Cross-Lingual Emotion Classification with Auxiliary and Attention Neural Networks

Lu Zhang, Liangqing Wu, Shoushan Li(✉), Zhongqing Wang, and Guodong Zhou

Natural Language Processing Lab, School of Computer Science and Technology,
Soochow University, Suzhou, China
{lzhang0107, lqwu}@stu.suda.edu.cn,
{lishoushan, wangzq, gdzhou}@suda.edu.cn

Abstract. In the literature, various supervised learning approaches have been adopted to address the task of emotion classification. However, the performance of these approaches greatly suffers when the size of the labeled data is limited. In this paper, we tackle this challenge from a cross-lingual sensoria where the labeled data in a resource-rich language (i.e., English in this study) is employed to improve the emotion classification performance in a resource-poor language (i.e., Chinese in this study). Specifically, we first use machine translation services to eliminate the language gap between Chinese and English data and then propose a joint learning framework to leverage both Chinese and English data, which develops auxiliary representations from several auxiliary emotion classification tasks. Furthermore, in our joint learning approach, we introduce an attention mechanism to capture informative words. Empirical studies demonstrate the effectiveness of the proposed approach to emotion classification.

Keywords: Sentiment analysis · Emotion classification · Attention mechanism

1 Introduction

Emotion classification aims to determine the involving emotion within a piece of text. With the tremendous growth of social media, such as Twitter and Facebook, emotion classification has drawn more and more attention. In the last decade, emotion classification has been proved to be invaluable in many applications, such as stock markets [1], online chat [2] and news classification [3].

Conventional approaches to emotion classification mainly conceptualize the task as a supervised learning problem where sufficient labeled data is essential for training the model. However, in most scenarios, the annotated corpus for emotion classification is scarce, and to obtain such labeled data is extremely costly and time-consuming. Some previous studies tackle this challenge by applying semi-supervised technique to make use of unlabeled data. For instance, Liu et al. [3] propose a co-training algorithm to improve the performance of emotion classification by leveraging the information in the unlabeled data. Li et al. [4] propose a two-view label propagation approach to emotion

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11108, pp. 429–441, 2018.
https://doi.org/10.1007/978-3-319-99495-6_36

E1:
Original: *今天大甩卖！我们去逛街吧~*
Translation: *There's a big sale on today! Let’s go shopping.*
E2:
Original: *最近总是七上八下。*
Translation: *Recently, I'm always in an unsettled state of mind.*

Fig. 1. Some examples in Chinese emotion corpus with their English translations

classification by exploiting two views, namely source text and response text in a label propagation algorithm (Fig. 1).

Instead of semi-supervised learning, we focus on addressing this issue from a cross-lingual view. On one hand, Chinese emotion corpus is limited but many English emotion corpora are freely available. On the other, the emotion involved in a given text may not be learned in an exact manner with the representation in Chinese. However, if we translate it into English, it becomes easier to determine the emotion. For instance, in **E1**, due to the lack of Chinese emotion corpus, “*大甩卖*” may not exist in the training set so that it cannot be correctly classified. But if we translate this word into English, i.e., “*big sale*”, then we can leverage English emotion corpus to make up for this. Similarly, in **E2**, “*七上八下*” is a Chinese idiom, which is difficult for machine to understand. However, if we translate it into English, i.e., “*an unsettled state of mind*”, the emotion expressed by this phrase can be understood more easily.

In this study, we propose a joint learning framework, namely, Aux-LSTM-Attention, which learns simultaneously from the labeled data from both resource-poor and resource-rich languages. First, machine translation services are used to translate Chinese emotion corpus into English corpus and also translate English emotion corpus into Chinese corpus. Then, we view the emotion classification task with original Chinese emotion corpus as a main task and the emotion classification tasks with additional corpora as auxiliary tasks. To perform joint learning, we share neural network layers from the auxiliary tasks into the main task. Consequently, the main task learns the emotion classification by using the knowledge from both the main and auxiliary tasks through the layer sharing. Furthermore, we utilize an attention mechanism [5, 6] to aggregate the representation of informative words into a vector for emotion prediction. Empirical studies demonstrate that the proposed joint learning approach significantly outperforms several baseline approaches to emotion classification.

