Bridging the Language Gap: Learning Distributed Semantics for Cross-Lingual Sentiment Classification

Guangyou Zhou[1], Tingting He[1], and Jun Zhao[2]

[1] School of Computer, Central China Normal University,
152 Luoyu Road, Wuhan 430079, China
[2] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
{gyzhou,tthe}@mail.ccnu.edu.cn, jzhao@nlpr.ia.ac.cn

Abstract. Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scare target language by exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces of the source language data and that of the target language data. To address this challenge, previous work in the literature mainly relies on machine translation engines or bilingual lexicons to directly adapt labeled data from the source language to the target language. However, machine translation may change the sentiment polarity of the original data. In this paper, we propose a new model which uses stacked autoencoders to learn language-independent distributed representations for the source and target languages in an unsupervised fashion. Sentiment classifiers trained on the source language can be adapted to predict sentiment polarity of the target language with the language-independent distributed representations. We conduct extensive experiments on English-Chinese sentiment classification tasks of multiple data sets. Our experimental results demonstrate the efficacy of the proposed cross-lingual approach.

Keywords: Cross-lingual, Sentiment Classification, Deep Learning.

1 Introduction

With the development of web 2.0, more and more user generated sentiment data have been shared on the web. They exist in the form of user reviews on shopping or opinion sites, in posts of blogs or customer feedback in different languages. These labeled user generated sentiment data are considered as the most valuable resources for the sentiment classification task. However, such resources in different languages are very imbalanced. Manually labeling each individual language is a time-consuming and labor-intensive job, which makes cross-lingual sentiment classification essential for this application.

Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scare target language by

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 138–149, 2014.

exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces of the source language data and that of the target language data. To address this challenge, previous work in the literature mainly relies on machine translation engines or bilingual lexicons to directly adapt labeled data from the source language to the target language [17,26,24,16,12,27]. Although the machine translation based approaches are intuitive and have advanced the task of cross-lingual sentiment classification, they have certain limitations. First, machine translation may change the sentiment polarity of the original data [9]. For example, the negative English sentence "it is too beautiful to be true" is translated to a positive sentence in Chinese "实在是太漂亮是真实的" by Google Translate (http://translate.google.com/), which literally means "it is too beautiful and true". Second, many sentiment indicative words cannot be learned from the translated labeled data due to the limited coverage of vocabulary in the machine translation results. Recently, Duh et al. [3] report a low overlap between the vocabulary of English documents and the documents translated from Japanese to English, and the experiments also show that vocabulary coverage has a strong correlation with sentiment classification accuracy. Third, translating all the sentiment data in one language into the other language is a time-consuming and labor-intensive job in reality.

In this paper, we propose a deep learning approach, which uses stacked autoencoders [2] to learn language-independent distributed representations of data for cross-lingual sentiment classification. Our model is firstly trained on a large-scale bilingual parallel data and then projects the source language and the target language into a bi-lingual space that fuses the two types of information together. The goal of our model is to learn distributed representations through a hierarchy of network architectures. The learned distributed representations can be used to bridge the gap between the source language and the target language. For example, if we have learned language-independent distributed representations English and Chinese sentiment data, then a classifier trained on labeled English sentiment data can be used to classify Chinese sentiment data.

The novelty of our approach lies in that we employs a deep learning approach to project the source language and the target language into a language-independent unified representations. Our work shares certain intuition with the mixture model for cross-lingual sentiment classification [9] and the bilingual word embeddings used in cross-lingual sentiment classification [11] and phrase-based machine translation [29]. A common property of these approaches is that a word-level alignment (extracted using GIZA++) of bilingual parallel corpus is leveraged [8,9,11,29]. In this paper, we only require alignment parallel sentences and do not rely on word-level alignments of bilingual corpus during training, which simplifies the learning procedure.

To evaluate the effectiveness of the proposed approach, we conduct experiments on the task of English-Chinese cross-lingual sentiment classification. The empirical results show the proposed approach is very effective for cross-lingual sentiment classification, and outperforms other comparison methods.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 presents our proposed learning distributed semantics for cross-lingual sentiment classification. Section 4 presents the experimental results. Finally, we conclude this paper in section 5.

