Learning Bilingual Sentiment Word Embeddings for Cross-language Sentiment Classification

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang
School of Computer Science and Technology
Dalian University of Technology, Dalian, P.R. China
{zhouhuiwei,huangdg}@dlut.edu.cn
{chenlong.415,shi_fl}@mail.dlut.edu.cn

Abstract

The sentiment classification performance relies on high-quality sentiment resources. However, these resources are imbalanced in different languages. Cross-language sentiment classification (CLSC) can leverage the rich resources in one language (source language) for sentiment classification in a resource-scarce language (target language). Bilingual embeddings could eliminate the semantic gap between two languages for CLSC, but ignore the sentiment information of text. This paper proposes an approach to learning bilingual sentiment word embeddings (BSWE) for English-Chinese CLSC. The proposed BSWE incorporate sentiment information of text into bilingual embeddings. Furthermore, we can learn high-quality BSWE by simply employing labeled corpora and their translations, without relying on large-scale parallel corpora. Experiments on NLP&CC 2013 CLSC dataset show that our approach outperforms the state-of-the-art systems.

1 Introduction

Sentiment classification is a task of predicting sentiment polarity of text, which has attracted considerable interest in the NLP field. To date, a number of corpus-based approaches (Pang et al., 2002; Pang and Lee, 2004; Kennedy and Inkpen, 2006) have been developed for sentiment classification. The approaches heavily rely on quality and quantity of the labeled corpora, which are considered as the most valuable resources in sentiment classification task. However, such sentiment resources are imbalanced in different languages. To leverage resources in the source language to improve the sentiment classification performance in the target language, cross-language sentiment classification (CLSC) approaches have been investigated.

The traditional CLSC approaches employ machine translation (MT) systems to translate corpora in the source language into the target language, and train the sentiment classifiers in the target language (Banea et al., 2008). Directly employing the translated resources for sentiment classification in the target language is simple and could get acceptable results. However, the gap between the source language and target language inevitably impacts the performance of sentiment classification. To improve the classification accuracy, multi-view approaches have been proposed. In these approaches, the resources in the source language and their translations in the target language are both used to train sentiment classifiers in two independent views (Wan, 2009; Gui et al., 2013; Zhou et al., 2014a). The final results are determined by ensemble classifiers in these two views to overcome the weakness of monolingual classifiers. However, learning language-specific classifiers in each view fails to capture the common sentiment information of two languages during training process.

With the revival of interest in deep learning (Hinton and Salakhutdinov, 2006), shared deep representations (or embeddings) (Bengio et al., 2013) are employed for CLSC (Chandar A P et al., 2013). Usually, paired sentences from parallel corpora are used to learn word embeddings across languages (Chandar A P et al., 2013; Chandar A P et al., 2014), eliminating the need of MT systems. The learned bilingual embeddings could easily project the training data and test data into a common space, where training and testing are performed. However, high-quality bilingual embeddings rely on the large-scale task-related parallel corpora, which are not always readily available. Meanwhile, though semantic similarities across languages are captured during bilingual embedding learning process, sentiment information of

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 430–440,
Beijing, China, July 26-31, 2015.

text is ignored. That is, bilingual embeddings learned from unlabeled parallel corpora are not effective enough for CLSC because of a lack of explicit sentiment information. Tang and Wan (2014) first proposed a bilingual sentiment embedding model using the original training data and the corresponding translations through a linear mapping rather than deep learning technique.

This paper proposes a denoising autoencoder based approach to learning bilingual sentiment word embeddings (**BSWE**) for CLSC, which incorporates sentiment polarities of text into the bilingual embeddings. The proposed approach learns BSWE with the original labeled documents and their translations instead of parallel corpora. The BSWE learning process consists of two phases: the unsupervised phase of semantic learning and the supervised phase of sentiment learning. In the unsupervised phase, sentiment words and their negation features are extracted from the source training data and their translations to represent paired documents. These features are used as inputs for a denoising autoencoder to learn the bilingual embeddings. In the supervised phase, sentiment polarity labels of documents are used to guide BSWE learning for incorporating sentiment information into the bilingual embeddings.

The learned BSWE are applied to project English training data and Chinese test data into a common space. In this space, a linear support vector machine (SVM) is used to perform training and testing. The experiments are carried on NLP&CC 2013 CLSC dataset, including book, DVD and music categories. Experimental results show that our approach achieves 80.68% average accuracy, which outperforms the state-of-the-art systems on this dataset. Although the BSWE are only evaluated on English-Chinese CLSC here, it can be popularized to many other languages.

The major contributions of this work can be summarized as follows:

- We propose bilingual sentiment word embeddings (BSWE) for CLSC based on deep learning technique. Experimental results show that the proposed BSWE significantly outperform the bilingual embeddings by incorporating sentiment information.
- Instead of large-scale parallel corpora, only the labeled English corpora and English-to-Chinese translations are required for BSWE learning. It is proved that in spite of the small-scale of training set, our approach outperforms the state-of-the-art systems in NLP&CC 2013 CLSC share task.
- We employ sentiment words and their negation features rather than all words in documents to learn sentiment-specific embeddings, which significantly reduces the dimension of input vectors as well as improves sentiment classification performance.

