Attention-based LSTM Network for Cross-Lingual Sentiment Classification

Xinjie Zhou, **Xiaojun Wan** and **Jianguo Xiao**
Institute of Computer Science and Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{zhouxinjie,wanxiaojun,xiaojianguo}@pku.edu.cn

Abstract

Most of the state-of-the-art sentiment classification methods are based on supervised learning algorithms which require large amounts of manually labeled data. However, the labeled resources are usually imbalanced in different languages. Cross-lingual sentiment classification tackles the problem by adapting the sentiment resources in a resource-rich language to resource-poor languages. In this study, we propose an attention-based bilingual representation learning model which learns the distributed semantics of the documents in both the source and the target languages. In each language, we use Long Short Term Memory (LSTM) network to model the documents, which has been proved to be very effective for word sequences. Meanwhile, we propose a hierarchical attention mechanism for the bilingual LSTM network. The sentence-level attention model learns which sentences of a document are more important for determining the overall sentiment while the word-level attention model learns which words in each sentence are decisive. The proposed model achieves good results on a benchmark dataset using English as the source language and Chinese as the target language.

1 Introduction

Most of the sentiment analysis research focuses on sentiment classification which aims to determine whether the users attitude is positive, neutral or negative. There are two classes of mainstreaming sentiment classification algorithms: unsupervised methods which usually require a sentiment lexicon (Taboada et al., 2011) and supervised methods (Pang et al., 2002) which require manually labeled data. However, both of these sentiment resources are unbalanced in different languages. The sentiment lexicon or labeled data are rich in several languages such as English and are poor in others. Manually building these resources for all the languages will be expensive and time-consuming. Cross-lingual sentiment classification tackles the problem by trying to adapt the resources in one language to other languages. It can also be regarded as a special kind of cross-lingual text classification task.

Recently, there have been several bilingual representation learning methods such as (Hermann and Blunsom, 2014; Gouws et al., 2014) for cross-lingual sentiment or text classification which achieve promising results. They try to learn a joint embedding space for different languages such that the training data in the source language can be directly applied to the test data in the target language. However, most of the studies only use simple functions, e.g. arithmetic average, to synthesize representations for larger text sequences. Some of them use more complicated compositional models such as the bi-gram non-linearity model in (Hermann and Blunsom, 2014) which also fail to capture the long distance dependencies in texts.

In this study, we propose an attention-based bilingual LSTM network for cross-lingual sentiment classification. LSTMs have been proved to be very effective to model word sequences and are powerful to learn on data with long range temporal dependencies. After translating the training data into the target language using machine translation

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 247–256,
Austin, Texas, November 1-5, 2016.

tools, we use the bidirectional LSTM network to model the documents in both of the source and the target languages. The LSTMs show strong ability to capture the compositional semantics for the bilingual texts in our experiments.

For the traditional LSTM network, each word in the input document is treated with equal importance, which is reasonable for traditional text classification tasks. In this paper, we propose a hierarchical attention mechanism which enables our model to focus on certain part of the input document. The motivation mainly comes from the following three observations: 1) the machine translation tool that we use to translate the documents will always introduce much noise for sentiment classification. We hope that the attention mechanism can help to filter out these noises. 2) In each individual language, the sentiment of a document is usually decided by a relative small part of it. In a long review document, the user might discuss both the advantages and disadvantages of a product. The sentiment will be confusing if we consider each sentence of the same contribution. For example, in the first review of Table 1, the first sentence reveals a negative sentiment towards the movie but the second one reveals a positive sentiment. As human readers, we can understand that the review is expressing a positive overall sentiment but it is hard for the sequence modeling algorithms including LSTM to capture. 3) At the sentence level, it is important to focus on the sentiment signals such as the sentiment words. They are usually very decisive to determine the polarity even for a very long sentence, e.g. "easy" and "nice" in the second example of Table 1.

"I felt it could have been a lot better with a little less comedy and a little more drama to get the point across. *However, its still a must see for any Jim Carrey fan.* "
"It is *easy* to read, it is *easy* to look things up in and provides a *nice* section on the treatments."

Table 1: Examples of the sentiment attention

In sum, the main contributions of this study are summarized as follows:

1) We propose a bilingual LSTM network for cross-lingual sentiment classification. Compared to the previous methods which only use weighted or arithmetic average of word embeddings to represent the document, LSTMs have obvious advantage to model the compositional semantics and to capture the long distance dependencies between words for bilingual texts.

