2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP) | 978-1-6654-7857-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSP54964.2022.9778497

A Text Structure-based Extractive And Abstractive Summarization Method

Jing Yan
Key Laboratory of Advanced Design and Intelligent Computing, Ministry Of Education, School of Software Engineering, Dalian University
Dalian, China
948893461@qq.com

Shihua Zhou*
Key Laboratory of Advanced Design and Intelligent Computing, Ministry Of Education, School of Software Engineering, Dalian University,
Dalian, China
* zhoushihua@dlu.edu.cn

***Abstract*—Extraction summarization and abstraction summarization have advantages and disadvantages, so how to better combine these two ways has become a difficult problem. To address this challenge, this paper proposed a new fusion method. The major novelty lies in the design of the new method. Wherein our approach, first, use the text structure information and the idea of the K-means algorithm to divide the text into regions, and then the main part and the non-main parts are determined according to the distribution of the subject words. Next, apply information extraction on the main part and text generation on the non-main parts. Finally, the two parts of the summarization are merged according to the sequence of the text. Experimental results show that the quality of the summarization is better than that of extraction summarization and abstraction summarization. In addition, to make the method more targeted in Chinese text processing, the Cw2vec model based on Chinese stroke information is used in the encoding process, and the experiment proves that the quality of summarization can be further improved.**

Keywords- text summarization; Seq2Seq; Cw2vec; the article structure

I. INTRODUCTION

On December 1, 2019, the first case of COVID-19 appeared in Wuhan. While COVID-19 affects people's physical health and even threatens people's lives, it has also brought huge economic losses to the government. Relying on the development of Internet technology, overwhelming text information about COVID-19 is more likely to be published on Websites and Apps, which has shown exponential growth in recent months. Text summarization is a natural language processing task. The long text is compressed, summarized, and then converted into concise, generalized short text, so the information is presented to the user in a timely and convenient manner, which greatly saves the user's browsing time and helps users obtain important information in a large number of epidemic text data.

Luhn of IBM proposed a text summary method based on high-frequency word scoring [1], which opened a precedent for Text Summarization. At present, there are two Text Summarization methods: Extraction-based and Summarization [2]. Extraction-based Summarization is to directly extract some important sentences from the long text. It is mainly represented by the Textrank algorithm [3]. Abstraction-based Summarization is to understand the original text and then rewrite the sentence. This method is mainly based on the Seq2Seq model [4]. Later incorporated the idea of an attention mechanism [5]. In recent years, more technologies have begun to emerge based on neural networks [6-9]. The pre-trained model Transformer also has good results [10].

Extraction-based Summarization has problems such as poor coherence between sentences and unclear target sentence subject matter, while Abstraction-based Summarization has rich semantic information and high sentence readability. Therefore, the method that combines the advantages of the two has naturally become the focus of scholars. At present, the method based on the extractive and Abstractive often divides The Summarization task into two sub-task modes, use the extractive method to locate and extract the key content, and then use the extractive method to rewrite the content. This method is ultimately based on Extraction, because the information of the full text is not utilized.

According to the advantages and disadvantages of extractive and Abstractive summarization techniques, this paper considers the article structure and proposes an extractive. In the article, especially in the news, the sentences at the beginning and the end of the paragraph contain more textual theme information, not only that, many Chinese texts have a strict writing structure, especially for news, usually divided into total-points, points-total, points-total-three categories, different parts contain the main idea of the strength of different. Therefore, according to the strength of the main idea, different sections of the article can be summarized in different ways.

When encoding sentences, the main word vector models used are based on Western languages, such as the Word2vec model, Bert model, etc. These Western languages are composed of Latin letters, which are completely different from Chinese writing. Chinese words have homophones. In scenarios such as typos, the Cw2vec model [11] can capture the semantic and morphological information of Chinese vocabulary through stroke information.

The algorithm calculates the distance between samples by Euclidean distance.

A. Seq2seq

The Seq2Seq [4] model, also known as the Sequence To Sequence model, its essential idea is that the sequence is converted into a fixed-length vector input into the encoder, and then the vector that needs to be decoded in the model is decoded and output through the decoder, basically it is constructed by the encoder-decoder framework, the encoder can be any model and data, the decoder can also be an arbitrary model, for the text summary, the traditional model needs to keep the input and output consistent, The Seq2Seq model can perform inputs and outputs with inconsistent lengths. Fig.1 is the structure of the Seq2Seq model.

At present, there are many Seq2Seq-based models, mainly based on RNN and LSTM, as well as very popular pre-trained models, this paper uses seq2Seq models based on LSTM and pre-trained model Transformer [10].

