ISSN: 2688-9323

Research on Keyword Generation of News Title based on Attention Mechanism

Yang Chen*

School of Management, Shanghai University, Shanghai, China

*cy8165@126.com

Abstract

With the continuous development of Internet technology, people have entered the era of big data, in which the text data is growing exponentially. It is difficult for people to quickly identify the information they need from the massive text information. Keywords are highly condensed text content, which can reflect the theme of text information and help people quickly understand the core content of the text. At the same time, keywords are widely used in various tasks in the field of natural language processing, such as text classification, search engine, recommendation system and so on. Therefore, keyword extraction technology is very important. But the traditional keyword extraction technology only depends on the shallow features of the text to extract important words and can only extract the content of the original text. In recent years, the keyword generation model based on neural network can better solve the above problems. However, there is still the problem that the generated keywords deviate from the original content. Therefore, many improvement methods have been proposed, including attention mechanism. This paper combines attention mechanism with Sequence to Sequence model to form a keyword generation model based on attention mechanism, and compares this model with traditional TF-IDF model, TextRank model and model without attention mechanism to study the advantages and feasibility of keyword generation model and attention mechanism in news headline keyword extraction task.

Keywords

Keyword Generation; Attention Mechanism; Sequence to Sequence Model.

1. Introduction

With the development of the Internet era, the number of online text information has shown an explosive growth. People can obtain text information from a variety of channels, such as the dynamic information released by users of social platforms (Weibo, Zhihu, Xiaohongshu, etc.), news published by major news websites, and academic papers in digital journals. The Internet has changed people's concept. Because of the convenience of the network, people can receive and disseminate information anytime and anywhere, from the traditional way of obtaining information through paper media to obtaining information through the network to grasp the latest social trends. However, the speed at which people receive and process information is far less than the speed at which current information is generated, which will lead to information overload. It is difficult for people to quickly find the information they need from the mass of text information. The key words are highly condensed text content, which can help people quickly understand the main idea and key content of text content, so that people can get twice the result with half the effort.

Keyword is a compact representation of text content. Keyword extraction of text can effectively express the text theme. The accuracy of keyword recognition has a great impact on text classification, text recommendation and text search. It is an important basic work in the field of

ISSN: 2688-9323

text mining and analysis. The application scenarios of text keyword extraction are very extensive, for example, search engines perform information retrieval based on keywords; In the summary system, keywords can be an important feature of sentence scoring. The more keywords a sentence contains, the more likely it is to become a summary sentence; In the recommendation system, keywords can be personalized and displayed to users to attract their attention; In addition, keyword extraction can also be used in many natural language processing tasks, such as text classification, text clustering, text emotion analysis, etc.

With the popularity of mobile devices, people can browse news anytime and anywhere. Considering the diversity of news, the speed of real-time updates, the number of news and the display efficiency of mobile devices, keywords are needed to help people quickly filter out the information they are interested in, reduce the cost of obtaining information, and improve the user experience. Therefore, it is very important to study a technology that can effectively extract keywords. This paper studies a Seq2Seq text generation model based on attention mechanism to generate keywords according to news headlines, and compares it with traditional machine learning methods to extract keywords.

2. Related Works

Keyword extraction technology mainly includes keyword extraction method and keyword generation method. Traditional text keyword extraction technology mainly focuses on keyword extraction method. The keyword extraction method selects important words from the original text as keywords, which is very different from the way humans generate keywords. The keywords generated in this way cannot accurately express the original text. The keyword generation method can generate words that do not appear in the original text by reorganizing the text information. In fact, at least 40% of the keywords given by humans do not come from the original text, so keyword extraction method is slowly developing towards keyword generation method.

Keyword extraction methods can be further divided into unsupervised methods and supervised methods. Unsupervised common methods include: methods based on simple statistics, such as Term Frequency Inverse Document Frequency (TF-IDF), which extracts keywords based on word frequency, such as TF-IDF, which extracts emotional words in emotional classification tasks, Alberto Purpura and others proposed that the best implementation method of text emotional classification in the field of e-commerce is dictionary based, However, it will be subject to two main limitations: out of vocabulary keywords (OOV) and non-cross domain use (1). They improved the traditional TF-IDF, and adopted a supervised method based on TF-IDF retrieval and polynomial linear regression elastic network regularization to extract the emotion dictionary; Graph based methods, such as the typical representative TextRank model in the graph model based on statistics, have put forward many improvement methods for the shortcomings of the TextRank model found in the research. For example, Hu Jiming et al. [2] used the word vector table trained by Word2vec in the process of sentence similarity calculation to transform each word in the policy text sentence into the corresponding word vector, Then, all word vectors in the sentence are averaged pooled, and the sentence is represented as a sentence vector. The TextRank model is improved to enhance the text representation effect; In addition, there are also unsupervised methods such as SingleRank model, LDA model, and language model.

