HiLL PUBLISHING

Advances in Computer and Communication, 2024, 5(2), 141-147
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Mongolian Automatic Text Summarization Method Based on Pre-trained Model and Improved TextRank

Yongshun Han, Qintu Si*, Siriguleng Wang

College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, Inner Mongolia, China.

How to cite this paper: Yongshun Han, Qintu Si, Siriguleng Wang. (2024) Mongolian Automatic Text Summarization Method Based on Pre-trained Model and Improved TextRank. *Advances in Computer and Communication*, 5(2), 141-147.
DOI: 10.26855/acc.2024.04.008

Received: March 22, 2024
Accepted: April 19, 2024
Published: May 16, 2024

***Corresponding author:** Qintu Si, College of Computer Science and Technology, Inner Mongolia Normal University, Hohhot, Inner Mongolia, China.

Abstract

At present, there is limited research on automatic Mongolian text summarization, especially using mainstream methods. The existing TextRank algorithm only considers the similarity between sentences, ignoring the characteristics of the sentence itself. In this paper, a Mongolian automatic text summarization method called IMNUBERT-mnTextRank, based on a pre-trained model and an enhanced TextRank algorithm, is proposed. The information from the Mongolian external knowledge base is incorporated into the TextRank algorithm in the form of sentence vectors to enhance the accuracy of similarity calculations between sentences. The process of calculating sentence weights is optimized by considering sentence features such as sentence position, similarity to the title, keyword coverage rate, and Mongolian conjunctions. Finally, the weight of each sentence is obtained through algorithm iteration. After sorting the sentences, the top two are selected for the summary. Experimental results show that, compared with the TextRank algorithm, the Rouge-1, Rouge-2, and Rouge-L indicators of the proposed method have improved by 0.183, 0.179, and 0.199, respectively. Consequently, the quality of the generated Mongolian summarization is enhanced.

Keywords

Text summarization, Mongolian, Pre-trained model, TextRank

1. Introduction

Faced with the rapid growth of Mongolian information, how to quickly and accurately extract key information from a large number of Mongolian information resources, reduce the reading burden of users, and improve the acquisition speed of Mongolian information has become a problem to be solved. Automatic text summarization technology is an important solution to solve this problem. Automatic text summarization can be generally divided into abstractive summarization and extractive summarization [1]. The former redescribes the text by mining the deep semantic information, and the latter generates the summary from the sentences in the original text. In the current research on automatic text summarization, the methods based on graph models are widely used in automatic text summarization tasks [2] because they do not need to manually label the training data set and have good performance. In 2004, Mihalcea [3] proposed TextRank algorithm based on the idea of Google's PageRank algorithm. Wang Xuxiang [4] proposed SW-TextRank, an automatic text summarization extraction algorithm based on improved TextRank, which improved the quality of the summary. Li Wei [5] improved the quality of Tibetan automatic text summarization by combining TextRank with word embedding. Zhu Bingbing [6] proposed pTextRank, an automatic text summarization algorithm based on clause extraction, which effectively solves the redundancy problem. Xu Fei [7] proposed an algorithm for abstract extraction of news articles based on the analysis of the structure of news articles.

Research on automatic Mongolian text summarization is still in its infancy. Under the guidance of natural language understanding based on full information theory, Li Chengcheng [8] proposed a domain-specific automatic summarization system for Mongolian text. At present, there are few researches on automatic Mongolian text summarization using mainstream methods. To this end, this paper uses the traditional Mongolian pre-trained model as an external knowledge base to construct sentence feature vectors, and introduces sentence feature information such as sentence position, sentence and title similarity, Keyword coverage rate, and Mongolian conjunctions on the basis of TextRank algorithm to improve the shortcomings of the original algorithm, and proposes the IMNUBERT-mnTextRank method. The experimental results show that the proposed algorithm has better performance, and the generated Mongolian summarization is of better quality.

2. Algorithm Design

In the task of automatic text summarization, TextRank algorithm splits the text information into multiple sentences and takes the sentences as network nodes, and the similarity between the sentences as the edges of the network graph, so as to build the network graph. Through iterative calculation, the importance score of each sentence is obtained and sorted to obtain the summary. The TextRank algorithm also has shortcomings. When extracting the summary, the TextRank algorithm relies on the calculation results of the similarity between sentences, and the accuracy of the similarity calculation between sentences affects the quality of the extracted summary to a certain extent. On the other hand, the TextRank algorithm does not consider the characteristics of the sentence itself and initializes all the sentence nodes uniformly during the iterative calculation, which leads to the fact that the TextRank algorithm cannot perceive the differences between each sentence well. To this end, this paper improves the TextRank algorithm and proposes IMNUBERT-mnTextRank.