B. *Cw2vec*

The word vector models are all based on Western languages. Therefore, how to effectively use Chinese character-specific information to train word vectors has become a hot topic [13, 14]. Cao et al. [11] believed that it was not enough to capture enough internal information of Chinese characters, and took advantage of the writing rules in the Chinese writing system, and designed a word embedding model Cw2vec using these stroke features, capture information of vocabulary hidden in Chinese strokes by using strokes.

The Cw2vec model divides Chinese strokes into 5 categories, namely horizontal, vertical, skimming, stroking, and folding, and assigns an integer ID to each of these five categories of strokes, which are 1, 2, 3, and 4, respectively., 5. The specific corresponding methods are shown in the table1 [11].

TABLE I. STROKES-ID [11]

Strokes name	Strokes	ID
horizontal	一	1
vertical	丨（亅）	2
skim	丿	3
Squeeze	㇏（丶）	4
Fold	㇄	5

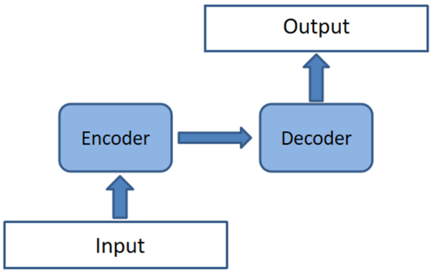


Figure 1. Structure diagram of the Seq2Seq model.

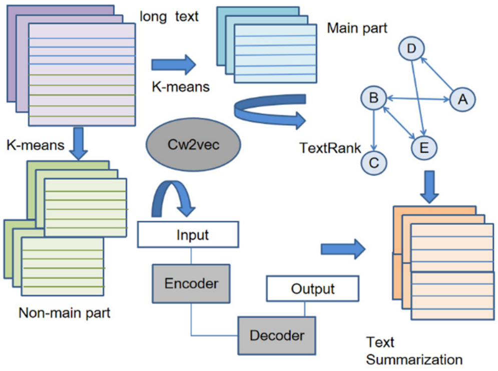


Figure 2. The structure of the algorithm.

II. AN EXTRACTIVE AND ABSTRACTIVE SUMMARIZATION ALGORITHM

To realize the integration of Extraction-based Summarization and Abstraction-based Summarization, divide chapters into parts, and then the main part is identified according to the subject words contained in each part. Then determine the structure of the article according to the location of the subject section. When generating the abstract, the extractive method Textrank is used to extract the main sentence for the main part of the article, and then the non-main parts are disposed by the abstractive method based on the seq2seq model of deep learning, finally, the two parts of the abstract are spliced according to the text order to form the final abstract. The system frame diagram is shown in the Fig.2.

A. *Divide Chapters into Parts*

The result of clustering using the K-means is that the samples in the category are messy and disordered. To ensure that the article is divided into three parts, the k-means algorithm is improved as follows: in the process of clustering, the center of the cluster is the midpoint, and the article is randomly divided into three parts according to the center point, and the distance sum of all samples in each category from the center is calculated, and readjusted. Center point and divide the area until the sum of the distances of all elements in all classes from the center point is the smallest. The internal distance of the three parts divided by the article is minimized, that is, the similarity between sentences within each part is maximized.

B. *Locate the Main Part*

After dividing the text, according to the distribution of the subject words, the area where the main idea part is located is determined. This paper using the TF-IDF (Term Frequency-Inverse Document Frequency) to count Subject words.

$$TF_{i,j}-IDF_i=\frac{n_{i,j}}{\sum_k n_{k,j}}-\log\frac{|D|}{1+|j:t_i\in d_j|} \tag{1}$$

$n_{i,j}$ is the number of occurrences, $TF_{i,j}$ is the frequency of occurrences. $|D|$ is the number of all documents, IDF_i is inverse document frequency.

When performing keyword statistics, instead of using the number of subject words, we use the density of subject words.

C. Coding with Cw2vec

To effectively use the stroke information hidden in the Chinese characters, the Chinese characters are split into stroke information, and the sentence sequence is converted into a stroke sequence according to the stroke-ID correspondence table, and then the stroke sequence information is used as input to start training the Cw2vec model. The trained Cw2vec model generates word embedding forms that incorporate stroke information.

D. Different Processing for Different Parts

When generating the abstract, the extractive method Textrank is used to extract the main sentence for the main part of the article, and then the non-main part is subjected to generative abstract processing based on the seq2seq model of deep learning. This paper applies the model based the LSTM and the pre-trained model Transformer [10].

