2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) | 978-1-7281-8028-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/IAEAC50856.2021.9390683

News Text Summarization Method based on BART-TextRank Model

Yisong Chen[1],Qing Song[2]

New Media Institute, Communication University of China, Beijing, China

961564847@qq.com, songqing@cuc.edu.cn

***Abstract*—With the rapid development of computer technology and network technology, the Internet has become the main place for people to release and obtain information. The amount of data on the network is growing explosively. The urgent problem is how to accurately obtain the real valuable information from a large amount of data. News report has always been the most important way for us to understand society and current affairs. Summarizing the main content of each news report in short language can help readers understand the content faster and save time. At present, automatic text summarization technology is used to generate news content summarization, which mainly includes extractive summarization and abstractive summarization. In this paper, we propose an improved strategy to solve the problem of topic deviation in abstractive summarization method. We combine TextRank with BART model. First, we use TextRank and BART to extract and generate summarization from news text. Then we splice the results of two methods to get the new text which improves the weight of the key sentences in the articles and make it more thematic. Finally, we input the above new texts enter the BART model again to get the final summarization. Experimental results show that compared with single BART model, the average recall scores of Rouge-1, Rouge-2 and Rouge-L are improved by 1.5%, 0.5% and 1.3% respectively.**

***Keywords*—TextRank; Thematic; BART; Automatic Text Summarization; Abstractive Summarization**

I. INTRODUCTION

With the rapid development of the Internet era, a large amount of data has accumulated on the network. How to extract useful information from the massive text has become an urgent problem in the era of big data. News report has always been one of the most important ways for the public to get to know the social development. With the emergence of a large number of news media organizations at home and abroad, a large number of long news reports on various topics have been accumulated on the Internet, and people spend less and less time in the busy and fast-paced work and life, so they will spend less and less energy on long news reading. Therefore, it is convenient for people to read the text in a short text, and it can improve the efficiency of automatic reading. Therefore, the automatic summarization technology for news text proposed in this paper is of great significance.

II. RELATED WORK

Automatic summarization technology can be roughly divided into extractive summarization and abstractive summarization[1], that is, to summarize and analyze a variety of text information, such as magazines, newspapers, news, papers, e-mail, etc., and finally get a short text can reflect the theme of the original text. Extractive summarization is based on the topic of the article, each sentence of the input article is scored, and the first few sentences with the highest score are extracted as the key sentences to form the abstract [2]. The TextRank algorithm proposed by Mihalcea et al. [3] in 2004 is the most classic one in extractive summarization. It introduces the graph model into the field of automatic summarization, calculates the key score of each sentence, and selects the top n sentences as the summary of the article. Zhang et al. [4] proposed a single document extractive summarization method based on the relationship between primary and secondary text. This method constructs a joint learning model of the relationship between primary and secondary text and text summarization based on neural network model. Compared with the current mainstream single document extractive summarization method, this method has significantly improved the rouge evaluation index; Jain et al. [5] proposed a method to extract a group of good features, and then used neural network to extract supervised abstracts, which is better for various online extractive text summarization generators; Sharaff et al. [6] proposed a fuzzy logic extractive text summarization model based on triangular membership function, which is in precision, The performance has been improved under various parameters such as recall and F-Measure. Shi et al. [7] proposed SumRank model. Considering the information features such as sentence position and keywords, based on TextRank algorithm, BERT pre-training model was used to get sentence vector, and MMR algorithm was used to remove redundancy to ensure the accuracy and diversity of results. Experimental results show that the model can effectively extract text summarization. However, extractive summarization technology lacks the understanding of the semantics between the words and sentences in the text, does not consider the relationship between the whole paragraph text structure. Only from the perspective of sentences, the abstracts generated are incoherent and less readable.

