2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)

Algorithm for Automatic Abstract Generation of Russian Text under ChatGpt System

Lina Hou
School of International Education
Heilongjiang University of Technology
Jixi, Heilongjiang, China
houlina927@163.com

2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) | 979-8-3503-1860-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCECE60827.2024.10548397

***Abstract*—In response to the problems of insufficient understanding of language characteristics, inaccurate semantic understanding, and inaccurate automatic summarization generation of Russian texts in current automatic summarization generation algorithms, this article would use the chatGpt system to study the algorithm for automatic summarization generation of Russian texts. The automatic summarization algorithm for Russian texts based on the ChatGPT system can effectively generate summaries for Russian texts. This algorithm utilizes ChatGPT's powerful natural language processing capabilities to deeply understand and analyze Russian texts, thereby generating concise and accurate summaries. The algorithm based on the ChatGPT system has an accuracy of over 97.64% in generating Russian text summaries, and the average accuracy of Russian text summarization generation in 10 experiments is 98.23%. This provides an effective solution for automatic summarization generation of Russian texts, and the effectiveness and reliability of the algorithm are verified through experiments.**

Keywords—chatgpt system, russian text, automatic summary generation, data preprocessing, bidirectional encoder representation from transformers

I. INTRODUCTION

With the development of internet technology, the amount of online information is exploding, especially the growth rate of textual information such as news and comments is astonishing. This leads to an excessive information load, requiring people to spend a lot of time and effort reading relevant information. In order to quickly filter information and ensure quality, while solving the problem of information overload, extracting key data from massive text information has become an urgent need. Due to the characteristics of massive, complex, diverse, and real-time nature of big data, the workload of traditional manual summarization is heavy, and the time cost increases sharply. Therefore, text summarization technology is favored by people. This article studies the ChatGPT system and investigates the algorithm for automatic summarization of Russian texts based on the ChatGPT system. This algorithm first preprocesses the input Russian text to extract key information from the text. It generates a concise summary of the text by using a certain algorithm to train the model. The Russian text automatic summarization generation algorithm based on the ChatGPT system mainly trains the dataset using the Transformer model to learn the structure and grammar rules of dialogues. The core idea of this method is to generate text through dialogue. Unlike traditional text summarization methods, ChatGPT can generate more diverse and flexible summaries while maintaining contextual coherence. For longer conversations, ChatGPT may encounter information loss issues when generating summaries. Therefore, in practical applications, it is necessary to combine other technical means to post-process and optimize the generated summaries.

II. RELATED WORK

When processing large amounts of Russian text data, automatic summarization generation technology is an efficient and convenient way to help quickly understand text content [1-2]. Angelina, B proposed an automatic training set generation and labeling algorithm based on the concept of single parent relationships, which is used to harvest the training sets of all ambiguous nouns, verbs, and adjectives presented in the synonym library, and then evaluate the quality of the obtained sets [3]. Zhang H's large-scale pre trained language model based on Transformer has become a new paradigm for natural language generation, allowing for the generation of more diverse and smoother texts. He targets different controllable text generation tasks that require different types of controlled constraints [4]. Shi T studied different sequence to sequence models from the perspectives of network structure, training strategies, and summary generation algorithms. He first proposed several models for language modeling and generation tasks, and also developed an open-source library, the Neural Abstract Text Summarizer Toolkit, for abstract text summarization [5]. Ma T. proposed a topic based automatic summarization method for Chinese short texts to obtain the score of each text corresponding to the topic in the original text data [6].

ChatGPT is a large-scale language model developed by OpenAI, which pre trains through a large amount of text data in an unsupervised learning manner and has the ability to understand and generate natural language [7-8]. Wu T provides a brief overview of the history, current situation, and potential future development of ChatGPT, which helps to provide a starting point for thinking about ChatGPT. From limited open access resources, he summarizes the core technologies of ChatGPT [9]. Dergaa, I used ChatGPT to analyze sources including reading and critical assessment, and identified relevant data to support research questions, believing that ChatGPT has the potential to improve academic writing and research efficiency [10].