The remainder of this paper is organized as followed. Section 2 gives a brief overview of related work. Section 3 proposes our joint learning framework on emotion classification with both resource-poor and resource-rich corpora. Section 4 evaluates the proposed approach before presenting the concluding remarks in Sect. 5.

2 Related Work

2.1 Cross-Lingual Sentiment Classification

Sentiment analysis is the field of analyzing people's opinions, sentiments, attitudes and emotions from the text they have published [7]. In previous studies, conventional approaches to sentiment analysis mainly focus on sufficient labeled data [8]. However, in most scenarios, there is insufficient labeled data and to manually label reliable corpus is not a trivial task. Cross-lingual addresses this issue in sentiment classification from a cross-language view.

Over the last decades, there has been a proliferation of work exploring various aspects of cross-lingual sentiment classification. Mihalcea et al. [9] generate resources for subjectivity annotations for a new language, by leveraging resources and tools available for English. Wan [10] uses machine translation services to eliminate the language gap between Chinese corpus and English corpus. Chinese features and English features are considered to be two independent views of the classification problem and a co-training algorithm is employed to make use of unlabeled Chinese data. Balamurali et al. [11] use WordNet synset identifiers as features of a supervised classifier. They leverage the linked WordNets of two languages to bridge the language gap. Prettenhofer and Stein [12] introduce the structural correspondence learning algorithm to learn a map between the source language and the target language. More recently, Zhou et al. [13] propose a bilingual document representation learning method for cross-lingual sentiment classification which directly learns the vector representation for documents in different languages.

Unlike all above studies, this work focuses on cross-lingual emotion classification. Compared to cross-lingual sentiment classification, cross-lingual emotion classification is more challenging due to the fact that the sentiment categories in two languages are the same while the emotion taxonomies in two languages might be different.

2.2 Emotion Classification

Our work is also related to emotion classification. Tokuhisa et al. [14] propose a data-oriented method for inferring the emotion of an utterance sentence in a dialog system. Bhowmick et al. [15] present a method for classifying news sentences into multiple emotion categories using multi-label KNN classification technique. Xu et al. [16] propose a coarse-to-fine analysis strategy for emotion classification which takes similarities to sentences in training set as well as adjacent sentences in the context into consideration. Yang et al. [17] introduce an Emotion-aware LDA model to build a domain-specific lexicon for predefined emotions. Felbo et al. [18] show how millions of readily available emoji occurrences on Twitter can be used to pertrain models to learn a richer emotional representation than traditionally obtained through distant supervision.

Unlike all above studies, our work is the first attempt to apply cross-lingual in emotion classification.

3 Our Approach

3.1 Machine Translation

In this section, we propose our joint learning approach to perform joint learning with both resource-poor and resource-rich data. In this study, we assume that Chinese is the resource-poor language and it has only a few labeled samples. English is the resource-rich language and it has many more labeled samples. In order to overcome the language gap, we translate one language into the other language with a machine translation tool. Specifically, we adopt *Baidu Translate*[1] for both English-to-Chinese translation and Chinese-to-English translation. Figure 2 shows the general framework of the machine translation in the training phase. We translate the labeled Chinese emotion corpus into English to set up a translated English view and translate English emotion corpus into Chinese to establish a translated Chinese view.

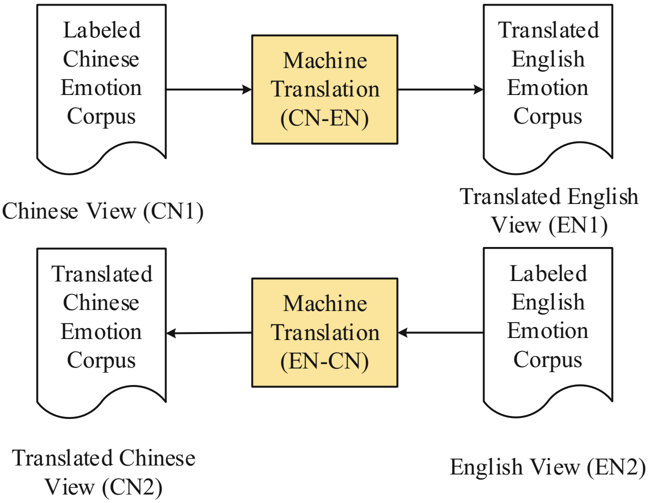
Labeled Chinese Emotion Corpus
Machine Translation (CN-EN)
Translated English Emotion Corpus
Chinese View (CN1)
Translated English View (EN1)
Translated Chinese Emotion Corpus
Machine Translation (EN-CN)
Labeled English Emotion Corpus
Translated Chinese View (CN2)
English View (EN2)