Abstractive summarization is a text that can summarize the content of the original text and have smooth sentences after understanding the content of the article through computer algorithm technology. It is not some existing paragraphs or sentences extracted from the source file, but a compressed interpretation of the main content of the document, which may use words that cannot be seen in the source document [8]. Sutskever et al. [9] first proposed a neural network-based sequence to sequence (seq2seq) model; Vaswani et al. [10] proposed a new simple network structure transformer, which is completely based on attention mechanism and provides a powerful algorithm tool for natural language processing; Tan et al. [11] proposed a new graph-based attention mechanism. Compared with the previous neural abstract model, this model can achieve considerable improvement. Hao et al. [12] proposed a feature enhanced seq2seq structure summarization model. The model uses two feature capture networks to improve the encoder and decoder in the traditional seq2seq structure, enhance the ability of the model to capture and store long-term features and global features, make the generated summarization information richer and smoother. Devlin et al. [13] put forward the pre-training model of Bert bidirectional transformer encoder, which makes full use of the context semantic information, and has achieved better results in many downstream tasks of natural language processing. Radford et al. [14] proposed GPT unidirectional transformer encoder model, which adopts pre-training model method and fine-tuning downstream tasks to process NLP tasks like BERT, thus improving the score of related downstream NLP tasks. Then, based on the two pre-training models, there are many fusion algorithm models to deal with the NLP task of automatic summarization. Tan et al. [15] proposed the BERT-PGN model to solve the problem of insufficient understanding of generative sentence context, integrated the pointer generation network into the best. The summary obtained from the experiment has improved the rouge-2 and rouge-4 indicators. For NLU tasks such as sentiment classification, named entity recognition, reading comprehension, etc., the above-mentioned BERT model has achieved good results. However, in NLP field, for sequence-to-sequence natural language generation tasks, such as machine translation, text summarization generation, dialogue generation, etc., only achieve suboptimal results. The strategy of BERT is to train its encoder for NLP to solve this problem, Kaitao Song et al. [16] trained the encoder and decoder separately and proposed the MASS model, which allows the encoder and decoder to learn at the same time in the pre-training stage. It is the first time to realize the unification of the BERT plus generation model, and the rouge score is improved compared with the BERT and other models.

However, the abstracts generated by machine learning or deep learning model may deviate from the theme of the article. The abstracts focus too much on the non key thematic sentence particles in the original article. For example, in a long news story that contains a question-and-answer style of characters' narration and events, the abstractive summarization may concentrate on summarizing the dialogue content of the characters and ignore the core feedback points of the whole event. The BART-TextRank model proposed in this paper introduces the key sentences extracted from the source of the data set, which makes the model better absorb the key sentence particles of the long article, makes the generated abstract more able to reflect the topic of the article and summarize the main idea of the article.

III. BART-TEXTRANK MODEL

Based on the original BART model, the proposed BART- TextRank model introduces the TextRank method to extract the key sentences from the data set, extends the abstractive summarization of BART model to multiple rounds. The summarization generated in the first round is fused with the extracted key sentence text, and then put the new dataset into BART model again to complete the second round of abstractive summarization. This method is divided into the following seven steps: the first step and the second step are to input the text sequence into the TextRank layer and the BART layer at the same time. After the third and fourth steps, we can get the TextRank summarization and the BART summarization. In the TextRank layer, we first vectorize the sentences to get the sentence vector, and then calculate the score of each sentence through the similarity matrix and graph model. Take the top sentences as the summarization. In the BART layer, [CLS] marks the beginning of the sentence and [SEP] marks the end of the sentence, with each word token in the middle. Then the word is embedded and converted into a vector, and then the predicted words are formed by a bidirectional encoder and a decoder from left to right to form a sentence as a summary. In the fifth step, the above two summarization results are combined to form a new data set. Currently, the enhanced data set contains the key sentence particles of the article. In the sixth step, put it into the BART layer again for processing, and in the seventh step, the final summarization is obtained. The network structure of the model is shown in Figure 1 below:

A. Key Sentence Extraction

The first step of BART-TextRank model is to input text sequence into TextRank layer to extract topic sentences and extract key sentence particles. TextRank algorithm is an extractive text summarization algorithm. The core of the algorithm is to introduce the concept of graph sorting. Its basic framework is derived from Google's PageRank algorithm. By dividing an article into several sentence level or word level group nodes into units, using each unit to establish the overall graph model, the similarity between units constitutes the connection nodes of the graph model Finally, the top n sentences are extracted as the key sentences to form the final summarization by sorting the accumulated weights.

