our BattRAE model.

*Gradient Calculation:* We compute the partial gradient for parameter $\theta_k$ as follows:

$$\frac{\partial J}{\partial \theta_k} = \frac{\partial E_{rec}(f, e)}{\partial \theta_k} + \frac{\partial E_{sem}(f, e)}{\partial \theta_k} + \lambda_k \theta_k \tag{15}$$

This gradient is fed into the toolkit libLBFGS[4] for parameter updating in our practical implementation.

## 4 Experiment

In order to examine the effectiveness of BattRAE in learning bilingual phrase embeddings, we carried out large scale experiments on NIST Chinese-English translation tasks.[5]

### 4.1 Setup

Our parallel corpus consists of 1.25M sentence pairs extracted from LDC corpora[6], with 27.9M Chinese words and

[2]The subscript $s$ and $t$ are used to denote the source and the target language.

[3]https://code.google.com/p/word2vec/

[4]http://www.chokkan.org/software/liblbfgs/

[5]Source code is available at https://github.com/DeepLearnXMU/BattRAE.

[6]This includes LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06.

| Method | MT06 | MT08 | AVG |
|---|---|---|---|
| *Baseline* | 31.55 | 23.66 | 27.61 |
| *BRAE* | 32.29 | 24.30 | 28.30 |
| *BCorrRAE* | 32.62 | 24.89 | 28.76 |
| *BattRAE* | **33.19**$^{\Uparrow **\uparrow}$ | **25.29**$^{\Uparrow **}$ | **29.24** |

Table 2: Experiment results on the MT 06/08 test sets. **AVG** = average BLEU scores for test sets. We highlight the best result in bold. "$\Uparrow$": significantly better than *Baseline* ($p < 0.01$); "$**$": significantly better than *BRAE* ($p < 0.01$); "$\uparrow$": significantly better than *BCorrRAE* ($p < 0.05$).

34.5M English words respectively. We trained a 5-gram language model on the Xinhua portion of the GIGAWORD corpus (247.6M English words) using *SRILM* Toolkit[7] with modified Kneser-Ney Smoothing. We used the NIST MT05 data set as the development set, and the NIST MT06/MT08 datasets as the test sets. We used minimum error rate training (Och and Ney 2003) to optimize the weights of our translation system. We used case-insensitive BLEU-4 metric (Papineni et al. 2002) to evaluate translation quality and performed the paired bootstrap sampling (Koehn 2004) for significance test.

In order to obtain high-quality bilingual phrases to train the BattRAE model, we used forced decoding (Wuebker, Mauser, and Ney 2010) (but without the leaving-one-out) on the above parallel corpus, and collected 2.8M phrase pairs. From these pairs, we further extracted 34K bilingual phrases as our development data to optimize all hyper-parameters using *random search* (Bergstra and Bengio 2012). Finally, we set $d_s$=$d_t$=$d_a$=$d_{sem}$=50, $\alpha$=0.125 (such that, $\beta$=0.875), $\lambda_L$=$1e^{-5}$, $\lambda_{rec}$=$\lambda_{att}$=$1e^{-4}$ and $\lambda_{sem}$=$1e^{-3}$ according to experiments on the development data. Additionally, we set the maximum number of iterations in the L-BFGS algorithm to 100.

## 4.2 Translation Performance

We compared BattRAE against the following three methods:

- *Baseline*: Our baseline decoder is a state-of-the-art bracketing transduction grammar based translation system with a maximum entropy based reordering model (Wu 1997; Xiong, Liu, and Lin 2006). The features used in this baseline include: rule translation probabilities in two directions, lexical weights in two directions, target-side word number, phrase number, language model score, and the score of the maximum entropy based reordering model.
- *BRAE*: The neural model proposed by Zhang et al. (2014). We incorporate the semantic distances computed according to BRAE as new features into the log-linear model of SMT for translation selection.
- *BCorrRAE*: The neural model proposed by Su et al. (2015) that extends BRAE with word alignment information. The structural similarities computed by BCorrRAE are integrated into the Baseline as additional features.