Fig. 2. Framework of the machine translation in the training phase

3.2 The Main Emotion Classification Task

Figure 3 illustrates the overall architecture of our Aux-LSTM-Attention approach which contains a main task and three auxiliary tasks. Specially, we consider the emotion classification task with original Chinese emotion corpus as the main task and the emotion classification tasks with other corpora as auxiliary tasks. The main idea of the proposed approach is to employ some auxiliary representations learned from the auxiliary tasks to assist the performance of the main task. Note that not all words contribute equally to representing the meaning of a post. Hence, instead of simply

[1] http://fanyi.baidu.com/translate.

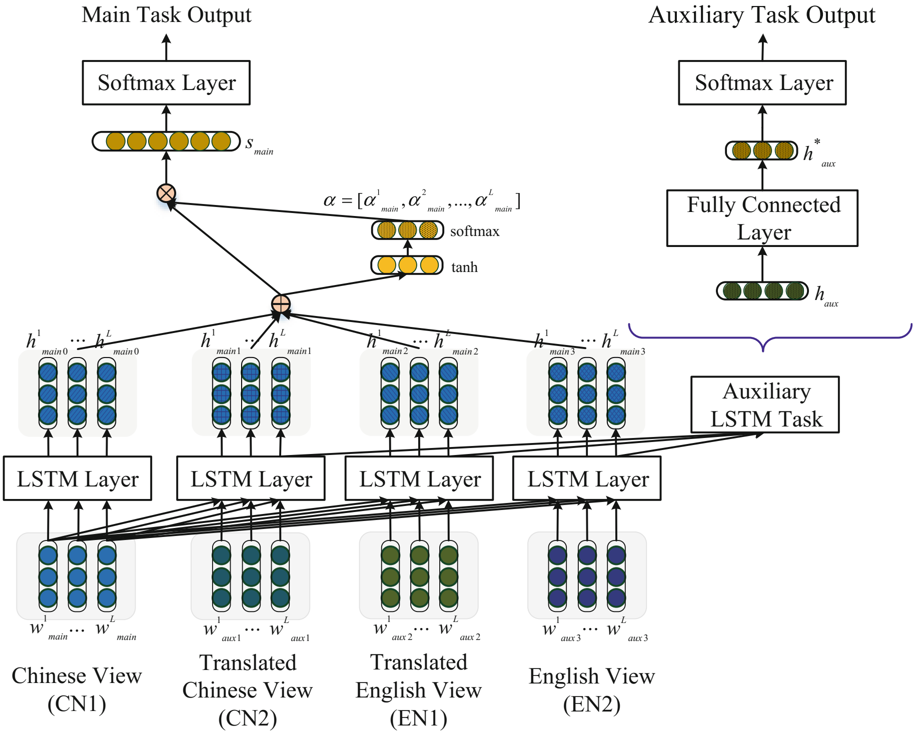
Main Task Output
Softmax Layer
s_{main}
$\alpha = [\alpha^1_{main}, \alpha^2_{main}, ..., \alpha^L_{main}]$
softmax
tanh
Auxiliary Task Output
Softmax Layer
h^*_{aux}
Fully Connected Layer
h_{aux}
$h^1_{main0} \cdots h^L_{main0}$
$h^1_{main1} \cdots h^L_{main1}$
$h^1_{main2} \cdots h^L_{main2}$
$h^1_{main3} \cdots h^L_{main3}$
Auxiliary LSTM Task
LSTM Layer
LSTM Layer
LSTM Layer
LSTM Layer
$w^1_{main} \cdots w^L_{main}$
$w^1_{aux1} \cdots w^L_{aux1}$
$w^1_{aux2} \cdots w^L_{aux2}$
$w^1_{aux3} \cdots w^L_{aux3}$
Chinese View (CN1)
Translated Chinese View (CN2)
Translated English View (EN1)
English View (EN2)

Fig. 3. The overall architecture of Aux-LSTM-Attention model

concatenating the representations from the main encoder layer and auxiliary sharing layers, we introduce an attention mechanism to produce an attention weight vector α and a weighted hidden representation s. To obtain the auxiliary representations, we adopt standard LSTM [19] layers as sharing layers between the networks from the main and auxiliary tasks.

