Research on Single Document Automatic Summarization Method Based on Hybrid Neural Network

Qiaohong Chen[1, a, *], Wen Dong[1, b], Qi Sun[1, c], Yubo Jia[1, d]

[1]*School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China*

[a] *chen_lisa@zstu.edu.cn*, [b] *dongw410@163.com*, [c] *sunqi@vip.sina.com*, [d] *jiayubo1964@163.com*

**corresponding author*

Keywords: hybrid neural network, automatic summarization, Convolutional Neural Network, Long Short-Term Memory Neural Network, deep learning.

Abstract: In order to extract the required information from the massive information, an automatic summarization method based on hybrid neural network is proposed to help people to browse and understand the document quickly, and improve the efficiency of automatic summarization. The method combines the Convolutional Neural Network that has high efficiency and small over-fitting phenomenon in the training process and the Long Short-Term Memory Neural Network model with good effect on sequence prediction. The model takes full account of the characteristics words, characteristic sentences, feature segment positions and other factors, and add a signal to the input of the Long Short-Term Memory. Experimental results show that compared with automatic summarization methods based on LSI model, LDA model, TextRank summarization algorithm, PCA summarization algorithm and Long Short-Term Memory Neural Network model, the proposed method based on hybrid neural network has a good effect on automatic summarization, and improves the quality of automatic summarization effectively.

1. Introduction

With the rapid growth of various kinds of information on the Internet, how to capture the key information on the Internet quickly and effectively has become an urgent problem to be solved. In order to grasp and understand a large amount of information more conveniently and directly, automatic summarization can compress long articles and extract texts that can represent the core contents of the original text. Obviously, the single document automatic summarization is for a single document, and the content of the text is extracted to form a compressed representation for presentation to the user. There are three main types of common document summary techniques [1], which are feature-based, lexical-based, and graph-based sorting methods. In recent years, the maturity of automatic summarization technology will provide great convenience for Internet users [2], and greatly saves the reading time.

Automatic summarization is mainly divided into four steps: text pretreatment, sentence similarity calculation, sentence weight calculation, and abstract sentence extraction. Hu B T et al. [3] proposed an automatic summarization method based on Recurrent Neural Network and applied it to the automatic generation of short texts, which achieved good results. However, this method is only

suitable for the generation of short text automatic digests, and it is not suitable for long-length documents. Nichols et al. [4] used the improved TF-IDF to characterize the similarity between sentences. This method clusters sentences by similarity to obtain multiple clusters containing several sentences, and then extracts from these clusters according to a certain proportion. The sentence forms a summary, and the microblog news corpus verifies that the method produces a summary that is more readable and consistent than the previous algorithm, but there is a problem that the generated summary contains bias. Blei DM et al. [5] proposed an LDA model, which can give the theme of each document in the document set as a probability distribution. By analysing some documents and extracting their themes, according to these themes the topic clustering can be performed. However, this method is more suitable for multi-document automatic summarization with common themes. Rush et al. [6] proposed a completely data-driven approach to the sentence summarization, the method uses a local attention model to study the input sentences to generate a final summarization of each word. Although the model is simple in structure, it is easy to carry out end-to-end training, and the author proposes to use Gigaword to construct many parallel sentence pairs, making it possible to use neural network training, but the grammar, accuracy and consistency of generating summarization remains to be improve. Based on collaborative training, Wang et al. [7] proposed an automatic summarization method based on SVM and NB, which can be used to test large scale features of corpus and select the most suitable features. The disadvantage is that to achieve the most appropriate effect, a large-scale training corpus is required for training [8].

In view of the above problem that the generated summarization is not accurate enough, the calculation method is slow, and the large-scale corpus is needed for training, this paper adopts a hybrid neural network-based method. The Convolutional Neural Network(CNN) used in this method has fewer training parameters than the fully connected forward neural network, which makes it efficient and difficult to over-fitting in the training process.Moreover, CNN can extract the semantic similarity between texts through the training of large-scale data sets [9-10]. Long Short-Term Memory Neural Network (LSTM) model is used to realize the representation of the text model, which can enrich the internal relations of the text and make it context-connected [11]. Integrating LSTM and CNN can improve the accuracy of automatic summarization significantly.

