Deep Transfer Learning for Social Media Cross-Domain Sentiment Classification

Chuanjun Zhao[1], Suge Wang[1,2(✉)], and Deyu Li[1,2]

[1] Shanxi University, Taiyuan 030006, Shanxi, China
zhaochuanjun@foxmail.com, {wsg,lidy}@sxu.edu.cn
[2] Key Laboratory of Computational Intelligence, Chinese Information Processing of Ministry of Education, Taiyuan 030006, Shanxi, China

Abstract. Social media sentiment classification has important theoretical research value and broad application prospects. Deep neural networks have been applied into social media sentiment mining tasks successfully with excellent representation learning and high efficiency classification abilities. However, it is very difficult to collect and label large scale training data for deep learning. In this case, deep transfer learning (DTL) can transfer abundant source domain knowledge to target domain using deep neural networks. In this paper, we propose a two-stage bidirectional long short-term memory (Bi-LSTM) and parameters transfer framework for short texts cross-domain sentiment classification tasks. Firstly, Bi-LSTM networks are pre-trained on a large amount of fine-labeled source domain training data. We fine-tune the pre-trained Bi-LSTM networks and transfer the parameters using target domain training data and continuing back propagation. The fine-tuning strategy is to transfer bottom-layer (general features) and retrain top-layer (specific features) to the target domain. Extensive experiments on four Chinese social media data sets show that our method outperforms other baseline algorithms for cross-domain sentiment classification tasks.

Keywords: Transfer learning · Long short-term memory · Parameters transfer · Cross-domain sentiment classification

1 Introduction

Sentiment analysis, also known as subjectivity analysis or opinion mining, is the process of analyzing, processing, summarizing, and reasoning the subjective texts. Sentiment analysis can also be subdivided into sentiment polarity analysis, subjective and objective analysis, emotional classification, and so on [7]. Individuals can express their sentiment about emergencies, public figures, and popular products through social media directly and quickly. Being an important research direction in sentiment mining, sentiment classification for short texts, usually from social media such as online reviews and Sina Weibo, has wide application

X. Cheng et al. (Eds.): SMP 2017, CCIS 774, pp. 232–243, 2017.
https://doi.org/10.1007/978-981-10-6805-8_19

prospects in the fields of public opinion analysis, consumer intention identification, and e-commerce commentary analysis. It can also provide quantitative and scientific decisions for government departments and enterprises [1].

Social media sentiment classification has always been a hotspot and difficult problem in natural language processing and artificial intelligence [19]. As we know, sentiment expression is domain-dependent, and different domains have different distributions. For example, "薄" (thin) expresses negative sentiment in hotel domain, while it expresses positive sentiment in notebook domain. Therefore, the classifier trained on the source domain may not be well adapted to the target domain. Deep neural networks (DNN) have achieved excellent results on sentiment classification tasks, but it requires massive training data, otherwise it is easy to over-fit [6]. Unfortunately, to collect and label massive domain-related samples require considerable time and efforts. Meanwhile, we have accumulated rich and fine-labeled data in traditional sentiment classification tasks, it is also extremely wasteful to discard the data completely. The goal of transfer learning is to learn the knowledge learned from the source domain to aid learning tasks about the target domain. It can take advantages of the commonality between different learning tasks to share the benefits of statistics and migration knowledge among tasks [17].

Deep transfer learning (DTL) approaches transfer deep neural networks which are trained on source domain to special target domain. It turns out to be successful in image recognition and natural language processing tasks [14]. Previous studies have proved that bottom layers can learn basic generic features, while top layers can learn data-specific and advanced features representation [4]. In other words, the features computed in higher layers of the network must depend greatly on the specific data set and tasks. In the context of deep learning, fine-tuning a deep network that pre-trained on the source domain data is a common strategy to learn task-specific features. The pre-training and fine-tuning strategies can be trained using existing data sets and adapted to target domain. In detail, it transfers bottom-layer (general) features and retrains (specific) top-layer features from the source domain to target domain.

