Combining Concept Graph with Improved Neural Networks for Chinese Short Text Classification

Jialu Liao[1,2], Fanke Sun[1,2], and Jinguang Gu[1,2(✉)]

[1] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China
simon@wust.edu.cn

[2] Hubei Province Key Laboratory of Intelligent Information Processing and Real-Time Industrial System, Wuhan 430065, China

Abstract. With the development of the Internet, network information is booming, and a large amount of short text data has brought more timely and comprehensive information to people. How to find the required information quickly and accurately from these pieces of information is the focus of the industry. Short text processing is one of the key technologies. Because of the sparse and noisy features of short texts, the traditional classification method can not provide good support. At present, the research on short text classification mainly focuses on two aspects: feature processing and classification algorithm. Most feature processing methods only use text literal information when performing feature expansion, which lacks the ability to discriminate the polysemy that is common in Chinese. In the classification algorithm, there are also problems such as insufficient input characteristics and insufficient classification effect. In order to improve the accuracy of Chinese short text classification, this paper proposes a method of Chinese short text classification based on improved convolutional recurrent neural network and concept graph, which achieves better classification results than existing algorithms.

Keywords: Chinese short text classification · Concept graph · Feature processing · Deep learning

1 Introduction

With the coming of mobile Internet era, short text has gradually become the most common form of text, which we call short text. Short texts are mainly produced in scenarios such as bullet-screen interaction, commodity evaluation and news headlines [1]. Enterprises need to understand users' real needs based on these information, so as to launch better services and products. The relevant government departments can also monitor public opinion based on these data. In order to eliminate useless, interference and even error information, we need to correctly use the increasingly mature technology to understand and analyze

X. Wang et al. (Eds.): JIST 2019, CCIS 1157, pp. 205–212, 2020.
https://doi.org/10.1007/978-981-15-3412-6_20

these massive data, and obtain valuable information accurately and efficiently. Short text has the characteristics of short text length, large number and large amount of interference information [2], which makes it impossible to provide sufficient context semantics, resulting in the difficulty of extracting effective text features from traditional algorithms for processing long text [3]. Therefore, how to design an effective model to extract the semantic information of short text and improve the classification accuracy of short text is the main challenge. In this paper, based on the characteristics of short text, the feature processing method of conceptual knowledge of word acquisition through concept graph is designed. Combining the advantages of the most popular neural network model, multiple neural network models are combined to improve the feature extraction effect; And from the three levels of characters, words and concepts, a three-channel Chinese short text classification model is constructed to improve the accuracy of short text classification.

The rest of this paper is organized as follows: Sect. 2 will give a brief analysis of current feature processing schemes and short text classification methods; Sect. 3 will introduce the three-channel Chinese text classification model designed in this paper, including how to combine the advantages of existing neural network models to generate an efficient feature extraction network; Sect. 4 will introduce the data set used in the experiment, compare the proposed algorithm with the existing algorithm using short Chinese text, and analyze the experimental results. The fifth section summarizes the work of the full text, and puts forward the prospect of the future research direction of short text feature processing methods and classification methods based on in-depth learning.

2 Research Progress of Text Classification

In the research of classification algorithms, with the deep learning method being applied in large scale in the field of natural language processing, the neural network method can learn from the text to the internal semantic features with the help of big data, and can complete the short text classification task better. Among them, Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) are the two most popular models in the current short text classification study. For CNN model, Kim [4] combined word vector with convolution neural network at first, and achieved good results in short text sentiment analysis. For RNN model, LSTM is the focus of research. Liang et al. [5] designed a tree structure LSTM with emotional polarity transfer sentiment analysis model.

In recent years, with the continuous development of natural language processing, some new methods have been proposed. Attention mechanism [6], proposed by Google, holds that the results of the model are determined by several or even one key feature, which gives researchers new inspiration. Some researchers apply attention mechanism to short text classification. Er et al. [7] proposed a convolution neural network based on attention pooling, constructed an intermediate vector representation of input text by using parallel LSTM, which was used as

the attention weight of the document eigenvector generated by convolution neural network. Finally, the processed text eigenvector was input into the classifier for classification.