2) We propose a hierarchical bilingual attention mechanism for our model. To the best of our knowledge, this is the first attention-based model designed for cross-lingual sentiment analysis.

3) The proposed framework achieves good results on a benchmark dataset from a cross-language sentiment classification evaluation. It outperforms the best team in the evaluation as well as several strong baseline methods.

2 Related Work

Sentiment analysis is the field of studying and analyzing peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions (Liu, 2012). The most common task of sentiment analysis is polarity classification which arises with the emergence of customer reviews on the Internet. Pang et al. (2002) used supervised learning methods and achieved promising results with simple unigram and bi-gram features. In subsequent research, more features and learning algorithms were tried for sentiment classification by a large number of researchers. Recently, the emerging of deep learning has also shed light on this area. Lots of representation learning methods has been proposed to address the sentiment classification task and many of them achieve the state-of-the-art performance on several benchmark datasets, such as the recursive neural tensor network (Socher et al., 2013), paragraph vector (Le and Mikolov, 2014), multi-channel convolutional neural networks (Kim, 2012), dynamic convolutional neural network (Blunsom et al., 2014) and tree structure LSTM (Tai et al., 2015). Very recently, Yang et al. (2016) proposed a similar hierarchical attention network based on GRU in the monolingual setting. Note that our work is independent with theirs and their study was released online after we submitted this study.

Cross-lingual sentiment classification is also a popular research topic in the sentiment analysis

community which aims to solve the sentiment classification task from a cross-language view. It is of great importance since it can exploit the existing labeled information in a source language to build a sentiment classification system in any other target language. Cross-lingual sentiment classification has been extensively studied in the very recent years. Mihalcea et al. (2007) translated English subjectivity words and phrases into the target language to build a lexicon-based classifier. Wan (2009) translated both the training data (English to Chinese) and the test data (Chinese to English) to train different models in both the source and target languages. Chen et al. (2015) proposed a knowledge validation method and incorporated it into a boosting model to transfer credible information between the two languages during training.

There have also been several studies addressing the task via multi-lingual text representation learning. Xiao and Guo (2013) learned different representations for words in different languages. Part of the word vector is shared among different languages and the rest is language-dependent. Klementiev et al. (2012) treated the task as a multi-task learning problem where each task corresponds to a single word, and the task relatedness is derived from co-occurrence statistics in bilingual parallel corpora. Chandar A P et al. (2014) and Zhou et al. (2015) used the autoencoders to model the connections between bilingual sentences. It aims to minimize the reconstruction error between the bag-of-words representations of two parallel sentences. Pham et al. (2015) extended the paragraph model into bilingual setting. Each pair of parallel sentences shares the same paragraph vector.

Compared to the existing studies, we propose to use the bilingual LSTM network to learn the document representations of reviews in each individual language. It has obvious advantage to model the compositional semantics and to capture the long distance dependencies between words. Besides, we propose a hierarchical neural attention mechanism to capture the sentiment attention in each document. The attention model helps to filter out the noise which is irrelevant to the overall sentiment.

3 Preliminaries

3.1 Problem Definition

Cross-language sentiment classification aims to use the training data in the source language to build a model which is adaptable for the test data in the target language. In our setting, we have labeled training data in English $L_{EN} = \{x_i, y_i\}_{i=1}^{N}$, where x_i is the review text and y_i is the sentiment label vector. $y_i = (1, 0)$ represents the positive sentiment and $y_i = (0, 1)$ represents the negative sentiment. In the target language Chinese, we have the test data $T_{CN} = \{x_i\}_{i=1}^{T}$ and unlabeled data $U_{CN} = \{x_i\}_{i=1}^{M}$. The task is to use L_{EN} and U_{CN} to learn a model and classify the sentiment polarity for the review texts in T_{CN}.

In our method, the labeled, unlabeled and test data are all translated into the other language using an online machine translation tool. In the subsequent part of the paper, we refer to a document and its corresponding translation in the other language as a pair of parallel documents.