In this paper, we propose a two-stage bidirectional long short-term memory (Bi-LSTM) and parameters transfer framework for short texts cross-domain sentiment classification tasks. There are two main advantages of our deep transfer learning framework: one is the powerful ability to capture variable length and n-gram context semantics of Bi-LSTM networks, the other is the ability to transfer knowledge from the source domain to target domain data with fine-tuning strategy. Firstly, we pre-train Bi-LSTM networks using a large number of fine-labeled source domain training samples. Then the Bi-LSTM networks are fine-tuned with limited target domain training data. In the parameters transfer process, bottom layers parameters are fine-tuned, and softmax layers parameters are retrained. Experimental results on four Chinese sentiment classification data sets show that our proposed method performs better than previous methods.

Our contributions in this paper can be summarized as follows.

- We introduce a novel Bi-LSTM and parameters transfer framework for cross-domain sentiment classification tasks. This framework can learn long-term dependence, word sequence semantic information and transfer knowledge from the source domain to target domain.
- We share bottom layers of Bi-LSTM networks and retrain top layers using a slight number of target domain training samples. This improves the effectiveness of cross-domain sentiment classification and generalization capabilities.
- Experiments demonstrate that our parameters transfer and fine-tuning schemes achieve state-of-the-art performance on Chinese short texts cross-domain classification tasks via deep transfer learning.

2 Deep Transfer Framework

In this section, we firstly introduce basic notations and problem formulation. Then we describe a deep transfer learning framework for cross-domain sentiment classification tasks in detail. Bidirectional LSTM networks are pre-trained on massive source domain training samples. Then pre-trained model parameters are transferred and fine-tuned with limited target domain data.

2.1 Notations and Problem Formulation

For a formal description of cross-domain sentiment classification tasks, $\mathcal{X} = \mathcal{R}$ denotes the instance space, $x = (x_1, x_2, \cdots, x_T)$ consists of a series of words x_i, $x \in \mathcal{X}$. $\mathcal{Y}$ is the label space for sentiment classification tasks, and $\mathcal{Y}_1 = \{very\ positive, positive, neutral, negative, very\ negative\}$ is the fine-grained sentiment classification label set, $\mathcal{Y}_2 = \{positive, negative\}$ is the binary sentiment classification label set. In this paper, x_i is a word2vec distributed representation, x_i is a d-dimensional feature vector, i.e., $x_i = (x_i^1, x_i^2, \cdots, x_i^d)$. For each instance (x, y), y is the sentiment label with x, $y \in \mathcal{Y}$.

$\mathcal{D}^S = \{(x_{s1}, y_{s1}), (x_{s2}, y_{s2}), \cdots, (x_{sm}, y_{sm})\}$ is the source domain training data set, the label space is $\{y_{s1}, y_{s2}, \cdots, y_{sm}\}$. The marginal probability distribution of source domain is $P_S(X)$. $\mathcal{D}^L = \{x_i, Y_i | 1 \leq i \leq n\}$ represents the target domain training set, $\mathcal{D}^U = \{x_i, Y_i | 1 \leq i \leq p\}$ represents the target domain testing set, $\mathcal{D}^T = \mathcal{D}^L \cup \mathcal{D}^U$ is the target domain data set. The distribution of target domain $P_T(X)$ is often different from $P_S(X)$.

There are two main transfer learning tasks: (i) transfer across domains: the data distributions between two domains are different, i.e., $P_S(X) \neq P_T(X)$, while the tasks are the same, i.e., $\mathcal{Y}^S = \mathcal{Y}^T$; (ii) transfer across tasks: both data distributions and tasks are different, i.e., $P_S(X) \neq P_T(X)$, $\mathcal{Y}^S \neq \mathcal{Y}^T$. In this paper, we verify our proposed framework on the above two tasks. The task of deep transfer learning can be formalized as follows: firstly, we learn pre-trained neural networks $f_S : \mathcal{D}^S \rightarrow \mathcal{Y}^S$, then transfer the neural networks $f_S \rightarrow f_D$ with fine-tuning the parameters weight of bottom layers and retraining the top layers on $\mathcal{D}^L$.