Among the feature processing methods based on conceptual graph we know, they are all for English corpus research. How to use knowledge graph to process Chinese short text features, and how to use neural network knowledge to construct an efficient feature extraction model to obtain more semantic knowledge to improve the classification effect of Chinese short text is still a subject that researchers need to continue to explore.

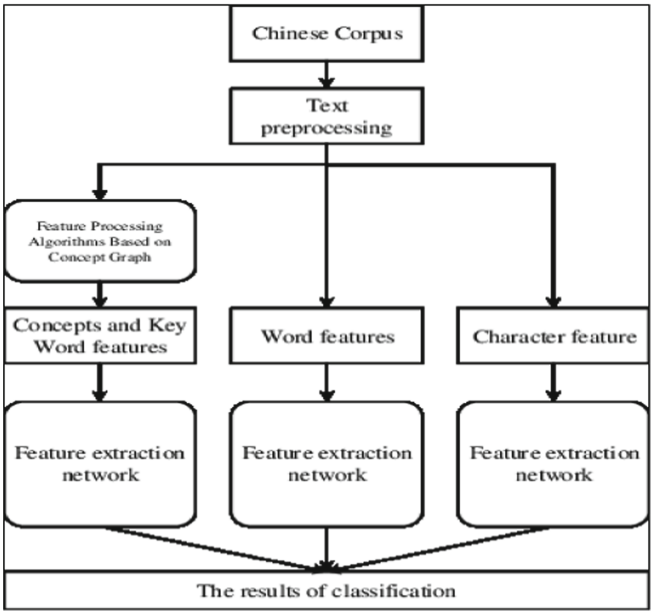

Fig. 1. Overall workflow diagram

3 Constructing Chinese Text Classification Model

3.1 The Process of Chinese Text Classification

The workflow of Chinese text classification is shown in Fig. 1. The research contents are as follows:

1. In order to solve the problem of single feature, this paper proposes a three-channel Chinese short text classification model, which extracts features from three levels: word feature, character feature and concept feature.

2. The Chinese short text classification model designed in this paper is an efficient feature extraction network structure, which combines the advantages of CNN model and Bi-LSTM model, and introduces the attention mechanism to enhance the attention of key words.
3. For the conceptual level features in the model, this paper uses Probase+, a conceptual graph system based on knowledge graph, to obtain candidate concepts of words, and constructs a semantic network from the perspective of the whole text by using concepts and original texts to analyze the correlation between them. In addition, with the help of feature iteration selection algorithm, keywords and concepts suitable for the current context are selected as the complement of short text features, which enriches the feature information of text.
4. The feature processing method and classification algorithm proposed in this paper are verified. By comparing with other popular algorithms and classification models, we can know that the feature processing algorithm and the Chinese short text classification model based on in-depth learning and concept graph can better accomplish the task of Chinese short text classification.

3.2 Three-Channel Chinese Short Text Classification Model

In natural language processing tasks, word vectors are generally used as vector representations of text, but this method has some limitations. If some professional words are not recognized well in the process of word segmentation, it will affect the accuracy of classification. Compared with thousands of words, the number of characters is much less than words, which is fast and efficient, and the accuracy does not decrease significantly when used alone. On the other hand, using character vectors as feature items can avoid the adverse effects when the segmentation effect is not good. Especially in short texts, the vocabulary is usually very small. If we can mine the relationship between words, the meaning of the text can be expressed more accurately. In this section, the character level signal [8] is used as a feature to form a text representation based on character vector features to assist the model to obtain more semantic information.

Whether the commonly used word features or character features mentioned above are processed on the basis of the original corpus [9], but it can not solve the problem of insufficient signals and many interference items provided by the short text itself. In order to solve this problem, it is a common practice to select appropriate feature extension methods to help understand the original semantics, and the method of introducing external knowledge graph is one of the solutions used in this paper. In this section, we will use the feature processing algorithm based on concept graph to get the conceptual set and keyword set of short text from the conceptual level. We hope that these two sets can also be added to text features as conceptual level signals to form a text representation based on conceptual vector features to solve the problem of "polysemy" that most models can't solve.