2. Single document automatic summarization model based on hybrid neural network

The single-document automatic summarization model architecture based on hybrid neural network designed in this paper is mainly divided into five steps: text preprocessing, CNN statement vector representation, LSTM document representation and sentence extraction, and generation summarization. The part of CNN obtains the vector representation of the sentence, which is used as an input to the LSTM to combine the sentences to generate a document representation. The overall model is shown in Figure1.

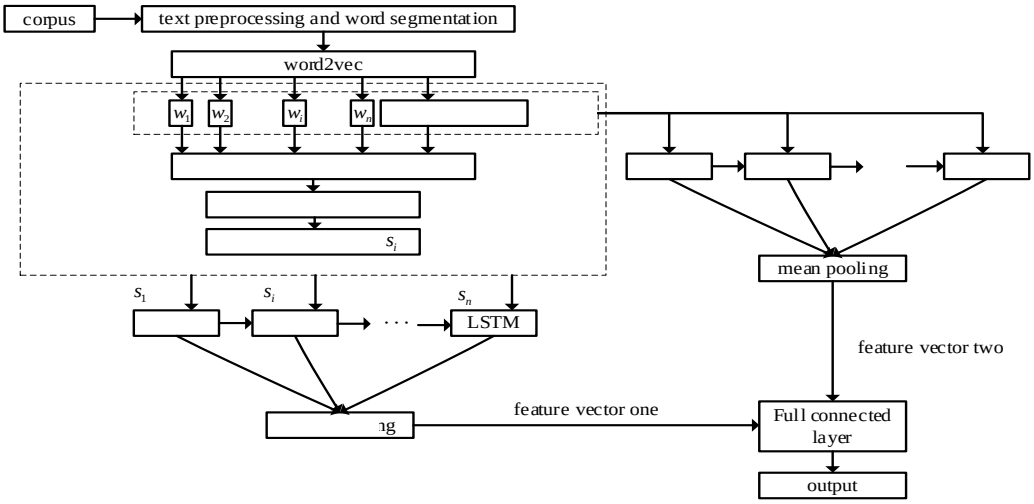

Figure 1 Hybrid neural network automatic summarization model.

The model combines a CNN that is efficient and difficult to over fitting during training and a LSTM that is sensitive to time series. After obtaining the corpus, we preprocess and segment the words and vectorize each word through the word2vec word turn, and then express each word in the whole sentence by CNN as a sentence vector representation. Then each sentence and each word in the sentence is input into two different LSTM models to get the matching degree between the sentence and the document. The sentence with the higher matching degree is extracted as the summarization of the document.

2.1. Text preprocessing

First, the preprocessing process of the original sentence is performed, including sentence segmentation, remove the stop words, punctuation marks, stems, etc. Clause the sentence by period, question mark, exclamation point, etc. Then use the jieba participle [12] tool to segment each sentence and remove the stop words, so that each document can be given a glossary.Words that are common in text and rarely express information about the degree of relevance of a document are called stop words, and mainly include English characters, numbers, mathematical characters, punctuation marks, and single Chinese characters with extremely high frequency. The preprocessing phase of the data is extremely important because it will be used during the coding phase and will directly affect the effect of the entire model [13-14].

2.2. CNN statement vector and document vector representation model

Convolutional Neural Networks (CNN) are widely used in image and speech modeling. Because of its multi-layered structure, the training process of CNN models is less prone to over-fitting and its training process is more efficient. Generally, better expected results can be achieved. The basic structure of the CNN model generally consists of two parts: the feature extraction layer and the feature mapping. Now with the development of CNN, it can achieve better results by applying it to natural language processing. The input in natural language processing tasks is a sentence or document expressed in matrix form, and the rows of the matrix represent a vector representation of a word or character. Each column of the matrix represents the number of words or characters. The model of CNN is shown in Figure 2.

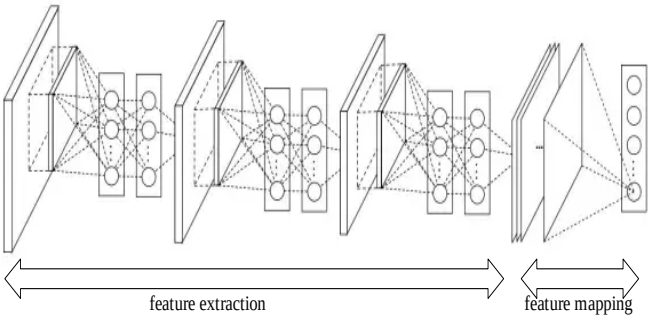

Figure 2 CNN model.