2 Related Work

In this section, we review the literature related to this paper from two perspectives: cross-language sentiment classification and embedding learning for sentiment classification.

2.1 Cross-language Sentiment Classification (CLSC)

The critical problem of CLSC is how to bridge the gap between the source language and target language. Machine translations or parallel corpora are usually employed to solve this problem. We present a brief review of CLSC from two aspects: machine translation based approaches and parallel corpora based approaches.

Machine translation based approaches use MT systems to project training data into the target language or test data into the source language. Wan (2009) proposed a co-training approach for CLSC. The approach first translated Chinese test data into English, and English training data into Chinese. Then, they performed training and testing in two independent views: English view and Chinese view. Gui et al. (2013) combined self-training approach with co-training approach by estimating the confidence of each monolingual system. Li et al. (2013) selected the samples in the source language that were similar to those in the target language to decrease the gap between two languages. Zhou et al. (2014a) proposed a combination CLSC model, which adopted denoising autoencoders (Vincent et al., 2008) to enhance the robustness to translation errors of the input.

Most recently, a number of studies adopt deep learning technique to learn bilingual representations with parallel corpora. Bilingual representations have been successfully applied in many NLP tasks, such as machine translation (Zou et al., 2013), sentiment classification (Chandar A P et al., 2013; Zhou et al., 2014b), text classification (Chandar A P et al., 2014), etc.

Chandar A P et al. (2013) learned bilingual representations with aligned sentences throughout two phases: the language-specific representation learning phase and the shared representation learning phase. In the language-specific representation learning phase, they applied autoencoders to obtain a language-specific representation for each entity in two languages respectively. In shared representation learning phase, pairs of parallel language-specific representations were passed to an autoencoder to learn bilingual representations. To joint language-specific representations and bilingual representations, Chandar A P et al. (2014) integrated the two learning phases into a unified process to learn bilingual embeddings. Zhou et al. (2014b) employed bilingual representations for English-Chinese CLSC. The work mentioned above employed aligned sentences in bilingual embedding learning process. However, in the sentiment classification process, only representations in the source language are used for training, and representations in the target language are used for testing, which ignores the interactions of semantic information between the source language and target language.

2.2 Embedding Learning for Sentiment Classification

Bilingual embedding learning algorithms focus on capturing syntactic and semantic similarities across languages, but ignore sentiment information. To date, many embedding learning algorithms have been developed for sentiment classification problem by incorporating sentiment information into word embeddings. Maas et al. (2011) presented a probabilistic model that combined unsupervised and supervised techniques to learn word vectors, capturing semantic information as well as sentiment information. Wang et al. (2014) introduced sentiment labels into Neural Network Language Models (Bengio et al., 2003) to enhance sentiment expression ability of word vectors. Tang et al. (2014) theoretically and empirically analyzed the effects of the syntactic context and sentiment information in word vectors, and showed that the syntactic context and sentiment information were equally important to sentiment classification.

Recent years have seen a surge of interest in word embeddings with deep learning technique (Bespalov et al., 2011; Glorot et al., 2011; Socher et al., 2011; Socher et al., 2012), which have been empirically shown to preserve linguistic regularities (Mikolov et al., 2013). Our work focuses on learning bilingual sentiment word embeddings (BSWE) with deep learning technique. Unlike the work of Chandar A P et al. (2014) that adopted parallel corpora to learn bilingual embeddings, we only use training data and their translations to learn BSWE. More importantly, sentiment information is integrated into bilingual embeddings to improve their performance in CLSC.

3 Bilingual Sentiment Word Embeddings (BSWE) for Cross-language Sentiment Classification

3.1 Denoising Autoencoder

It has been demonstrated that the denoising autoencoder could decrease the effects of translation errors on the performance of CLSC (Zhou et al., 2014a). This paper proposes a deep learning based approach, which employs the denoising autoencoder to learn the bilingual embeddings for CLSC.

A denoising autoencoder is the modification of an autoencoder. The autoencoder (Bengio et al., 2007) includes an encoder f_θ and a decoder $g_{\theta'}$. The encoder maps a d-dimensional input vector $\mathbf{x} \in [0,1]^d$ to a hidden representation $\mathbf{y} \in [0,1]^{d'}$ through a deterministic mapping $\mathbf{y} = f_\theta(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$, parameterized by $\theta = \{\mathbf{W}, \mathbf{b}\}$. $\mathbf{W}$ is a weight matrix, $\mathbf{b}$ is a bias term, and $\sigma(x)$ is the activation function. The decoder maps $\mathbf{y}$ back to a reconstructed vector $\hat{\mathbf{x}} = g_{\theta'}(\mathbf{y}) = \sigma(\mathbf{W}^T\mathbf{y} + \mathbf{c})$, parameterized by $\theta' = \{\mathbf{W}^T, \mathbf{c}\}$, where $\mathbf{c}$ is the bias term for reconstruction.