With the rapid development of machine learning and deep learning, supervised keyword extraction methods are getting better and better. The general patterns based on supervised keyword extraction can be roughly divided into two categories. The first category regards keyword extraction as a classification task, that is, candidate words are divided into keyword and non-keyword categories through the classification model. Wang et al. used Support Vector

ISSN: 2688-9323

Machine (SVM) to filter keywords [3], and judged whether words in the text were keywords according to features, including word frequency and location information of words. The second type of supervised keyword extraction method regards it as a sequence tagging task, that is, using the sequence tagging model to learn the relationship between words in the sentence sequence with tagged keywords, and then tag the unlabeled sentence sequence to extract the keywords in the sentence. With the increase of the number of network layers, the performance of the neural network model becomes more and more powerful, which also adds new impetus to the keyword extraction task. Zhang et al. proposed a recurrent neural networks (RNN) model with two hidden layers [4]. This model captures word information through the first layer RNN network, and then labels keywords in the sequence through the second layer RNN network. Because the gradient is easy to disappear in the reverse propagation of RNN network in a long sequence, Basaldella et al. replaced the first layer of RNN network with a bidirectional long-short term memory (Bi LSTM) model.

The traditional keyword extraction model has three shortcomings: first, most of them can only extract words from the original text; Second, it mainly depends on the shallow features of the text to extract important words, so it is difficult to mine and make full use of the potential semantic information behind the text. Third, the unsupervised method is simple and easy to operate, but the accuracy rate is not high. The supervised method process is standardized, and the accuracy is high, but it needs a lot of annotation data support. On the whole, the current mainstream keyword extraction method research is more inclined to the supervised method, and the model in the research is also shifted from the classification model to the sequence annotation model.

In recent years, the keyword generation model based on neural network has been proved to be able to overcome the limitations of the above keyword extraction methods. The technology of keyword generation is different from that of keyword extraction, and the research progress related to keyword generation is relatively slow. This is due to the difficulty of keyword generation technology and the lack of mature programs to promote the research in this area. Thanks to the development of deep learning technology, keyword generation research has been greatly developed in recent years.

In 2014, Cho et al. put forward the earliest RNN based Seq2Seq model [6], which has brought great changes to machine translation tasks and provided new opportunities for keyword generation research. One of the advantages of Seq2Seq model is that it can solve the problem that the length of the original sequence is not equal to that of the target sequence, which makes a breakthrough in natural language processing tasks with inconsistent input and output lengths. Seq2Seq model can solve the problem that the length of the original text is inconsistent with that of the translated text in machine translation, which is very similar to the scenario of keyword generation. Therefore, some researches have proposed to use Seq2Seq model to solve the keyword generation task and achieved some results. Seq2Seq model can generally be disassembled into encoder module and decoder module. The former is responsible for encoding and compressing the information of the document, and the latter is responsible for decompressing it, so as to realize the mapping relationship between two unequal length sequences. At present, the mainstream keyword generation methods are based on this. In order to solve the problem of information transmission in long sequences, Meng et al. used the Gated Recurrent Unit (GRU) network [7] to replace the simple RNN network in the research of keywords in academic texts generated based on the Seq2Seq model. Zhang et al. proposed to add attention mechanism [8] to solve the problem of semantic coverage and semantic relevance.

The research related to keyword generation only appeared more intensively after 2017, which is closely related to the development of the deep learning model. The main factors affecting the keyword generation results are divided into two aspects: first, the semantic coverage of generated keywords. Because the simple Seq2Seq model does not consider the similarity

ISSN: 2688-9323

between keywords, the generated keywords cover each other semantically to a high extent; The second is the problem of labeling data. The deep learning model is data-driven, and the size of the data directly affects the generation result of the model, which limits the application scenarios of the model. Therefore, the current research on keyword generation also focuses on these two aspects.

