Automatic Classification of Long Mongolian Text Based on Word Embedding Distributed Representation Deep Learning Model

Gang JIN
School of Computer Science and Information Engineering in Huhhot Minzu College, Huhhot, 010000, China

Abstract

With the development of information technology of Mongolian language and the publication of international standard of Mongolian language coding, Mongolian electronic texts have sprung up in recent years. The research of automatic classification technology in the field of Mongolian information processing has received much concern compared to the time-consuming effort of manual processing and classification. The current research on long Mongolian text automatic classification mainly relies on traditional algorithms such as Naive Bayes and Support Vector Machine, as well as sparse feature representation methods based on word frequency like TF-IDF or One-Hot. However, in terms of the characteristics of Mongolian language itself, Mongolian sentences in news texts are usually long, and there are some problems such as difficulty in extracting semantic features and poor classification effect by using traditional algorithms. In this paper, the combination model of CNN and RNN (CNN+RNN Model) which is characterized by word embedding distributed expression is proposed to carry out the experiment of automatic classification of long Mongolian text which takes news text as data. Experimental results show that compared with CNN model or RNN model used alone, the combination model is superior to the model used alone in various indexes such as precision rate, loss rate, accuracy rate, recall rate and F1 value, and can effectively process longer Mongolian text.

Keywords: *Mongolian Text; Automatic Classification; CNN+RNN Model*

I. Introduction

Text classification is a very basic and important technology in natural language processing technology, which can be applied to emotion analysis, topic classification, dialogue behavior classification, event prediction and many other application systems. Text classification is the process of assigning one or more predefined labels to a given text through a classifier. Among it, classifiers are all kinds of algorithms designed to automatically classify texts by computers. These algorithms are usually Machine Learning algorithms with sample features as input, and the process of obtaining sample features is particularly important in this process. However, text features are different from images, audio and other data, and various linguistic features such as words, grammar, vocabulary and semantics are needed. Especially in the modern information explosion era, it is a difficult and challenging task to extract the corresponding features from massive electronic texts.

Before the emergency of Deep Learning algorithm, the research of text classification mainly relies on traditional algorithms such as Naive Bayes and Support Vector Machine. In addition to some linguistic features designed manually, these algorithms mainly use sparse feature representation methods based on word frequency, such as TF-IDF or One-Hot. However, these feature representation methods usually ignore the sequence structure or contextual semantic information of text data, especially for Mongolian. For example, Mongolian sentences in news and other texts are usually long, so it is difficult to extract semantic features and the classification effect is poor by using traditional algorithms. Compared with the traditional algorithms mentioned above, the deep learning algorithm can effectively and automatically acquire sample features, avoiding manual design of rules or features. This has made a great contribution to reducing the complexity of feature engineering, but the problem of long-distance semantic dependence is still a thorny problem in the process of feature acquisition of deep learning algorithms.

In view of this, this paper puts forward a combination model of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which is characterized by word embedding and distributed representation, and implements an automatic classification experiment of long Mongolian texts with news texts as data, aiming at capturing the features of long sentences in texts by using convolutional neural networks, and capturing semantic dependencies by using recurrent neural networks as the main classifier.

II. Related research

2.1 Mongolian text classification research

The research of Mongolian text classification is still in the initial stage, and the traditional algorithms such as Naive Bayes and Support Vector Machine are mainly used in the current research of Mongolian text classification. For example, He Hui (2006) applied a method of active learning with support vector machine to the automatic classification of Mongolian texts, which solved the problem caused by excessive acquisition cost of Mongolian training samples[1]. Su Dao (2007) used TF-IDF and other four feature representation methods on the basis of extracting Mongolian word stems as features, and adopted K-nearest neighbor and support vector machine algorithm. The experimental results showed that the recall rate reached 87.5% and the accuracy rate reached 63.6%. [2]. In addition to the above research, Ayana (2009) and Gong Zheng et al. (2011) studied the methods of extracting Mongolian stop words by Entropy and Union Entropy in the fields of text classification and information retrieval [3-4].