E. Combine Summaries

After treating different parts in different ways, the two parts of the summarization are spliced according to the text order to form the final summarization.

F. Data Sets

At present, the data sets used by text summarization technology are mainly divided into manually written DUC data sets and Gagiword data sets. This article uses the second data set, crawled the three thousand news of the new coronary pneumonia on People's Daily Online.

III. Results & Discussion

To evaluate the quality of the method combined extraction-based summarization and abstraction-based summarization, three experiments were conducted in this paper, The ROUGE method is used as the evaluation method in this paper [15].

Firstly, to verify that the fusion method is superior to the separate method, the method in this article is compared with the Extractive and Abstractive methods, respectively. The Textrank algorithm, the Sq2seq model based on LSTM, and the transformer based on the pre-trained model are used in this paper. The article is represented by Textrank, LSTM, and Transformer. The algorithms of the two fusion methods are represented by LSTM+Textrank and Transformer+Textrank, respectively.

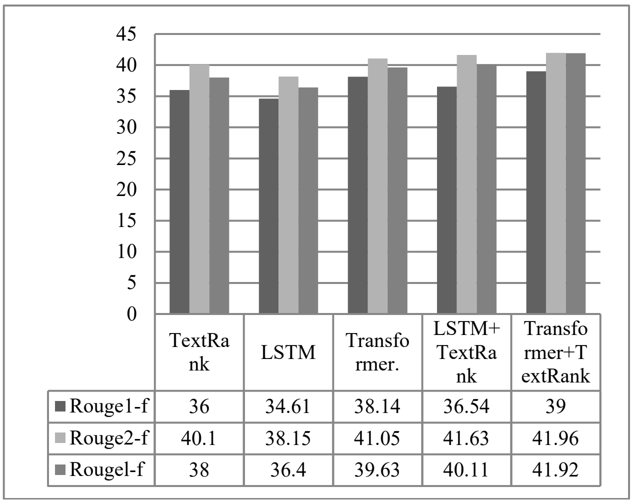
Secondly, to verify the validity of the blended stroke information, after the above experiments, Cw2vec based on the stroke information is used in the coding stage, and the resulting abstract is evaluated for quality. The two algorithms incorporating stroke information are represented by LSTM+Textrank* and Transformer+Textrank*, respectively.

Finally, to verify the accuracy of the positioning of the subject part, the text uses a cross-validation method, using the TextRank algorithm for extraction of the two non-subject parts, generative processing of the remaining two parts, and after the abstract is stitched, the abstract quality is evaluated. The two cross-validation methods are represented by Text1, Text2, Text3, and Text4 respectively.

The results are shown in Figures. In Fig.3, the Rouge values of the two methods in this paper are all higher than those of the extractive method and the two abstractive algorithms, and the fusion method using the pre-training model has a higher score.

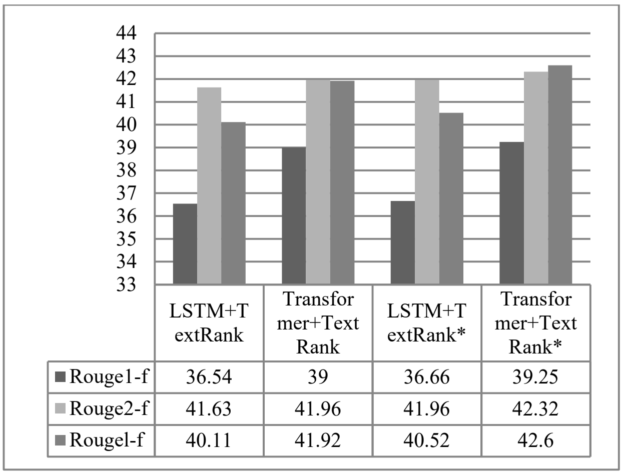
In Fig.4, the scores of the method using the Cw2vec model are higher. There was an average improvement of 0.18 on Rouge-1, 0.35 on Rouge-2, and 0.55 on Rouge-l.

In Fig.5, the fusion method is higher than the other two test algorithms in Rouge values.