2 Related Work

2.1 Monolingual Sentiment Classification

Sentiment classification has gained wide interest in natural language processing (NLP) community. Methods for automatically classifying sentiments expressed in products and movie reviews can roughly be divided into supervised and unsupervised (or semi-supervised) sentiment analysis. Supervised techniques have been proved promising and widely used in sentiment classification [13,14,7]. However, the performance of these methods relies on manually labeled training data. In some cases, the labeling work may be time-consuming and expensive. This motivates the problem of learning robust sentiment classification via unsupervised (or semi-supervised) paradigm.

The most representative way to perform semi-supervised paradigm is to employ partial labeled data to guide the sentiment classification [4,18,6]. However, we do not have any labeled data at hand in many situations, which makes the unsupervised paradigm possible. The most representative way to perform unsupervised paradigm is to use a sentiment lexicon to guide the sentiment classification [22,20,28] or learn sentiment orientation of a word from its semantically related words mined from the lexicon [15]. Sentiment polarity of a word is obtained from off-the-shelf sentiment lexicon, the overall sentiment polarity of a document is computed as the summation of sentiment scores of the words in the document. All these work focuses on monolingual sentiment classification, we point the readers to recent books [14,7] for an in-depth survey of literature on sentiment classification.

2.2 Cross-Lingual Sentiment Classification

Cross-lingual sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of data in a label-scare target language by exploiting labeled data from a label-rich language. The fundamental challenge of cross-lingual learning stems from a lack of overlap between the feature spaces of the source language data and that of the target language data.

To bridge the language gap, previous work in the literature mainly relies on machine translation engines or bilingual lexicons to directly adapt labeled data from the source language to the target language. Banea et al. [1] employed the machine translation engines to bridge the language gap in different languages for multilingual subjectivity analysis. Wan [23] and Wan [24] proposed to use ensemble methods to train Chinese sentiment classification model on English labeled data and their Chinese translations. English labeled data are first translated into

Chinese, and then the bi-view sentiment classifiers are trained on English and Chinese labeled data respectively. Pan et al. [12] proposed a bi-view non-negative matrix tri-factorization (BNMTF) model for cross-lingual sentiment classification problem. They employed machine translation engines so that both training and test data are able to have two representations, one in source language and the other in target language. The proposed model is derived from the non-negative matrix factorization models in both languages in order to make more accurate prediction. Prettenhofer and Stein [16] proposed a cross-lingual structural correspondence learning (CL-SCL) method to induce language-independent features. Instead of using machine translation engines to translate labeled text, the authors first selected a subsect of pivot features in the source language to translate them into the target language, and then use these pivot pairs to induce cross-lingual representations by modeling the correlations between pivot features and non-pivot features in an unsupervised fashion. Recently, Xiao and Guo [27] used the similar idea with [16] for cross-lingual sentiment classification. Instead of in an fully unsupervised fashion, Xiao and Guo [27] performed representation learning in a semi-supervised manner by directly incorporating discriminative information with respect to the target prediction task. In this paper, we propose a deep learning approach, which uses stacked autoencoders [2] to learn language-independent distributed representations of data instead of machine translation engines.

Another group of works propose to use an unlabeled parallel corpus to induce language-independent representations [8,9]. They assume parallel sentences in the corpus should have the same sentiment polarity and labeled data in both the source and target languages are available. However, this method requires labeled data in both the source and target language, which are not always readily available [9]. Meng et al. [9] proposed a generative cross-lingual mixture model (CLMM) to learn previously unseen sentiment words from the large bilingual parallel data. A common property of this approach is that a word-level alignment (extracted using GIZA++) of bilingual parallel corpus is leveraged [9]. In this paper, we only require alignment parallel sentences and do not rely on word-level alignments of bilingual corpus during training, which simplifies the learning procedure.

3 Learning Distributed Semantics for Cross-Lingual Sentiment Classification

3.1 Model Formulation

Recently, parallel data in multiple languages provides an alternative way for multiview representations, as parallel texts share their semantics, and thus one language can be used to ground the other. Some work has exploited this idea to learn distributed representations at the word level [11,29]. A common property of these approaches is that a word-level alignment (extracted using GIZA++) of bilingual parallel corpus is leveraged [11,29]. In this paper, we only require