3.2 RNN and LSTM

Recurrent neural network (RNN) (Rumelhart et al., 1988) is a special kind of feed-forward neural network which is useful for modeling time-sensitive sequences. At each time *t*, the model receives input from the current example and also from the hidden layer of the network's previous state. The output is calculated given the hidden state at that time stamp. The recurrent connection makes the output at each time associated with all the previous inputs. The vanilla RNN model has been considered to be difficult to train due to the well-known problem of vanishing and exploding gradients. The LSTM (Hochreiter and Schmidhuber, 1997) addresses the problem by re-parameterizing the RNN model. The core idea of LSTM is introducing the "gates" to control the data flow in the recurrent neural unit. The LSTM structure ensures that the gradient of the long-term dependencies cannot vanish. The detailed architecture that we use in shown in Figure 1.

4 Framework

In this study, we try to model the bilingual texts through the attention based LSTM network. We first

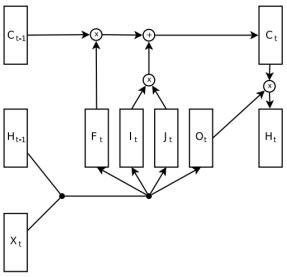

Figure 1: The LSTM architecture. The image is adopted from (Jozefowicz et al., 2015).

describe the general architecture of the model and then describe the attention mechanism used in it.

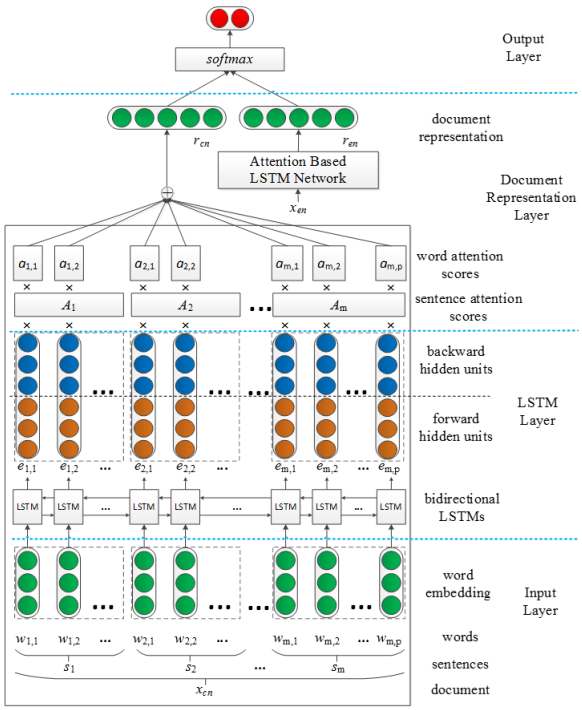

Figure 2: The architecture of the proposed framework. The inputs x_{cn} and x_{en} are parallel documents. Due to space limit, we only illustrate the attention based LSTM network in Chinese language. For the English document x_{en}, the network architecture is the same as the Chinese side but has different model parameters.

4.1 Architecture

The general architecture of our approach is shown in Figure 2. For a pair of parallel documents x_{cn} and x_{en}, each of them is sent into the attention based LSTM network. The English-side and Chinese-side architectures are the same but have different parameters. We only show the Chinese-side network in the figure due to space limit. The whole model is divided into four layers. In the input layer, the documents are represented as a word sequence where each position corresponds to a word vector from pre-trained word embeddings. In the LSTM layer, we get the high-level representation from a bidirectional LSTM network. We use the hidden units from both the forward and backward LSTMs. In the document representation layer, we incorporate the attention model into the network and derive the final document representation. At the output layer, we concatenate the representations of the English and Chinese documents and use the softmax function to predict the sentiment label.

Input Layer: The input layer of the network is the word sequences in a document x which can be either Chinese or English. The document x contains several sentences $\{s_i\}_{i=1}^{|x|}$ and each sentence is composed of several words $s_i = \{w_{i,j}\}_{j=1}^{|s_i|}$. We represent each word in the document as a fixed-size vector from pre-trained word embeddings.