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

With respect to the two neural baselines BRAE and BCorrRAE, we used the same training data as well as the same methods as ours for hyper-parameter optimization, except for the dimensionality of word embeddings, which we set to 50 in experiments.

Table 2 summaries the experiment results of BattRAE against the other three methods on the test sets. BattRAE significantly improves translation quality on all test sets in terms of BLEU. Especially, it achieves an improvement of up to *1.63* BLEU points on average over the Baseline. Comparing with the two neural baselines, our BattRAE model obtains consistent improvements on all test sets. It significantly outperforms BCorrRAE by 0.48 BLEU points, and BRAE by almost 1 BLEU point on average. We contribute these improvements to the incorporation of clues and interactions at different levels of granularity since neither BCorrRAE nor BRAE explore them.

## 4.3 Attention Analysis

Observing the significant improvements and advantages of BattRAE over BRAE and BCorrRAE, we would like to take a deeper look into how the bidimensional attention mechanism works in the BattRAE model. Specifically, we wonder which words are highly weighted by the attention mechanism. Table 3 shows some examples in our translation model. We provide phrase structures learned by RAE and visualize attention weights for these examples.

We do find that the BattRAE model is able to learn what is important for semantic similarity computation. The model can recognize the correspondence between "*yige*" and "*same*", "*yanzhong*" and "*serious concern*", "*weizhi*" and "*so far*". These word pairs tend to give high semantic similarity scores to these translation instances. In contrast, because of incorrect translation pairs "*taidu*" (*attitude*) vs. "*very critical*", "*yinwei*" (*because*) vs. "*the part*", "*wenti*" (*problem*) vs. "*hong kong*", the model assigns low semantic similarity scores to these negative instances. These indicate that the BattRAE model is indeed able to detect and focus on those semantically related parts of bilingual phrases.

Further observation reveals that there are strong relations between phrase structures and attention weights. Generally, the BattRAE model will assign high weights to words subsumed by many internal nodes of phrase structures. For example, we find that the correct translation of "*wenti*" actually appears in the corresponding target phrase. However, due to errors in learned phrase structures, the model fails to detect this translation. Instead, it finds an incorrect translation "*hong kong*". This suggests that the quality of learned phrase structures has an important impact on the performance of our model.

# 5 Related Work

Our work is related to *bilingual embeddings* and *attention-based neural networks*. We will introduce previous work on these two lines in this section.

## 5.1 Bilingual Embeddings

The studies on bilingual embeddings start from bilingual word embedding learning. Zou et al. (2013) use word align-

| Type | Phrase Structure | | Attention Visualization | |
|---|---|---|---|---|
| | *Src* | *Tgt* | *Src* | *Tgt* |
| *Good* | ((*yi*, *ge*), *zhongguo*) | (*to*, (*the*, (*same*, *china*))) | *yi ge zhongguo* | *to the same china* |
| | ((*yanzhong*, *de*), *shi*) | (((*serious*, *concern*), *is*), *the*) | *yanzhong de shi* | *serious concern is the* |
| | (*jiezhi*, (*muqian*, *weizhi*)) | ((*so*, *far*), (*this*, *year*)) | *jiezhi muqian weizhi* | *so far this year* |
| *Bad* | (*yanjin*, (*de*, *taidu*)) | ((*be*, (*very*, *critical*)), *of*) | *yanjin de taidu* | *be very critical of* |
| | (*zhuyao*, (*shi*, *yinwei*)) | ((*on*, (*the*, *part*)), *of*) | *zhuyao shi yinwei* | *on the part of* |
| | (*cunzai*, (*de*, *wenti*)) | (*problems*, (*of*, (*hong*, *kong*))) | *cunzai de wenti* | *problems of hong kong* |

Table 3: Examples of bilingual phrases from our translation model with both *phrase structures* and *attention visualization*. For each example, important words are highlighted in dark red (with the highest attention weight), red (the second highest), light red (the third highest) according to their attention weights. *Good* = good translation pair, *Bad* = bad translation pair, judged according to their semantic similarity scores.

