Sentiment Classification Method for Chinese and Vietnamese Bilingual News Sentence Based on Convolution Neural Network

Xiaoxu He[1,2], Shengxiang Gao[1,2(✉)], Zhengtao Yu[1,2], Shulong Liu[1,2], and Xudong Hong[1,2]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
gaoshengxiang.yn@foxmail.com
[2] Key Lab of Computer Technologies Application of Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China

Abstract. Sentiment classification of the Chinese and Vietnamese news is important to analysis public opinion of Chinese and Vietnamese. For cross-lingual sentiment analysis, this paper studies the method of sentiment analysis for Chinese and Vietnamese bilingual news sentence based on convolution neural network (CNN). Firstly, collect Chinese and Vietnamese news data from Internet, use the data collected to train word vector. Secondly, employ the pre-trained word vector and bilingual lexicon to project the Chinese news labeled data to bilingual word vector space, the problem of the lack of annotated corpus in Vietnamese is solved in a certain level. Finally, using convolution neural network to make cross-lingual sentiment classification. Experiment result shows that the F score of Vietnamese news classification has reached 86%. Compared with the traditional methods, the proposed method does not need to construct sentiment lexicon and perform feature extraction, and the performance of cross-lingual sentiment classification is also remarkable.

Keywords: Sentiment analysis · Chinese news · Vietnamese news
Convolution neural network

1 Introduction

With the rapid developing of Internet and the rising of Web2.0, the text on the network is increasing. It is very important to realize the analysis and utilization of the Internet text. Opinion Mining [1] is generated in this context, the purpose is to automate the analysis of text views or emotional information [2]. In the early years, sentiment analysis for critical texts are mainly studied in this domain. For example, Liu divided sentiment analysis into three classifications: document level, sentence level and factor level [3], Sentence level sentiment classification assumes that the emotion expressed by a sentence is consistent, and that the object of the emotion is only one (sentence level sentiment classification can not achieve fine-grained recognition). The comment text obviously does not satisfy the above hypothesis, such as hotel comment sentence "the location of the hotel is far away, but the environment and the service is very good."

K. Deng et al. (Eds.): ICMIR 2018, AISC 856, pp. 1230–1239, 2019.
https://doi.org/10.1007/978-3-030-00214-5_151

The sentiment of this sentence is positive, but the evaluation of the position is negative. The news text differs from the comment text that a piece of news usually only discusses an event, and the emotion expressed in the news is usually fixed, and there are few different emotions at the same time in a sentence. Based on the above considerations, it is reasonable to analyze the sentimental orientation of news texts. The purpose of this paper is to analyze the sentimental orientation of Chinese and Vietnamese news sentences, and find whether the sentence is positive or negative (two classification).

At present, in the research of sentiment classification, the following main algorithms are proposed, which are dictionary based method [4–6], machine learning based method [7, 8] and depth learning based method [9–13]. In the dictionary based method, we first need to construct an emotion dictionary and classify the sentences according to the emotion dictionary. For example, Mihalcea uses two types of bilingual dictionaries to construct subjective word dictionary [4] with bilingual alignment, and then constructs a rule-based sentiment classifier based on the bilingual subjective lexicon. The main idea of machine learning based method is to extract emotion features, and then classify them by classification algorithms. For example, Pang extracts the unigram of text as a feature, and uses SVM algorithm and Naive Bayes Algorithm to divide the text into two classifications: positive and negative [7]. In recent years, deep learning has become a hot topic in the field of artificial intelligence. Compared with traditional machine learning algorithms, deep learning has the main advantage that it does not require tedious feature extraction steps. In the early days, deep learning achieved better results in image processing. With further exploration, deep learning has been well applied in the field of Natural Language Processing. For example, Kim [11] proposed the convolutional neural network applied to sentence classification task, the experimental results show that this method achieves better effect compared with the traditional machine learning algorithms, and the convolutional neural network classifier does not need artificial feature extraction.