III. RUSSIAN TEXT DATA COLLECTION AND PREPROCESSING

A. Data Preparation

Choosing a representative and diverse Russian text dataset is crucial for improving the algorithm's generalization ability. The training corpus used in this article consists of hundreds of thousands of Russian articles, covering the basic Russian corpus of Wikipedia and economic and political news obtained through web crawlers. The Russian corpus can be downloaded directly from the official website as a binary file without conversion. This corpus covers a wide range of nouns and is the foundation of word vector models. In order to make the model more professional, it can add articles from different fields by yourself, and the more articles, the better. This article mainly focuses on the field of political news, using web

crawlers to crawl multiple political news articles from major news websites and local government news websites, and adding them to the corpus.

B. Data Preprocessing

This article preprocesses Russian text, including word segmentation, removal of stop words, and stem extraction, in order to convert the text into a format that the model can process. During the processing, it is necessary to ensure that the main information of the text is preserved. Figure 1 shows the process of Russian text preprocessing: first, special characters are removed, then various open-source word segmentation tools are used to segment the Russian text, followed by data deduplication and removal of stop words.

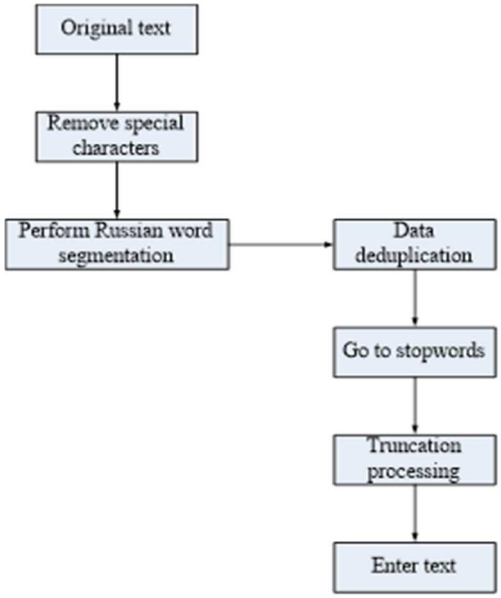

Fig. 1. Data preprocessing process

There may be some difficulties in collecting Russian corpus: (1) Traditional characters appear in some texts, as the model cannot distinguish between simplified and traditional characters. It would treat the simplified and traditional versions of a word as two different words, thereby affecting the accuracy of the model; (2) Some compound words have no actual meaning, and in natural language processing, these words are called stop words. If this message is included in the training model, it would affect the quality of the entire model; (3) The segmentation content cannot remove punctuation marks. Punctuation marks have the same impact on the model as stop words, so punctuation marks must be removed before training. For preprocessing Russian text sets, the main steps can be as follows: first, convert symbols and simple numbers. Download a simple and easy-to-use list of the most commonly used Russian words, create a new Python file, convert it into a script, and save it to the file. Continuing with Russian word segmentation, this article uses a third-party Python extension package to do Russian word segmentation, supporting three segmentation modes that are suitable for text analysis. The next is to remove punctuation marks, create markers, and pass on the segmentation results in the second step.

IV. Model Training under ChatGPT System

A. Model Construction

Traditional word models use unsupervised training methods to train continuous word vectors. However, the problem of representing polysemous words has always been a problem in the field of word vectors, and traditional models cannot distinguish their meanings in many contexts when encoding polysemous words. Therefore, this article would use the ChatGPT system to study the new language pre learning model, Bidirectional Encoder Representation from Transformers (BERT), and achieve good results in multiple language tasks. Meanwhile, BERT obtains features at different levels of different processes from pre trained methods to better recognize the features of words. In order to enable the model to obtain more semantic information, it can be considered to combine the BERT sentence vector with the content vector on the improved decoder side to obtain a new vector. Due to the fact that the training model first generates character level vectors, for Russian, combining character level vectors with mean values can obtain a complete word vector.

Calculate the hidden state vector of the decoder and the vocabulary distribution that jointly participates in the current time step using the following formula:

$$Q_{vocab} = softmax\left(W^{'}(W[r_t, f_t'] + a) + a'\right) \quad (1)$$

Among them, $W^{'}$, W, a, a' are the learning parameters, and Q_{vocab} represents the probability of publishing all words in the vocabulary list.