Aiming at the problems in the current research of keyword generation, this paper studies the news title keyword generation model based on attention mechanism, uses Seq2Seq model with attention mechanism to generate news title keywords, and sets up contrast experiments to verify the improvement effect of attention mechanism on obtaining text context features, solving the problems that puzzle keyword generation such as semantic coverage and semantic relevance.

3. Dataset and Proposed Method

3.1. Dataset

The experimental data used in this paper is from the public dataset. The original data is 2.5 million news data. The news sources cover 63000 news media. Each data in the dataset contains seven fields: news id, keywords, title, description, source, time, and content. The research problem of this paper is the generation of news title keywords, so only the content of the title field and keyword field of each data is needed, and the title and keyword form the input and output data of the Seq2Seq model.

Table 1. Dataset examples

news_id	keywords	title	desc	source	time	content
610130831	guide\admission ticket	The off-season ticket of the Forbidden City is 40 yuan, and the " Illegal tour guide" is 140 yuan	A netizen's microblog recently reported that "Illegal guides" appeared at the ticket office…	Xinhuanet	03-22 12:00	Recently, a netizen microblog reported that "black guides" appeared at the ticket office of the Wumen Square in the Forbidden City…
410648968	sea horse\ grille	2015 Hippocampus M3 will be launched in April and pre sold for 60000 to 80000 yuan	It was recently learned that the 2015 Haima M3 will be officially launched in April…	Fan Yang	04-01 11:04	A few days ago, it was learned that the 2015 Haima M3 will be officially launched in April, and 10 models will be launched…

ISSN: 2688-9323

After obtaining the original data, preliminary screening of the data is required. Due to the limitation of experimental equipment, it is impossible to use all the original data for the experiment. The experimental data used in this experiment is randomly sampled from 2.5 million news data, totaling 76797 news data. Later, the data was further processed. Through observation of the experimental data, it was found that there was data with the same title field and keyword field in the data, which accounted for 78% of the entire experimental data. Such data would seriously affect the effect of the text generation task. Therefore, this part of data was removed from the experimental data, and 16920 news data remained after screening. The title and keyword content of 16920 news data were extracted to form the final experimental sample, and the training set and test set were divided according to the proportion of 90% and 10%, with 15228 pieces of data in the training set and 1692 pieces of data in the test set. An example of raw data is shown in Table 1.

3.2. Proposed Method

Long-Short Term Memory (LSTM) is a variant of RNN. During the operation of RNN internal structure, the calculation under the current time step is to splice the activation value output from the previous time step with the current input value, and output the activation value of the current time step through the tanh activation function. This disadvantage is that RNN units can only focus on the state of the previous near time. If the input text is too long, it will lead to the loss of gradient due to the long back propagation path, and eventually lead to the loss of long-distance text information. LSTM is proposed to solve the problem of long text information memory. The internal structure of LSTM is more complex than that of RNN, adding new computing processes, introducing memory cells and gating mechanisms, and using memory cells to store information, and using input gates, forgetting gates, and output gates to maintain and control information updates. Besides the tanh function, the activation function used is also sigmoid function, adding summation operation to reduce the possibility of gradient disappearance and gradient explosion.

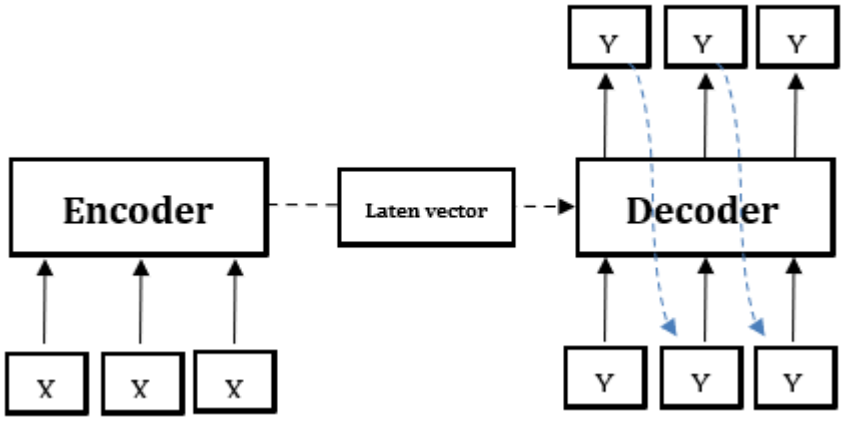

Fig 1. Seq2Seq model

The basic Seq2Seq model generally consists of two parts, namely encoder and decoder. In natural language processing, the encoder and decoder generally use RNN or its variants GRU and LSTM. The general feedforward neural network lacks the concept of time sequence, while the cyclic neural network can obtain the sequence information in data such as text and voice, because the output of the cyclic neural network depends on the current input and previous output information. The encoder and decoder in the basic Seq2Seq model are equivalent to the RNN language model. The encoder represents the text as a vector, then obtains the text information through various time steps, and finally obtains the final output hidden layer vector. The final hidden layer vector of the encoder will be the initial hidden layer state of the decoder.