Through the process of encoding and decoding, the parameters θ and θ' of the autoencoder will be trained by gradient descent to minimize the loss function. The sum of reconstruction cross-entropies across the training set is usually used as the loss function:

$$l(x) = -\sum_{i=1}^{d}[\mathbf{x}_i \log \hat{\mathbf{x}}_i + (1-\mathbf{x}_i)\log(1-\hat{\mathbf{x}}_i)] \quad (1)$$

A denoising autoencoder enhances robustness to noises by corrupting the input $\mathbf{x}$ to a partially destroyed version $\tilde{\mathbf{x}}$. The desired noise level of the input $\mathbf{x}$ can be changed by adjusting the destruction fraction ν. For each input $\mathbf{x}$, a fixed number νd (d is the dimension of $\mathbf{x}$) of components are selected randomly, and their values are set to 0,

while the others are left untouched. Like an autoencoder, the destroyed input $\tilde{\mathbf{x}}$ is mapped to a latent representation $\mathbf{y} = f_\theta(\tilde{\mathbf{x}}) = \sigma(\mathbf{W}\tilde{\mathbf{x}} + \mathbf{b})$. Then $\mathbf{y}$ is mapped back to a reconstructed vector $\hat{\mathbf{x}}$ through $\hat{\mathbf{x}} = g_{\theta'}(\mathbf{y}) = \sigma(\mathbf{W}^T\mathbf{y} + \mathbf{c})$. The loss function of a denoising autoencoder is the same as that of an autoencoder. Minimizing the loss makes $\hat{\mathbf{x}}$ close to the input $\mathbf{x}$ rather than $\tilde{\mathbf{x}}$.

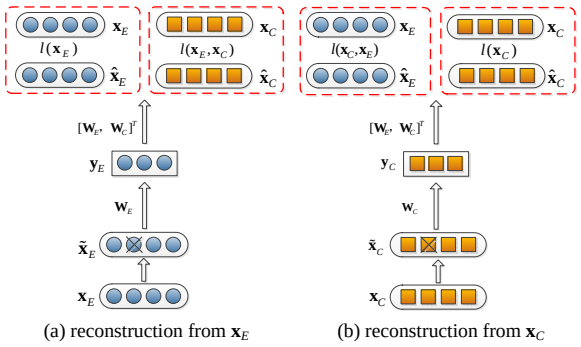
Our BSWE learning process can be divided into two phases: the unsupervised phase of semantic learning and the supervised phase of sentiment learning. In the unsupervised phase, a denoising autoencoder is employed to learn the bilingual embeddings. In the supervised phase, the sentiment information is incorporated into the bilingual embeddings based on sentiment labels of documents to obtain BSWE.

3.2 Unsupervised Phase of the Bilingual Embedding Learning

In the unsupervised phase, the English training documents and their Chinese translations are employed to learn the bilingual embeddings (Sentiment polarity labels of documents are not employed in this phase). Based on the English documents, 2,000 English sentiment words in MPQA subjectivity lexicon[1] are extracted by the Chi-square method (Galavotti et al., 2000). Their corresponding Chinese translations are used as Chinese sentiment words. Besides, some sentiment words are often modified by negation words, which lead to inversion of their polarities. Therefore, negation features are introduced to each sentiment word to represent its negative form.

We take into account 14 frequently-used negation words in English such as *not* and *none*; 5 negation words in Chinese such as 不 (*no*/*not*) and 没有 (*without*). A sentiment word modified by these negation words in the window [-2, 2] is considered as its negative form in this paper, while sentiment word features remain the initial meaning. Negation features use binary expressions. If a sentiment word is not modified by negation words, the value of its negation features is set to 0. Thus, the sentiment words and their corresponding negation features in English and Chinese are adopted to represent the document pairs $(\mathbf{x}_E, \mathbf{x}_C)$.

We expect that pairs of documents could be forced to capture the common semantic information of two languages. To achieve this, a denoising autoencoder is used to perform the reconstructions of paired documents in both English and Chinese. Figure 1 shows the framework of bilingual embedding learning.



(a) reconstruction from $\mathbf{x}_E$ (b) reconstruction from $\mathbf{x}_C$

Figure 1: The framework of bilingual embedding learning.

For the corrupted versions $\tilde{\mathbf{x}}_E$ ($\tilde{\mathbf{x}}_C$) of the initial input vector $\mathbf{x}_E$ ($\mathbf{x}_C$), we use the sigmoid function as the activation function to extract latent representations:

$$\mathbf{y}_E = f_\theta(\tilde{\mathbf{x}}_E) = \sigma(\mathbf{W}_E\tilde{\mathbf{x}}_E + \mathbf{b}) \quad (2)$$

$$\mathbf{y}_C = f_\theta(\tilde{\mathbf{x}}_C) = \sigma(\mathbf{W}_C\tilde{\mathbf{x}}_C + \mathbf{b}) \quad (3)$$

where $\mathbf{W}_E$ and $\mathbf{W}_C$ are the language-specific word representation matrices, corresponding to English and Chinese respectively. Notice that the bias $\mathbf{b}$ is shared to ensure that the produced representations in two languages are on the same scale.