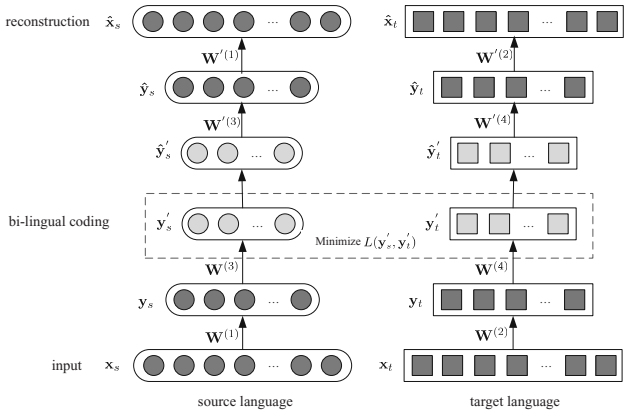

Fig. 1. Denoising stacked autoencoders (DAEs) trained on large-scale parallel sentence pairs $(\mathbf{x}_s, \mathbf{x}_t)$. Input to the model are binary bag-of-words vector representations obtained from the source language and the target language. The model minimize the distance between the sentence level bi-lingual coding of bitext $L(\mathbf{y}'_s, \mathbf{y}'_t)$ as well as the reconstruction errors from the source language and the target language.

alignment parallel sentences and do not rely on word-level alignments of bilingual corpus during training, which simplifies the learning procedure.

Given a large-scale parallel sentence pairs $(\mathbf{x}_s, \mathbf{x}_t)$, we would like to use it to learn distributed representations in both languages that are aligned. The idea is that a shared representation of two parallel sentences would be forced to capture the common information between two languages. Figure 1 shows the model architecture. For each sentence with binary bag-of-words representation $\mathbf{x}_s$ in the source language and an associated binary bag-of-words representation $\mathbf{x}_t$ for the same sentence in the target language, we use the hyperbolic tangent function as the activation function for an encoder f_θ and a decoder $g_{\theta'}$. The weights of each autoencoder are tied, i.e., $\mathbf{W}'^{(1)} = \mathbf{W}^{(1)}$ in Figure 1. We employ denoising stacked autoencoders (DAEs) for pre-training the sentences in each language. For example in Figure 1, let $\widetilde{\mathbf{x}}_s$ and denote the corrupted versions of the initial input vector $\mathbf{x}_s$, we have the following high-level latent representations: $\mathbf{y}_s = f_{\theta_s}(\widetilde{\mathbf{x}}_s) = s(\mathbf{W}^{(1)}\widetilde{\mathbf{x}}_s + \mathbf{b}^{(1)})$, $\mathbf{y}'_s = f_{\theta_s}(\mathbf{y}_s) = s(\mathbf{W}^{(3)}\mathbf{y}_s + \mathbf{b}^{(3)})$. Essentially, the same steps repeat for the input vector $\mathbf{x}_t$.

During the decoding phase, we want to be able to perform a reconstruction of the original sentence in any of the languages. In particular, given a representation in any language, we'd like a decoder $g_{\theta'_s}$ that can perform a reconstruction in the source language and another decoder $g_{\theta'_t}$ that can perform a reconstruction in the target language. Given the reconstruction layers, we have $\hat{\mathbf{y}}'_\mathbf{s} = g_{\theta'_s} = s(\mathbf{W}'^{(5)}\hat{\mathbf{y}} + \mathbf{b}'^{(5)})$, $\hat{\mathbf{y}}_s = g_{\theta'_s} = s(\mathbf{W}'^{(3)}\hat{\mathbf{y}}' + \mathbf{b}'^{(3)})$, and $\hat{\mathbf{x}}_s = g'_{\theta_s} = s(\mathbf{W}'\hat{\mathbf{y}}_s + \mathbf{b}'^{(1)})$. Essentially, the same steps repeat for the reconstruction process of $\hat{\mathbf{x}}_t$.

The encoder/decoder decomposition allows us to learn a mapping within each language and across the languages. Specially, for a given parallel sentence pair $(\mathbf{x}_s, \mathbf{x}_t)$, we can train the model to (1) reconstruct $\mathbf{x}_s$ from itself (loss $L(\mathbf{x}_s, \hat{\mathbf{x}}_s)$);