On the basis of previous studies, this paper proposes a sentiment classification method based on the convolution neural network for Chinese and Vietnamese Bilingual News sentence. This method uses the Chinese Vietnamese bilingual dictionary to realize cross language sentiment analysis, and on this basis, uses the convolution neural network to realize the sentiment classification of Chinese and Vietnamese news sentences. In addition, this paper compares the difference between the use of word ID as input (one-hot vector) and the use of word vector (word2vec) as input, and experiments show that using word vector as input has achieved good results.

2 Sentiment Classification Algorithm for Chinese and Vietnamese Bilingual News Sentences Based on CNN

In order to classify the Chinese and Vietnamese Bilingual News sentences, a kind of cross language sentiment classification algorithm based on convolutional neural network is studied in this paper, as shown in Fig. 1. First, we collect Chinese and Vietnamese news texts, use participle tools to segment Chinese and Vietnamese news texts, and use word2vec tools to train Chinese word vectors and Vietnamese word vectors. Then, according to the word vectors obtained by training and bilingual dictionaries, the

Chinese News sentence is represented as a bilingual vector matrix, which is used as the input of the convolution neural network model. Finally, the convolution neural network classification model is constructed, and the labeled Chinese sentences are used as training sets to train the model. The algorithm only needs to annotate Chinese News sentences, and uses the Chinese Vietnamese bilingual dictionary to realize the sentiment classification of Vietnamese news sentences.

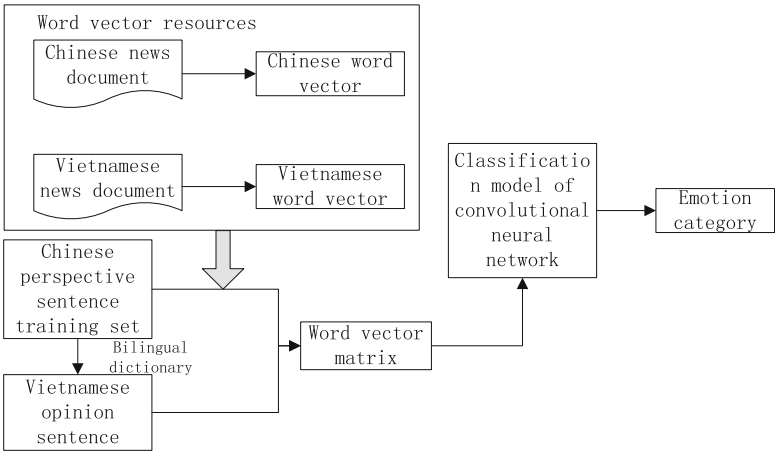

Fig. 1. The frame word of Chinese and Vietnamese news sentence sentiment classification based on convolution neural network

3 A Classification Model Based on Convolution Neural Network

In order to realize the sentiment classification of Chinese and Vietnamese Bilingual News sentences, this paper studies a Chinese and Vietnamese bilingual sentiment classification algorithm based on the convolution neural network. The algorithm model is shown in Fig. 2. The sentiment classification model based on convolution neural network consists of four layers, which are input layer, convolution layer, max-pooling layer and output layer. The functions of each layer are explained as follows.

First layer: the input layer. the main purpose of this layer is to represent the Chinese News sentence as the bilingual word vector that the computer can recognize and process. Due to the lack of Vietnamese corpus, different input strategies are used in model training and model testing. In the model training stage, according to the bilingual dictionary, the Chinese News sentence that has marked the sentiment classification (positive sign is 1, negative mark is 0) is mapped to Vietnamese, and the Vietnamese expression of Chinese News sentence is obtained. In the model testing phase, the Vietnamese unclassified news sentence is mapped to Chinese for processing. We use the above strategy to ensure that each language has two perspectives.