For the latent representations in either language, we would like two decoders to perform reconstructions in English and Chinese respectively. As shown in Figure 1(a), for the latent representation $\mathbf{y}_E$ in English, one decoder is used to map $\mathbf{y}_E$ back to a reconstruction $\hat{\mathbf{x}}_E$ in English, and the other is used to map $\mathbf{y}_E$ back to a reconstruction $\hat{\mathbf{x}}_C$ in Chinese such that:

$$\hat{\mathbf{x}}_E = g_{\theta'}(\mathbf{y}_E) = \sigma(\mathbf{W}_E^T\mathbf{y}_E + \mathbf{c}_E) \quad (4)$$

$$\hat{\mathbf{x}}_C = g_{\theta'}(\mathbf{y}_E) = \sigma(\mathbf{W}_C^T\mathbf{y}_E + \mathbf{c}_C) \quad (5)$$

where $\mathbf{c}_E$ and $\mathbf{c}_C$ are the biases of the decoders in English and Chinese, respectively. Similarly, the same steps repeat for the latent representation $\mathbf{y}_C$ in Chinese, which are shown in Figure 1(b).

The encoder and decoder structures allow us to learn a mapping within and across languages. Specifically, for a given document pair $(\mathbf{x}_E, \mathbf{x}_C)$, we can learn bilingual embeddings to reconstruct $\mathbf{x}_E$ from itself (loss $l(\mathbf{x}_E)$), reconstruct $\mathbf{x}_C$ from itself (loss $l(\mathbf{x}_C)$), construct $\mathbf{x}_C$ from

[1] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon

$\mathbf{x}_E$ (loss $l(\mathbf{x}_E, \mathbf{x}_C)$), construct $\mathbf{x}_E$ from $\mathbf{x}_C$ (loss $l(\mathbf{x}_C, \mathbf{x}_E)$) and reconstruct the concatenation of $\mathbf{x}_E$ and $\mathbf{x}_C$ ($[\mathbf{x}_E, \mathbf{x}_C]$) from itself (loss $l([\mathbf{x}_E, \mathbf{x}_C], [\hat{\mathbf{x}}_E, \hat{\mathbf{x}}_C])$). The sum of 5 losses is used as the loss function of bilingual embeddings:

$$\begin{aligned} L =& l(\mathbf{x}_E) + l(\mathbf{x}_C) + l(\mathbf{x}_E, \mathbf{x}_C) + l(\mathbf{x}_C, \mathbf{x}_E) \\ &+ l([\mathbf{x}_E, \mathbf{x}_C], [\hat{\mathbf{x}}_E, \hat{\mathbf{x}}_C]) \end{aligned} \tag{6}$$

3.3 Supervised Phase of Sentiment Learning

In the unsupervised phase, we have learned the bilingual embeddings, which could capture the semantic information within and across languages. However, the sentiment polarities of text are ignored in the unsupervised phase. Bilingual embeddings without sentiment information are not effective enough for sentiment classification task. This paper proposes an approach to learning BSWE for CLSC, which introduces a supervised learning phase to incorporate sentiment information into the bilingual embeddings. The process of supervised phase is shown in Figure 2.

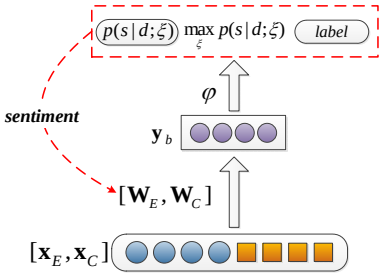


Figure 2: The supervised learning process.

For paired documents $[\mathbf{x}_E, \mathbf{x}_C]$, the sigmoid function is adopted as the activation function to extract latent bilingual representations $\mathbf{y}_b = \sigma([\mathbf{W}_E, \mathbf{W}_C][\mathbf{x}_E, \mathbf{x}_C] + \mathbf{b})$, where $[\mathbf{W}_E, \mathbf{W}_C]$ is the concatenation of $\mathbf{W}_E$ and $\mathbf{W}_C$.

The latent bilingual representation $\mathbf{y}_b$ is used to obtain the positive polarity probability $p(s = 1|d; \xi)$ of a document through a sigmoid function:

$$p(s = 1|d; \xi) = \sigma(\varphi^T \mathbf{y}_b + b_l) \tag{7}$$

where φ is the logistic regression weight vector and b_l is the bias of logistic regression. The sentiment label s is a Boolean value representing sentiment polarity of a document: $s = 0$ represents negative polarity and $s = 1$ represents positive polarity. Parameter $\xi^* = \{[\mathbf{W}_E, \mathbf{W}_C]^*, \mathbf{b}^*, \varphi^*, b_l^*\}$ is learned by maximizing the objective function according to the sentiment polarity label s_i of document d_i:

$$\xi^* = \arg\max_{\xi} \sum_{i=1} \log p(s_i|d_i; \xi) \tag{8}$$

Through the supervised learning phase, $[\mathbf{W}_E, \mathbf{W}_C]$ is optimized by maximizing sentiment polarity probability. Thus, rich sentiment information is encoded into the bilingual embeddings.

The following experiments will prove that the proposed BSWE outperform the traditional bilingual embeddings significantly in CLSC.

3.4 Bilingual Document Representation Method (BDR)

Once we have learned BSWE $[\mathbf{W}_E, \mathbf{W}_C]$, whose columns are representations for sentiment words, we can use them to represent documents in two languages.