In order to handle improperly registered login words, a pointer mechanism has been introduced. From the feature vector, the decoder hides the input vector from the state vector, which can control the decoder to generate a new word from the set vocabulary or copy previously appeared words from the original text. The specific formula is as follows:

$$Q_{gen} = \mu\left(M_f^H * f_a' + M_s^H s_h + M_y^H y_h + a_{ptr}\right) \quad (2)$$

Among them, M_f^H, M_s^H, M_y^H, a_{ptr} are the learning parameters.

The calculation formula for the probability Q(r) of the current predicted word r is as follows:

$$Q(r) = Q_{gen} * Q_{vocab}(r) + (1 - Q_{gen}) \sum_{i:R_r} b_i^h \quad (3)$$

B. Automatic Summary Generation

In order to combine the Russian article sentences obtained through key sentences into a summary, it is necessary to design a summary generation framework to improve the quality of the summary. Unlike traditional extractive summarization methods, this framework combines key sentences organically instead of simply connecting them. The main approach to improving the quality of abstracts is to consider two aspects: improving the fluency of abstracts and enhancing the criticality of abstracts. Therefore, the abstract generation framework designed in this article aims to further improve the fluency of the abstract and verify its criticality, ultimately applying it in the construction phase of the abstract generation framework. The process of processing key sentences in the abstract generation framework is shown in Figure 2.

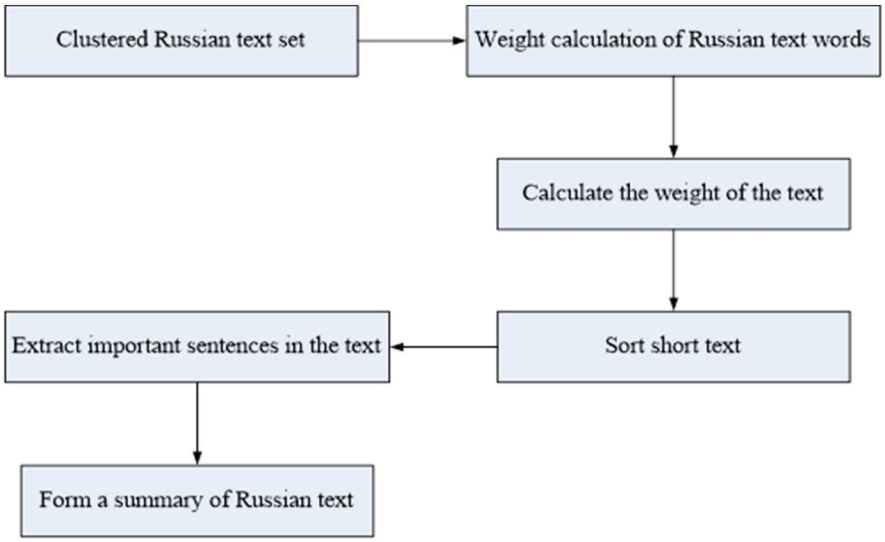

Fig. 2. Schematic diagram of Russian text abstract extraction

C. Methods to Improve Abstract Quality

There are two main methods to improve the quality of automatic summaries: one is to improve the fluency between summary statements; The second is to improve the accuracy of the key information contained in the abstract. For improving fluency, this article uses the ChatGPT system to construct a generative summary model, which can significantly improve the fluency of the summary and make key sentences have a certain degree of correlation. Inspired by the idea of extractive summarization, it can assign weights to sentences based on their position in the original text, thereby ranking key sentences in a top-down order according to structural order, reducing ambiguity and improving fluency. To improve the accuracy of key information, the criticality of a sentence can be measured not only by its similarity with other sentences, but also by the number of keywords it contains.