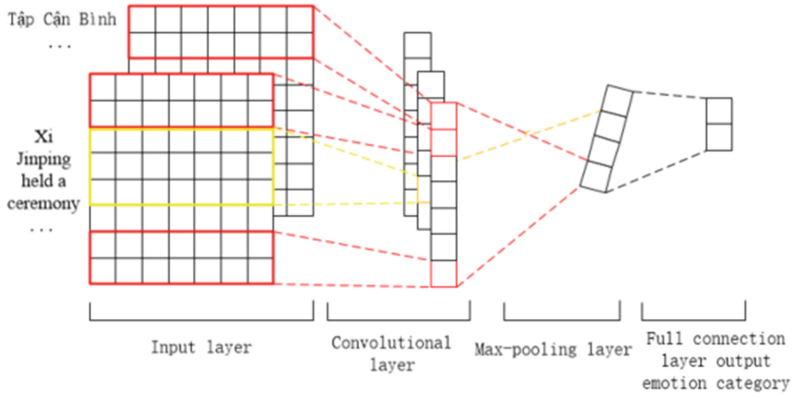

Fig. 2. Chinese and Vietnamese news sentence sentiment classification based on convolution neural network

The sentence "Xi Jinping held a ceremony to welcome the Vietnamese Communist Party General Secretary Ruan Fuzhong's visit to China " is shown as an example in the Fig. 2. First, the Chinese News sentence is mapped to Vietnamese using a bilingual dictionary. Then, we use Chinese segmentation tool and Vietnamese word segmentation tool to segment Chinese sentence and Vietnamese sentence. Finally, according to the pre trained word vectors, the sentences after segmentation are represented as matrices, and each row of the matrix represents the word vector of a word. Suppose that the sentence contains M words, the word vector length is n, and the size of the matrix is m * n.

Second layer: the convolution layer. The main job of this layer is to extract features from the word vector representation of sentences using convolution operations. $x_i \in R^N$ represent an n-dimensional word vector of the i-th word in Chinese sentences or Vietnamese sentences. And a sentence containing m words can be expressed as:

$$x_{1:m} = x_1 \oplus x_2 \oplus \ldots \oplus x_m \tag{1}$$

Wherein, $\oplus$ represents connection operations, such as (1,1,1,1) $\oplus$ (2,2,2,2) = (1,1, 1,1,2,2,2,2).Convolution operation refers to the sliding window with length h, from which the $x_{1:m}$ intercept vector is used as the input of the convolution layer activation function. After the activation function is calculated, the output is the feature to be extracted. The following is shown:

$$c_i = f(W \cdot x_{i:i+h-1} + b) \tag{2}$$

Function f represents activation function, and ReLU function is chosen as activation function in this paper. W and b represent the weight matrix and the offset unit, which are the parameters of the neural network, and can be trained by the neural network. C_i represents the characteristic calculated according to the activation function f and the input vector $x_{i:i+h-1}$.

The above formula describes the length of sliding window h at position $x_{i:i+h-1}$ on the operation, when the sliding window from the beginning of the sentence slide to the end, can obtain a set of input, denoted as ($x_{1:h}$, $x_{2:1+h}$,…, $x_{m-h+1:m}$), then we can get a group according to the input and output, said (C_1, C_2,…, C_{m-h+1}), a group of output is called feature mapping. It should be noted that the input layer of the model contains two channels, and each sliding window needs to glide through two channels when calculating the feature mapping.

Third layer: max-pooling layer, the main purpose of this layer is to process the feature mapping of the upper layer, and obtain the final feature vector. Specifically, the maximum value of the feature map is extracted, and the idea is that the maximum in each feature map represents the most important feature of the mapping. In addition, a major advantage of the max-pooling layer is that a fixed length feature vector can be obtained with the aid of this layer without requiring attention to the length of the original input sentence.