Since CNN cannot directly perform the convolution operation on the text after the above segmentation, it needs to be converted into a word vector first. The word-to-vector model used in this paper is word2vec. The word2vec tool converts the result of the word segmentation into a word vector, and then the obtained word vector is used as the input of CNN. Through continuous training, eventually we can get vector representation. After performing the convolution operation, the model can obtain the feature combination representation of the adjacent words. The combination of these features indicates that after the local maximum pooling operation, the combination that can strongly characterize the word features will be selected and performed several times. After the convolution and maximum pooling operations, the vector representation of the original statement can be generated, and the dimensions of the veator is fixed. The CNN-based sentence representation model is shown in Figure 3.

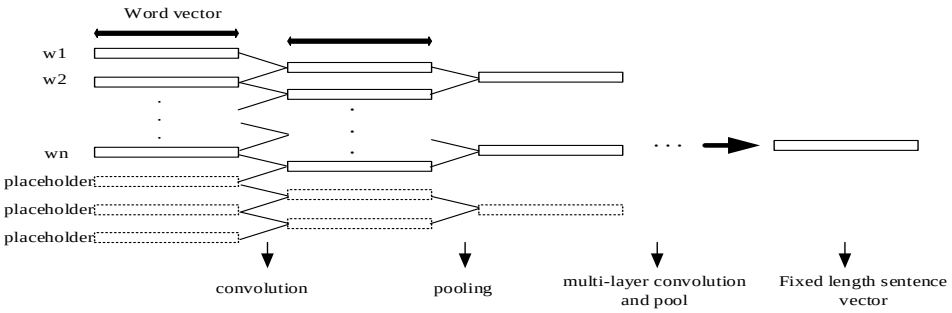

Figure 3 CNN-based sentence representation model.

The reason for adding placeholders is to ensure that each sentence can have the same length. In an article, the length of words in sentences is often inconsistent, which leads to inconsistent node sizes entered during CNN training, and the training process will be extremely difficult. In this paper, the number of input nodes of CNN is the number of word vectors in the longest sentence in the article. For the sentences whose length is less than the number, the vector of all zeros is used to represent the placeholder vector. In the training process of CNN, the effect of placeholders will gradually weaken, so this idea has worked very well. The word vector in each sentence is trained by CNN to finally get a vector representation of each sentence.

2.3. Improved LSTM sentence extraction model

LSTM is an improvement on Recurrent Neural Networks (RNN), which replaces the hidden layer nodes of the RNN with a memory unit to learn long-term dependency information. The LSTM memory unit is divided into four parts, namely a memory cell, an input gate, a forget gate, and an output gate. Memory cells are used to store and update historical information, and the remaining

three gates are used to protect and control cellular status. It is mainly composed of a sigmoid function and a point multiplication operation. The value of the sigmoid function is between 0 and 1. The point multiplication operation determines how much information can be transmitted. When it is 0, it does not transmit. When it is 1, it transmits all. The structure of the LSTM memory unit is shown in Figure 4.

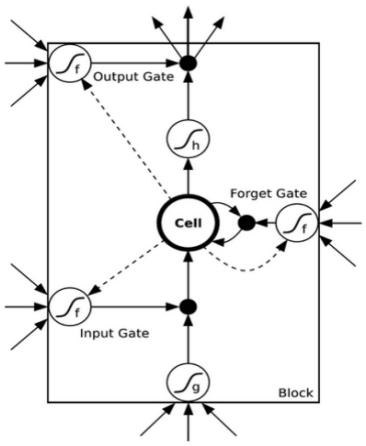

Figure 4 LSTM memory unit structure.

Based on this structure, a cell state signal is added to the input of each gate to form an improved LSTM model that can improve the performance of the original LSTM model. The formula for this improved LSTM is as follows:

$$f_t = \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i) \tag{2}$$

$$C_t = f_t * C_{t-1} + (1 - f_t) * \tilde{C}_t \tag{3}$$

$$o_t = \sigma(W_o \cdot [C_{t-1}, h_{t-1}, x_t] + b_o) \tag{4}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

$$h_t = o_t * \tanh(C_t) \tag{6}$$

Where σ, tanh are the sigmoid function and the hyperbolic tangent function, respectively, W, b, x, h are the weight matrix, the offset vector, and the input and output of the memory unit. f_t, o_t, i_t, C_t, C_{t-1} are forgotten gates, input gates, output gates, candidate values, and new cell states. h_t represents the final output.