ments to connect embeddings of source and target words. To alleviate the reliance of bilingual embedding learning on parallel corpus, Vulić and Moens (2015) explore document-aligned instead of sentence-aligned data, while Gouws et al. (2015) investigate monolingual raw texts. Different from the abovementioned corpus-centered methods, Kočiský et al. (2014) develop a probabilistic model to capture deep semantic information, while Chandar et al. (2014) testify the use of autoencoder-based methods. More recently, Luong et al. (2015b) jointly model context co-ocurrence information and meaning equivalent signals to learn high quality bilingual embeddings.

As phrases have long since been used as the basic translation units in SMT, bilingual phrase embeddings attract increasing interests. Since translation equivalents share the same semantic meaning, embeddings of source/target phrases can be learned with information from their counterparts. Along this line, a variety of neural models are explored: multi-layer perceptron (Gao et al. 2014), RNN encoder-decoder (Cho et al. 2014) and recursive autoencoders (Zhang et al. 2014; Su et al. 2015).

The most related work to ours are the bilingual recursive autoencoders (Zhang et al. 2014; Su et al. 2015). Zhang et al. (2014) represent bilingual phrases with embeddings of root nodes in bilingual RAEs, which are learned subject to transformation and distance constraints on the source and target language. Su et al. (2015) extend the model of Zhang et al. (2014) by exploring word alignments and correspondences inside source and target phrases. A major limitation of their models is that they are not able to incorporate clues of multiple levels of granularity to learn bilingual phrase embeddings, which exactly forms our basic motivation.

### 5.2 Attention-Based Neural Networks

Over the last few months, we have seen the tremendous success of attention-based neural networks in a variety of tasks, where learning alignments between different modalities is a key interest. For example, Mnih et al. (2014) learn image objects and agent actions in a dynamic control problem. Xu et al. (2015) exploit an attentional mechanism in the task of image caption generation. With respect to neural machine translation, Bahdanau et al. (2014) succeed in jointly learning to translate and align words. Luong et al. (2015a) further evaluate different attention architectures on translation. Inspired by these works, we propose a *bidimensional attention network* that is suitable in the bilingual context.

In addition to the abovementioned neural models, our model is also related to the work of Socher et al. (2011a) and Yin and Schütze (2015) in terms of multi-granularity embeddings. The former preserves multi-granularity embeddings in tree structures and introduces a dynamic pooling technique to extract features directly from an attention matrix. The latter extends the idea of the former model to convolutional neural networks. Significantly different from their models, we introduce a bidimensional attention matrix to generate attention weights, instead of extracting features. Additionally, our attention-based model is also significantly different from the tensor networks (Socher et al. 2013) in that the attention-based model is able to handle variable-length inputs while tensor networks often assume fixed-length inputs.

We notice that the very recently proposed attentive pooling model (dos Santos et al. 2016) which also aims at modeling mutual interactions between two inputs with a *two-way attention* mechanism that is similar to ours. The major differences between their and our work lie in the following four aspects. First, we perform a transformation ahead of attention computation in order to deal with language divergences, rather than directly compute the attention matrix. Second, we calculate attention weights via a sum-pooling approach, instead of max pooling, in order to preserve all interactions at each level of granularity. Third, we apply our bidimensional attention technique to recursive autoencoders instead of convolutional neural networks. Last, we aim at learning bilingual phrase representations rather than question answering. Most importantly, our work and theirs can be seen as two independently developed models that provide different perspectives on a new attention mechanism.

## 6 Conclusion and Future Work

In this paper, we have presented a bidimensional attention based recursive autoencoder to learn bilingual phrase representations. The model incorporates clues and interactions across source and target phrases at multiple levels of granularity. Through the bidimensional attention network, our model is able to integrate them into bilingual phrase embeddings. Experiment results show that our approach significantly improves translation quality.