Formally, the middle representation of the main task is generated from both the main encoder layer and auxiliary sharing layers, i.e.

$$h_{main0} = \text{LSTM}_{main}(T_{main}) \tag{1}$$

$$h_{main1} = \text{LSTM}_{aux1}(T_{main}) \tag{2}$$

$$h_{main2} = \text{LSTM}_{aux2}(T_{main}) \tag{3}$$

$$h_{main3} = \text{LSTM}_{aux3}(T_{main}) \tag{4}$$

where $T_{main} = \{w^1_{main} \ldots w^L_{main}\}$ represents the input sequence from the Chinese emotion corpus (CN1). h_{main0} means the representation for the classification model via the main encoder layer. While h_{main1}, h_{main2} and h_{main3} mean the representations for the classification model via the auxiliary sharing layers.

With these main and auxiliary representations, we compute an attention weight vector $\alpha = [\alpha_{main}^1, \alpha_{main}^2, \ldots, \alpha_{main}^L]$ as follows:

$$m_{main}^i = \tanh(W_m \cdot [h_{main0}^i \oplus h_{main1}^i \oplus h_{main2}^i \oplus h_{main3}^i] + b_m) \tag{5}$$

$$\alpha_{main}^i = \text{softmax}(m_{main}^i) = \frac{\exp(m_{main}^i)}{\sum_{t=1}^{L} m_{main}^t} \tag{6}$$

where $1 \le i \le L$ and L is the length of the input sequence. h_{main0}^i, h_{main1}^i, h_{main2}^i and h_{main3}^i represent the i-th word in h_{main0}, h_{main1}, h_{main2} and h_{main3} respectively. $\oplus$ denotes the concatenate operator. W_m is an intermediate matrix and b_m is an offset value.

Then, we compute the final sample representation as a weighted sum of the word annotations:

$$s_{main} = \sum_{i=1}^{L} \alpha_{main}^i \cdot [h_{main0}^i \oplus h_{main1}^i \oplus h_{main2}^i \oplus h_{main3}^i] \tag{7}$$

To perform emotion classification, a softmax layer is followed to transform s_{main} to conditional probability distribution:

$$p(y_{main}|T_{main}) = \text{softmax}(W_{main} \cdot s_{main} + b_{main}) \tag{8}$$

where $p(y_{main}|T_{main})$ is the output of the main task, W_{main} is the weight vector to be learned and b_{main} is the bias term.

3.3 The Auxiliary Emotion Classification Task

The representations of three auxiliary tasks are generated from corresponding auxiliary sharing layers respectively, i.e.,

$$h_{aux1} = \text{LSTM}_{aux1}(T_{aux1}) \tag{9}$$

$$h_{aux2} = \text{LSTM}_{aux2}(T_{aux2}) \tag{10}$$

$$h_{aux3} = \text{LSTM}_{aux3}(T_{aux3}) \tag{11}$$

where $T_{aux1} = \{w_{aux1}^1 \ldots w_{aux1}^L\}$, $T_{aux2} = \{w_{aux2}^1 \ldots w_{aux2}^L\}$, $T_{aux3} = \{w_{aux3}^1 \ldots w_{aux3}^L\}$ mean the input sequences from English-to-Chinese (CN2), Chinese-to-English (EN1) and English (EN2) emotion corpora respectively. h_{aux1}, h_{aux2} and h_{aux3} are the outputs from three auxiliary sharing layers respectively.

Then, a fully-connected layer followed by a dropout layer is leveraged to gain a feature vector for classification, i.e.,

$$h_{aux1}^* = dense(h_{aux1}) \cdot D(p_{aux1}^*) \tag{12}$$

$$h^*_{aux2} = dense(h_{aux2}) \cdot D(p^*_{aux2}) \tag{13}$$

$$h^*_{aux3} = dense(h_{aux3}) \cdot D(p^*_{aux3}) \tag{14}$$

where $dense(\cdot)$ denotes the output of the fully-connected layer. D defines the dropout operation and p^* is the dropout probability.