After obtaining the vector representation of the sentence and the vector representation of each word in the sentence, this paper uses the LSTM model to extract the document summarization. The traditional cyclic neural network RNN has too many layers, which will cause the gradient of the model parameters to disappear during the training process. LSTM is proposed to solve this problem. The hidden layer of the RNN is implemented by the LSTM memory unit. Instead, this will greatly improve the applicability of the model. The LSTM model uses a memory module to record historical information, and then updates and utilizes historical information through input gates, forgetting gates, and output gates.

The input to the LSTM is the sequence of sentence vectors and document vectors generated in the previous step. The idea of this paper is to model the vector representation of sentences and documents with an LSTM model and get a vector representation of the two. Then, through vector splicing, the method of logistic regression is used to predict, so that the experimental effect is greatly improved. The model is shown in Figure 5.

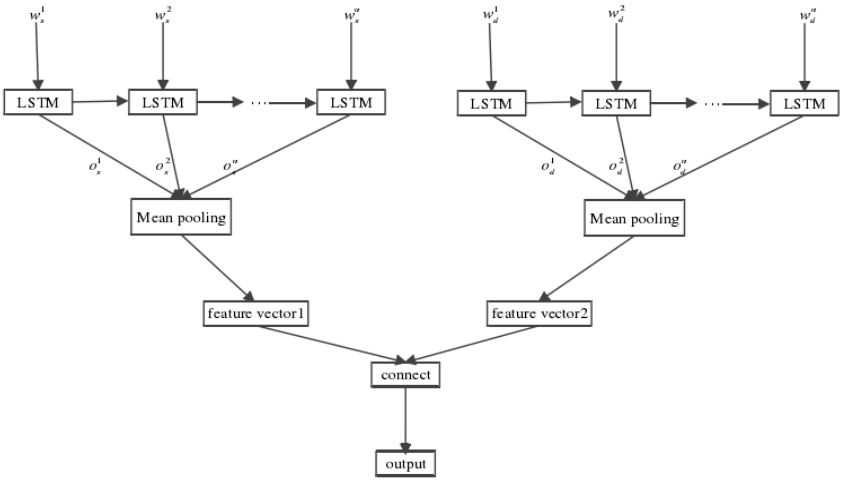

Figure 5 LSTM sentence extraction model.

In Figure 5, w_s^i represents the word vector corresponding to the ith word of the sentence. w_d^j represents the vector representation of the jth sentence in the original document, o_s^j represents the output vector obtained after the ith word is input to the LSTM, and o_d^j represents the output vector obtained after the jth word in the document is input to the LSTM. This model can be used to obtain summarization automatically.. In the implementation process, the hidden layer and the mean pooling layer of the model are set to two layers. When the model training is finished, the test will be performed as a summarization of several sentences with high probability of matching.

2.4. Evaluation standard

This paper adopts the Rouge evaluation method [15]. The main idea of this evaluation method is to compare the automatic summarization generated by the system with the artificially generated standard summarization, and by counting the basic units (n-grams, word sequence and word pair) overlapping between the two to evaluate the quality of the summarization. Improve the stability and robustness of the evaluation system through the comparison of multi-expert manual summarization. The evaluation method mainly includes ROUGE-N, ROUGE-L, ROUGE-W,etc. Among them, ROUGE-N represents the recall value of N co-occurring words, and its calculation formula is as follows:

$$ROUGE\text{-}N = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \tag{7}$$

Where n represents the length of $n\text{-}gram$, which is n continuous words, $Count_{match}(gram_n)$ represents the maximum number of $n\text{-}gram$ that appear simultaneously in a candidate summarization and reference summarization set, and $Count(gram_n)$ represents the number of

$n\text{-}gram$ in the reference summarization set. It can be seen from the equation that ROUGE-N is a way of recall rate. The numerator is the number of $n\text{-}gram$ in the reference summarization in the candidate summarization, and the denominator is the number of all $n\text{-}gram$ in the reference summarization. The Rouge scoring function for multiple documents can be expressed as:

$$ROUGE\text{-}N_{multi} = avg_i Rouge - N(r_i, S) \tag{8}$$

Where r_i is the reference summarization and S is the system generated summarization.