The fourth layer: the output layer. the main task of this layer is to calculate the emotion category based on the feature vector of the upper layer. This layer is the full connection layer, using the sigmoid function as the activation function, and the result is processed by softmax to obtain the probability. Let K dimension be characterized as X_k, then:

$$y_i = \frac{e^{(W^T \cdot X_k + b)}}{1 + e^{(W^T \cdot X_k + b)}} \tag{3}$$

Among them, W^T represents the transpose of vector W, and W and b represent the parameters of the fully connected layer network, which represents the output value of the output layer neuron. The output layer of the model contains only two neurons. In order to get the probability that the sentence is positive or negative (positive by 1 and negative by 0), further processing is needed, as shown in the following formula.

$$\hat{P}(y = 1 | s_{cn}, s_{vn}) = \frac{e^{y_1}}{\sum_{i=1}^{2} e^{y_i}} \tag{4}$$

$$\hat{P}(y = 0 | s_{cn}, s_{vn}) = \frac{e^{y_0}}{\sum_{i=1}^{2} e^{y_i}} \tag{5}$$

S_{cn}, S_{vn} expresses Chinese sentences and Vietnamese sentences respectively.

4 Experiment and Result Analysis

4.1 Experimental Design

Using the Chinese Vietnamese news sentence as the experimental data set, in order to test the proposed model, the sentiment polarity of a part of Chinese News sentence and

Vietnamese news sentence is marked artificially. For example, the sentence "Xi Jinping held a ceremony to welcome the Vietnamese Communist Party General Secretary Ruan Fuzhong's visit to China. " to the positive polarity, the label is 1, annotation data are shown in Table 1. In addition, the Chinese Vietnamese bilingual dictionary consists of 256752 word pairs.

Table 1. Information of experiment data

Language	Number	Average length	Emotional polarity (positive)	Emotional polarity (negative)
Chinese	3000	33	1430	1570
Vietnamese	3000	45	1320	1680

4.2 Evaluation of Method

The accuracy, recall rate and F value were used to evaluate the experiment. Let PR denote positive prediction for positive numbers, PN expressed negative positive predictive data, MR said negative predictive negative number, MN expressed negative predictive positive number, while the accuracy rate, recall rate and F-measure calculation formula is as follows:

$$precise = \frac{PR}{PR + MN} \tag{6}$$

$$recall = \frac{PR}{PR + PN} \tag{7}$$

$$F_1 = 2 \cdot \frac{precise \cdot recall}{precise + recall} \tag{8}$$

4.3 Experiment Results and Analysis

In order to obtain the word vector in Chinese and Vietnamese, first use crawlers collected a large number of Chinese and Vietnamese unlabeled corpus, including about 1000000 Chinese sentences and 500000 Vietnamese sentences, from Chinese and Vietnamese major news sites. After word segmentation, gensim is used to train the word vector, and the word vector dimension is set to 100. In order to evaluate the effect of word vectors, the first five most similar words are selected for testing, and the results are shown in Table 2. It can be observed from the table that the training of Chinese word vector and Vietnamese word vector has good effect on word similarity calculation. In the experiment, it is necessary to pay attention to the Chinese News annotation corpus used in the training phase of the model. At the model testing stage, we used Vietnamese unlabeled corpus.

In order to compare the effects of different word vectors on the results, the one-hot word vectors and the word2vector word vectors are used as input respectively, and the

Table 2. Chinese word vector and Vietnamese word vector result

Chinese words	The 5 most similar words	Vietnamese words	The 5 most similar words
Jinping Xi	Ministry of Foreign Affairs, China, Yi Wang, Keqiang Li, Summit	NguyễnPhúTrọng (Zhongfu Yuan)	NguyễnChíVịnh (Nguyen Chi Vinh), PhùngQuangThanh (PhungQuangThanh), VIệt (Vietnam), BiểnĐông (east sea), PhạmBình Minh (Pham Minh Minh)
China	Nation, America, foreign, country, inland	Việt Nam (Vietnam)	Việt (Vietnam), HàNội (Hanoi), Mỹ (Inch), HồChí Minh (Ho Chi Ming), LiênHiệpQuốc (United Nation)
economy	Build, militar, market, medical care, investment	kinhtế (economy)	văn hóa (civilization), quân đội (military), cải cách (reform), buôn bán (business), thị trường (market)

word vectors are shown in Fig. 3. Since the parameters of the neural network are numerous, it is necessary to set these parameters according to the task before training the convolution neural network model. The parameters of this paper are set as follows.