2.2 Bidirectional LSTM Pre-training

For sentiment classification tasks, Bi-LSTM (actually using forward and backward LSTM) can capture variable length and bidirectional n-gram context information [18]. Background topics and sentiment indicators of social media texts could be far away from the target aspect. The traditional bag-of-words based machine learning methods could not distinguish the implicit or hidden dependency in long conversations. However, the memory cell in LSTM can settle long distance dependency problems. The sequence of words in a sentence plays an important role in sentiment expression. Such as (I am very upset today.) and (I am not very happy today.) express different sentiment intensities. Compared with convolution neural networks, Bi-LSTM focuses on the reconstruction of the adjacent position, so it is more suitable for the sequence structure of language modeling.

Although Bi-LSTM model has achieved good results in sentiment classification tasks, it needs large number of related training samples, otherwise it is very prone to over-fit. However, to collect and annotate a large-scale domain-related data set require considerable time and efforts. Existing sentiment classification tasks have accumulated a large number of fine-labeled sentiment classification data [16]. An intuitive idea is to use these source domain data to assist target domain sentiment classification tasks. Bi-LSTM networks are firstly pre-trained on source domain data and then parameters are transferred into target domain.

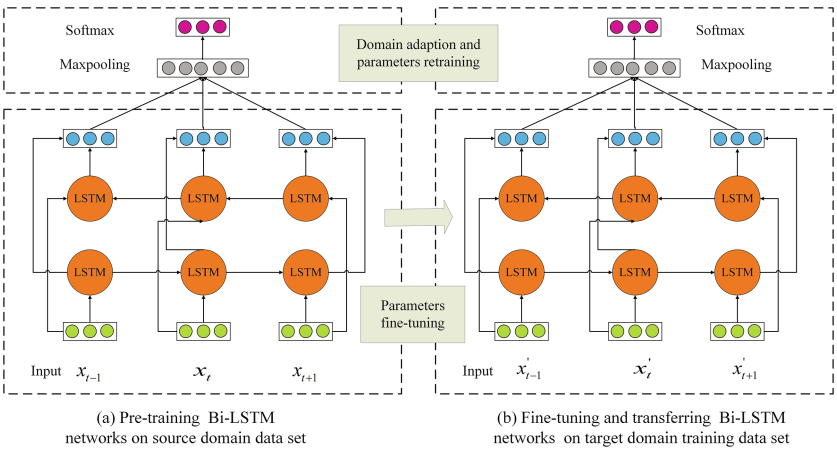


Fig. 1. Flow chart of Bi-LSTM networks pre-training and fine-tuning processes. This framework can be divided into two parts: (a) Pre-training Bi-LSTM networks on source domain data set. (b) Fine-tuning Bi-LSTM networks and transferring parameters on target domain training data set. The bottom layers parameters weight are transformed to target domain, while top layers are randomized and adapted to target domain

Figure 1(a) shows the process of six layers Bi-LSTM networks pre-training on $\mathcal{D}^S$. We treat each word x_i as a time node. The input units are a sequence

of words $x = (x_1, x_2, \cdots, x_T)$, $x \in \mathcal{D}^S$. Then the word sequence layer is entered into the forward hidden sequence $\overrightarrow{h}$ and the backward hidden sequence $\overleftarrow{h}$.

A LSTM memory cell consists of a memory cell c_t, an input gate i_t, a forget gate f_t, and an output gate o_t. The input gate (current cell matters) can be formalized as: $i_t = \sigma(W^{xi}x_t + W^{hi}h_{t-1} + W^{ci}c_{t-1} + b^i)$. Forget gate (gate 0, forget past): $f_t = \sigma(W^{xf}x_t + W^{hf}h_{t-1} + W^{cf}c_{t-1} + b^f)$. Output gate (how much cell is exposed): $o_t = \sigma(W^{xo}x_t + W^{ho}h_{t-1} + W^{co}c_{t-1} + b^o)$. New memory cell: $c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W^{hc}x_t + W^{hc}h_{t-1} + bc)$. A d-dimensional hidden state: $h_t = o_t \odot \tanh(c_t)$. where $x = (x_1, x_2, \cdots, x_T)$ is the input feature sequence, σ is the logistic function. The symbol $\odot$ represents the element-wise operation. W is the weight matrix and the superscript indicates the matrix between two different gates.