3. Evaluation criteria and analysis of experimental results

3.1. Data preparation

The experiment selected the text data of Sina Weibo as the experimental corpus. After preprocessing and denoising, the experiment obtained a Chinese microblog corpus containing microblogs and original text. The data of 6852 microblogs was selected and divided into 4,796 training data and 2056 test data . The process of data processing mainly includes: removing special characters, removing emoticons, replacing the contents in parentheses, replacing date tags, replacing hyperlink tags, replacing full-width English tags, replacing numbers, etc. After the text is preprocessed, the training corpus is prepared: the original text of the microblog is taken as input, and the target sequence to be predicted is the summarizationt of the microblog. After the corpus is prepared, to segment the text and remove the stop words.

3.2. Feature processing

There are many words in the results of the word segmentation that have no positive effect on the generation of automatic summarization. The existence of these words will increase the dimension of the feature vector, so consider deleting these words. The basic method of this paper is to use the top 3000 words with high to low occurrence as feature words. The method of selecting these feature words is TF-IDF model. This method can express the weight $w_{i,j}$ of each word in the sentence by the following formula.

$$w_{i,j} = TF_{i,j} * IDF_i \tag{9}$$

Where $TF_{i,j} = \dfrac{f_{i,j}}{\sum_z f_{z,j}}$ represents the word frequency of sentence s_j, $f_{i,j}$ represents the number of occurrences of keyword k_i in sentence s_j and $\sum_z f_{z,j}$ represents the number of all words in sentence s_j. $IDF_i = \log \dfrac{N}{c_i}$ represents the inverse document frequency of k_i, c_i represents the number of sentences with the word k_i, and N represents the total number of documents. After obtaining the weight of each word, the top 3000 words with the weights from high to low are used as feature words, and then the word2vec tool is used to train the feature words in each sentence into a fixed-dimensional vector. The basic idea is to pass the training maps each word into a K-dimensional real number vector (K is a hyperparameter in the model) and judges the semantic similarity between the words by the distance(such as cosine similarity, Euclidean distance, etc.) between the words . It uses a three-layer neural network, input layer - hidden layer - output layer. The advantage of using word2vec is that it makes extensive use of the context and semantic

information of the word. In this paper, the dimension of the word vector is set to 200, the sliding window size is set to 5 during training, and the words with word frequency less than 5 are filtered out directly.

3.3. Analysis of results

In the process of hybrid neural network model to achieve Chinese automatic summarization, this paper uses several summarization methods to conduct comparative experiments: mainly including the common summarization method and the LSTM model method in deep learning method. The first step of the hybrid neural network model proposed in this paper is to obtain the representation of the word vector, and then obtain the vector representation of the sentence through CNN. Finally, the matching degree of the sentence and the article can be obtained by using the LSTM model of the sentence vector and the word vector. Then select the sentence with high matching degree as the summarization of the document.When using CNN to obtain the sentence vector representation, the three-layer convolution pooling layer is used, and the number of input neurons of the model is 200. For the number of words in the sentence less than 200, complement by adding placeholders. The CNN output is also a fixed-length vector that represents the vector representation of the input sentence. The word vector dimension in the model used to obtain the matching degree of the sentence by LSTM is still 200. When the model training is finished, the test is performed to input several sentences with high matching probability as the summarization of the article.

After determining the experimental parameters, this paper determines the following methods for comparison experiments, which are automatic summarization algorithm based on LSI model, automatic summarization algorithm based on LDA model, TextRank based summarization generation algorithm, PCA based summarization extraction algorithm and automatic summarization algorithm based on LSTM. The verification set selected by the experiment is 1000 articles, and the obtained summarization are analyzed for different features of the LSI. For the LDA topic model, the choice of the number of topics will directly affect the final effect of the experiment. A series of comparative experiments were carried out on the parameter's selection of LSI and LDA models. The results of comparative analysis are shown in Table 1. At the same time, the line graphs corresponding thereto are given in Figure 6 and Figure 7, so that the final parameters of the LSI and LDA models can be determined. When using the LSTM model for automatic retrieval of summarization, this paper use a single sequence based on LSTM model to combine documents and sentences into a sequence and use an LSTM model to model the merged statements of the two. Experiments show that this method can improve the efficiency of automatic summarization, and it is helpful to raise the ROUGE value. The resulting summarization is also more contextual.

Table 1 Experimental results table based on LSI and LDA models.