$$onehot(经济) = [0,0,0,0,0,0,0,0,0,0,...,0,0,0,1,000,0,...,0,0,0,0,0,0,0,0,0,0,0,0]_{1\times|D|}$$
$$word2vector(经济) = [\ 1.67981553,\text{-}0.82601517,...,1.64033353,...,\text{-}0.15920518\ ]_{1\times100}$$

Fig. 3. The comparison of one-hot word vector and word2vector word vector

Sliding window: the sliding window is generally set to 1–5. If the parameter is too small, the input of the model will lack the semantic information. Or if the parameter is too large, the model will be too complex and difficult to train. Because the text is processed by word segmentation in the input layer, using word vectors instead of character vectors and the unlabeled corpus is widely used in the process of training vector in the word, the unlabeled corpora contain semantic information and reflect semantic information into the word vector. Therefore, the sliding window is set to 1, 2, and 3, and different sliding windows are configured with 100 different feature mappings (corresponding to different neurons in the neural network).

Mini-batch size: the traditional gradient descent algorithm is used to train the model.

Each iteration of the parameter needs to observe all the training samples, and the overhead is great. Using Mini-batch strategy for training, training process faster, better results. According to the size of the training set, this article sets mini-batch size to 30.

Loss function: cross entropy loss function.

Activation function: in this paper, the ReLU function is used as the activation function in the convolution layer, and the output layer uses the sigmoid function as the activation function.

The Dropout: Dropout parameter is used to prevent the overfitting of the model. The dropout layer is added after the input layer and the max-pooling layer, and the dropout value is set to 0.2.

Using the above parameters to train the model, in addition, to illustrate the effectiveness of the proposed method, this paper compares it with the dictionary based approach [4] and the support vector machine [7]. Method for word vector one-hot is marked M1 and method for word2vector word vector is marked as M2 in the input layer, The dictionary based method is marked as C1, and the SVM based method is marked as C2 (select unigram as feature). To achieve the above method, half off cross validation is carried out to obtain the accuracy, recall and F values of different methods, as shown in the following Figs. 4, 5, and 6.