Given an English training document d_E containing 2,000 sentiment word features $s_1, s_2, \cdots, s_{2,000}$ and 2,000 corresponding negation features, we represent it as the TF-IDF weighted sum of BSWE:

$$\phi_{d_E} = \sum_{i=1}^{4,000} TF - IDF(s_i)\mathbf{W}_{E.,s_i} \tag{9}$$

Similarly, for its Chinese translation d_C containing 2,000 sentiment word features $t_1, t_2, \cdots, t_{2,000}$ and 2,000 corresponding negation features, we represent it as:

$$\phi_{d_C} = \sum_{j=1}^{4,000} TF - IDF(t_j)\mathbf{W}_{C.,t_j} \tag{10}$$

We propose a bilingual document representation method (**BDR**) in this paper, which represents each document d_i with the concatenation of its English and Chinese representations $[\phi_{d_E}, \phi_{d_C}]$. BDR is expected to enhance the ability of sentiment expression for further improving the classification performance. Such bilingual document representations are fed to a linear SVM to perform sentiment classification.

4 Experiment

4.1 Experimental Settings

Data Set. The proposed approach is evaluated on NLP&CC 2013 CLSC dataset[2] [3]. The dataset con-

[2]http://tcci.ccf.org.cn/conference/2013/dldoc/evsam03.zip
[3]http://tcci.ccf.org.cn/conference/2013/dldoc/evdata03.zip

sists of product reviews on three categories: book, DVD, and music. Each category contains 4,000 English labeled data as training data (the ratio of the number of positive and negative samples is 1:1) and 4,000 Chinese unlabeled data as test data.

Tools. In our experiments, Google Translate[4] is adopted for both English-to-Chinese and Chinese-to-English translation. ICTCLAS (Zhang et al., 2003) is used as Chinese word segmentation tool. A denoising autoencoder is developed based on Theano system (Bergstra et al., 2010). BSWE are trained for 50 and 30 epochs in unsupervised phase and supervised phases respectively. SVM^{light} (Joachims, 1999) is used to train linear SVM sentiment classifiers

Evaluation Metric. The performance is evaluated by the classification accuracy for each category, and the average accuracy of three categories, respectively. The category accuracy is defined as:

$$Accuracy_c = \frac{\#system\_correct_c}{\#system\_total_c} \quad (11)$$

where c is one of the three categories, and $\#system\_correct_c$ and $\#system\_total_c$ stand for the number of being correctly classified reviews and the number of total reviews in the category c, respectively.

The average accuracy is shown as:

$$Average = \frac{1}{3}\sum_c Accuracy_c \quad (12)$$

4.2 Evaluations on BSWE

In this section, we evaluate the quality of BSWE for CLSC. The dimension of bilingual embeddings d is set to 50, and destruction fraction ν is set to 0.2.

Effects of Bilingual Embedding Learning Methods

We first compare our unsupervised bilingual embedding learning method with the parallel corpora based method. The parallel corpora based method uses the paired documents in the parallel corpus[5] to learn bilingual embeddings, while our method only uses the English training documents and their Chinese translations (Sentiment polarity labels of documents are not employed here). The Boolean feature weight calculation method is adopted to represent documents for bilingual embedding learning and BDR is employed to represent training data and test data for sentiment classification. To represent the paired documents in the parallel corpus, 27,597 English words and 31,786 Chinese words are extracted for bilingual embedding learning. Our method only needs 2,000 English sentiment words, 2,000 Chinese sentiment words, and their negation features, which significantly reduces the dimension of input vectors.

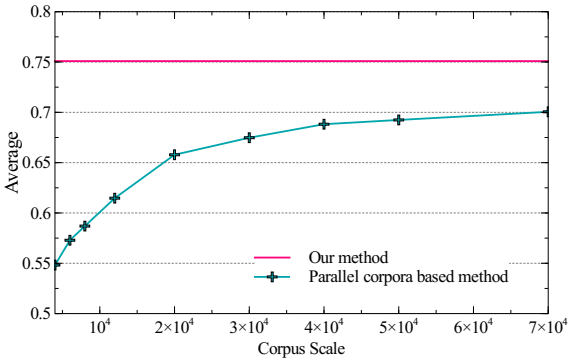


Figure 3: Our unsupervised bilingual embedding learning method vs. Parallel corpora based method.

The average accuracies on NLP&CC 2013 test data of the two bilingual embedding learning methods are shown in Figure 3. As can be seen from Figure 3, when the corpus scales of the two methods are the same (4,000 paired documents), our method (75.09% average accuracy) surpasses the parallel corpora method (54.82% average accuracy) by about 20%. With the scale of the parallel corpora increasing, the performance of parallel corpora based method is steadily improved. However, the performance is not as good as our bilingual embedding learning method. Though the document number of the parallel corpus is up to 70,000 , the average accuracy is only 70.05%. It is proved that our method is more suitable for learning bilingual embeddings for cross-language sentiment classification than the parallel corpora based method.

Effects of Feature Weight in Bilingual Embeddings

In this part, we compare the Boolean and TF-IDF feature weight calculation methods in bilingual embedding learning process.