In the future, we would like to exploit different functions to compute semantically matching scores (Eq. (6)), and other neural models for the generation of phrase structures. Additionally, the bidimensional attention mechanism can be used in convolutional neural network and recurrent neural network. Furthermore, we are also interested in adapting our model to semantic tasks such as paraphrase identification and natural language inference.

## Acknowledgements

The authors were supported by National Natural Science Foundation of China (Grant Nos. 61303082, 61403269, 61622209 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

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Bergstra, J., and Bengio, Y. 2012. Random search for hyper-parameter optimization. *JMLR* 281–305.

Chandar A P, S.; Lauly, S.; Larochelle, H.; Khapra, M.; Ravindran, B.; Raykar, V. C.; and Saha, A. 2014. An autoencoder approach to learning bilingual word representations. In *Proc. of NIPS*. Curran Associates, Inc. 1853–1861.

Chiang, D. 2007. Hierarchical phrase-based translation. *Comput. Linguist.* 33:201–228.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proc. of EMNLP*, 1724–1734.

dos Santos, C.; Tan, M.; Xiang, B.; and Zhou, B. 2016. Attentive Pooling Networks. *ArXiv e-prints*.

Gao, J.; He, X.; Yih, W.-t.; and Deng, L. 2014. Learning continuous phrase representations for translation modeling. In *Proc. of ACL*, 699–709.

Gouws, S.; Bengio, Y.; and Corrado, G. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, 748–756. JMLR.org.

Koehn, P.; Och, F. J.; and Marcu, D. 2003. Statistical phrase-based translation. In *Proc. of NAACL*, 48–54.

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proc. of EMNLP*, 388–395.

Kočiský, T.; Hermann, K. M.; and Blunsom, P. 2014. Learning bilingual word representations by marginalizing alignments. In *Proc. of ACL*, 224–229.

Luong, M.; Pham, H.; and Manning, C. D. 2015a. Effective approaches to attention-based neural machine translation. *Proc. of EMNLP* 1412–1421.

Luong, T.; Pham, H.; and Manning, C. D. 2015b. Bilingual word representations with monolingual quality in mind. In *Proc. of VSM-NLP*, 151–159.

Mnih, V.; Heess, N.; Graves, A.; and kavukcuoglu, k. 2014. Recurrent models of visual attention. In *Proc. of NIPS*. Curran Associates, Inc. 2204–2212.

Och, F. J., and Ney, H. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29:19–51.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, 311–318.

Socher, R.; Huang, E. H.; Pennin, J.; Manning, C. D.; and Ng, A. Y. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proc. of NIPS*, 801–809.

Socher, R.; Pennington, J.; Huang, E. H.; Ng, A. Y.; and Manning, C. D. 2011b. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. of EMNLP*, 151–161.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proc. of NIPS*. Curran Associates, Inc. 926–934.

Su, J.; Xiong, D.; Zhang, B.; Liu, Y.; Yao, J.; and Zhang, M. 2015. Bilingual correspondence recursive autoencoder for statistical machine translation. In *Proc. of EMNLP*, 1248–1258.

Vulić, I., and Moens, M.-F. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. In *Proc. of ACL-IJCNLP*, 719–725.

Wu, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics, Volume 23, Number 3, September 1997*.

Wuebker, J.; Mauser, A.; and Ney, H. 2010. Training phrase translation models with leaving-one-out. In *Proc. of ACL*, 475–484.

Xiong, D.; Liu, Q.; and Lin, S. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proc. of ACL*, 521–528.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. of ICML*, 2048–2057.

Yin, W., and Schütze, H. 2015. Multigrancnn: An architecture for general matching of text chunks on multiple levels of granularity. In *Proc. of ACL-IJCNLP*, 63–73.

Zhang, J.; Liu, S.; Li, M.; Zhou, M.; and Zong, C. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proc. of ACL*, 111–121.

Zou, W. Y.; Socher, R.; Cer, D.; and Manning, C. D. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP*, 1393–1398.