Based on the above two points, this paper hopes to combine the acquired character signals and conceptual signals to assist the word signals, so that the

classification model can get more semantic information. Based on this idea, this paper designs a three-channel short text classification model. By establishing three feature matrices of word vector, character vector and concept vector, they are input into the convolutional recurrent neural network model based on attention mechanism, and their respective features are extracted. Then, the cascade fusion is completed through the full connection layer. Finally, the classifier completes the classification. This multi-channel structure of neural network overcomes the shortcomings of insufficient signals and greater randomness in short text data, and can obtain more abundant short text semantic features. Through the method of multi-feature fusion, the information in text can be obtained more comprehensively, in order to make the model have stronger representation ability for text signals.

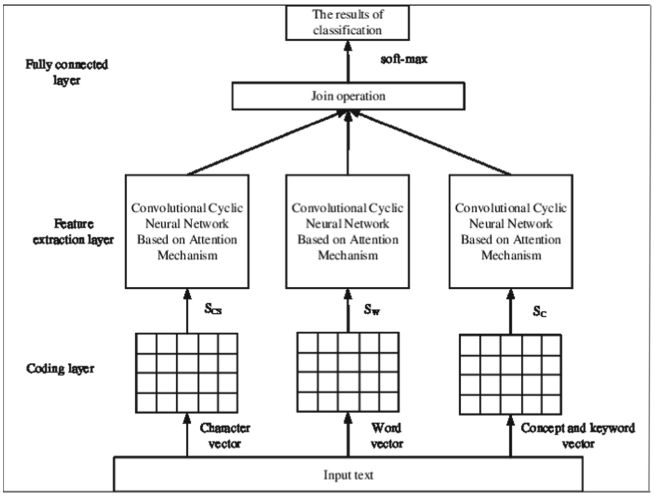

Fig. 2. Three-channel short text classification model

The overall model is shown in Fig. 2. The three-channel model contains three parallel network structures, and their characteristics are computed by parallel computing. The first network deals with character vectors of short text; the second network deals with word vectors; and the third network deals with keywords and concept vectors. The whole model is divided into three layers: coding layer, feature extraction layer and full connection layer. The input module layer and feature extraction layer have one in each channel, and the full connection layer is shared by three channels.

3.3 Feature Processing Based on Concept Graph

Because of the characteristics of short text itself, useful information is very limited, and even there will be many interference words unrelated to the main body

of the text, so it is necessary to obtain more semantic signals (e.g concepts) for feature expansion and ambiguity elimination. At the same time, we need to take these signals and the original text as a whole, find the relationship between them, so that they can interact with each other, so as to help the computer obtain the true semantics of the short text to the greatest extent.

This paper uses Microsoft concept graph System Probase+ [10], which can realize the conceptual operation of short text. At present, the core of most conceptual selection work is mapping short text words to candidate concepts in knowledge graph one by one to obtain conceptual words and their probabilities.

The flow chart of feature processing algorithm is divided into three stages: The first stage is feature extension based on Probase+ concept graph. The main work of this stage is to generate candidate concept sets by using words in short text, which can be obtained directly from concept graph system. The second stage is to construct a word-concept semantic network by using the content of word set and candidate concept set. The third stage is to use the constructed semantic network for feature selection calculation and to screen out the final keywords and concepts set. Through this algorithm, the number of effective features increases rapidly, which greatly enriches the feature set of short text.

4 Experiments and Assessment

4.1 Experimental Data Set

The short text data used in this experiment are from the Chinese news headline classification task in NLPCC 2017 [11] shared task. The corpus contains a total of 228,000 manually annotated Chinese news headlines, of which the training set contains 156,000 data, the verification set contains 36,000 data, and the remaining 36,000 are test sets. It is divided into 18 categories, such as sports, tourism, military, social and historical.

4.2 Data Preprocessing

Before the experiment, the data sets need to be processed as follows:

(1) Because the data set is a Chinese corpus, it is necessary to use Jieba tool to segment the whole data set to form a segmented word data set.
(2) We find that the result of deleting stop words in this data set is relatively poor, which means that deleting stop words in very short modeled text may damage the original syntactic structure or even the semantic representation. Therefore, this paper does not perform the operation of deleting stop words in this data set.
(3) Since the classification model designed in this paper contains three channels, the input data of each channel need to be processed.
(4) In order to make the computer understand the text, we need to train the skip-gram model of word2vec tool to get 200-dimensional word vector.