2.1 Sentence similarity calculation

In the task of automatic text summarization, the weights between the sentence nodes in the network graph of the TextRank algorithm play a crucial role, which directly affects the final score of each sentence and thus has a profound impact on the quality of the generated text summarization. The calculation of these weights depends on the similarity between sentences, so the accuracy of similarity calculation is crucial to the performance of the algorithm. In order to improve the accuracy of sentence similarity calculation, external language knowledge bases are widely used to represent more knowledge of the language itself in the form of sentence vectors, such as Word2Vec, Doc2Vec, BERT, etc. Among them, BERT [9] uses deep bidirectional language representation, can consider context information, generate dynamic sentence vectors, and performs better in capturing the semantic relationship between sentences. By combining external knowledge bases, sentences are represented as vectors, and calculating the cosine similarity of sentence vectors becomes a more reliable and accurate method. This helps TextRank to better understand the semantic relationship between sentences in the text, and ensures that the weights between sentence nodes can more accurately reflect their similarity, thereby improving the quality of text summarization. The cosine similarity is calculated by formula (1). x and y represent sentence vectors, n is the dimension of the sentence vector, x_i is the value in the i'th dimension of x, y_i is the value in the i'th dimension of y.

$$\cos(x, y) = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}$$

In this paper, the Mongolian pre-trained model (IMNUBERT) trained in our laboratory is used as the external language knowledge base, and the Mongolian language knowledge is integrated into the TextRank algorithm in the form of sentence vectors to improve the accuracy of sentence similarity calculation. The acquisition process of sentence vectors is shown in FIG 1.

2.2 Sentence feature weight calculation

TextRank algorithm is widely used to evaluate the importance of a sentence in an article, but its limitation is that it assigns the same initial weight value to all sentence nodes in the initial stage. This generalized approach fails to fully

consider the importance of different sentences in the text. Therefore, this paper introduces four new sentence features: sentence position, sentence-title similarity, Keyword coverage rate, and Mongolian conjunctions. By comprehensively considering these four new sentence features, the initial weight of each sentence node is recalculated to more accurately reflect the importance of the sentence in the article.

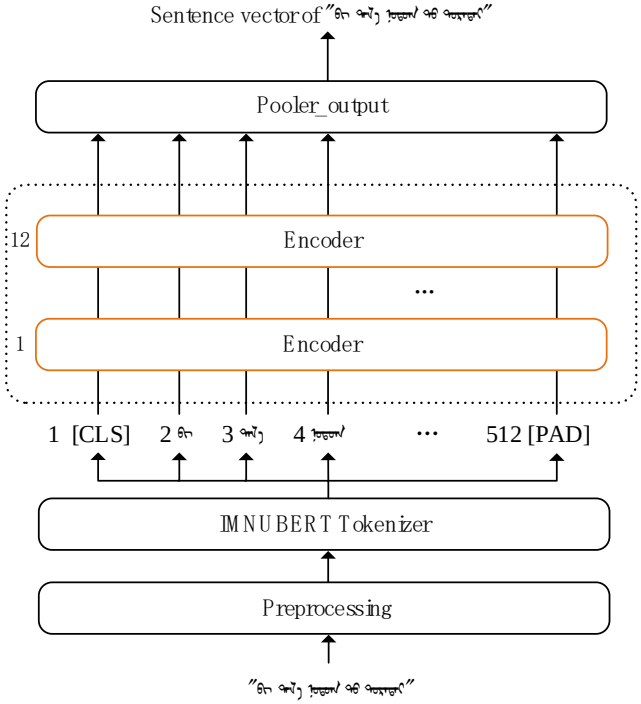


Figure 1. Sentence vector acquisition process.