V. Evaluation of Russian Abstract Generation Algorithm on ChatGPT System

A. Evaluation Indicators

Sentences in artificial abstracts should be filtered out from the original text and their order of appearance should be determined in the original text. To reduce the impact of content, it is necessary to manually create a lot of content and calculate the average overlap between them and automatic content. The calculation formula is as follows:

$$avg_Q = \frac{\sum_{i=1}^{m} Q_i}{m} \tag{4}$$

Among them, m represents a total of m manual abstracts, and Q_i represents the overlap between the ith first manual abstract and the automatic abstract. The average overlap is the result of dividing the total overlap by m manual abstracts.

The Recall-Oriented Understudy for Gisting Evaluation (ROUGR) has become one of the standard methods in summary evaluation techniques. It improves the stability of the evaluation system by dividing the samples to be compared and manually generated standard samples.

$$Rouge - N = \frac{\sum s \in (Ref\ Sum) \sum n-gram \in s^{Count_{match(n-gram)}}}{\sum s \in (Ref\ Sum) \sum n-gram \in s^{Count_{(n-gram)}}} \tag{5}$$

Among them, n represents the length of $\mathrm{n-gram}$, $\mathrm{Count}_{\mathrm{match(n-gram)}}$ represents the number of occurrences of $\mathrm{n-gram}$ in both manual and automatic summaries, and $\mathrm{Count}_{\mathrm{(n-gram)}}$ represents the number of occurrences of $\mathrm{n-gram}$ in manual summaries. The n value is determined based on the length of the automatic summary. By calculating the overlap of characters, the higher the overlap, the larger the ROUGE value, and the better the summary effect.

B. Experimental Analysis

This article uses the general evaluation method ROUGE in the abstract as the evaluation criterion. In order to better demonstrate the superiority of the ChatGPT system in generating Russian text automatic summaries, this paper compares the experimental results with the Russian text automatic summary generation algorithm based on Term Frequency Inverse Document Frequency (TF-IDF), deep learning (DL), TextRank, and Convolutional Neural Network (CNN). This article selects the average evaluation results of all categories under ROUGE. The specific summary evaluation results are shown in Table 1.

TABLE I. Comparison of Evaluation Results for Russian Text Abstracts

Algorithm	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-N
ChatGPT system	0.678	0.389	0.447	0.596
TF-IDF	0.432	0.145	0.329	0.411
DL	0.561	0.288	0.371	0.439
TextRank	0.489	0.243	0.314	0.408
CNN	0.607	0.296	0.402	0.473

As shown in Table 1, it can be observed that the evaluation results of the Russian text automatic summarization generation algorithm based on the ChatGPT system are much higher than other algorithms. This is because the word weights calculated in the algorithm studied in this article not only consider word frequency, TF-IDF values, and the relationships between words, but also integrate the left and right entropy values, coverage, and determine that the weights of words should not be equal. It should be determined by the

weight of the corresponding words, effectively improving the accuracy of short text similarity calculation.

Russian text as a whole is very complex, and the accuracy of abstracts generated by automatic generation algorithms is very low. The text content is very messy, and the Russian text automatic summary generation algorithm based on the ChatGPT system effectively improves the accuracy of Russian text automatic summary generation. In order to further demonstrate the superiority of the algorithm studied in this paper, a comparative study was conducted on the Russian text automatic summarization generation algorithms constructed using CNN, DL, and TF-IDF. The specific comparison results are shown in Figure 3.