LSI			LDA		
Features	ROUGE-2	ROUGE-3	Topics	ROUGE-2	ROUGE-3
2	0.1925	0.1734	2	0.2012	0.1753
10	0.2453	0.2356	3	0.1874	0.1655
15	0.2511	0.2365	4	0.1743	0.1551
50	0.2512	0.2369	5	0.2005	0.1702
100	0.2522	0.2371	6	0.2201	0.1898
200	0.2523	0.2372	7	0.2255	0.1944
500	0.2522	0.2372	8	0.2306	0.2088

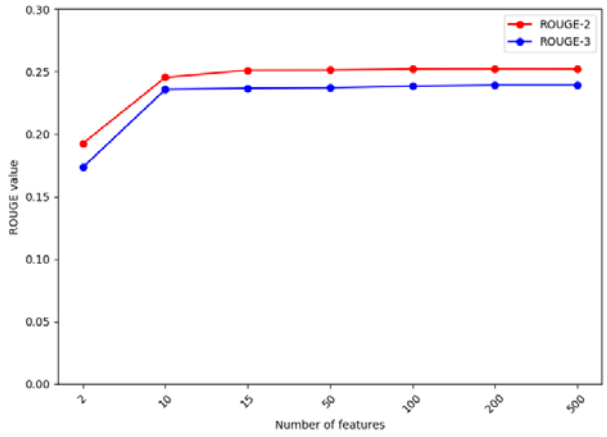

Figure 6 Comparison of different feature quantity selection experiments based on LSI model.

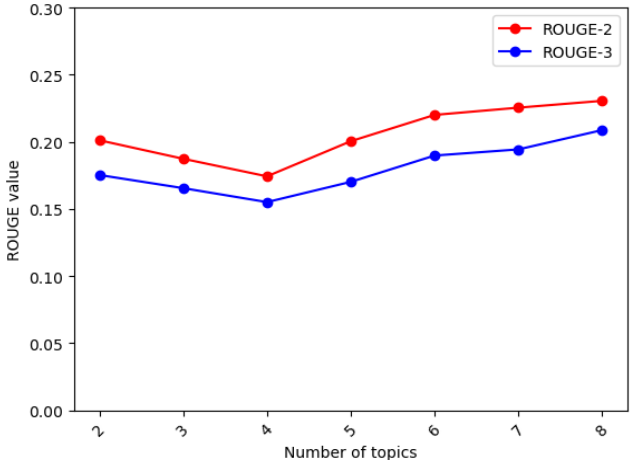

Figure 7 Comparison of different subject number selection experiments based on LDA model.

As can be seen from Table 1, Figure 6, and Figure 7, when the feature number of the LSI is set to 200, the ROUGE value of the experimental result is the highest, and as the number of selected features increases, the final experimental results do not have much effect, so the number of features is set to 200 during the test. As can be seen from the chart, the choice of the number of topics does have a great influence on the quality of the summarization, and in this experiment, when the number of topics is set to 8, the experiment works best, so in the test, we take this parameter to test. The comparison between the algorithm in this paper and the above proposed algorithm is shown in Table 2.

Table 2 Experimental results table of the algorithm and other algorithms.

Methods	ROUGE-2	ROUGE-3
LSI	0.2523	0.2372
LDA	0.2306	0.2088
LSTM	0.2534	0.2413
TextRank	0.2461	0.2256
PCA	0.2335	0.2038
Proposed algorithm	0.2624	0.2445

The corresponding line chart is shown in Figure 8.

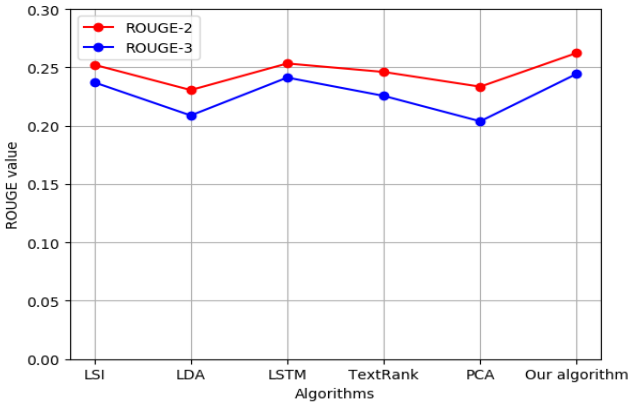

Figure 8 ROUGE values for different algorithms.