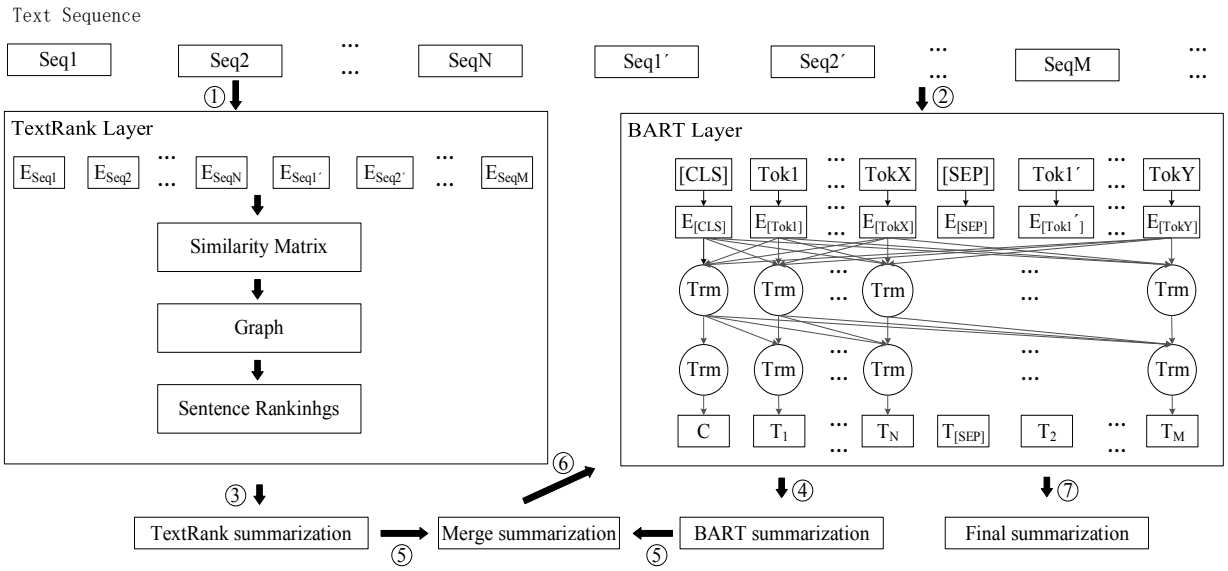

Fig. 1. The diagram of BART-TextRank model network structure

In the TextRank layer, the input text sequence is vectorized to get the sentence level word embedding vector firstly. The sentences in the text are regarded as nodes, and the processed articles are modeled as a graph model. Through a similarity matrix function, the edge values between sentences, that is, between nodes, are calculated. This function is used to weight the edges in graph model. The higher the similarity between sentences, the more important the value of the edges in graph.

TexRank determines the similarity degree of two sentences in a shared text. This overlapping calculation method is very simple, that is, dividing the number of common lexical markers between them by the length of each marker to avoid long sentences. The principle of the algorithm can be expressed as follows:

Definition 1. For two sentence S_i,S_j, the set of n words in a sentence is expressed as $S_i = w_1^i, w_2^i, \dots, w_n^i$. The similarity calculation of two sentences is defined as the follow formula.

$$Sim(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \quad (1)$$

Through this process, we build a dense graph model of a given text containing each sentence, that is, the graph graph modeling operation in the TextRank layer. After the graph model is constructed, the weight of the connecting edge is calculated. The formula for calculating the cumulative weight ws of each node is as follows:

Definition 2. W_{ij} represents the weight of the connecting edge of two nodes in the graph model, which means the similarity between sentences,$In(V_i)$ represents a collection of points that other nodes point to.,$Out(V_i)$ represents a collection of points that point to other nodes.d is damping coefficient, means the probability of a node jumping to another node in the graph model. According to experience, the value is generally 0.85.

$$WS(V_i) = (1-d) + d\sum_{V_j \in In(V_i)} \frac{W_{ji}}{\sum_{V_k \in Out(V_j)} W_{jk}} WS(V_j) \quad (2)$$

After iterative convergence, the final accumulated weight of a node includes the shared weight of itself and other nodes. In the whole operation process, the iterative calculation inputs global text information. If the final weight of a node is high, it not only means that it is a key sentence, but also means that there are more sentences pointing to it. Therefore, the sentence represented by the node is the most reflective of the text theme. Finally, the ranking of sentences is obtained by the importance of sentences, and the top n sentences are selected to form the summarization.

B. Multiple Rounds of Abstractive Summarization Generation

1) First Round Abstractive Summarization

When the text sequence is input into TextRank layer, it is also input into BART layer to get BART summarization. By combining the two models of BERT and GPT, Google team proposed the BART pre-training model [17] in October 2019. By integrating the two models and fine-tuning, it has achieved better results in processing natural language downstream tasks. The framework of BART model is shown in Figure 2 below:

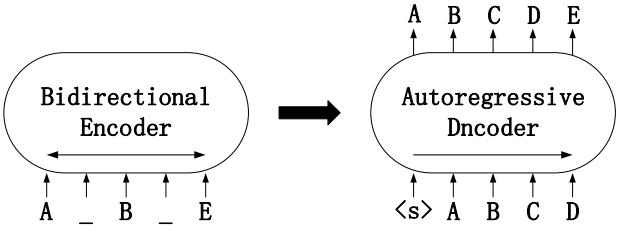

Fig. 2. The framework of BART model

As shown in the figure 2, compared with Bert, Bart's encoder makes some changes. Its input does not need to align the output of the decoder one by one. That is to say, it allows the text to be destroyed by any noise, the mask replaces the original word order of the text, or the text segment is rotated and omitted to destroy the input text. The original document in the figure 2 is [A,B,C,D,E], the text [C,D] is masked before encoding, and an additional mask is inserted before B, finally the damaged document [A( )B( )E] is put into the decoder. Firstly, the damaged text is encoded by a bidirectional encoder, and then the original document is reconstructed by a left to right autoregressive decoder. In BERT, the mask mechanism is used to replace any token in the sequence, because the traditional recurrent neural network processing method tends to process the last time information of the whole time line, and is not sensitive to the initial input information, so the encoder uses bidirectional coding [16]. Since the token to be masked is predicted separately, the input prediction sequence is sequential. This prediction method is effective for some tasks that allow to use the information after the I position to predict the I position. However, it is not effective for text generation tasks such as summary generation, because the prediction result needs to use the previous information.

2) Topic Abstractive Summarization Generation

The processed TextRank summarization and BART summarization are merged one by one, that is, each text in the newly synthesized text sequence contains the key sentence particles extracted from the source text, and the preliminary concise smooth summarization obtained from the generative model, which makes the merged data set fuller and fuller, and covers the whole of the source text more key sentence level particles are added, and then the synthetic text data set is sent to the BART layer for the second summary generation, so that the key sentences extracted by TextRank can penetrate into the summarization generated by the first round of BART, and help the model refine the weight of key sentences again, making the final summarization more thematic than the original.