(2) reconstruct $\mathbf{x}_t$ from itself (loss $L(\mathbf{x}_t, \hat{\mathbf{x}}_t)$); and (3) distance between the sentence level encoding of the bitext (loss $L(\mathbf{y}'_s, \mathbf{y}'_t)$). The overall objective function is therefore the weight sum of these errors over a set of binary bag-of-words input vectors $\mathcal{C} = \{(\mathbf{x}_s^{(1)}, \mathbf{x}_t^{(1)}), \cdots, (\mathbf{x}_s^{(n)}, \mathbf{x}_t^{(n)})\}$:

$$J(\mathbf{x}_s, \mathbf{x}_t; \theta, \theta') = \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}} \Big\{ L(\mathbf{x}_s, \hat{\mathbf{x}}_s) + L(\mathbf{x}_t, \hat{\mathbf{x}}_t) + L(\mathbf{y}'_s, \mathbf{y}'_t) + \frac{\lambda}{2}(\|\theta\|_2 + \|\theta'\|_2) \tag{1}$$

where L is a loss function, such as cross-entropy. $\theta = \{\theta_s, \theta_t\}$ and $\theta' = \{\theta'_s, \theta'_t\}$ are the set of all model parameters. Note that we use tied weights for the stacked autoencoder, i.e., $\mathbf{W}^{(1)} = \mathbf{W}'^{(1)}$. In our experiments, we also add the constraints $\mathbf{b}^{(1)} = \mathbf{b}^{(2)}$, $\mathbf{b}^{(3)} = \mathbf{b}^{(4)}$, $\mathbf{b}'^{(1)} = \mathbf{b}'^{(2)}$ and $\mathbf{b}'^{(3)} = \mathbf{b}'^{(4)}$ before the nonlinearity across encoders, to encourage the encoders in both languages to produce representations on the same scale.

3.2 Learning Algorithm

Let $\theta = \{\theta_s, \theta_t\} = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}, \mathbf{W}^{(4)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}, \mathbf{b}^{(3)}, \mathbf{b}^{(4)}\}$ and $\theta' = \{\theta'_s, \theta'_t\} = \{\mathbf{W}'^{(1)}, \mathbf{W}'^{(2)}, \mathbf{W}'^{(3)}, \mathbf{W}'^{(4)}, \mathbf{b}'^{(1)}, \mathbf{b}'^{(2)}, \mathbf{b}'^{(3)}, \mathbf{b}'^{(4)}\}$ be the set of our model parameters, then the gradient becomes:

$$\frac{\partial L}{\partial \theta} = \sum_{(\mathbf{x}_s, \mathbf{x}_t) \in \mathcal{C}} \frac{\partial J(\mathbf{x}_s, \mathbf{x}_t; \theta, \theta')}{\partial \theta} + \lambda\theta. \tag{2}$$

The gradient can be computed efficiently via backpropagtion. Since the derivation of the minimization of the distance between the sentence-level bi-lingual coding of bitext and the reconstruction errors can also modify the matrices $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}^{(3)}$ and $\mathbf{W}^{(4)}$, the above objective is not necessarily continuous and a step in the gradient descent direction may not necessarily decrease the objective. However, we find that L-BFGS run over the unlabeled parallel data to minimize the objective works well in practice, and that convergence is smooth, with the algorithm typically finding a good solution quickly.

3.3 Cross-Lingual Sentiment Classification

Once we have learned the parameters θ and θ', we can transform the binary bag-of-words representation of the training data from the source language into the bi-lingual coding space using the learned parameter θ, and then train a simple sentiment classification model using a linear support vector machine (SVM) [5]. For each of the test data from the target language, we also transform its bag-of-words representations into the bi-lingual coding space using the learned parameter θ' and then predict the sentiment polarity of the test data using the trained classification model.

Table 1. Statistics of data sets used in this paper

	MPQA	NTCIR-EN	NTCIR-CH
Positive	1,471 (30%)	528 (30%)	2,378 (55%)
Negative	3,487 (70%)	1,209 (70%)	1,916 (44%)
Total	4,958	1,737	4,294

4 Experiments

4.1 Experimental Setup

In this section, we conduct experiments for cross-lingual sentiment classification. We focus on the two common cross-lingual sentiment classification settings. In the first setting, no labeled data in the target language are available. This task has realistic significance, since in some situations we need to quickly develop a sentiment classifier for languages that we do not have labeled data in hand. In this case, we classify documents in the target language using only labeled data in the source language. In the second setting, we have some labeled data in the target language. In this case, a more reasonable method is to make full use the labeled data in the source language and the target language to build the sentiment classification model. In our experiments, for each setting, we consider two cases, one is English as the source language and Chinese as the target language, another is Chinese as the source language and English as the target language.

4.2 Data Set

For cross-lingual sentiment classification, we use the benchmark data set described in [8,9]. The labeled data sets consist of two English data sets and one Chinese data set.