ISSN: 2688-9323

At each time step, the decoder will form the probability distribution of the word list by passing the hidden layer vector output from the time step through the softmax layer, and the output value is the word with the highest probability of the current time step. The hidden layer vectors and words output from each time step will be transferred to the next time step as input for the calculation of the next time step. The basic structure of Seq2Seq model is shown in Figure 1.

The core goal of attention mechanism is to select the information that is more critical to the current task goal from a large amount of information. Applying it to the Seq2Seq model is to find the key information in the input text. The model encoder will output a hidden layer vector h. In the model without attention mechanism, the hidden layer vector h received by the decoder will participate in the calculation of each time step of the decoder intact. However, as the input sequence length increases, the hidden layer vector h will lose some information. In the model with attention mechanism, each time step will have its own hidden layer vector h (h0, h1, h2). The hidden layer vector h of each time step will calculate an attention value for each input word through the hidden layer and softmax layer to judge which information in the original text should be paid attention to. The attention mechanism is shown in Figure 2.

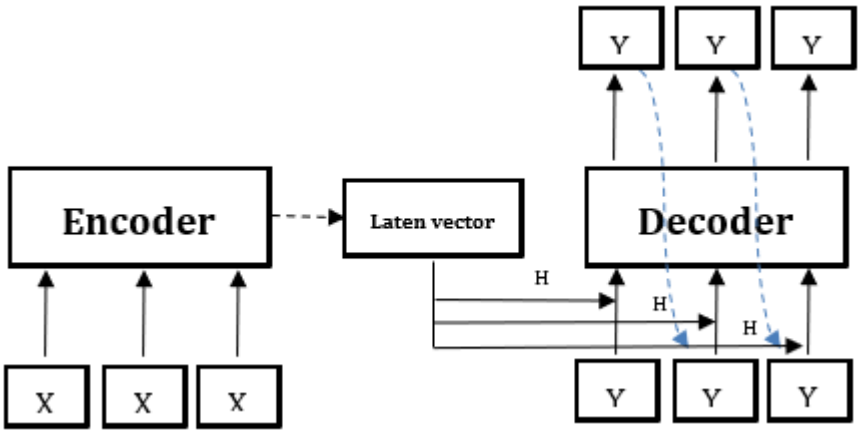

Fig 2. Attention mechanism

TextRank is a sorting algorithm for text, which is an improvement of PageRank algorithm, and was later applied to keyword extraction and text summarization. TextRank does not need to train corpus, but directly extracts keywords from the text itself. It performs well in keyword extraction tasks, and is one of the keyword extraction algorithms widely used at present. Its basic idea is to use voting mechanism to sort the important components of the text by dividing the text into words and building a graph model. The main steps are: first, divide the document into sentences, then divide the sentences into words, and remove the stop words to get the candidate word set. Then, the undirected graph model of the text is constructed, and the edges between the two construction points are obtained according to the collinear relationship. Then, according to the node weight formula, the weights of each node are iterated until convergence. Finally, sort the node weights and find the most important Top N words as keywords. In the experiment of this study, we use Python third-party library jieba to implement TextRank algorithm to extract keywords in news titles.

TF-IDF is a weighted technology commonly used in information retrieval and text mining. TF-IDF is a statistical method to evaluate the importance of a word or word to a document set or one of the documents in a corpus. The importance of a word increases proportionally with the number of times it appears in the document, but decreases inversely with the frequency of its appearance in the corpus. The main idea is: if a word appears frequently in one article and rarely appears in other articles, it is considered that the word or phrase has good classification ability and is suitable for classification. It can be considered that the word is more important

ISSN: 2688-9323

than other words in this article. In the experiment of this study, we use the Python third-party library jieba to implement TF-IDF algorithm to extract keywords from news headlines.