Bi-LSTM networks compute the forward layer as the forward hidden sequence $\overrightarrow{h}$ from $t = 1$ to T, the backward layer as backward hidden sequence $\overleftarrow{h}$ by iterating from $t = T$ to 1, and the output layer y as the output sequence $y = (y_1, y_2, \cdots, y_T)$.

$$\overrightarrow{h}_t = H(W_{x\overrightarrow{h}}x_t + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}) \tag{1}$$

$$\overleftarrow{h}_t = H(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}) \tag{2}$$

$$y_t = W_{\overrightarrow{h}y}\overrightarrow{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y \tag{3}$$

Where H is the LSTM block transition function. These six weight matrices $W_{x\overrightarrow{h}}, W_{x\overleftarrow{h}}, W_{\overrightarrow{h}\overrightarrow{h}}, W_{\overrightarrow{h}y}, W_{\overleftarrow{h}\overleftarrow{h}}$, and $W_{\overleftarrow{h}y}$ are repeated at every time. It is worth noting that there is no flow of information between the forward and backward hidden layers, which ensures that the expansion is non-cyclic.

And we wish to predict sentiment label y from the label space $\mathcal{Y}$. $z = (Max(y_i))_{i=1}^T$ is the maxpooling over all the time steps results. $p(y|x)$ is predicted by softmax classifier that takes the Bi-LSTM average output z as input:

$$p(y|x) = softmax(W^z z + b^z) \tag{4}$$

$$y = \arg\max p(y|x) \tag{5}$$

In the parameter update rules of **Adagrad**, the learning rate η varies with each iteration according to the historical gradient. Assume that at an iteration time t, $g_{t,i} = \nabla_\theta J(\theta_i)$ is the gradient of the objective function to the parameter.

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_t + \varepsilon}} \cdot g_{t,i} \tag{6}$$

Where $G_t \in R^{d \times d}$ is a diagonal matrix, $\varepsilon = e^{-8}$ is a smoothing item to prevent G_t from being equal to 0.

We use mean squared logarithmic error (MSLE) as the loss function:

$$\varepsilon = \frac{1}{n}\sum_{i=1}^{n}(\log(Y+1) - \log(y+1))^2 \tag{7}$$

Where Y represents the true label of x, and y is the prediction label.

2.3 Fine-Tuning and Parameters Transfer

Motivation: As we all know, sentiment classification is a domain-dependent issue. The Bi-LSTM model that trained on source domain may not be necessarily well suited to target domain. The distributions between the source and target domains may be not precisely the same. In this case, to label target domain training samples is time-consuming and laborious. Besides this, the amount of training data is not normally adequate for retraining new neural networks. On account of this, the well-trained Bi-LSTM model requires a domain-adaption process. Therefore, fine-tuning and parameters transfer are just an ideal choice. Previous experiments have verified that fine-tuning performs better than the model which only trains on limited target domain samples.

Transfer Bottom Layers: Figure 1(b) shows the domain adaptation and parameters transfer processes. We pre-train Bi-LSTM networks with a low initial learning rate η and high dropout rate on $\mathcal{D}^S$. The bottom layers parameters weight of pre-trained Bi-LSTM networks W^S are $W_{x\overrightarrow{h}}$, $W_{\overrightarrow{h}\overrightarrow{h}}$, $b_{\overrightarrow{h}}$, $W_{x\overleftarrow{h}}$, $W_{\overleftarrow{h}\overleftarrow{h}}$, $b_{\overleftarrow{h}}$, $W_{\overrightarrow{h}y}$, $W_{\overleftarrow{h}y}$, and b_y. We use target domain training data $\mathcal{D}^L$ as fine-tuning source data. Then W_S is fine-tuned with a high initial learning rate η and low dropout rate from $\mathcal{D}^L$ by back propagation algorithm. We use layer-by-layer feature transference to transfer bottom layers parameters weight W_S. This is motivated by the observation that the general features of Bi-LSTM networks contain more generic features that should be useful to target domain. We do not wish to distort them too quickly or too much, so we keep learning rate low and dropout rate decay really high.

Retrain Top Layers: It is possible to fine-tune some of earlier layers fixed (due to over-fitting concerns) and retrain some higher-level portion of the networks. The later layers of the Bi-LSTM become more specific to the details of the classes contained in the target domain data set. The top-layer features depend greatly on the chosen special data set and tasks, so called as specific features. The full connection layer (softmax classifier) of transferred Bi-LSTM networks is replaced and retrained. The softmax layer parameters weight W^z and b^z are initialized randomly, and then retrained on target domain training data set $\mathcal{D}^L$. We remove the output layer, and then use the entire network as a fixed feature extractor for target domain data set. Therefore, our framework can be applied into transfer across domains and transfer across tasks problems. These pre-trained networks demonstrate a strong ability to generalize to new data set via transfer learning.