Once obtaining the representations of these auxiliary tasks, we feed them into a softmax layer respectively to perform emotion classification:

$$p(y_{aux1}|T_{aux1}) = \text{softmax}(W_{aux1} \cdot h^*_{aux1} + b_{aux1}) \tag{15}$$

$$p(y_{aux2}|T_{aux2}) = \text{softmax}(W_{aux2} \cdot h^*_{aux2} + b_{aux2}) \tag{16}$$

$$p(y_{aux3}|T_{aux3}) = \text{softmax}(W_{aux3} \cdot h^*_{aux3} + b_{aux3}) \tag{17}$$

where $p(y_{aux1}|T_{aux1})$, $p(y_{aux2}|T_{aux2})$ and $p(y_{aux3}|T_{aux3})$ are the outputs of the auxiliary tasks respectively. W_{aux1}, b_{aux1}, W_{aux2}, b_{aux2}, W_{aux3} and b_{aux3} are the parameters f or softmax layers.

3.4 Joint Learning

The model can be trained in an end-to-end manner where the objective loss function is a linear combination of the main task and auxiliary tasks:

$$\begin{aligned} J(\theta) = & -\lambda_1 \cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y^j_{main} \cdot \log p(y^j_{main}|T^i_{main}) - \lambda_2 \cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y^j_{aux1} \cdot \log p(y^j_{aux1}|T^i_{aux1}) \\ & -\lambda_3 \cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y^j_{aux2} \cdot \log p(y^j_{aux2}|T^i_{aux2}) - \lambda_4 \cdot \sum_{i=1}^{N}\sum_{j=1}^{C} y^j_{aux3} \cdot \log p(y^j_{aux3}|T^i_{aux3}) \\ & + \frac{l}{2}\|\theta\|_2^2 \end{aligned} \tag{18}$$

where y^j_{main}, y^j_{aux1}, y^j_{aux2} and y^j_{aux3} are the ground-truth labels from the main task and auxiliary tasks. N is the total quantity of training samples. C is the category number. l is a L_2 regularization to bias parameters and θ denotes all parameters. λ_1, λ_2, λ_3 and λ_4 are the weight parameters to balance the importance of losses between the main task and auxiliary tasks and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 1$. We take Adadelta [20] as the optimizing algorithm with a learning rate of 1.0. All the matrix and vector parameters are initialized with a uniform distribution in $\left[-\sqrt{6/(r+c)}, \sqrt{6/(r+c)}\right]$, where r and c are the rows and columns of the matrices.

4 Experiment

In this section, we systematically evaluate the performance of our approach to emotion classification.

4.1 Experimental Settings

- **Data Settings:** In order to assess the performance of the proposed approach, we use the Chinese emotion corpus constructed by Yao et al. [21]. This corpus consists of 14,000 instances, of which 7,407 instances express emotions. Seven basic emotions are defined as candidate categories, namely *anger*, *happiness*, *sadness*, *fear*, *like*, *surprise* and *disgust*. In addition, we use the dataset of SemEval 2018 Task1 as the English emotion corpus. It contains a lot of tweets and corresponding emotion categories, i.e., *anger*, *joy*, *sadness* and *fear*. Table 1 illustrates the distribution of these two datasets. As to enlarge the corpora mentioned above, we can translate one language into the other language. The Chinese data is much imbalanced and we extract a balanced dataset for each emotion category in Chinese. Due to the fact that the number of instances in *fear* category is too small, we decide to set the number of instances in *surprise* category as the basis to avoid contingency. We use 80% of instances as training data and the remaining 20% as test data. Furthermore, we set aside 10% of the training data as development data to fine tune the parameters in learning algorithm.

Table 1. Emotion categories and distribution on two corpora

Emotion	#Sentences in Chinese Corpus	#Sentences in English Corpus
anger	669	1901
happiness	1460	1816
sadness	1173	1733
fear	148	2452
like	2203	–
surprise	362	–
disgust	1392	–

- **Word Segmentation and Representations:** FudanNLP[2] is employed to segment each Chinese post into words and we learn distributed representation of each word with word2vec[3] (The skip-gram model is used) on each dataset. The vector dimension is set to be 100 and the window size is set to be 5.
- **Hyper-parameters:** The hyper-parameters in our approach are tuned according to the performance in the development data. The size of units in LSTM layer is 128 and all models are trained by mini-batch of 32 instances. λ_1 is set to be 0.5, λ_2, λ_3 and λ_4 are the same as each other.
- **Evaluation Metric:** We use *Macro-F1 (F)* and *Accuracy* to measure the divergences between predicted labels and ground-truth labels. Besides, *t*-test is used to determine whether the performance difference is statistically significant.