4. Experiment and Result Analysis

4.1. Experiment Setting

The experimental model used in this paper is the Seq2Seq model with attention mechanism. First, 16920 pieces of text data are preprocessed, and the text data is segmented using the Python third-party library jieba. The uppercase letters in the data are converted to lowercase, and <start> and <end> marks are added to the beginning and end of each data title and key word as the marks of the beginning and end prediction of the model. The vocabulary constructed in the experiment contains 37810 words, the embedded dimension of word vector is set to 64, and the number of hidden neurons is set to 128. When training the experimental data set, the batch sample size is set to 64, the Adam optimizer is used to optimize the model, and the learning rate is set to 3e-6. During the model training, the convergence of the model on the data set is confirmed, that is, the cross-entropy loss value will no longer decline, the training is stopped, and the model is saved.

4.2. Contrast Experiment

The Seq2Seq model with attention mechanism adopted in this paper is compared with the Seq2Seq model without attention mechanism, TextRank keyword extraction model, TF IDF keyword extraction model. TextRank and TF IDF are unsupervised keyword extraction methods, and Seq2Seq model is keyword generation methods.

The Seq2Seq model without attention mechanism has the same parameters as the Seq2Seq model with attention mechanism except the learning rate. The only difference is whether it has attention mechanism, so as to compare the impact of attention mechanism on the model effect. Its learning rate is set as 8e-5.

TextRank and TF-IDF keyword extraction are implemented through the jieba library. Both methods input the title list and return the keyword list. At the same time, these two unsupervised extraction methods cannot determine the number of keywords generated independently, so the topk parameter needs to be set to specify how many keywords are returned. By observing the distribution of the number of keywords in the experimental data, 93.3% of the data have 2 keywords, so set the topk parameter to 2, that is, each news title returns two keywords. The number distribution of keywords is shown in Figure 3.

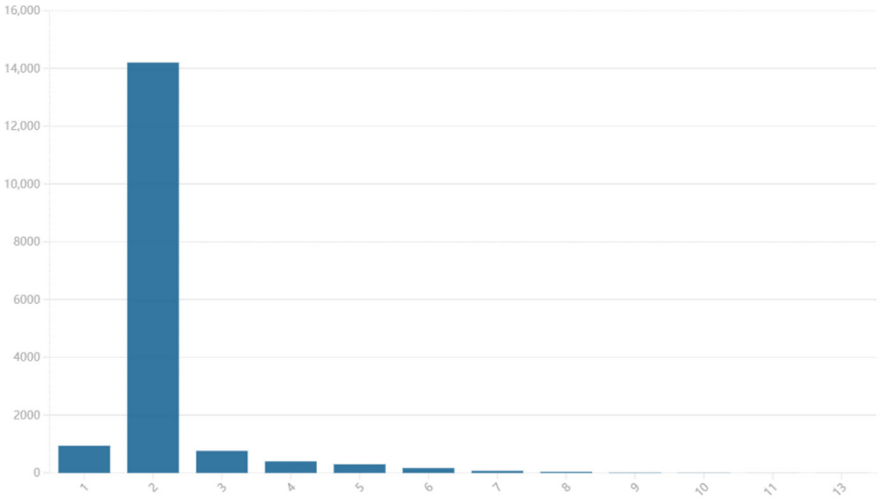

Fig 3. Keywords length distribution

ISSN: 2688-9323

4.3. Metrics

The evaluation index selected in this paper is Bilingual Evaluation Understudy (BLEU). BLEU is a common evaluation index for text generation tasks in natural language processing. At first, this index was proposed for machine translation, and later it is widely used to evaluate the differences between sentences generated by the model and actual sentences. The BLEU value range is 0.0 to 1.0. The larger the BLEU value, the higher the matching degree of the two sentences. Its advantages are easy to calculate, easy to understand, irrelevant to language, highly relevant to human evaluation results, and widely used by academia and industry. The implementation of BLEU method is to calculate the N-grams model of the sentences generated by the model and the actual sentences respectively, and then calculate them by counting the number of matches. This paper will calculate BLEU-1, BLEU-2, BLEU-3 and BLEU-4 respectively to compare the effects of the experimental model. The experimental results are shown in Table 2.