2.2 Research on text classification based on neural network

There are mainly two kinds of neural networks available in the field of natural language processing, namely feedforward networks and recurrent networks. Yoon Kim et al. (2014) applied convolution neural network, a special structure of feedforward network, to text classification, and achieved remarkable results [5]. The CNN model proposed by Kim et al. is a shallow convolutional neural network for text classification, while Conneau et al. (2017) proposed a character-level deep convolutional neural network (VDCNN)[6]. Compared with the shallow convolutional neural network, the classification effect of deep convolutional networks has been significantly improved. Recurrent neural networks can express sequences of arbitrary length as vectors of fixed length, and pay attention to the structural attributes of inputs. Long Short-Term Memory (LSTM) is a variant of cyclic neural network. It was proposed by Schmidhuber in 1997. Its structure has three gate functions: forget gate, input gate and output gate. Yoshua Bengio et al. (2014) pointed out that LSTM can save long time series information and plays an important role in mining deep semantics [7]. After LSTM, Cho et al. (2014) proposed another RNN model, Gated Recurrent Unit (GRU), which has fewer parameters than LSTM and performs better in terms of convergence [8]. A one-way RNN can only use information in one direction, while a two-way RNN can use information in two directions at the same time in the past and the future, so that the final classification result is more accurate. In this regard, Schuster et al. (1997) proposed a bidirectional recurrent neural network. The model contains two hidden layers in different directions, which can fully learn the context information of the text [9]. In addition, the combination of various neural networks will also produce effective models. For example, Xiao and Cho (2016) applied the combination model of convolution network and recurrent network to the task of text classification, and achieved good results [10]. Adji, B. D., et al. (2017) proposed a TopicRNN model combining latent topic model and recurrent neural network, aiming at capturing local (grammatical) dependencies and global (semantic) dependencies by using RNN [11]. Felbo et al.(2017) proposed a two-layer model combined Bi-LSTM with Attention, which has a good application in the effect of emoticons on text emotion classification [12].

III. Classification model

To get the features of long text better, this paper uses a combination model of convolution network and recurrent

neural network characterized by word embedding. The model structure is as follows (Figure 1):

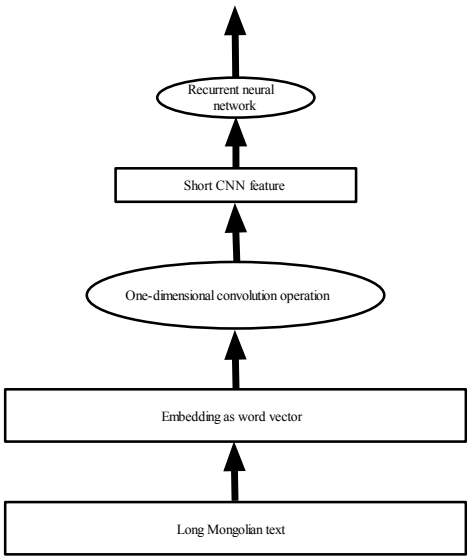

Figure 1: Structure of the model used in this paper

As shown in Figure 1, in this paper, every word in Mongolian news text is embedded into a fixed-length vector by word2vec model, and each text is represented as a word vector matrix. On this basis, a short CNN feature sequence is obtained by using one-dimensional convolution operation. Finally, the CNN feature sequence is transmitted to the recurrent neural network, and the final category probability is obtained.

For the one-dimensional convolution operation used in the task of text classification, it refers to that N vectors are stacked on top of each other to obtain an n×d sentence matrix, and then l different k×d matrices are used to slide on the sentences. The corresponding elements are multiplied with the corresponding sentence matrix segments to perform the matrix convolution operation (Figure 2). After convolution operation, the most significant information in the whole window position is obtained by pooling operation. The most common pooling operation is max-pooling, which extracts the maximum value from each dimension (Figure 3).

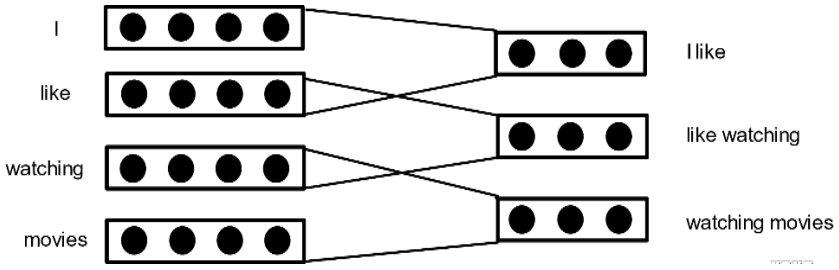

Figure 2: convolution operation with size k=2 and output 3-dimensional (l=3)

Convolutional neural network mainly focuses on the local features of text, while recurrent neural network can express sequences of arbitrary length as fixed-length vectors and pay attention to the structural attributes of input, which is very effective in obtaining the statistical laws of linear input. Therefore, this paper extracts local features from long Mongolian news texts by convolution operation, transforms them into short sequences, and then inputs the short sequences to the recurrent neural network. That is, the maximum pooled vector is used as the input of the recurrent neural network.