2.2.1 Sentence position

Sentence position plays a crucial role in the process of abstract extraction. According to the results of scientific research, the first sentence is selected as the abstract in 85% of the manual abstract extraction, and the last sentence is selected close to 7% [10]. Especially in news and other types of articles, the first sentence often explains the main content of the article, and the last sentence usually summarizes the article reasonably. Therefore, this paper dynamically assigns sentence weight according to the position of the sentence in the article. That is, the earlier sentences are given more weight because they are more likely to contain key information. As the sentence's position in the article structure decreases, the weight gradually decreases. At the same time, the final sentence is also given a certain weight. The sentence position feature weights are calculated by formula (2). i represents the sentence sequence number, x represents the total number of sentences in the article, and e is the weight adjustment threshold, which is set to 1 in this paper.

$$W_{p,i} = \begin{cases} 1, & i = 0 \\ e - i \times \frac{e}{x}, & 1 \le i \le x-1 \\ 0.5, & i = x \end{cases} \tag{2}$$

2.2.2 Sentence and title similarity

The title plays a crucial role in the article, especially in the news content. It is the facade of the article and has the function of reflecting the main idea and content of the article. Therefore, the similarity between the sentence in the article and the title becomes a key indicator of the importance of the sentence. A higher similarity score means that the sentence is more relevant to the main content of the article and should receive a higher weight. This measure of relatedness helps ensure that the generated text summary better reflects the main points of the article. In this paper, cosine similarity is used to calculate the similarity between the sentence and the title. Different similarity results will

lead to different weights for each sentence.

2.2.3 Keyword coverage rate

The topic content of an article is often reflected by keywords. The more keywords appear in a sentence, the higher their importance. In the text preprocessing stage, the TF-IDF method is used to extract keywords from the text, and these keywords are used to calculate the Keyword coverage rate of each sentence. The weight of a sentence is positively correlated with the keyword coverage rate, that is, the higher the coverage rate is, the more keywords the sentence contains, the greater its weight is. Its weight is calculated by formula (3). $keywords(s_i)$ denotes the number of keywords contained in the sentence s_i and $len(s_i)$ denotes the number of words contained in the current sentence.

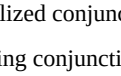
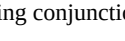
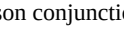
$$W_{k,i} = \frac{keywords(s_i)}{len(s_i)} \tag{3}$$

2.2.4 Mongolian conjunctions

Mongolian conjunctions have no specific lexical meaning, only grammatical conjunction meaning. Among the Mongolian conjunctions, the sentences containing general conjunctions, turning conjunctions and reason conjunctions are often summaries of the previous text or lead to more important sentences. Therefore, such sentences should be given higher weight. The weight is calculated by formula (4).

$$W_{c,i} = \begin{cases} 1, & si \text{ include mongolian conjunctions} \\ 0, & si \text{ not include mongolian conjunctions} \end{cases} \tag{4}$$

Table 1. Some Mongolian conjunctions

Types of Mongolian conjunctions	Mongolian conjunctions	Meaning
Generalized conjunctions		Namely, so, in fact
Turning conjunctions		but, yet, however
Reason conjunctions		therefore, because, due to

2.3 IMNUBERT-mnTextRank

In this paper, the method of weighting sentence features described above is called mnTextRank algorithm. The calculated by formula (5).

$$W_i = (W_p + W_t + W_k + W_c) \times l \tag{5}$$

W_p is the sentence position feature weight; W_t is the feature weight of sentence and title similarity; W_k is the feature weight of Keyword coverage rate; W_c is the Mongolian conjunctions feature weight; l is the weight value amplification coefficient, which is used to enlarge the sentence weight value calculated, which is conducive to the iterative calculation and convergence of the algorithm, and is set to 1000 in this paper. On the basis of mnTextRank algorithm, the method of introducing the Mongolian external language knowledge base is called IMNUBERT-mnTextRank. The implementation process of this method is as follows.

Input: The original Mongolian text that needs to be abstracted.

Output: Extracted Mongolian summary.

Step 1: Perform data preprocessing on the original Mongolian text.

Step 2: Vectorize the sentence representation by the Mongolian IMNUBERT pre-trained model.

Step 3: Calculate the sentence similarity matrix by vector cosine similarity.

Step 4: Calculate sentence feature weights such as sentence position, sentence, and title similarity, Keyword coverage rate, and Mongolian conjunctions.

Step 5: Update the inter-sentence similarity matrix.

Step 6: Build a text network graph according to the sentence similarity matrix**,** and calculate it iteratively by TextRank algorithm.

Step 7: Select the Top-2 sentences according to the sentence score ranking, and adjust the sentence order according

to the sentence position information in the original text.

Step 8: Output the summary.