[2] https://github.com/FudanNLP/fnlp/.

[3] https://github.com/dav/word2vec/.

4.2 Experimental Results

In this section, we report the experimental results of our joint learning approach to emotion classification. For thorough comparison, we provide selected baseline approaches. In addition, we also implement some state-of-the-art approaches in sentiment classification to emotion classification.

- **LSTM (CN1):** This method applies the standard LSTM model using only the Chinese emotion corpus for emotion classification.
- **CNN-Tensor (CN1) [22]:** This is a state-of-the-art approach to sentiment classification, which appeals to tensor algebra and uses low-rank n-gram tensors to directly exploit interactions between words already at the convolution stage. It applies only the Chinese emotion corpus for emotion classification.
- **Attention-LSTM (CN1) [23]:** This is a state-of-the-art approach to aspect-level sentiment classification, which leverages the attention mechanism to concentrate on different parts of a sentence. Note that we ignore aspect embedding and use sentence representations from LSTM to yield an attention weight vector directly. It applies only the Chinese emotion corpus for emotion classification.
- **LSTM (CN1 + EN1):** This method combines the results of LSTM (CN1) and LSTM (EN1) by averaging the probabilities. It applies both Chinese and Chinese-to-English emotion corpora for emotion classification. This is an ensemble approach by Wan [24] which is proposed to deal with cross-lingual sentiment classification. Since the categories in the Chinese corpus and the English corpus are different, this approach could not be directly applied to combine all corpora (i.e., CN1 + CN2 + EN1 + EN2).
- **LSTM (CN1 + CN2):** This method simply merges Chinese and English-to-Chinese emotion samples in corresponding categories and applies the standard LSTM model for emotion classification.
- **Aux-LSTM (CN1 + CN2):** It applies the Aux-LSTM model with both Chinese and English-to-Chinese emotion corpora for emotion classification. It simply concatenates the representations from the main and auxiliary task.
- **Aux-LSTM (CN1 + EN1):** It applies the Aux-LSTM model with both Chinese and Chinese-to-English emotion corpora for emotion classification. It simply concatenates the representations from the main and auxiliary task.
- **Aux-LSTM (CN1 + EN2):** It applies the Aux-LSTM model with both Chinese and English emotion corpora for emotion classification. It simply concatenates the representations from the main and auxiliary task.
- **Aux-LSTM (CN1 + CN2 + EN1 + EN2):** It applies the Aux-LSTM model with all Chinese, English-to-Chinese, Chinese-to-English and English emotion corpora for emotion classification. It simply concatenates the representations from the main and auxiliary tasks.
- **Aux-LSTM-Attention (CN1 + CN2 + EN1 + EN2):** It applies the Aux-LSTM model with attention on all Chinese, English-to-Chinese, Chinese-to-English and English emotion corpora for emotion classification.

Table 2 shows the results of different approaches to Chinese emotion classification. From the table, we can see that all Aux-LSTM models consistently outperform the

Table 2. Performance comparison of different approaches to emotion classification

	Macro-F1	*Accuracy*
LSTM (CN1)	0.39087	0.36255
CNN-Tensor (CN1)	0.41468	0.40125
Attention-LSTM (CN1)	0.44048	0.41573
LSTM (CN1 + EN1)	0.40278	0.37669
LSTM (CN1 + CN2)	0.42063	0.40619
Aux-LSTM (CN1 + CN2)	0.45238	0.43472
Aux-LSTM (CN1 + EN1)	0.45833	0.44773
Aux-LSTM (CN1 + EN2)	0.45040	0.43806
Aux-LSTM (CN1 + CN2 + EN1 + EN2)	0.46429	0.43795
Aux-LSTM-Attention (CN1 + CN2 + EN1 + EN2)	**0.49802**	**0.49355**

baseline approaches whichever corpus is employed, which verifies the effectiveness of the proposed Aux-LSTM model. These results encourage to incorporate other-language labeled data to improve the performance of emotion classification. However, with the increase in number of additional corpora, it could not bring about remarkable results any more. Hence, instead of simply concatenating the representations from the main encoder layer and auxiliary sharing layers, we introduce an attention mechanism to provide insight into which words contribute to the emotion classification decision. Among all these approaches, our Aux-LSTM-Attention model performs best, which suggests sharing additional corpora and utilizing attention mechanism to capture the informative words. Significance test shows that the improvement of Aux-LSTM-Attention model over the other approaches is significant (*p-value* < 0.05).