Table 1 shows the classification accuracy with

[4]http://translate.google.cn/

[5]http://www.datatang.com/data/45485

Category	book	DVD	music	Average
Boolean	76.22%	74.30%	74.75%	75.09%
TF-IDF	76.65%	77.60%	74.50%	76.25%

Table 1: The classification accuracy with the Boolean and TF-IDF methods.

the Boolean and TF-IDF methods. Generally, the TF-IDF method performs better than the Boolean method. The average accuracy of the TF-IDF method is 1.16% higher than the Boolean method, which illustrates that the TF-IDF method could reflect the latent contribution of sentiment words to each document effectively. The TF-IDF weight calculation method is exploited in the following experiments. Notice that sentiment information is not yet introduced in the bilingual embeddings here.

Effects of Sentiment Information in BSWE

Incorporating sentiment information in the bilingual embeddings, the performance of bilingual embeddings (without sentiment information) and BSWE (with sentiment information) is compared in Figure 4.

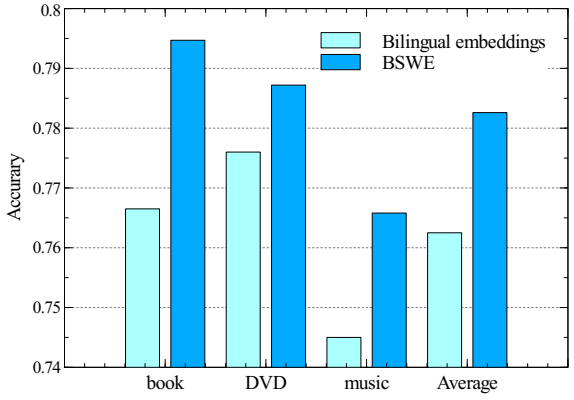


Figure 4: Performance comparison of the bilingual embeddings and BSWE.

As can be seen from Figure 4, by encoding sentiment information in the bilingual embeddings, the performance in book, DVD and music categories significantly improves to 79.47%, 78.72% and 76.58% respectively (2.82% increase in book, 1.12% in DVD, and 2.08% in music). The average accuracy reaches 78.26%, which is 2.01% higher than that of the bilingual embeddings. The experimental results indicate the effectiveness of sentiment information in the bilingual embedding learning. The BSWE learning approach is employed for CLSC in the following experiments.

Effects of Bilingual Document Representation Method

In this experiment, our bilingual document representation method (BDR) is compared with the following monolingual document representation methods.

En-En: This method represents training and test documents in English only with $\mathbf{W}_E$. English training documents and Chinese-to-English translations of test documents are both represented with $\mathbf{W}_E$.

Cn-Cn: This method represents training and test documents in Chinese only with $\mathbf{W}_C$. English-to-Chinese translations of training documents and Chinese test documents are both represented with $\mathbf{W}_C$.

En-Cn: This method represents English training documents with $\mathbf{W}_E$, while represents Chinese test documents with $\mathbf{W}_C$. Chandar A P et al. (2014) employed this method in their work.

BDR: This method adopts our bilingual document representation method, which represents training and test documents with both $\mathbf{W}_E$ and $\mathbf{W}_C$.

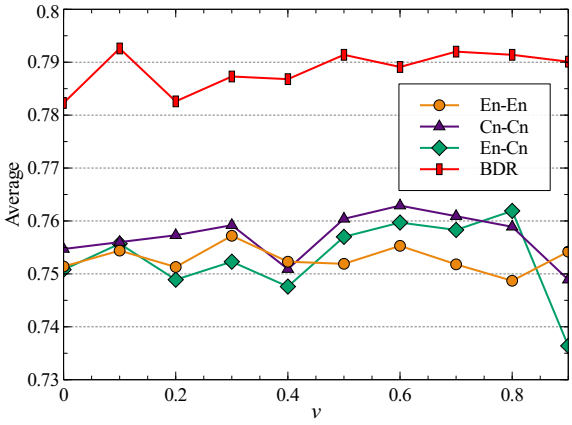


Figure 5: Effects of bilingual document representation method (BDR).

Figure 5 shows the average accuracy curves of different document representation methods with different destruction fraction ν. We vary ν from 0 to 0.9 with an interval of 0.1.

From Figure 5 we can see that En-En, Cn-Cn, and En-Cn get similar results. BDR performs constantly better than the other representation methods throughout the interval [0, 0.9]. The absolute superiority of BDR benefits from the enhanced ability of sentiment expression.

Meanwhile, when the input $\mathbf{x}$ is partially de-

stroyed (ν varies from 0.1 to 0.9), the performance of En-En, Cn-Cn and En-Cn remains stable, which illustrates the robustness of the denoising autoencoder to corrupting noises. In addition, the average accuracies of BDR in the interval $\nu \in [0.1, 0.9]$ are all higher than the average accuracy under the condition $\nu = 0$ (78.23%). Therefore, adding noises properly to the training data could improve the performance of BSWE for CLSC.