3 Experiment and Analysis

3.1 Data Sets and Experiment Setup

We use four Chinese social media sentiment classification data sets to validate our deep transfer learning framework. Hotel (H) and Notebook (N) data sets

are collected from Jingdong shopping website (https://www.jd.com/). Weibo (W) data set is collected from COAE 2015 (https://www.ccir2015.com/). Fine-grained data set electronic (E) including 8000 samples is collected from COAE 2011 task 3 (https://www.ccir2011.com/). The detail of four data sets can be seen in Table 1.

Table 1. The detail of four sentiment classification data sets

Data set	Very positive	Positive	Neutral	Negative	Very negative
Hotel (H)	*	2000	*	2000	*
Notebook (N)	*	2000	*	2000	*
Weibo (W)	*	5000	*	5000	*
Electronic (E)	801	453	1139	2295	3311

We use THULAC tool (http://thulac.thunlp.org/) to get the word segmentation. After this, we use Glove vectors of 100 dimension to train the distributed word vector [12] with all source domain and target domain texts. We use 5-fold cross validation method to extract 20% target domain randomly as the target domain training data, the rest composes the target domain testing data. Back-propagation through time (BPTT) method with AdaGrad initial learning rate of 0.5 and dropout rate 0.7 on source domain, initial learning rate of 0.8 and dropout rate 0.3 on target domain, epoch number as 5, hidden layer units as 64, and mini-batch size of 20 are used to train our model. Our model is implemented by Keras deep learning library (https://keras.io/). We utilize accuracy to evaluate the baselines and our proposed framework.

3.2 Baselines and Our Framework

(1) **Active learning:** an instance-based transfer method with active learning for cross-domain sentiment classification which was proposed by Li et al. [10]. We follow original settings as bag-of-words and binary vectors representation, maximum entropy classifier, and 20% target domain data as the initial labeled data.
(2) **Multi-instance:** a hybrid strategy which combined transfer learning, deep learning and multi-instance learning which was proposed by Dimitrios et al. [9]. We use 3 epochs, mini-batch size of 50, objective function of SGD iterations with 1050 iterations and a learning rate of $\alpha = 0.0001$.
(3) **BLPT:** our proposed Bi-LSTM networks and parameters transfer method.

Three strategies are used to evaluate our framework and shown as follows:

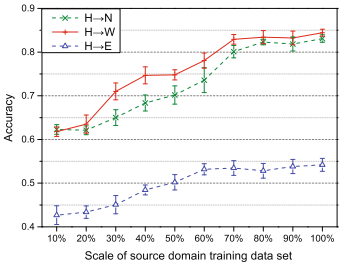
BLPT-random: BLPT method with randomly initialized vectors;

BLPT-fixed: BLPT method with fixed word vectors which are trained by Glove method;

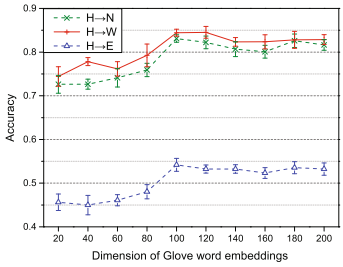
BLPT-tuned: BLPT method with Glove word vectors and updated in the training process.