To better understand why our joint learning approach is so effective, we calculate the standard precision (*P*), recall (*R*) and F-score (*F*) in each category. Table 3 demonstrates these specific results. For clarity, we only report the results of LSTM (CN1) and Aux-LSTM-Attention (CN1 + CN2 + EN1 + EN2). From Table 3, we can see that our joint learning approach is obviously superior to LSTM (CN1) in almost every category, especially in the *like* category and *surprise* category. The performance of our approach

Table 3. Comparative results with standard precision (*P*), recall (*R*) and F-score (*F*) in each category

	LSTM (CN1)			Aux-LSTM-Attention (CN1 + CN2 + EN1 + EN2)		
	P	*R*	*F*	*P*	*R*	*F*
anger	0.551	0.681	0.609	0.610	0.694	0.649
happiness	0.393	0. 611	0.478	0.603	0.528	0.563
sadness	0.337	0. 486	0.398	0.494	0.542	0.517
fear	0.409	0. 500	0.450	0.419	0.431	0.425
like	0.346	0. 125	0.184	0.524	0.458	0.489
surprise	0.233	0. 139	0.174	0.408	0.556	0.471
disgust	0.333	0. 194	0.246	0.444	0.278	0.342

is consistent in each category while LSTM (CN1) fluctuates widely. It indicates that our approach is more effective for predicting emotion in Chinese emotion text.

4.3 Case Study

In Fig. 4, we list two examples from the test set which have not been correctly inferred by LSTM (CN1) model due to the limitation of Chinese emotion corpus. In **E3**, "骗子 *(liar)*" expresses a strong sentiment signal to the emotion *anger* but it does not exist in the training set so that this sample could not be correctly classified by LSTM (CN1). However, when we translate it into English, i.e., *liar* and it can be found in the additional English emotion corpus, our Aux-LSTM model can work well. In **E4**, "炫耀 *(show off)*" is not contained in the training set either. But if the corresponding translation "*show off*" can be found in the additional English emotion corpus, then we can leverage our Aux-LSTM model to predict the correct emotion (*happiness*) easily.

E3: Ground-truth label: *anger*	
Original: *刚刚收到短信…这绝对是骗子！* Translation: *Just received a text message…This is definitely a* ***liar****!*	
LSTM (CN1)	Aux-LSTM (CN1+CN2+EN1+EN2)
× (*surprise*)	√ (*anger*)

E4: Ground-truth label: *happiness*	
Original: *来炫耀！入手了表哥千里迢迢捎来的…哪怕惧怕昆虫的人也该去入一本。* Translation: *To* ***show off****! Started the cousin's journey…even those who are afraid of insects should go to a book.*	
LSTM (CN1)	Aux-LSTM (CN1+CN2+EN1+EN2)
× (*like*)	√ (*happiness*)

Fig. 4. Examples of emotion classification

4.4 Visualization of Attention

Figure 5 shows the attention visualization for a post in the test set. The color depth indicates the importance degree of corresponding word - the darker the shade, the more important the word. Obviously, the attention mechanism obtains the important elements which carry strong sentiment signals from the whole post dynamically, such as "*terrible*" and "*exaggerated*".

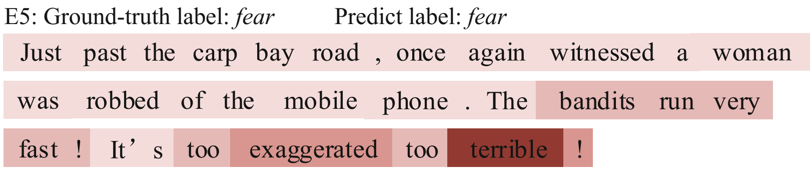
E5: Ground-truth label: *fear*
Predict label: *fear*
Just past the carp bay road , once again witnessed a woman
was robbed of the mobile phone . The bandits run very
fast ! It' s too exaggerated too terrible !