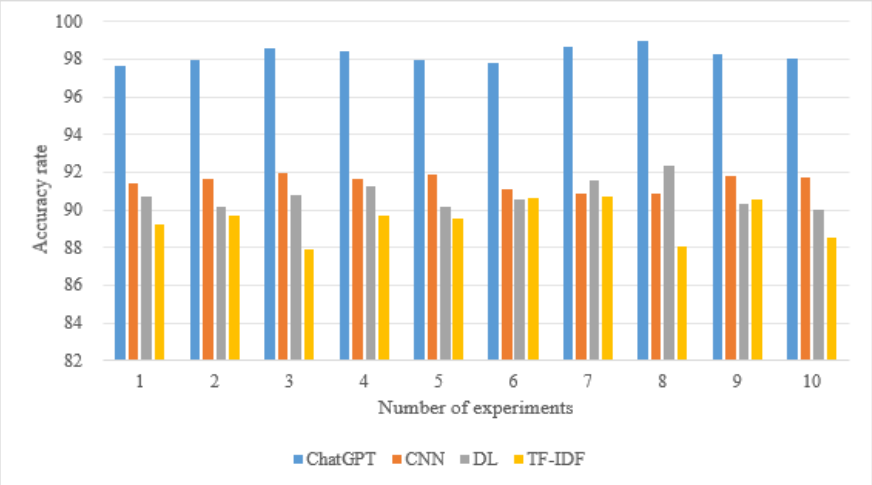

Fig. 3. Comparison of accuracy in Russian abstract generation using different algorithms

As shown in Figure 3, the accuracy of the Russian text automatic generation algorithm studied in this article for abstract generation is much higher than that of algorithms constructed by other methods. Among them, the algorithm based on the ChatGPT system has an accuracy of over 97.64% in generating Russian text summaries. The accuracy of algorithms based on CNN, DL, and TF-IDF for Russian text summarization generation is below 91.97%, 92.34%, and 90.73%, respectively. At the same time, 10 experiments were conducted, and the algorithm based on the ChatGPT system studied in this paper achieved an average accuracy of 98.23% for Russian text summarization, which was 6.75%, 7.45%, and 8.77% higher than the average accuracy of algorithms based on CNN, DL, and TF-IDF, respectively.

VI. Conclusions

The Russian text automatic summarization generation algorithm based on the ChatGPT system has achieved automatic summarization generation of a large amount of Russian text data. The main contribution of this algorithm is to use the ChatGPT system to deeply understand Russian texts and mine contextual information, thereby improving the accuracy and readability of abstracts. In addition, the algorithm adopts a rule-based approach for summary generation, which can generate concise and smooth summaries, improving the efficiency of summary generation. Due to the fact that the generation of abstracts is based on dialogue, the quality and relevance of dialogue play a decisive role in the final summary generation. In practical applications, it is necessary to combine other technical means to post-process and optimize the generated summaries. In summary, the Russian text automatic summarization generation algorithm based on the ChatGPT system is a promising natural language processing technology, which provides new ideas and methods for the processing and analysis of Russian text.

References

[1] Sengupta, J., Alzbutas, R. and Kl, H., 2022, October. An Empirical Analysis on Detection and Recognition of Intra-Cranial Hemorrhage (ICH) using 3D Computed Tomography (CT) images. In *2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon)* (pp. 1-5). IEEE.

[2] Tang Juxiang, Sun Yihui, Liao Xiao, Liu Jianguo, & Yu Juan. (2021). Research on the word extraction method of Russian text with multi-strategy fusion. Chinese science and technology terminology, 23(3), 59-62.

[3] Angelina, B., & Loukachevitch, N. (2020). All-words word sense disambiguation for Russian using automatically generated text collection. Cybernetics and Information Technologies, 20(4), 90-107.

[4] Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys, 56(3), 1-37.

[5] Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C. K. (2021). Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science, 2(1), 1-37.

[6] Ma, T., Wang, H., Zhao, Y., Tian, Y., & Al-Nabhan, N. (2020). Topic-based automatic summarization algorithm for Chinese short text. Mathematical Biosciences and Engineering, 17(4), 3582-3600.

[7] Sharma, S., & Yadav, R. (2022). Chat GPT–A Technological Remedy or Challenge for Education System. Global Journal of Enterprise Information System, 14(4), 46-51.

[8] Vishnu, C., Khandelwal, J., Mohan, C.K. and Reddy, C.L., 2023. EV AA-Exchange Vanishing Adversarial Attack on LiDAR Point Clouds in Autonomous Vehicles. *IEEE Transactions on Geoscience and Remote Sensing*.

[9] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica, 10(5), 1122-1136.

[10] Lateef Haroon PS, A., Patil, S.N., Bidare Divakarachari, P., Falkowski-Gilski, P. and Rafeeq, M.D., 2024. An optimized system for sensor ontology meta-matching using swarm intelligent algorithm. Internet Technology Letters, p.e498.