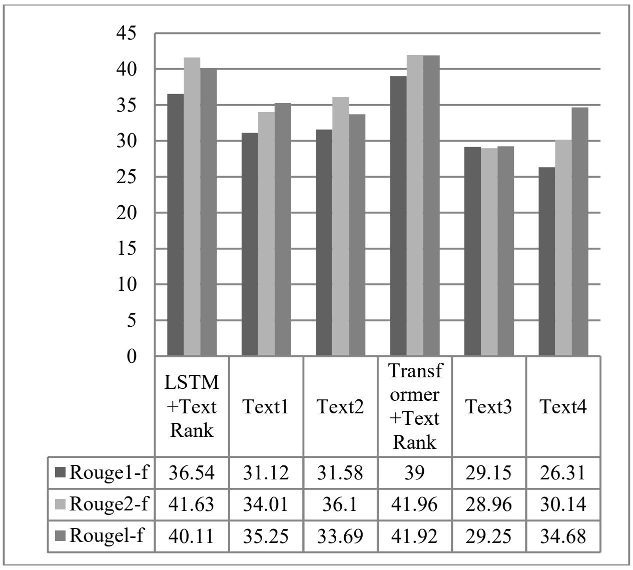
	TextRank	LSTM	Transformer.	LSTM+TextRank	Transformer+TextRank
■ Rouge1-f	36	34.61	38.14	36.54	39
■ Rouge2-f	40.1	38.15	41.05	41.63	41.96
■ Rougel-f	38	36.4	39.63	40.11	41.92

Figure 3. ROUGE values for the five methods.



	LSTM+TextRank	Transformer+TextRank	LSTM+TextRank*	Transformer+TextRank*
■ Rouge1-f	36.54	39	36.66	39.25
■ Rouge2-f	41.63	41.96	41.96	42.32
■ Rougel-f	40.11	41.92	40.52	42.6

Figure 4. ROUGE values for the four methods.



	LSTM +Text Rank	Text1	Text2	Transformer +Text Rank	Text3	Text4
■Rouge1-f	36.54	31.12	31.58	39	29.15	26.31
■Rouge2-f	41.63	34.01	36.1	41.96	28.96	30.14
■Rougel-f	40.11	35.25	33.69	41.92	29.25	34.68

Figure 5. ROUGE values for the six methods.

IV. CONCLUSIONS

By evaluating the quality of the generated summary, the following conclusions can be drawn:

By using the chapter structure of the article, the combination of extraction and generative methods is realized, and the results are proved to be better than the extraction and generative methods through experimental verification.

After improving the coding process of the model, it is proved that after using Cw2vec, the method can effectively use the stroke information hidden in the text, which improves the quality of the abstract to a certain extent.

Through cross-verification of the structure of the chapter, the accuracy of the positioning of the main point of the article is proved.

REFERENCES

[1] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of research and development, 2(2), 159-165.

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

[3] Mihalcea, R., Tarau, P. (2004) Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona, Spain. pp. 404-411.

[4] Sutskever, I., Vinyals, O., Le, Q.V. (2014) Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. Cambridge, MA. 3104-3112.

[5] Rush, A. M., Chopra, S., Weston, J. (2015) A neural attention model for abstractive sentence summarization. https://arxiv.53yu.com/abs/1509.00685.

[6] S e, A., Liu, P. J., Manning, C. D. (2017) Get to the point: Summarization with pointer-generator networks. https://arxiv.53yu.com/abs/1704.04368.

[7] Paulus, R., Xiong, C., Socher, R. (2017) A deep reinforced model for abstractive summarization. https://arxiv.53yu.com/abs/1705.04304.

[8] Liu, L., Lu, Y., Yang, M., Qu, Q., Zhu, J., Li, H. (2018) Generative adversarial network for abstractive text summarization. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1). New York.

[9] Celikyilmaz, A., Bosselut, A., He, X., Choi, Y. (2018) Deep communicating agents for abstractive summarization. https://arxiv.53yu.com/abs/1803.10357.

[10] Lee, H., Choi, Y., Lee, J. H. (2020) Attention history-based attention for abstractive text summarization. In Proceedings of the 35th Annual ACM Symposium on Applied Computing. New York. pp. 1075-1081.

[11] Cao, S., Lu, W., Zhou, J., Li, X. (2018, April) cw2vec: Learning Chinese word embeddings with stroke n-gram information. In: Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1). New York.

[12] Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29(3), 433-439.

[13] Yin, R., Wang, Q., Li, P., Li, R., Wang, B. (2016) Multi-Granularity Chinese Word Embedding. In: Empirical methods in natural language processing. Austin, Texas.

[14] Xu, J., Liu, J., Zhang, L., Li, Z., Chen, H. (2016) Improve Chinese Word Embeddings by Exploiting Internal Structure. In: North American chapter of the association for computational linguistics. San Diego, California.

[15] Lin C. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In: Meeting of the association for computational linguistics. Barcelona, Spain.