MPQA-EN (Labeled English Data): The multi-perspective question answering (MPQA-EN) corpus [25] consists of newswire documents manually labeled with subjectivity information. Following the literature [8], we also discard the sentences with both positive and negative strong expressions.

NTCIR-EN (Labeled English Data) and **NTCIR-CH (Labeled Chinese Data):** The NTCIR opinion analysis task [19] provides sentiment labeled news data in Chinese and English. The sentences with a sentiment polarity agreed to by at least two annotators are extracted. In this paper, we use the Chinese data from NTCIR-6 as our Chinese labeled data, the English data from NTCIR-6 and NTCIR-7 as our English labeled data. The Chinese sentences are segmented using the Stanford Chinese word segmenter [21].

The statistics of the data sets are shown in Table 1. In our experiments, we evaluate four settings of the data: (1) MPQA-EN → NTCIR-CH; (2) NTCIR-EN → NTCIR-CH; (3) NTCIR-CH → MPQA-EN; and (4) NTCIR-CH → NTCIR-EN, where the word before an arrow corresponds with the source language and the word after an arrow corresponds with the target language.

To learn the parameters θ and θ', we use the Chinese-English parallel corpus [10]. As mentioned earlier, unlike the previous work [8,9,11], we do not use any word alignment between these parallel sentences. Specifically, we segment the Chinese sentences using the Stanford Chinese word segmenter [21] and remove all punctuations from the parallel sentences.

4.3 Model Architecture

Our model has many hyper-parameters, we set these parameters empirically as follows: the source language autoencoder (see Figure 1, left side) and the target language autoencoder (see Figure 1, right side) consist of 1000 hidden units which are then mapped to the second hidden layer with 500 units (the corruption parameter is set to $\nu = 0.5$). The 500 source language and the 500 target language hidden units are fed to a bi-lingual autoencoder containing 500 latent units. We use the model described above and the language-independent representations obtained from the output of the bi-lingual latent layer for the cross-lingual task. Note that some performance gains could be expected if these parameters are optimized on the development set.

4.4 Baseline Methods

In our experiments, we compare our proposed DAEs with the following baseline methods:

SVM: This method learns a SVM classifier for each language given the monolingual labeled data. In this paper, SVM-light [5] is used for all the SVM-related experiments.

MT-SVM: This method employs Google Translate (http://translate.google.com) to translate the labeled data from the source language (e.g., English) to the target language (e.g., Chinese) and uses the translated results to train a SVM classifier for the target language.

MT-Cotrain: This method is based on a co-training framework described in [24]. For easy description, we assume that the source language is English while the target language is Chinese. First, two monolingual SVM classifiers are trained on English labeled data and Chinese data translated from English labeled data. Second, the two classifiers make prediction on Chinese unlabeled data and their English translation, respectively. Third, the most confidently predicted English and Chinese documents are added to the training set and the two monolingual SVM classifier are re-trained on the expanded training set. Following the literature [9], we repeat the second and third steps 100 times to obtain the final classifiers.

Joint-Train: This method uses English labeled data and Chinese labeled data to obtain initial parameters for two maximum entropy classifiers, and then conduct

Table 2. Sentiment classification accuracy for Chinese only using English labeled data. Improvements of different methods over baseline MT-SVM are shown in parentheses.

	Method	MPQA-EN → NTCIR-CH	NTCIR-EN → NTCIR-CH
1	SVM	N/A	N/A
2	MT-SVM	54.33	62.34
3	MT-Cotrain	59.11 (+4.78)	65.13 (+2.79)
4	Joint-Train	N/A	N/A
5	CLMM	71.52 (+17.19)	70.96 (+8.62)
6	DRW	72.27 (+17.94)	71.63 (+9.29)
7	**DAEs**	**72.85 (+18.52)**	**72.21 (+9.87)**

EM-iterations to update the parameters to gradually improve the agreement of the two monolingual classifiers on the unlabeled parallel data [8].

CLMM: This method proposes a generative cross-lingual mixture model (CLMM) [9] and learns previously unseen sentiment words from the large-scale bilingual parallel data to improve the vocabulary coverage.

DRW: This is the state-of-the-art method for cross-lingual sentiment classification described in [11]. This method learns distributed representations of words via multitask and word alignment for cross-lingual sentiment classification.

4.5 Cross-Lingual Sentiment Classification Only Using Source Language Labeled Data

In this section, we investigate cross-lingual sentiment classification towards the case that we have only labeled data from the source language. The first set of experiments are conducted on using only English labeled data to build sentiment classifier for Chinese sentiment classification. This is a challenging task since we do not have any Chinese labeled data in hand.