Fig. 5. Attention visualization

5 Conclusion

In this paper, we address the corpus scarce challenge in emotion classification from a cross-lingual view and propose a joint learning framework, namely Aux-LSTM-Attention, to perform emotion classification when both resource-poor and resource-rich corpora exist. Specially, we employ sharing layers to develop auxiliary representations for the main task. Furthermore, an attention mechanism is utilized to capture the informative words. Empirical studies show that our joint learning approach successfully improves the performance of emotion classification by using the labeled data from a different language. Moreover, empirical studies demonstrate that our approach outperforms several strong baseline approaches to emotion classification.

In our future work, we would like to explore tackling the corpus scarcity challenge by using the labeled data from multiple languages. Furthermore, we will attempt to apply our approach to other natural language processing tasks in which the annotated corpus is limited.

Acknowledgements. This research work has been partially supported by two NSFC grants, No. 61331011 and No. 61672366.

References

1. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
2. Galik, M., Rank, S.: Modeling emotional trajectories of individuals in an online chat. In: MATES, pp. 96–105 (2012)
3. Liu, H., Li, S., Zhou, G., Huang, C., Li, P.: Joint modeling of news reader's and comment writer's emotions. In: ACL, pp. 511–515 (2013)
4. Li, S., Xu, J., Zhang, D., Zhou, G.: Two-view label propagation to semi-supervised reader emotion classification. In: COLING, pp. 2647–2655 (2016)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR, abs/1409.0473, 2014
6. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: NAACL-HLT, pp. 1480–1489 (2016)
7. Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers. 1–167 (2012)
8. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: EMNLP, pp. 79–86 (2002)
9. Mihalcea, R., Banea, C., Wiebe, J.: Learning multilingual subjective language via cross-lingual projections. In: ACL, pp. 976–983 (2007)
10. Wan, X.: Co-training for cross-lingual sentiment classification. In: ACL, pp. 235–243 (2009)
11. Balamurali, A., Aditya, J., Pushpak, B.: Cross-lingual sentiment analysis for indian languages using linked wordnets. In: COLING, pp. 73–82 (2012)
12. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: ACL, pp. 1118–1127 (2007)
13. Zhou, X., Wan, X., Xiao, J.: Cross-lingual sentiment classification with bilingual document representation learning. In: ACL, pp. 1403–1412 (2016)

14. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the web. In: COLING, pp. 881–888 (2008)
15. Bhowmick, P., Basu, A., Mitra, P., Prasad, A.: Multi-label text classification approach for sentence level news emotion analysis. In: PReMI, pp. 261–266 (2009)
16. Xu, J., Xu, R., Lu, Q.: Coarse-to-fine sentence-level emotion classification based on the intra-sentence features and sentential context. In: CIKM, pp. 2455–2458 (2012)
17. Yang, M., Peng, B., Chen, Z., Zhu, D., Chow, K.: A topic model for building fine-grained domain-specific emotion lexicon. In: ACL, pp. 421–426 (2014)
18. Felbo, B., Mislove, A., SØgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domian representations for detecting sentiment, emotion and sarcasm. In: EMNLP, pp. 1615–1625 (2017)
19. Graves, A.: Generating sequences with recurrent neural networks. CoRR, abs/1308.0850, 2013
20. Zeiler, M.: ADADELTA: an adaptive learning rate method. CoRR, abs/1212.5701 (2012)
21. Yao, Y., Wang, S., Xu, R., Liu, B., Gui, L., Lu, Q., Wang, X.: The construction of an emotion annotated corpus on microblog. J. Chin. Inf. Process. **28**(5), 83–91 (2014)
22. Lei, T., Barzilay, R., Jaakkola, T.: Modeling CNNs for text: non-linear, non-consecutive convolutions. In: EMNLP, pp. 1565–1575 (2015)
23. Wang, Y., Huang, M., Zhao, L., Zhu, X.: Attention-based LSTM for aspect-level sentiment classification. In: EMNLP, pp. 606–615 (2016)
24. Wan, X.: Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In: EMNLP, pp. 553–561 (2008)