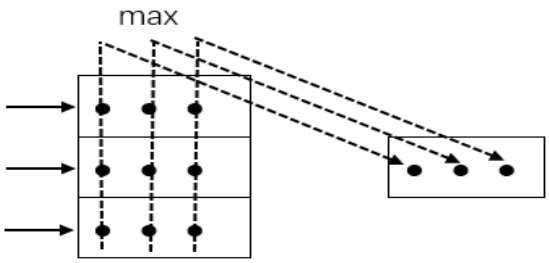

Figure 3: Maximum pooling processing

Recurrent neural network can be regarded as a function that returns a single d_{out}-dimensional vector y_n with ordered n d_{in}-dimensional vector sequences as inputs, which can be expressed as $y_n=RNN(x_{1:n})$. However, the main feature of the recurrent neural network, which is different from other feedforward networks, is that the current state contains the information of the whole past sequence. As shown in Figure 4, the current output h_t depends on the state S_{t-1} of the previous output h_{t-1}, which can be expressed as $h_t=f(h_{t-1}W_h+x_tW_x+b)$, where W is the weight and b is the bias.

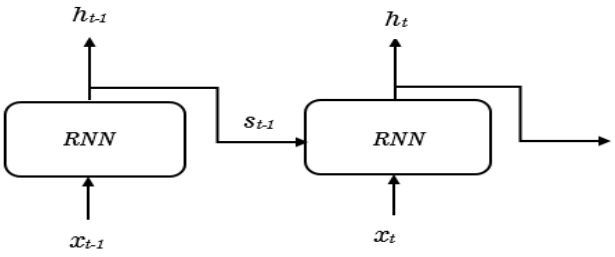

Figure 4: Recurrent neural network

IV. Data and experiments

4.1 Data

In this paper, 4287 texts were automatically collected from the Mongolian news website (http://www.mgyxw.net), which were manually identified as 13 topics including conference, economy, education and entertainment based on the original classification of the website. In the preprocessing stage, punctuation marks and characters other than Mongolian were deleted, and then these 13 types of texts were divided into training set, validation set and test set in a ratio of 8:1:1. And the text was padding to a fixed length of 800 words as the input data of this experiment.

4.2 Experiment

The experiment was verified by multiple adjustments to obtain the optimal model of structure and main parameters (see Table 1) with 6 layers depth. After the 800-word text was input into the input layer of the model, every word in the text was transformed into a 100-dimensional vector (see Figure 5)[13] through the word vector pre-trained by the skip-gram model of word2vec tool in the second layer, i.e., the embedded layer, and an 800×100 text matrix was obtained. After that, one-dimensional convolution operation and maximum pooling operation were implemented in the third and fourth layers.

Table 1: Main parameters of CNN+RNN (GRU) model

Layer	Output Shape	Param #
Input Layer	(None, 800)	0
Embedding	(None, 800, 100)	16897500

Conv1D	(None, 798, 512)	154112
Max-Pooling	(None, 159, 512)	0
GRU	(None, 128)	246144
Dense	(None, 13)	1677
Total params: 17,299,433		
Trainable params: 401,933		
Non-trainable params: 16,897,500		

0.013849 -0.921053 0.143123 -0.241046 0.246790 0.090617 0.311626 0.283276
0.255109 -0.943476 0.087172 -0.583919 0.403438 -0.279123 0.124959 -0.23621
0.405919 -0.490708 0.693280 -0.427098 0.829895 0.635154 -0.212724 0.337965 0
0.324504 0.305134 0.385833 -0.654688 0.073390 -0.249758 -0.606530 0.229941
-0.680939 -0.596396 0.479828 -0.616804 0.330790 -1.024642 -0.033150 -0.396
0.108147 -1.014338 0.263324 0.180038 0.060821 -0.169670 -0.158272 0.111374
0.280282 -0.728655 0.393279 -0.362324 0.498548 -0.229354 -0.671014 -0.295508
-0.047050 -1.163425 -0.007539 -0.682207 0.161660 -0.218782 -0.408049 -0.10

Figure 5: 100-dimensional Mongolian word vector

In the convolution operation, the sliding window size was set to 3, and the output dimension was set to 512. After the maximum pooling, the data with length of 159 and output dimension of 512 was obtained and then input to the fifth-layer recurrent network. In this paper, GRU (Gated Recurrent Unit), a simplified version of the long short-term memory model, was selected as the recurrent network, and 256-dimensional vectors were output, among which dropout and recurrent dropout were used to optimize the network. Finally, the 256-dimensional vector was calculated and the probabilities of 13 categories were output by softmax function. In the process of compiling the above model, categorical cross entropy was used as the loss function, and optimized with Adam optimizer.