3.3 Experimental Results

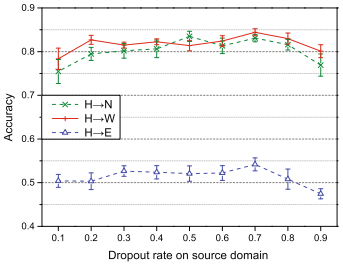
Performance with Different Parameters: We compare **BLPT-tuned** performances with different parameters, "scale of source domain training data set", "dimension of Glove word embeddings", "dropout rate", and "epoch" respectively on H→N, H→W, and H→E tasks. Figure 2a and b show the accuracy performances with different scales of source domain and word embeedings dimension under fixed dropout rate as 0.7 and epoch number as 5. We can find that more source domain training data generally performs better. The accuracy grows when the dimension of Glove word embeddings changes from 20 to 100, while the impact of the dimension on the results is not particularly obvious when dimension changes from 100 to 200. We fix the scale of source domain as 100%, word embeddings dimension as 100, and compare the impact of dropout rate and epoch of Bi-LSTM model in Fig. 2c and d. Dropout can prevent the neural network overfitting effectively, and we find that increasing dropout rate does not lead to significant improvements. A good classification performance is achieved when the epoch is 5 or 6 in Bi-LSTM networks, and our model may be over-fitting when the epoch is larger than 7.



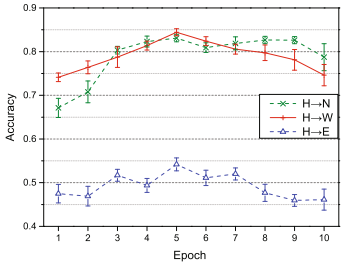
(a) Performance with different scales of source domain training data



(b) Performance with different dimension of Glove word embeddings



(c) Performance with different dropout rates



(d) Performance with different epochs

Fig. 2. The performance of transferred Bi-LSTM model trained with different source domain training scales, word embeddings dimension, dropout rates, and epoch sizes

Table 2. Mean accuracy ± standard deviation (%) results of 12 cross-domain sentiment classification tasks

Task	Active learning	Multi-instance	BLPT-random	BLPT-fixed	BLPT-tuned
H→N	80.8 ± 0.5	82.1 ± 0.1	79.8 ± 0.5	82.3 ± 0.4	**83.1** ± 0.9
H→W	80.6 ± 0.8	82.8 ± 0.5	81.1 ± 0.7	83.0 ± 0.5	**84.4** ± 0.8
H→E	51.3 ± 0.7	53.2 ± 0.7	51.3 ± 1.4	53.1 ± 1.2	**54.2** ± 1.5
N→H	81.7 ± 0.6	83.1 ± 0.6	80.1 ± 1.2	83.7 ± 0.3	**85.7** ± 0.6
N→W	80.4 ± 0.4	82.0 ± 0.4	81.8 ± 0.4	82.0 ± 0.8	**84.8** ± 0.5
N→E	51.0 ± 0.9	52.1 ± 0.8	51.8 ± 0.4	54.8 ± 0.4	**55.3** ± 0.5
W→H	82.1 ± 0.4	83.6 ± 0.9	81.8 ± 1.3	82.1 ± 0.5	**84.3** ± 0.4
W→N	80.9 ± 0.6	82.3 ± 0.5	82.0 ± 0.6	84.7 ± 0.5	**85.9** ± 1.3
W→E	56.8 ± 0.5	55.8 ± 0.7	54.5 ± 0.8	57.8 ± 1.3	**59.2** ± 0.7
E→H	82.3 ± 1.2	81.8 ± 0.8	82.1 ± 0.7	84.4 ± 0.4	**84.9** ± 1.2
E→N	82.0 ± 0.8	82.4 ± 1.3	82.0 ± 0.4	83.2 ± 1.2	**85.0** ± 0.7
E→W	81.5 ± 1.1	82.1 ± 0.4	81.1 ± 0.8	83.0 ± 0.4	**83.5** ± 0.5
Average	74.3	75.3	74.1	76.2	**77.5**

Comparing Results: Table 2 gives the mean accuracy of 12 cross-domain sentiment classification tasks on four data sets. From Table 2, we can find that:

(i) Comparing with **Active learning** and **Multi-instance** methods, our proposed framework **BLPT-tuned** generally performs better. This proves the excellent feature presentation ability and good generalization of transferred Bi-LSTM for short texts cross-domain sentiment classification tasks.
(ii) In contrast with **BLPT-random** and **BLPT-fixed** methods, our transformed Bi-LSTM networks through tuned word embeedings improve 3.4% and 1.3% respectively. The word embeddings are updated in the supervised learning process, so its semantics is more clear and the classification performance is better.
(iii) Our work can be readily adapted into transfer across domains and transfer across tasks problems. Comparing with binary sentiment classification, fine-grained sentiment classification is a more detail and difficult task. The accuracies of H→E, N→E, and W→E tasks are significantly lower than other tasks.