LSTM Layer: In each individual language, we use bi-directional LSTMs to model the input sequences. In the bidirectional architecture, there are two layers of hidden nodes from two separate LSTMs. The two LSTMs capture the dependencies in different directions. The first hidden layers have recurrent connections from the past words while second one's direction of recurrent of connections is flipped, passing activation backwards in the texts. Therefore, in the LSTM layer, we can get the forward hidden state $\overrightarrow{h}_{i,j}$ from the forward LSTM network and the backward hidden state $\overleftarrow{h}_{i,j}$ from the backward LSTM network. We represent the final state at position (i, j), i.e. the j-th word in the i-th sentence of the document, with the concatenation of $\overrightarrow{h}_{i,j}$ and $\overleftarrow{h}_{i,j}$.

$$h_{i,j} = \overrightarrow{h}_{i,j} \parallel \overleftarrow{h}_{i,j}$$

It captures the compositional semantics in both directions of the word sequences.

Document Representation Layer: As described above, different parts of the document usually have different importance for the overall sentiment. Some

sentences or words can be decisive while the others are irrelevant. In this study, we use a hierarchical attention mechanism which assigns a real value score for each word and a real value score for each sentence. The detailed strategy of our attention model will be described in the next subsection.

Suppose we have the sentence attention score A_i for each sentence $s_i \in x$, and the word attention score $a_{i,j}$ for each word $w_{i,j} \in s_i$, both of the scores are normalized which satisfy the following equations,

$$\sum_i A_i = 1 \quad and \quad \sum_j a_{i,j} = 1$$

The sentence attention measures which sentence is more important for the overall sentiment while the word attention captures sentiment signals such as sentiment words in each sentence. Therefore, the document representation r for document x is calculated as follows,

$$r = \sum_i [A_i \cdot \sum_j (a_{i,j} \cdot h_{i,j})]$$

Note that many LSTM based models represent the word sequences only using the hidden layer at the final node. In this study, the hidden states at all the positions are considered with different attention weights. We believe that, for document sentiment classification, focusing on some certain parts of the document will be effective to filter out the sentiment-irrelevant noise.

Output Layer: At the output layer, we need to predict the overall sentiment of the document. For each English document x_{en} and its corresponding translation x_{cn}, suppose the document representations of them are obtained in previous steps as r_{en} and r_{cn}, we simply concatenate them as the feature vector and use the softmax function to predict the final sentiment.

$$\hat{y} = softmax(r_{cn} \parallel r_{en})$$

4.2 Hierarchical Attention Mechanism

For document-level sentiment classification task, we have shown that capturing both the sentence and word level attention is important. The general idea is inspired by previous works such as Bahdanau et al. (2014) and Hermann et al. (2015) which have successfully applied the attention model to machine translation and question answering. Bahdanau et al. (2014) incorporated the attention model into the sequence to sequence learning framework. During the decoding phase of the machine translation task, the attention model helps to find which input word should be "aligned" to the current output. In our case, the output of the model is not a sequence but only one sentiment vector. We hope to find the important units in the input sequence which are influential for the output.

We propose to learn a hierarchical attention model jointly with the bilingual LSTM network. The first level is the sentence attention model which measures which sentences are more important for the overall sentiment of a document. For each sentence $s_i = \{w_{i,j}\}_{j=1}^{|s_i|}$ in the document, we represent the sentence via the final hidden state of the forward LSTM and the backward LSTM, i.e.

$$s_i = \vec{h}_{i,|s_i|} \parallel \overleftarrow{h}_{i,1}$$

We use a two-layer feed-forward neural network to predict the attention score of s_i

$$\hat{A}_i = f(s_i; \theta_s)$$

$$A_i = \frac{\exp(\hat{A}_i)}{\sum_j \exp(\hat{A}_j)}$$

where f denotes the two-layer feed-forward neural network and θ_s denotes the parameters in it.

At the word level, we represent each word $w_{i,j}$ using its word embedding and the hidden state of the bidirectional LSTM layer, i.e. $h_{i,j}$. Similarly, we use a two-layer feed forward neural network to predict the attention score of $w_{i,j}$,

$$e_{i,j} = w_{i,j} \parallel \vec{h}_{i,j} \parallel \overleftarrow{h}_{i,j}$$

$$\hat{a}_{i,j} = f(e_{i,j}; \theta_w)$$

$$a_{i,j} = \frac{\exp(\hat{a}_{i,j})}{\sum_j \exp(\hat{a}_{i,j})}$$

where θ_w denotes the parameters for predicting word attention.