Table 2 shows the accuracy of the baseline systems as well as the proposed model (DAEs). As seen from the table, our proposed approach DAEs outperforms all baseline methods for Chinese sentiment classification only using the labeled English data. Specifically, our proposed approach improves the accuracy, compared to MT-SVM, by 18.52% and 9.87% (row 2 vs. row 7) on Chinese in the first setting and in the second setting, respectively. Meanwhile, the accuracy of MT-SVM on NTCIR-EN → NTCIR-CH is much better than that on MPQA-EN → NTCIR-CH. The reason may be that NTCIR-EN and NTCIR-CH cover similar topics. Besides, we also observe that using a parallel corpus instead of machine translations can improve the classification accuracy (row 2 and row 3 vs. row 5, row 6 and row 7). Moreover, Our proposed DAEs outperforms CLMM and DRW (row 5 and row 6 vs. row 7, the comparisons are mildly significant with t-test (p-value < 0.08)). The reason may be that our method can effectively learn sentence-level distributed representations rather than using the off-the-shelf word alignment tools (e.g., GIZA++) to bridge the language gap.

Table 3. Sentiment classification accuracy for English only using Chinese labeled data. Improvements of different methods over baseline MT-SVM are shown in parentheses.

	Method	NTCIR-CH → MPQA-EN	NTCIR-CH → NTCIR-EN
1	SVM	N/A	N/A
2	MT-SVM	52.47	58.51
3	MT-Cotrain	58.63 (+6.16)	63.72 (+5.21)
4	Joint-Train	N/A	N/A
5	CLMM	68.29 (+15.82)	69.15 (+10.64)
6	DRW	70.85 (+18.38)	72.57 (+14.06)
7	**DAEs**	**71.42 (+18.95)**	**73.38 (+14.87)**

Table 4. Sentiment classification accuracy for Chinese by using English and Chinese labeled data. Improvements of different methods over baseline SVM are shown in parentheses.

	Method	MPQA-EN → NTCIR-CH	NTCIR-EN → NTCIR-CH
1	SVM	80.58	80.58
2	MT-SVM	54.33 (-26.25)	62.34 (-18.24)
3	MT-Cotrain	80.93 (+0.35)	82.28 (+2.79)
4	Joint-Train	83.42 (+2.84)	83.11 (+2.53)
5	CLMM	83.02 (+2.44)	82.73 (+2.15)
6	DRW	83.54 (+2.96)	83.26 (+2.68)
7	**DAEs**	**83.81 (+3.23)**	**83.59 (+3.01)**

The second set of experiments are conducted on using only Chinese labeled data to build sentiment classifier for English sentiment classification. Table 3 shows the sentiment classification accuracy for English using only Chinese labeled data. From this table, we have the similar observations as in Table 2.

4.6 Cross-Lingual Sentiment Classification Using Source Language and Target Language Labeled Data

The third set of experiments are conducted on using both English labeled data and Chinese labeled data to build the Chinese sentiment classifier. We conduct 5-fold cross validation on Chinese labeled data and use the similar settings with [9].

Table 4 shows the average accuracy of baseline systems as well as our proposed DAEs. From this table, we can see that SVM performs significantly better than MT-SVM. The reason may be that we use the original Chinese labeled data instead of translated Chinese labeled data. We also find that all four methods which employ the unlabeled parallel corpus, namely MT-Cotrain, Joint-Train, CLMM and DAEs, still show improvements over the baseline SVM. Moreover, our proposed DAEs outperforms than DRW and obtains the state-of-the-art accuracy on both data sets. This again validates that learning sentence-level distributed representations is better than using word alignment tools for cross-lingual sentiment classification. Due to limited space, we do not present the

experimental results for English and some other related discussions, we will leave these works for further research.

5 Conclusion

In this paper, we present a model that uses stacked autoencoders to learn distributed representations through a hierarchy of network architectures. The learned distributed representations can be used to bridge the gap between the source language and the target language. To evaluate the effectiveness of the proposed approach, we conduct experiments on the task of English-Chinese cross-lingual sentiment classification. The empirical results show the proposed approach is effective for cross-lingual sentiment classification, and outperforms other comparison methods.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 61303180, No. 61272332 and No. 61333018), the Beijing Natural Science Foundation (No. 4144087), CCF Opening Project of Chinese Information Processing, and also Sponsored by CCF-Tencent Open Research Fund. We thank the anonymous reviewers for their insightful comments.