3.4 Discussions

(1) Deep transfer learning can obtain good performance with abundant source domain data and limited target domain data. Neural networks have achieved excellent results and need large scale training data to train the parameters weight. It is relatively rare to have a data set of sufficient size which is required for the depth of networks. In real applications, source domain

data set is relatively large in size and similar in content compared to target domain data set. Deep transfer learning depends on the scale of source domain and target domain training data, and similarity degree between source domain and target domain. It can bring feature representation of deep neural networks to a new domain. Since we have limited target domain training data, we can fine-tune the full network that trained on source domain. It is common to pre-train Bi-LSTM networks on a very large data set and then use trained parameters weights either as an initialization or a fixed feature extractor for the task of interest.

(2) Deep transfer learning including parameters transfer and fine-tuning strategies helps the training process for the target domain better. We fine-tune some of the earlier layers under lower learning rate, and retrain some higher-level portion of the networks. This is motivated by the observation that earlier features of Bi-LSTM contain more generic features, while the fully connected softmax layer becomes progressively more specific to particular data set. We can share pre-trained parameters to a new model to speed and optimize model learning to avoid learning from scratch and time-consuming training. Fine-tuning enables us to bring the power of pre-trained models to target domain with insufficient data. This can effectively exploit powerful generalization capabilities of deep neural networks, and eliminate the need to redesign complex models. Our approach solves the problem of over-fitting of deep neural model such as Bi-LSTM networks on limited samples. Besides this, our approach achieves a significant improvement of average accuracy and generalization across domains.

4 Related Work

4.1 Cross-Domain Sentiment Classification

There has been a lot of work on the issue of cross-domain sentiment classification tasks. Researchers have gradually begun to use transfer learning (TL) techniques to solve cross-domain sentiment classification tasks. The existing work can be divided into four parts: instance-based, feature-based, parameter-based, and relational-based [10]. For instance-based transfer, previous studies mainly focus on selecting valuable samples from source domain which can be used to assist the target domain sentiment classification. Feature-based transfer is to find the correlation features (shared features) between source domain and target domain, and to construct the unified feature representation space of cross-domain data. Parameter-based methods discover shared parameters or priors between the source domain and target domain models, which can benefit for transfer learning. Relational-based methods build mapping of relational knowledge from different domains. Tan et al. [15] attempted to tackle domain-transfer problem by combining source domain labeled examples with target domain unlabeled ones. The basic idea was to use source domain trained classifier to label some informative unlabeled examples in the new domain, and retrain the base classifier over these selected examples. Spectral feature alignment (SFA) was presented by

Pan et al. [13] to discover a robust representation for cross-domain data by fully exploiting the relationship between the domain-specific and domain-independent words via simultaneously co-clustering them in a common latent space. Li et al. [10] performed active learning for cross-domain sentiment classification by actively selecting a small amount of labeled data in the target domain.

4.2 Deep Transfer Learning

Deep transfer learning (DTL) focuses on adapting knowledge from an auxiliary source domain to a target domain with little or without any label information to construct neural networks model of good generalization performance [9]. It is an approach in which a deep model is trained on a source problem, and then reused to solve a target problem [5]. DTL usually trains deep neural networks in source domain, and transfer and fine-tune the parameters weight to the target domain. This strategy has been proved to improve cross-domain classification results effectively [11]. A source-target selective joint fine-tuning scheme was introduced by Ge et al. [2] for improving the performance of deep learning tasks with insufficient training data. Dimitrios et al. [9] combined transfer learning, deep learning and multi-instance learning, and reduced the need for laborious human labelling of fine-grained data when abundant labels were available at the group level. Chetak et al. [8] proposed a ensemble methodology to reduce the impact of selective layer based transference and provide optimized framework to work for three major transfer learning cases. Xavier et al. [3] studied the problem of domain adaptation for sentiment classifiers, whereby a system was trained on labeled reviews from one source domain but was meant to be deployed on another. Then a meaningful representation for each review was extracted in an unsupervised fashion.