4.3 Training of the Proposed Model

The proposed model is trained in a semi-supervised manner. In the supervised part, we use the cross entropy loss to minimize the sentiment prediction error between the output results and the gold standard labels,

$$L_1 = \sum_{(x_{en}, x_{cn})} \sum_i -y_i \log(\hat{y}_i)$$

where x_{en} and x_{cn} are a pair of parallel documents in the training data, y is the gold-standard sentiment vector and $\hat{y}$ is the predicted vector from our model.

The unsupervised part tries to minimize the document representations between the parallel data. Following previous research, we simply measure the distance of two parallel documents via the Euclidean Distance,

$$L_2 = \sum_{(x_{en}, x_{cn})} \|r_{en} - r_{cn}\|^2$$

where x_{en} and x_{cn} are a pair of parallel documents from both the labeled and unlabeled data.

The final objective function is a weighted sum of L_1 and L_2,

$$L = L_1 + \alpha \cdot L_2$$

where α is the hyper-parameter controlling the weight. We use Adadelta (Zeiler, 2012) to update the parameters during training. It can dynamically adapt over time using only first order information and has minimal computational overhead beyond vanilla stochastic gradient descent.

In the test phase, the test document in T_{CN} is sent into our model along with the corresponding machine translated text in T_{EN}. The final sentiment is predicted via a softmax function over the concatenated representation of the bilingual texts as described above.

5 Experiment

5.1 Dataset

We use the dataset from the cross-language sentiment classification evaluation of NLP&CC 2013.[1] The dataset contains reviews in three domains including book, DVD and music. In each domain, it has 2000 positive reviews and 2000 negative reviews in English for training and 4000 Chinese reviews for test. It also contains 44113, 17815 and 29678 unlabeled reviews for book, DVD and music respectively.

5.2 Implementation Detail

We use Google Translate[2] to translate the labeled data to Chinese and translate the unlabeled data and test data to English. All the texts are tokenized and converted into lower case.

In the proposed framework, the dimensions of the word vectors and the hidden layers of LSTMs are set as 50. The initial word embeddings are trained on both the unlabeled and labeled reviews using `word2vec` in each individual language. The word vectors are fine-tuned during the training procedure. The hyper-parameter a is set to 0.2. The dropout rate is set to 0.5 to prevent overfitting. Ten percent of the training data are randomly selected as validation set. The training procedure is stopped when the prediction accuracy does not improve for 10 iterations. We implement the framework based on theano (Bastien et al., 2012) and use a GTX 980TI graphic card for training.

5.3 Baselines and Results

To evaluate the performance of our model, we compared it with the following baseline methods:

LR and **SVM:** We use logistic regression and SVM to learn different classifiers based on the translated Chinese training data. We simply use unigram features.

MT-PV: Paragraph vector (Le and Mikolov, 2014) is considered as one of the state-of-the-art monolingual document modeling methods. We translate all the training data into Chinese and use paragraph vector to learn a vector representation for the training and test data. A logistic regression classifier is used to predict the sentiment polarity.

Bi-PV: Pham et al. (2015) is one the state-of-the-art bilingual document modeling methods. It extends the paragraph vector into bilingual setting.

[1] The dataset can be found at http://tcci.ccf.org.cn/conference/2013/index.html. NLP&CC is an annual conference specialized in the fields of Natural Language Processing (NLP) and Chinese Computing (CC) organized by Chinese Computer Federation (CCF).

[2] http://translate.google.com/

Each pair of parallel sentences in the training data shares the same vector representation.

BSWE: Zhou et al. (2015) proposed the bilingual sentiment word embedding algorithm based on denoising autoencoders. It learns the vector representations for 2000 sentiment words. Each document is then represented by the sentiment words and the corresponding negation words in it.

H-Eval: Gui et al. (2013) got the highest performance in the NLP&CC 2013 cross-lingual sentiment classification evaluation. It uses a mixed CLSC model by combining co-training and transfer learning strategies.

A-Eval: This is the average performance of all the teams in the NLP&CC 2013 cross-lingual sentiment classification evaluation.

The attention-based models **EN-Attention, CN-Attention** and **BI-Attention**: Bi-Attention is the model described in the above sections which concatenate the document representations of the English side and the Chinese side texts. EN-Attention only translates the Chinese test data into English and uses English-side attention model while CN-Attention only uses the Chinese side attention model.