The experimental comparison shows that the classification accuracy, loss rate, precision rate, recall rate and F1 value of the combination model proposed in this paper are superior to using convolution network or recurrent network alone (Table 2, Figure 6-7).

Table 2: Comparison of CNN, RNN(GRU) and CNN+RNN(GRU)

Model	Accuracy rate	Loss rate	Accuracy rate	Recall rate	F1
CNN	90.65%	0.380	0.91	0.91	0.90
RNN (GRU)	90.65%	0.307	0.91	0.91	0.91
CNN+RNN (GRU)	91.59%	0.274	0.92	0.92	0.92

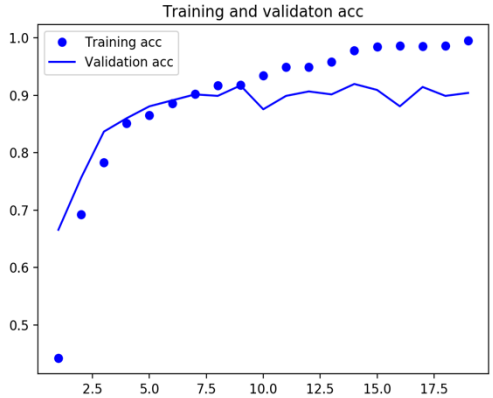

Figure 6: Training and validation acc. The abscissa is Epochs, the ordinate is accuracy.

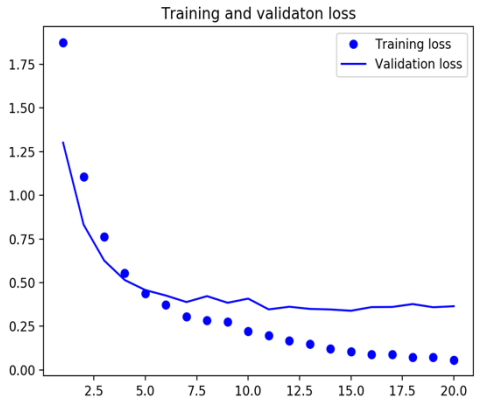

Figure 7: Training and validation loss. The abscissa is Epochs, the ordinate is loss rate.

V. Summary and outlook

In this paper, a combination model of Convolutional Neural Network (CNN) and recurrent neural Network (RNN) is proposed for the classification of Mongolian long texts. Experimental results show that the combination model proposed in this paper surpasses the single model in accuracy, loss rate, precision rate, recall rate and F1 value and effectively solves the classification problem of dealing with long Mongolian texts. What needs further improvement in the follow-up research is that only one kind of corpus was used as experimental data in this experiment. So in the future research, it is necessary to expand the amount and types of data, consolidate empirical research, and provide powerful data support for further in-depth study of text representation methods, feature extraction methods and various neural network variants.

Acknowledgement

This research is part of Doctoral Fund Project of Hohhot Minzu College (HM-BS-202003).

References

[1] He hui, Wang Junyi (2006) research on active support vector machine and its application in Mongolian text classification, journal of inner Mongolia university (natural science edition), pp. 560-563
[2] Su Dao (2007) Research on Mongolian Text Classification Technology and System Design and Implementation, Master's Thesis, Inner Mongolia University
[3] Ayana (2009) The Influence of Mongolian Discontinued Vocabulary and Stem Extraction on Mongolian Text Classification, Master's Thesis, Inner Mongolia University
[4] Gong Zheng, Guan Gaowa (2011) A Comparative Study of Mongolian stop words and English stop words,Journal of Chinese Information Processing , pp. 35-38
[5] Kim,Y. (2014) Convolutional Neural Networks for Sentence Classification. *EMNLP*, pp. 1761-1751
[6] Conneau, A. et al (2017) Very Deep Convolutional Networks for Text Classification, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.
[7] Good Fellow, I., Bengio, Y. and A. Courville (2017) Deep Learning, People's Post and Telecommunications Press.
[8] Cho et al. (2014) On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, arXiv: 1409.1259v2.
[9] Schuster, M. and Paliwal, K. K. (1997) Bidirectional Recurrent Neural Networks. IEEE Transactions on Signal Processing, 45(11) : 2673-2681.

[10] Xiao Yijun and Cho, K. Efficient Character-Level Document Classification by Combining Convolution and Recurrent Layers. *CoRR*, abs/1602.00367.
[11] Adji, B. D., Chong W., Jianfeng, G., and J. Paisley. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. ICLR, 2017.
[12] Felbo, B. et al. (2017) Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. arXiv: 1708.00524.
[13] Jin Gang. Mongolian word vector training based on Word2vec [J], Journal of Inner Mongolia University, 2018, (5):13-18.