5 Conclusions and Future Work

In this paper, we propose a deep transfer learning framework for Chinese short texts cross-domain sentiment classification tasks. Our work takes advantages of transfer learning, deep neural networks, and fine-tuning strategies. Firstly bidirectional LSTM networks are pre-trained on source domain data. Then we use transfer learning strategy to transfer and fine-tune Bi-LSTM networks on target domain training samples. We use extra massive source domain training data to enhance the performance of current learning task, including generalization accuracy, learning efficiency and comprehensibility. Experiments on four data sets show that our pre-training and fine-tuning schemes achieve better performances than previous methods. In the future, we intend to use multiple source domains training data and ensemble the final results. We will also consider attention-based RNN models for further improving the sequential representation of short texts.

Acknowledgments. This work was supported by: National Natural Science Foundation of China (61573231, 61632011, 61672331, 61432011); Shanxi Province Graduate Student Education Innovation Project (2016BY004, 2017BY004).

References

1. Zhao, C., Wang, S., Li, D.: Fuzzy sentiment membership determining for sentiment classification. In: Proceedings of ICDMW 2014, pp. 1191–1198. IEEE (2014)
2. Ge, W., Yu, Y.: Borrowing treasures from the wealthy: deep transfer learning through selective joint fine-tuning. arXiv preprint arXiv:1702.08690 (2017)
3. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: a deep learning approach. Proc. ICML **2011**, 513–520 (2011)
4. Guan, L., Zhang, Y., Zhu, J.: Segmenting and characterizing adopters of e-books and paper books based on Amazon book reviews. In: Li, Y., Xiang, G., Lin, H., Wang, M. (eds.) SMP 2016. CCIS, vol. 669, pp. 85–97. Springer, Singapore (2016). doi:10.1007/978-981-10-2993-6_7
5. Haaren, J.V., Kolobov, A., Davis, J.: Todtler: two-order-deep transfer learning. In: Proceedings of AAAI 2015 on Artificial Intelligence, pp. 3007–3015 (2015)
6. Huang, M., Cao, Y., Dong, C.: Modeling rich contexts for sentiment classification with LSTM. arXiv preprint arXiv:1605.01478 (2016)
7. Wang, S., Li, D., Zhao, L., Zhang, J.: Sample cutting method for imbalanced text sentiment classification based on BRC. Knowl. Based Syst. **37**, 451–461 (2013)
8. Kandaswamy, C., Silva, L.M., Alexandre, L.A., Santos, J.M.: Deep transfer learning ensemble for classification. In: Rojas, I., Joya, G., Catala, A. (eds.) IWANN 2015. LNCS, vol. 9094, pp. 335–348. Springer, Cham (2015). doi:10.1007/978-3-319-19258-1_29
9. Kotzias, D., Denil, M., Blunsom, P., de Freitas, N.: Deep multi-instance transfer learning. arXiv preprint arXiv:1411.3128 (2014)
10. Li, S., Xue, Y., Wang, Z., Zhou, G.: Active learning for cross-domain sentiment classification. Proc. IJCAI **2013**, 2127–2133 (2013)
11. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. Proc. ICML **2015**, 97–105 (2015)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Proc. NIPS **2013**, 3111–3119 (2013)
13. Pan, S.J., Ni, X., Sun, J.T., Yang, Q., Chen, Z.: Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of WWW 2010, pp. 751–760. ACM (2010)
14. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755 (2016)
15. Tan, S., Wu, G., Tang, H., Cheng, X.: A novel scheme for domain-transfer problem in the context of sentiment analysis. In: Proceedings of ACM 2007, pp. 979–982. ACM (2007)
16. Tang, D., Qin, B., Feng, X., Liu, T.: Target-dependent sentiment classification with long short term memory. arXiv preprint arXiv:1512.01100 (2015)
17. Tang, J., Lou, T., Kleinberg, J., Wu, S.: Transfer learning to infer social ties across heterogeneous networks. ACM Trans. Inf. Syst. (TOIS) **34**(2), 7 (2016)
18. Wang, J., Yu, L.C., Lai, K.R., Zhang, X.: Dimensional sentiment analysis using a regional CNN-LSTM model. Proc. ACL **2016**, 225–230 (2016)
19. Wang, S., Li, D., Song, X., et al.: A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Syst. Appl. **38**(7), 8696–8702 (2011)