Method	Domains			Average
	book	DVD	music	
LR	0.765	0.796	0.741	0.767
SVM	0.779	0.814	0.707	0.767
MT-PV	0.753	0.799	0.748	0.766
Bi-PV	0.785	0.820	0.753	0.796
BSWE	0.811	0.816	0.794	0.807
A-Eval	0.662	0.660	0.675	0.666
H-Eval	0.785	0.777	0.751	0.771
EN-Attention	0.798	0.827	0.808	0.811
CN-Attention	0.820	**0.840**	0.809	0.823
BI-Attention	**0.821**	0.837	**0.813**	**0.824**

Table 2: Cross-lingual sentiment prediction accuracy of our methods and the comparison approaches.

Table 2 shows the cross-lingual sentiment classification accuracy of all the approaches. The first kind baseline algorithms are based on traditional bag-of-word features. SVM performs better than LR on book and DVD but gets much worse result on music. The second kind baseline algorithms are based on deep learning methods which learn the vector representations for words or documents. MT-PV achieves similar results with LR. Bi-PV improves the accuracy by about 0.03 using both the bilingual documents. While MT-PV and Bi-PV directly learn document representations, BSWE learns the embedding for the words in a bilingual sentiment lexicon. It gets higher accuracy than both Bi-PV and MT-PV which shows that the sentiment words are very important for this task.

Our attention based models achieve the highest prediction accuracy among all the approaches. The results show that CN-Attention always outperforms EN-Attention. The combination of the English-side and Chinese-side model brings improvement to both the book and music domains and yields the highest average prediction accuracy. The attention-based models outperform the algorithms using traditional features as well as the existing deep learning based methods. Compared to the highest performance in the NLP&CC evaluation, we improve the average accuracy by about 0.05.

5.4 Influence of the Attention Mechanism

In this study, we propose a hierarchical attention mechanism to capture the sentiment-related information of each document. In table 3, we show the results of models with different attention mechanisms. All the models are based on the bilingual bi-directional LSTM network as shown in Figure 2. LSTM is the basic bilingual bi-directional LSTM network. LSTM+SA considers only sentence-level attention while LSTM+WA considers only word-level attention. LSTM+HA combines both word-level and sentence-level attentions. From the results, we can observe that LSTM+HA outperforms the other three methods, which proves the effectiveness of the hierarchical attention mechanism. Besides, the word-level attention shows better performance than the sentence-level attention.

Method	Average Accuracy
LSTM	0.811
LSTM+SA	0.814
LSTM+WA	0.821
LSTM+HA	0.824

Table 3: Comparison of different attention mechanisms

We also conduct a case study using the examples in Table 1. We show the visualized word attention

using a heat map in Figure 3 by drawing the attention of each word in it. The darker color reveals higher attention scores while the lighter part has little importance. We can observe that our model successfully identifies the important units of the sentence. The sentiment word "easy" gets much higher attention score than the other words. The word "nice" gets the third highest score in the sentence right after the two "easy". Note that our attention mechanism considers both the word embedding vector and the hidden state vectors. Therefore, the same word "easy" gets different scores in different positions.

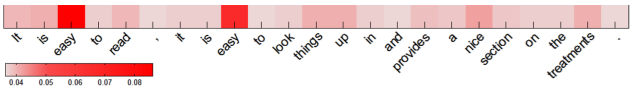

Figure 3: Attention visualization for a review sentence

5.5 Influence of the Word Embeddings

For the deep learning based methods, the initial word embeddings used as the inputs for the network usually play an important role. We study four different settings called *rand*, *static*, *fine-tuned* and *multi-channel*, respectively. In *rand* setting, the word embeddings are randomly initialized. The *static* setting keeps initial embedding fixed while the *fine-tuned* setting learns a refined embedding during the training procedure. *Multi-channel* is the combination of *static* and *fine-tuned*. Two same word vectors are concatenated to represent each word. During the training procedure, half of it is fine-tuned while the rest is fixed. Note that *fine-tuned* is the embedding setting that we use in our model.

Embedding Settings	Domains			Average
	book	DVD	music	
rand	0.789	0.786	0.746	0.774
static	0.804	0.810	0.784	0.799
fine-tuned	0.821	0.837	0.813	0.824
multi-channel	0.822	0.835	0.806	0.821

Table 4: Performance of our model with four different word embedding settings

Table 4 shows the performance of our model in these settings. *Rand* gets the lowest accuracy among them. The *fine-tuned* word embeddings perform better than *static* which fits the results in previous study (Kim, 2012). *Multi-channel* gets similar results with *fine-tuned* on DVD and music but is a bit lower on book. We also find that using pre-trained word embeddings helps the model to converge much faster than random initialization.