IV. EXPERIMENT AND ANALYSIS

A. Experimental Steps

For extractive summarization, the result only extracts a few sentences which are most suitable for the topic in the source text, which can not summarize the meaning of the whole text, and has the problem of unsmooth sentences, which is not in line with people's reading habits. The strategy adopted in this paper is to add the TextRank extracted results to the input of BART model generated summarization, and to generate the summarization by artificially guiding the text to be closer to the topic direction. The operation process is shown in the figure 3 below:

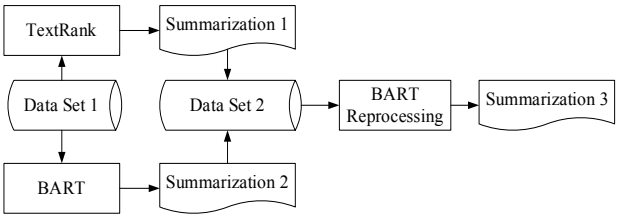

Fig. 3. The diagram of experiment steps

In this paper, CNN / Daily Mail data set is used. Firstly, the data set is processed by TextRank and BART to get two different styles of summarization 1 and summarization 2. Summarization 1 is to extract the most representative sentences of the article as the guiding clue, and summarization 2 is to generate the sentences that summarize the content of the article, and then the two are combined into data set 2. At this time, there are some key sentence clues with guiding theme in the dataset and then more thematic summarization can be generated by BART. The idea of mixing the two different styles of results as a new data set perfectly fits the structure of BART's encoder. For the input text, the model accepts the data processed by arbitrary noise, that is to say, it allows any processing such as destroying and modifying the input source to increase the diversity of the input end and improve the training effect of the model. The mixed new data set means that the original article will be changed according to the data set, some new data after processing are spliced, and the spliced data is the closest sentence from the original article, which makes the whole data segment reflect the more value of the original text and make it more efficient meanwhile.

B. Experimental Data

This experiment uses the official CNN / Daily Mail data set as a single text summarization corpus, and each news article provides several artificially extracted sentences according to the order of the original sentences as a standard summary reference. The original data set contains about 28W news articles, which can be processed into training text, test text and verification text in TXT format. In this experiment, 13760 test news articles are selected for the experiment.

C. Evaluating Indicator

Scientific evaluation indicator affects the research and development of the whole field of natural language processing. At present, the most commonly used evaluation indicator in the field of automatic text summarization is ROUGE, which evaluates the quality of the summarization generated by calculating the different basic units overlapped between the summarization generated by the model and the standard manual summarization. In this paper, the average recall rate of rouge-1 average-r, rouge-2 average-r and rouge-l average-r is used as the evaluation indicator to evaluate and analyze the experimental results.

D. Experimental Environment and Parameter Settings

In this paper, the experimental hardware environment is archlinux system, 16g RAM, single GTX-1650super (GPU) graphics card. The model used in the experiment is bart.large.cnn.The experiment parameters are set as batch size 10, hidden layer 768, beam search parameter 3 and maximum sequence length of 140, minimum sequence length of 55, and the total running time is about 6.5 hours. The ratio of sentences extracted from TextRank layer is 0.1, because the dataset used in this paper is long news texts, so the number of sentences extracted should not be too many, and a certain number should be controlled.

E. Experimental Results and Analysis

The experiment steps of this paper are as follows. Firstly, 13760 pieces of data are separately processed by TextRank and BART to get their respective summarization, then the average recall scores of rouge-1, rouge-2 and rouge-l are calculated by ROUGE scoring system, and then the corresponding scores are calculated by ROUGE scoring system combined with the optimized strategy. In addition, this paper also sets up a set of additional comparative experiments, which directly merge the summarization extracted by TextRank with the source data set, and then send them to the BART layer to generate the final summarization, and calculate their corresponding scores. The experiment results are shown in TABLE 1 below:

TABLE I. EXPERIMENT RESULTS

Indicator	*TextRank*	*TextRank-source*	*BART*	*TextRank-BART*
Rouge-1 Average-R	21.882	36.727	38.33	39.882
Rouge-2 Average-R	9.35	16.648	18.07	18.467
Rouge-L Average-R	19.769	33.773	35.53	36.813