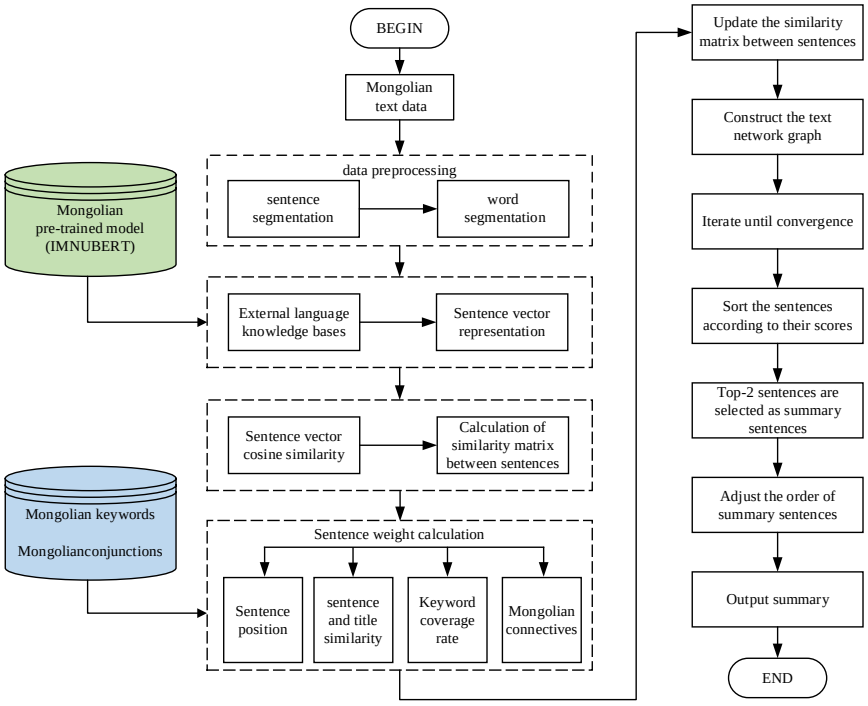


Figure 2. IMNUBERT-mnTextRank process.

3. Experiments

3.1 Experimental data and evaluation criteria

At present, most of the public text summarization datasets are composed [11] of "article-title". However, Mongolian automatic text summarization research is still in its infancy, and there is no public datasets. Therefore, this paper adopts the "article-title" method to collect news data from existing Mongolian news websites and uses the title of the news as the reference summary to construct a dataset of 1000 Mongolian automatic text summarization. The evaluation standard of automatic summarization is Rouge [12], an automatic evaluation method. In this paper, three evaluation indexes Rouge-1, Rouge-2 and Rouge-L are selected to evaluate the quality of the abstract generated by the algorithm.

3.2 Experimental results

In order to verify the effectiveness of mnTextRank algorithm proposed in this paper, experiments are carried out by introducing sentence position, sentence, and title similarity, Keyword coverage rate, and Mongolian conjunctions. The experimental results are shown in Table 2.

Table 2. Experimental results of mnTextRank algorithm

Algorithm	Rouge-1	Rouge-2	Rouge-L
TextRank	0.272	0.162	0.173
TextRank+ Sentence position	0.385	0.276	0.266
TextRank+ Sentence and title similarity	0.382	0.273	0.257
TextRank+ Keyword coverage rate	0.308	0.188	0.205
TextRank+ Mongolian conjunctions	0.285	0.177	0.183

The experimental results show that the introduction of any sentence feature on the basis of TextRank algorithm can improve the quality of the summary, which verifies the effectiveness of mnTextRank algorithm. In order to verify the influence of different Mongolian external language knowledge bases on the algorithm, we use different Mongolian external language knowledge bases to conduct experiments on the basis of mnTextRank algorithm.

Table 3. Combines the experimental results of different external language knowledge bases

Algorithm	Rouge-1	Rouge-2	Rouge-L
mnTextRank	0.398	0.287	0.269
Word2vec+mnTextRank	0.393	0.282	0.271
CINO+mnTextRank	0.416	0.303	0.293
IMNUBERT+mnTextRank	0.458	0.345	0.374

Among them, Word2vec is a model trained in our laboratory with a dimension of 100. CINO is Chinese minority pre-trained language model [13], the CINO-base-v2 version is used in this paper. The experimental results show that when the IMNUBERT pre-trained model trained by our laboratory is integrated into the mnTextRank algorithm as a Mongolian external language knowledge base, the quality of the summary is the best. In addition, we conduct comparative experiments with different algorithms to verify the effectiveness of the algorithm proposed in this paper. Lead3 takes the first three sentences in the article as the summary. TF-IDF calculates the weight of each word in the sentence to indicate the importance of the current sentence. TextRank is the algorithm without any improvement. mnTextrank is the algorithm combined with sentence feature weight proposed in this paper. IMNUBERT-mnTextRank is the final algorithm proposed in this paper. The experimental results are shown in Table 4.