5.6 Influence of Vector Sizes

In our experiment, we set the size of the hidden layers in both the forward and backward LSTMs the same as the size of the input word vectors. Therefore, the dimension of the document representation is twice of the word vector size. In Figure 4, we show the performance of our model with different input vector sizes. We use the vector size in the following set $\{10, 25, 50, 100, 150, 200\}$. Note that the dimensions of all the units in the model also change with that.

We can observe from Figure 4 that the prediction accuracy for the book domain keeps steady when the vector size changes. For DVD and music, the performance increases at the beginning and becomes stable after the vector size grows larger than 50. It shows that our model is robust to a wide range of vector sizes.

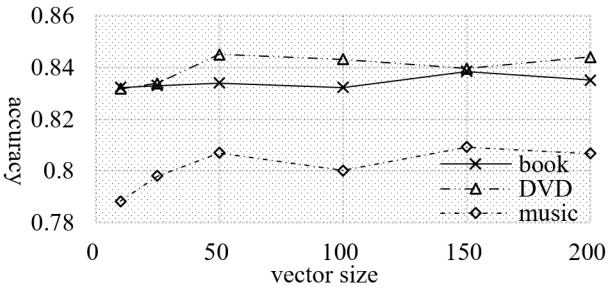

Figure 4: Performance with different vector sizes

6 Conclusion

In this paper, we propose an attention based LSTM network for cross-language sentiment classification. We use the bilingual bi-directional LSTMs to model the word sequences in the source and target languages. Based on the special characteristics of the sentiment classification task, we propose a hierarchical attention model which is jointly trained with the LSTM network. The sentence level attention

enables us to find the key sentences in a document and the word level attention helps to capture the sentiment signals. The proposed model achieves promising results on a benchmark dataset using Chinese as the source language and English as the target language. It outperforms the best results in the NLPC&CC cross-language sentiment classification evaluation as well as several strong baselines. In future work, we will evaluate the performance of our model on more datasets and more language pairs. The sentiment lexicon is also another kind of useful resource for classification. We will explore how to make full usages of these resources in the proposed framework.

Acknowledgments

The work was supported by National Natural Science Foundation of China (61331011), National Hi-Tech Research and Development Program (863 Program) of China (2015AA015403, 2014AA015102) and IBM Global Faculty Award Program. We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian Goodfellow, Arnaud Bergeron, Nicolas Bouchard, David Warde-Farley, and Yoshua Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*.

Phil Blunsom, Edward Grefenstette, and Nal Kalchbrenner. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*.

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

Qiang Chen, Wenjie Li, Yu Lei, Xule Liu, and Yanxiang He. 2015. Learning to adapt credible knowledge in cross-lingual sentiment analysis. In *Proceedings of 52rd Annual Meeting of the Association for Computational Linguistic*.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2014. Bilbowa: Fast bilingual distributed representations without word alignments. *arXiv preprint arXiv:1410.2455*.

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

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of 52rd Annual Meeting of the Association for Computational Linguistic*, pages 58–68.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yoon Kim. 2012. Convolutional neural networks for sentence classification. In *Proceedings of EMNLP 2014*, pages 1746–1751.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1759–1774.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

B Liu. 2012. Sentiment analysis and opinion mining: Synthesis lectures on human language technologies, vol. 16. *Morgan & Claypool Publishers, San Rafael*.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Hieu Pham, Minh-Thang Luong, and Christopher D Manning. 2015. Learning distributed representations for multilingual text sequences. In *Proceedings of NAACL-HLT*, pages 88–94.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1988. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics.

Min Xiao and Yuhong Guo. 2013. Semi-supervised representation learning for cross-lingual text classification. In *Proceedings of EMNLP 2013*, pages 1465–1475.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489. Association for Computational Linguistics.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Huiwei Zhou, Long Chen, Fulin Shi, and Degen Huang. 2015. Learning bilingual sentiment word embeddings for cross-language sentiment classification. In *Proceedings of 52rd Annual Meeting of the Association for Computational Linguistic*, pages 430–440.