Because of the high dimension of word vector representation, it is necessary to index the characters or words in the data set, replace the words with real numbers, and even the whole text. This method can greatly increase the speed of computer text processing.

4.3 Experimental Analysis

In this experiment, the single-channel classification model with three granularities of characters, words and concepts, the two-channel classification model of characters and words, and the three-channel classification model based on word vector data and concept vector data are used respectively. This experiment is mainly evaluated by Accuracy index.

Table 1. Comparison results of different granularity data

Model	Character vector	Word vector	Concept vector	Character+Word two-channel	Three-channels
CNN	74.02%	76.61%	64.53%	78.08%	79.75%
LSTM	73.56%	75.57%	60.34%	76.60%	79.01%
A-BiLSTM	76.52%	77.63%	65.32%	79.96%	80.34%
CRAN	73.89%	77.01%	66.21%	78.39%	79.83%
C-LSTM	76.83%	78.52%	68.63%	79.78%	81.01%
The model in this paper	78.01%	79.34%	68.01%	80.87%	82.24%

The three-channel method of words, characters and concepts designed in this paper combines the features of word vectors, word vectors and concept vectors, and further overcomes the shortcomings of inadequate signal and large randomness in short text data. By means of multi-feature fusion, more abundant feature information can be obtained, so that the classification model can have enough information for feature extraction in training. From Table 1, it can be seen that the classification model designed in this paper can achieve the best classification results under the conditions of single channel or multi-channel. Compared with the single channel classification method, our method has improved by about 3%, and compared with the double channel classification method of words and characters, the correct rate has increased by 1%–2%, Compared with the mixed model in reference [11], the accuracy is improved by 0.5%. Which proves that the three channel classification method designed in this paper is more effective.

5 Conclusion

With the development of mobile Internet, short text analysis and research has become a hot issue. As a key technology of natural language processing, short

text classification has great research value. Traditional text classification methods depend on a large amount of text information, so it is impossible to obtain semantic information by statistical methods. Therefore, this paper mainly studies and improves short text classification from two aspects: feature processing and classification algorithm. Experiments verify the effectiveness of the proposed method.

Acknowledgment. This work was partially supported by a grant from the NSF (Natural Science Foundation) of China under grant number 61673304 and U1836118, the Key Projects of National Social Science Foundation of China under grant number 11&ZD189.

References

1. Zhou, T., Chen, M., Yu, J., Terzopoulos, D.: Attention-based natural language person retrieval. In: Computer Vision and Pattern Recognition Workshops (2017)
2. Yu, B., Zhang, L., School of Management: Chinese short text classification based on CP-CNN. Appl. Res. Comput. **35**, 1001–1004 (2018)
3. Zhang, D., Xu, H., Su, Z., Xu, Y.: Chinese comments sentiment classification based on word2vec and SVM^{perf}. Expert Syst. Appl. **42**(4), 1857–1863 (2014)
4. Kim, Y.: Convolutional neural networks for sentence classification. Eprint Arxiv (2014)
5. Li, S., Yan, Z., Wu, X., Li, A., Zhou, B.: A method of emotional analysis of movie based on convolution neural network and bi-directional LSTM RNN. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 156–161 (2017)
6. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: 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, pp. 1480–1489 (2016)
7. Zhang, Y., Er, M.J., Wang, N., Pratama, M.: Attention pooling-based convolutional neural network for sentence modelling. Inf. Sci. Int. J. **373**(C), 388–403 (2016)
8. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
9. Wang, J., Wang, Z., Zhang, D., Yan, J.: Combining knowledge with deep convolutional neural networks for short text classification. In: Twenty-Sixth International Joint Conference on Artificial Intelligence (2017)
10. Liang, J., Xiao, Y., Wang, H., Zhang, Y., Wang, W.: Probase+: inferring missing links in conceptual taxonomies. IEEE Trans. Knowl. Data Eng. **29**(6), 1281–1295 (2017)
11. Lu, Z., Liu, W., Zhou, Y., Hu, X., Wang, B.: An effective approach for Chinese news headline classification based on multi-representation mixed model with attention and ensemble learning. In: Huang, X., Jiang, J., Zhao, D., Feng, Y., Hong, Y. (eds.) NLPCC 2017. LNCS (LNAI), vol. 10619, pp. 339–350. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73618-1_29