In TABLE 1, TextRank is the result of summarization 1, BART is the result of summarization 2, and TextRank-BART is the result of summarization 3. It can be seen from the table that after the optimization strategy of the data set, the final results showed that the Average-R values of Rouge-1, Rouge -2 and Rouge -L are improved. Compared with BART, Rouge -1 / Average-R is improved by 1.5, Rouge-2 / Average-R is improved by 0.5, Rouge-L / Average-R is improved by 1.3, which shows that the average recall rate of binary words and the longest common subsequence matching is improved under Rouge score.It is proved that after the introduction of TextRank summarization into the dataset, the number of overlapping basic units between the final summarization and the standard summarization is increased according to the above three calculation methods, the experiment results are closer to the standard summarization set, and the summarization of the full text understanding topic is more fully and accurately. It also shows that the optimization strategy of this paper has a certain improvement in the final recall rate compared with the two separate automatic summarization methods.

However, the score of additional comparative experiment is lower than BART model and TextRank-BART model proposed in this paper, which proves that adding the summarization extracted by TextRank to the source data set alone can not improve the score. This method only improves the frequency of key sentences in the data set, and does not extract the key sentences from the real semantic point of view, so we need to put the source data set into the BART layer to generate a summarization, and then merge it with the summarization extracted by TextRank to achieve the effect, which proves the rationality and effectiveness of the model proposed in this paper.

V. CONCLUSIONS

Modern Internet as the media and entertainment industry, has more and more influence on people's work and life. The huge user group makes the Internet accumulate numerous overloaded information. The smooth, concise and automatic summarization in line with the theme of the original text is becoming more and more important for improving the efficiency of people's daily business processing and browsing information. It can not only eliminate the redundancy of long text news, but also provide accurate decision-making analysis assistance for people to judge public opinion information, so that the summarization can better serve human society and make people better adapt to the current fast-paced life.

In this paper, we propose a strategy of adding TextRank extractive algorithm to the data set to generate summarization, combining with BART pre-training model to generate summarization. We use the human guided algorithm to generate more thematic summarization of the original article, and realize an optimization strategy of enhancing the ability to guide the data set to generate summarization. The experiment results show that the average recall scores of Rouge-1, Rouge-2 and Rouge-L summarization are improved, which can better summarize the ideas and main contents of the original text.

REFERENCES

[1] Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review, 2017, 47(1): 1-66.

[2] Nenkova A, McKeown K. A survey of text summarization techniques[M]//Mining text data. Springer, Boston, MA, 2012: 43-76.

[3] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. 2004: 404-411.

[4] Ying Z , Zhongqing W , Hongling W . Single Document Extractive Summarization with Satellite and Nuclear Relations[J]. Journal of Chinese Information Processing, 2019.

[5] Jain A, Bhatia D, Thakur M K. Extractive text summarization using word vector embedding[C]//2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 2017: 51-55.

[6] Sharaff A, Khaire A S, Sharma D. Analysing Fuzzy Based Approach for Extractive Text Summarization[C]//2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019: 906-910.

[7] SHI Yuan-bing,ZHOU Jun,WEI Zhong. TextRank-based Chinese Automatic Summarization Method[J].Communications Technology,2019,52(09):2233-2239.

[8] Nallapati R, Zhou B, Gulcehre C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond[J]. arXiv preprint arXiv:1602.06023, 2016.

[9] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[11] Tan J, Wan X, Xiao J. Abstractive document summarization with a graph-based attentional neural model[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1171-1181.

[12] Hao Z, Ji J, Xie T, et al. Abstractive Summarization Model with a Feature-Enhanced Seq2Seq Structure[C]//2020 5th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS). IEEE, 2020: 163-167.

[13] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[14] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[15] Jinyuan T, Yufeng D, Ruihua Q, et al. CCDM2020+ 62: Automatic summary generation of Chinese news text based on BERT-PGN model [J]. Journal of Computer Applications, 2020: 0-0.

[16] Song K, Tan X, Qin T, et al. Mass: Masked sequence to sequence pre-training for language generation[J]. arXiv preprint arXiv:1905.02450, 2019.

[17] Lewis M, Liu Y, Goyal N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv preprint arXiv:1910.13461, 2019.