Table 4. Experimental results of different algorithms

Algorithm	Rouge-1	Rouge-2	Rouge-L
Lead3	0.342	0.238	0.218
TF-IDF	0.326	0.207	0.215
TextRank	0.275	0.166	0.175
mnTextRank	0.398	0.287	0.269
IMNUBERT-mnTextRank	0.458	0.345	0.374

Experimental results show that Lead3 has a high Rouge value, but there is a redundancy problem. TF-IDF method can perceive some key information in the article through the word frequency. TextRank algorithm only considers the similarity between sentences, without considering some features of the sentence itself, so the effect is poor. mnTextRank algorithm, which comprehensively considers the weight of sentence features on the basis of TextRank algorithm, has a higher improvement in the quality of the summary. On the basis of the mnTextRank algorithm, the IMNUBERT pre-trained model is integrated into the algorithm in the form of an external language knowledge base to further improve the quality of the summary. Compared with the TextRank algorithm, the summary generated by the IMNUBERT-mnTextRank method proposed in this paper has the three evaluation indicators Rouge-1, Rouge-2, and Rouge-L improved by 0.183, 0.179, and 0.199, respectively. The effectiveness of the IMNUBERT-mnTextRank method proposed in this paper is verified.

4. Conclusion

In this paper, we propose IMNUBERT-mnTextRank, an automatic Mongolian text summarization method. This method integrates the information of the Mongolian external knowledge base into the TextRank algorithm to improve the accuracy of similarity calculation between sentences. The quality of the generated summary is improved by calculating the weight of sentence features such as sentence position, sentence and title similarity, Keyword coverage rate, and Mongolian conjunctions. The experimental results show that the Mongolian summarization generated by

the proposed IMNUBERT-mnTextRank method has better quality. In the next work, more Mongolian summarization datasets are constructed, and a better Mongolian automatic text summarization method is studied through neural network to improve the quality of the summarization.

Acknowledgments

We thank Associate Professor Si Qintu and Professor Wang Siriguleng for their guidance and help in the research process.

Funding

This work was supported by the Natural Science Foundation of Inner Mongolia (2022MS06002), the Open Project Foundation of Inner Mongolia Discipline Inspection and Supervision Big Data Laboratory (IMDBD202109), the Science and Technology Plan Project of Inner Mongolia Autonomous Region (2021GG0139), the Innovation Fund for Postgraduates of Inner Mongolia Normal University (CXJJS22139), the Basic Scientific research fund of Inner Mongolia Normal University (2022JBXC018).

References

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

[2] Li JP, Zhang C, Chen XJ. A Survey on Automatic Text Summarization [J]. Journal of Computer Research and Development, 2021, 58(01):1-21.

[3] Mihalcea R, Tarau P. TextRank: Bringing order into texts [C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. ACL, 2004: 404-411,

[4] Wang Y X, Han B, Gao R. Automatic Extraction of Text Summarization Based on Improved TextRank [J]. Computer Applications and Software, 2021, 38(06):155-160.

[5] Li W, Yan X D, Xie X Q. An Improved TextRank for Tibetan Summarization [J], Journal of Chinese Information Processing, 2020, 34(9): 36-43.

[6] Zhu B B, Luo F, Luo Y J. An Automatic Text Summarization Algorithm Based on Clause Extraction [J/OL] Journal of East China University of Science and Technology, 2024, 1-7.

[7] Xu F, Peng J J, Iu J. Automatic News Summarization Model Based on Multi-feature TextRank [J]. Computer Systems & Applications, 2023, 32(02):242-249. DOI:10.15888/j.cnki.csa.008913.

[8] Li C C. Study of Mongolian automatic text summarization based on natural language understanding [D]. Beijing University of Posts and Telecommunications, 2005.

[9] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. Proceedings of NAACL-HLT 2019, pages 4171-4186. Minneapolis, Minnesota, June 2-June 7, 2019.

[10] Baxendale PB. Machine-made index for technical literature—An experiment. IBM Journal of Research and Development, 1958, 2(4): 354-361. [doi:10.1147/rd.24.0354] arXiv preprint arXiv:1810.04805, 2018.

[11] Yan XD, Wang YQ, Huang S. A dataset of Tibetan text summarization [J]. Chinese Journal of Scientific Data, 2012, 7(02):43-49.

[12] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. 2004.

[13] Yang Z Q, Xu Z H, Cui Y M, et al. CINO: A Chinese minority pre-trained language model [C]//Proceedings of the 29th International Conference on Computational Linguistics (COLING). Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022:3937-3949.