Table 2. Results

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
TF-IDF	0.250	0.129	0.088	0.072
TextRank	0.181	0.087	0.063	0.053
Seq2Seq	0.018	0.005	0.003	0.003
Seq2Seq+Attention	**0.902**	**0.881**	**0.856**	**0.563**

4.4. Results

According to the experimental results in Table 2, the Seq2Seq model with attention mechanism studied in this paper achieved the best results in the four BLEU scores. From the specific results, we can see that for the keyword extraction method, TF-IDF model is better than TextRank model in the news title keyword extraction task. For the keyword generation method, this paper uses Seq2Seq model to achieve the text generation task. It can be seen that the model with attention mechanism is much better than the model without attention mechanism. In fact, during the experiment, Seq2Seq model without attention mechanism could not generate meaningful keywords after convergence. No matter what the title of the model input is, only single and repeated words will be output. It can be seen that the attention mechanism plays a role in keyword generation, indicating that paying attention to important information in the text has a great impact on keyword generation. In general, the Seq2Seq keyword generation method based on attention mechanism is far superior to the traditional TF IDF and TextRank keyword extraction methods. The keyword generation effect of Seq2Seq model with attention mechanism is shown in Table 3. It can be seen from the examples that the generated keywords are meaningful and can well reflect the meaning of the news title, and the model can also generate words that are not in the original news title but can explain the meaning of the original title, which makes the generated keywords more flexible and diverse, and more accurately summarize the meaning of the title, which is the advantage of the keyword generation method.

Table 3. Seq2Seq with attention generation examples

News title	Keywords
Chongqing: a family shopping in a shoe shop, children covering adults' theft	Shoe shop\theft
Methods of tennis training with players	Tennis
Should managers and subordinates keep a distance?	Subordinate care

ISSN: 2688-9323

5. Conclusion

This paper studies the problem of keyword generation of news headlines based on attention mechanism, and introduces the general mode and difference between keyword extraction method and keyword generation method. For keyword generation task, Seq2Seq model is used as the basic model of the generative method, and LSTM is selected to process the text information. In addition, attention mechanism is introduced to better obtain important information in the text, and is compared with Seq2Seq model without attention mechanism. The results show that, on the news title dataset selected in this paper, attention mechanism is of great significance to Seq2Seq model and greatly improves the model effect. This paper also compares the keyword generation method with the keyword extraction method. The results show that the generation method is better than the extraction method, and the generation method can overcome the disadvantage that the extraction method can only extract the original text, and generate words that are not included in the original text but have meaning. Keyword extraction, as a basic task in the field of natural language processing, has a very wide range of application scenarios. The keyword generation method based on the attention mechanism brings new ideas for keyword extraction. In the future, you can try to add data sets, add theme features and other information to the documents to be extracted to further improve the effect of the model.

References

[1] Alberto Purpura, Chiara Masiero, Gianmaria Silvello, et al. Supervised lexicon extraction for emotion classification[C]// Proc. of the 19th World Wide Web Conference. San Francisco CA, USA: Association for Computing Machinery, 2019:1071-1078.

[2] Hu Ji-Ming, FU Wen-lin, QIAN Wei, TIAN Pei-Lin. Policy text Classification Model combining topic Model and attention mechanism [J]. Information Theory and Practice, 2021. (Jiming Hu, Wenlin Fu, Wei Qian, Peilin Tian. Research on policy text classification model based on topic model and attention mechanism [J]. Information Studies: Theory & Application, 2021.

[3] Wang J, Peng H. Keyphrases extraction from web document by the least squares support vector machine [C]// Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiegne, France: IEEE Computer Society, 2005:293-296.

[4] Zhang Q, Wang Y, Gong Y. et al. Keyphrase extraction using deep recurrent neural networks on twitter [C] //Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, USA: Association for Computational Linguistics, 2016:836-845.

[5] Basaldella M, Antolli E, Serra G. et al. Bidirectional LSTM recurrent neural network for keyphrase extraction [C]//Proc. of the 14th Italian Research Conference on Digital Libraries, Udine, Italy: Springer, 2018:180-187.

[6] Cho K, van Merrienboer B, Gülçehre Ç. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar: ACL, 2014:1724-1734.

[7] Meng R, Zhao S, Han S. et al. Deep keyphrase generation[C]//Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada: Association for Computational Linguistics, 2017:582-592.

[8] Zhang Y, Xiao W. Keyphrase generation based on deep Seq2Seq model [J]. IEEE Access, 2018, 6: 46047-46057.