4.3 Influences of Dimension d and Destruction Fraction ν

Figure 6 shows the relationship between accuracies and dimension d of BSWE as well as that between accuracies and destruction fraction ν in autoencoders in different categories. Dimension of embeddings d varies from 50 to 500, and destruction fraction ν varies from 0.1 to 0.9.

As shown in Figure 6, the average accuracies generally move upward as dimension of BSWE increasing. Generally, the average accuracies keep higher than 80% with ν varying from 0.1 to 0.5 as well as dimension varying from 300 to 500. When $\nu = 0.1$ and $d = 400$, the average accuracy reaches the peak value 80.68% (category accuracy of 81.05% in book, 81.60% in DVD, and 79.40% in music). The experimental results show that in BSWE learning process, increasing the dimension of embeddings or properly adding noises to the training data helps improve the performance of CLSC. In this paper, we only evaluate BSWE when dimension d varies from 50 to 500. However, there is still space for further improvement if d continues to increase.

4.4 Comparison with Related Work

Table 2 shows comparisons of the performance between our approach and some state-of-the-art systems on NLP&CC 2013 CLSC dataset. Our approach achieves the best performance with an 80.68% average accuracy. Compared with the recent related work, our approach is more effective and suitable for eliminating the language gap.

Chen et al. (2014) translated Chinese test data into English and then gave different weights to sentiment words according to the subject-predicate component of sentiment words. They got 77.09% accuracy and took the 2nd place in NLP&CC 2013 CLSC share task. The machine translation based approach was limited by the translation errors.

System	book	DVD	music	Average
Chen et al. (2014)	77.00%	78.33%	75.95%	77.09%
Gui et al. (2013)	78.70%	79.65%	78.30%	78.89%
Gui et al. (2014)	80.10%	81.60%	78.60%	80.10%
Zhou et al. (2014a)	80.63%	80.95%	78.48%	80.02%
Our approach	**81.05%**	**81.60%**	**79.40%**	**80.68%**

Table 2: Performance comparisons on the NLP&CC 2013 CLSC dataset.

Gui et al. (2013; 2014) and Zhou et al. (2014a) adopted the multi-view approach to bridge the language gap. Gui et al. (2013) proposed a mixed CLSC model by combining co-training and transfer learning strategies. They achieved the highest accuracy of 78.89% in NLP&CC CLSC share task. Gui et al. (2014) further improved the accuracy to 80.10% by removing noise from the transferred samples to avoid negative transfers. Zhou et al. (2014a) built denoising autoencoders in two independent views to enhance the robustness to translation errors in the inputs and achieved 80.02% accuracy. The multi-view approach learns language-specific classifiers in each view during training process, which is difficult to capture the common sentiment information of the two languages. Our approach integrates the bilingual embedding learning into a unified process, and outperforms Chen et al. (2014), Gui et al. (2013), Gui et al. (2014) and Zhou et al. (2014a) by 3.59%, 1.79%, 0.58%, and 0.66% respectively. The superiority of our approach benefits from the unified bilingual embedding learning process and the integration of semantic and sentiment information.

5 Conclusion and Future Work

This paper proposes an approach to learning BSWE by incorporating sentiment information into the bilingual embeddings for CLSC. The proposed approach learns BSWE with the labeled documents and their translations rather than parallel corpora. In addition, BDR is proposed to enhance the sentiment expression ability which combines English and Chinese representations. Experiments on the NLP&CC 2013 CLSC dataset show that our approach outperforms the previous state-of-the-art systems as well as traditional bilingual embedding systems. The proposed BSWE are only evaluated on English-Chinese CLSC in this paper, but it can be popularized to other languages.