It can be seen from Table 2 and Figure 8 above that the performance of the traditional automatic summarization generation method is ideal for LSI, and its ROUGE-2 value and ROUGE-3 value are 0.2523 and 0.2372, using the LSTM model to generate automatic summarization ,the ROUGE-2 and ROUGE-3 are 0.2534 and 0.2265, and the ROUGE values of this paper's algorithm are 0.2624 and 0.2445. It can be seen from comparison that the proposed algorithm model is compared with other algorithms, both the ROUGE-2 value and the ROUGE-3 value are larger, indicating that the proposed algorithm has a significant improvement effect on the generation of automatic summarization. At the same time, due to the combination of CNN and LSTM models, the accuracy and consistency of the summarization are effectively improved, which makes the generated summarization more readable and the central content of the presentation more clearly.

4. Conclusion

In this paper, a neural network is used to fuse the two methods proposed by the predecessors, and a new hybrid neural network CNN-LSTM model is proposed. The model first divides the text sentences by traditional methods, and then uses the word-to-vector tool word2vec to convert the words in the sentences into vector representations. The word vector of the entire sentence is then input into the CNN model to obtain a vector representation of the sentence. Finally, the sentence vector and the word vector are respectively passed into the LSTM model, and the output result is input into the logistic regression model. The obtained result is used to judge the matching degree between the sentence and the article, and the sentence with higher matching degree is selected as the summarization of the document. The experimental results show that compared with automatic summarization methods based on LSI model, LDA model, TextRank summarization algorithm, PCA summarization algorithm and LSTM model，the proposed method based on hybrid neural network has a good effect on automatic summarization. Improved the intelligence and quality of summarization. It can be used in practice to generate automatic summarization.

Acknowledgements

This work is supported by Zhejiang Provincial Natural Science Foundation of China (No. LY17E050028).

References

[1] Hu, X., Lin, W., Wang,C., et al. Overview of automatic text abstract technology[J]. Journal of Information, 2010, 29(8): 144-147.

[2] Wu, X. F., Zong, C. Q. A CRF automatic abstracting method based on LDA[J]. Chinese Journal of Information Science, 2009, 23(6): 39-45.

[3] Hu, B. T., Chen, Q. C., Zhu, F. Z. LCSTS: A large scale chinese short text summarization dataset[J]. Proceedings of Conference on Empirical Methods in Natural Language Processing, 2015: 1967-1972.

[4] Nichols, J., Mahmud, J., Drews, C. Summarizing sporting events using Twitter. Acm International Conference on Intelligent User Interfaces ACM, 2012: 189-198.

[5] Blei, D. M., Ng, A. Y., Jordan, M. I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[6] Rush, A. M., Chopra, S., Weston, J. A. A neural attention model for abstractive sentence summarization[J]. Proceedings of Conference on Empirical Methods in Natural Language Processing, 2015: 379-389.

[7] Wong, K. F., Wu, M., Li, W. Extractive summarization using supervised and semi-supervised learning. International Conference on Coling, 2008: 985-992.

[8] Majak, M., Zolnierek, A., Wegrzyn, K., et al. Tweet classification framework for detecting events related to health problems. Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017, 2017: 453-461.

[9] Niu, X. X., Suen, C. Y. A novel hybrid CNN-SVM classifier for recognizing handwritten digits[J]. Pattern Recognition, 2012, 45(4): 1318-1325.

[10] Razavian, A. S., Azizpour, H., Sullivan, J., et al. CNN features Off-the-Shelf: An astounding baseline for recognition. Computer Vision and Pattern Recognition Workshops IEEE, 2014: 512-519.

[11] Sundermeyer, M., Schluter, R., Ney, H. LSTM Neural Networks for Language Modeling. Interspeech, 2012: 601-608.

[12] Mikolov, T., Chen, K., Corrado, G., et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.

[13] Hovy, E., Lin, C. Y., Zhou, L., et al. Automated summarization evaluation with basic elements[C]. Conference on Language Resources and Evaluation, 2003.

[14] Radev, D. R., Zhang, W. Webinessence: A Personalized Web-Based Multi-Document Summarization and Recommendation System[C]. NAACL Workshop on Automatic Summarization, 2014: 79-88.

[15] Lin, C. Y. ROUGE: A Package for Automatic Evaluation of summaries. The Workshop on Text Summarization Branches Out, 2004: 10.