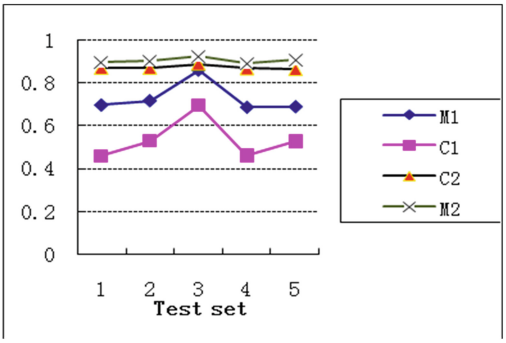

Fig. 4. Precision

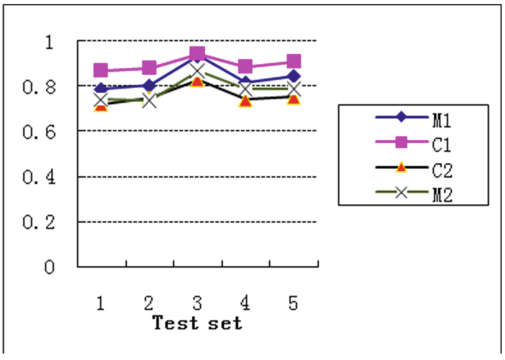

Fig. 5. Recall

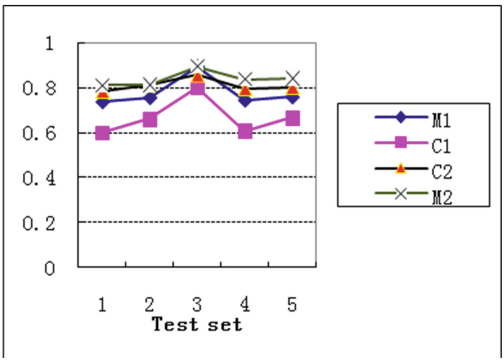

Fig. 6. F score

As can be seen from the diagram, using the pre trained word2vector word vector as input, the model achieves better results than using the one-hot word vector as input. This is because the pre - trained word vectors contain large amounts of semantic information in unlabeled corpora, whereas one-hot, a high-dimensional sparse word vector, does not have this function. In addition, the method of using word2vector word vector is better than dictionary based method on the whole. Although the dictionary based method achieves higher recall rate, the single rate of accuracy is poor. It is proved that the method presented in this paper not only has the function of cross linguistic analysis, but also achieves good results in various evaluation indexes.

5 Conclusions

In order to achieve Chinese and Vietnamese Bilingual News sentence sentiment classification, this paper proposes a cross language classification algorithm based on convolution neural network. The method uses a bilingual dictionary as a cross language bridge, and uses the word vector tool to map the natural language text into the vector space as the input layer of the convolution neural network. First of all, in order to get better word vectors, a large number of Chinese and Vietnamese news corpora are collected as word vectors for training corpus. Secondly, the Chinese Vietnamese bilingual dictionary is used to map Chinese into Vietnamese, and get the two perspectives of the text. It is trained as the two channel of the neural network. With the help of bilingual dictionaries, the model can not only classify the sentiment in Chinese, but also classify the sentiment in Vietnamese. The experimental results show that the model has achieved good results.

Acknowledgements. This work was supported by National Nature Science Foundation (Grant Nos. 61472168, 61672271, 61732005), and Science and Technology Innovation Talents Fund Project of Ministry of Science and Technology (Grant No. 2014HE001),Innovation Team Project of Yunnan Province (Grant No. 2014HC012).

References

1. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: International Conference on World Wide Web, pp. 519–528. ACM (2003)
2. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retrieval **2** (1–2), 1–135 (2008)
3. Liu, B.: Sentiment analysis: Mining opinions, sentiments, and emotions. Comput. Linguist. **42**(3), 1–4 (2016)
4. Mihalcea, R., Banea, C., Wiebe, J.M.: Learning multilingual subjective language via cross-lingual projections. In: Meeting of the Association of Computational Linguistics (2007)
5. Maks, I., Vossen, P.: A lexicon model for deep sentiment analysis and opinion mining applications. Decis. Support Syst. **53**(4), 680–688 (2012)
6. Taboada, M., Brooke, J., Tofiloski, M., et al.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of Emnlp, pp. 79–86 (2002)
8. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25–26 July 2004, Barcelona, Spain, pp. 412–418. DBLP (2004)
9. Poria, S., Cambria, E., Gelbukh, A.: Aspect extraction for opinion mining with a deep convolutional neural network. Knowl. Based Syst. **108**, 42–49 (2016)
10. Gao, Y., Rong, W., Shen, Y., et al.: Convolutional Neural Network based sentiment analysis using Adaboost combination. In: International Joint Conference on Neural Networks, pp. 1333–1338 (2016)
11. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751 (2014)
12. Santos, C.N.D., Gattit, M.: Deep convolutional neural networks for sentiment analysis of short texts. In: International Conference on Computational Linguistics (2014)
13. Severyn, A., Moschitti, A.: Twitter sentiment analysis with deep convolutional neural networks. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 959–962. ACM (2015)