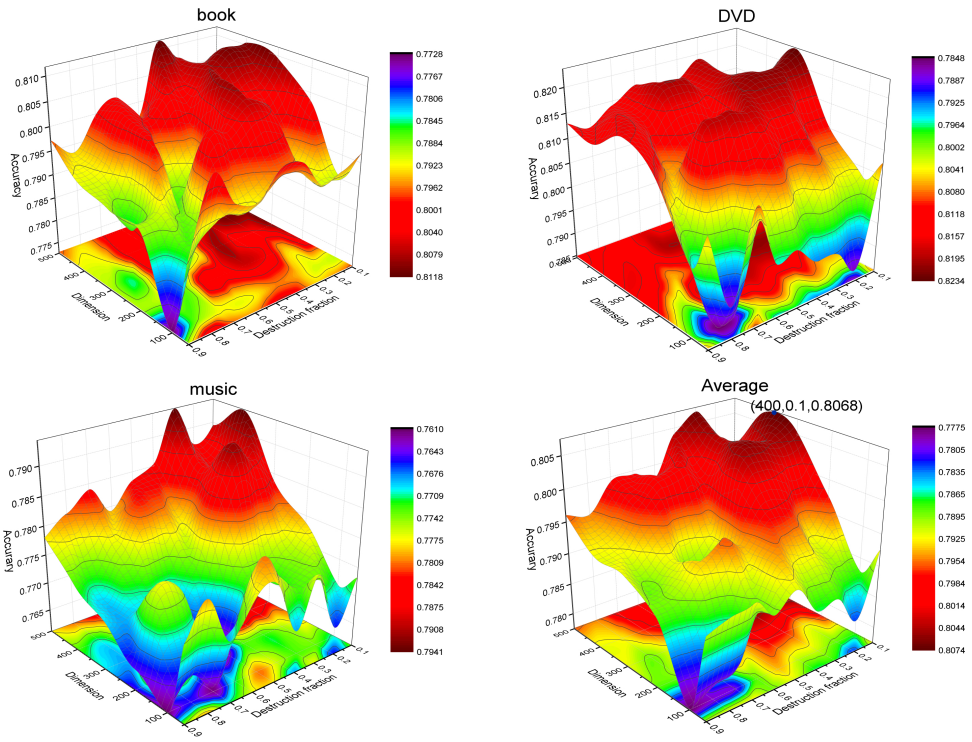


Figure 6: The relationship between accuracies and dimension d as well as that between accuracies and destruction fraction ν.

Both semantic and sentiment information play an important role in sentiment classification. In the following work, we will further investigate the relationship between semantic and sentiment information for CLSC, and balance their functions to optimize their combination for CLSC.

Acknowledgments

We wish to thank the anonymous reviewers for their valuable comments. This research is supported by National Natural Science Foundation of China (Grant No. 61272375).

References

Carmen Banea, Rada Mihalcea, Janyce Wiebe and Samer Hassan. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 127-135. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, vol 3: 1137-1155.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *Proceedings of Advances in Neural Information Processing Systems 19 (NIPS 06)*, pages 153-160. MIT Press.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8): 1798-1828. IEEE.

James Bergstra, Olivier Breuleux, Frederic Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, Yoshua Bengio. 2010. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*.

Dmitriy Bespalov, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. 2011. Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the Conference on Information and Knowledge Management*, pages 375-382. ACM.

Sarath Chandar A P, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar and Amrita Saha. 2013. Multilingual deep learning. In *Deep Learning Workshop at NIPS 2013*.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh M Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems*, pages 1853-1861.

Qiang Chen, Yanxiang He, Xule Liu, Songtao Sun, Min Peng, and Fei Li. 2014. Cross-Language Sentiment Analysis Based on Parser (in Chinese). *Acta Scientiarum Naturalium Universitatis Pekinensis*, 50 (1): 55-60.

G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science*, vol 313: 504-507.

Luigi Galavotti, Fabrizio Sebastiani, and Maria Simi. 2000. Feature Selection and Negative Evidence in Automated Text Categorization. In *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of 28th International Conference on Machine Learning*, pages 513-520.

Lin Gui, Ruifeng Xu, Jun Xu, Li Yuan, Yuanlin Yao, Jiyun Zhou, Qiaoyun Qiu, Shuwei Wang, Kam-Fai Wong, and Ricky Cheung. 2013. A mixed model for cross lingual opinion analysis. In *Proceedings of Natural Language Processing and Chinese Computing*, pages 93-104. Springer Verlag.

Lin Gui, Ruifeng Xu, Qin Lu, Jun Xu, Jian Xu, Bin Liu, and Xiaolong Wang. 2014. Cross-lingual Opinion Analysis via Negative Transfer Detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 860-865. Association for Computational Linguistics.

Thorsten Joachims. 1999. Making large-Scale SVM Learning Practical. Universität Dortmund.

Alistair Kennedy and Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2): 110-125.

Shoushan Li, Rong Wang, Huanhuan Liu, and Chu-Ren Huang. 2013. Active learning for cross-lingual sentiment classification. In *Proceedings of Natural Language Processing and Chinese Computing*, pages 236-246. Springer Verlag.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 142-150. Association for Computational Linguistics.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, pages 746-751. Association for Computational Linguistics.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79-86. ACM.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271-278. Association for Computational Linguistics.

Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 129-136. Bellevue.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic Compositionality through Recursive Matrix-Vector Spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1201-1211. Association for Computational Linguistics.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistic*, pages 1555-1565. Association for Computational Linguistics.

Xuewei Tang and Xiaojun Wan. 2014. Learning Bilingual Embedding Model for Cross-language Sentiment Classification. In *Proceedings of 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 134-141. IEEE.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096-1103. ACM.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235-243. Association for Computational Linguistics.

Yuan Wang, Zhaohui Li, Jie Liu, Zhicheng He, Yalou Huang, and Dong Li. 2014. Word Vector Modeling for Sentiment Analysis of Product Reviews. In *Proceedings of Natural Language Processing and Chinese Computing*, pages 168-180. Springer Verlag.

Huaping Zhang, Hongkui Yu, Deyi Xiong, and Qun Liu. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In *2nd SIGHAN workshop affiliated with 41th ACL*, pages 184-187. Association for Computational Linguistics.

Guangyou Zhou, Tingting He, and Jun Zhao. 2014b. Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification. In *Proceedings of Natural Language Processing and Chinese Computing*, pages 138-149. Springer Verlag.

Huiwei Zhou, Long Chen, and Degen Huang. 2014a. Cross-lingual sentiment classification based on denoising autoencoder. In *Proceedings of Natural Language Processing and Chinese Computing*, pages 181-192. Springer Verlag.

Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. 2013. Bilingual Word Embedding for Phrase-Based Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1393-1398.