References

1. Banea, C., Mihalcea, R., Wiebe, J., Hassan, S.: Multilingual Subjectivity Analysis Using Machine Translation. In: Proceedings of EMNLP, pp. 127–135
2. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Proceedings of NIPS. Universite De Montreal, Montreal Quebec
3. Duh, K., Fujino, A., Nagata, M.: Is Machine Translation Ripe for Cross-lingual Sentiment Classification? In: Proceedings of ACL, pp. 429–433
4. Goldberg, A.B., Zhu, X.: Seeing Stars when There Aren'T Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. In: Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, pp. 45–52
5. Joachims, T.: Making Large-scale Support Vector Machine Learning Practical. Advances in Kernel Methods, pp. 169-184
6. Li, S., Wang, Z., Zhou, G., Lee, S.Y.M.: Semi-supervised Learning for Imbalanced Sentiment Classification. In: Proceedings of IJCAI, pp, pp. 1826–1831
7. Liu, B.: Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies
8. Lu, B., Tan, C., Cardie, C., Tsou, B.K.: Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In: Proceedings of ACL, pp. 320–330
9. Meng, X., Wei, F., Liu, X., Zhou, M., Xu, G., Wang, H.: Cross-lingual Mixture Model for Sentiment Classification. In: Proceedings of ACL, pp. 572–581
10. Munteanu, D.S., Marcu, D.: Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. Comput. Linguist., 477–504
11. Klementiev, A., Titov, I., Bhattarai, B.: Inducing Crosslingual Distributed Representations of Words. In: Proceedings of COLING

12. Pan, J., Xue, G.-R., Yu, Y., Wang, Y.: Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS, vol. 6634, pp. 289–300. Springer, Heidelberg (2011)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In: Proceedings of EMNLP, pp. 79–86
14. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. In: Found. Trends Inf. Retr., pp. 1–135
15. Peng, W., Park, D.H.: Generate Adjective Sentiment Dictionary for Social Media Sentiment Analysis Using Constrained Nonnegative Matrix Factorization. In: Proceedings of ICWSM
16. Prettenhofer, P., Stein, B.: Cross-language Text Classification Using Structural Correspondence Learning. In: Proceedings of ACL, pp. 1118–1127
17. Shanahan, J.G., Grefenstette, G., Qu, Y., Evans, D.A.: Mining Multilingual Opinions through Classification and Translation. In: Proceedings of CIKM
18. Sindhwani, V., Melville, P.: Document-word co-regularization for semi-supervised sentiment analysis. In: Proceedings of ICDM
19. Seki, Y., Evans, D.K., Ku, L.-W., Chen, H.-H., Kando, N., Lin, C.-Y.: Overview of Opinion Analysis Pilot Task at NTCIR-6. In: Proceedings of the Workshop Meeting of the National Institute of Informatics (NII) Test Collection for Information Retrieval Systems (NTCIR), pp. 265–278
20. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based Methods for Sentiment Analysis. Comput. Linguist. 267-307
21. Tseng, H.: A conditional random field word segmenter. In: Fourth SIGHAN Workshop on Chinese Language Processing
22. Turney, P.D.: Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL, pp. 417–424
23. Wan, X.: Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis. In: Proceedings of EMNLP, pp. 553–561
24. Wan, X.: Co-training for Cross-lingual Sentiment Classification. In: Proceedings of ACL-IJCNLP, pp. 235–243
25. Wiebe, J., Cardie, C.: Annotating expressions of opinions and emotions in language. Language Resources and Evaluation. Language Resources and Evaluation (formerly Computers and the Humanities)
26. Wu, K., Wang, X., Lu, B.-L.: Cross language text categorization using a bilingual lexicon. In: Proceedings of IJCNLP
27. Xiao, M., Guo, Y.: Semi-Supervised Representation Learning for Cross-Lingual Text Classification. In: Proceedings of ACL, pp. 1465–1475
28. Zhou, G., Zhao, J., Zeng, D.: Sentiment Classification with Graph Co-Regularization. In: Proceedings of COLING, pp. 1331–1340
29. Zou, W.Y., Socher, R., Cer, D.M., Manning, C.D.: Bilingual Word Embeddings for Phrase-Based Machine Translation. In: Proceedings of ACL, pp. 1393–1398