Keyword extraction method for machine reading comprehension based on natural language processing

Ruiheng Li1, Xuan Zhang2*, Chengdong Li2, Zhongju Zheng3, Zihang Zhou4, Yuyin Geng5

1 School of Data Science and Software Engineering, Qingdao University, Qingdao, Shangdong, 266071, China
2 School of Data Science, City University of Macao, Macao, China
3 School of Finance and Economics, Henan Polytechnic University, Jiaozuo, Henan, 434003, China
4 School of Electrical and Electronic Engineering, Hubei University of Technology, Wuhan, Hubei, 430068, China
5 School of Vocational and Technical, Hubei University of Technology, Wuhan, Hubei, 430068, China
*Corresponding author’s e-mail: D20091100189@cityu.mo

Abstract. The traditional keyword extraction method has the problem that the accuracy of identification and extraction decreases when the amount of data is increasing. This paper introduces natural language processing technology and designs a keyword extraction method for machine reading comprehension. In this process, considering that the information in the text of machine reading comprehension may come from the web page, we can refer to the NLP system of natural language algorithm. Firstly, analyze the content of a single document, then analyze the short string level, and finally organize the preprocessing of natural language text information, so as to achieve the preprocessing of machine reading and comprehension of Chinese text. At the same time, the SKEA keyword extraction algorithm system is designed to merge text words with the same meaning, locate the location information of document resources, and identify the annotated words in the resource base. After the annotation processing is completed, the annotation information is extracted. In addition, a comparative experiment is presented to prove the effectiveness of the proposed method compared with the traditional method.

1. Introduction
The related research of natural language processing still belongs to the research content of artificial intelligence field in China. Through effective processing of natural language, linguistic content, mathematical content and computer content can be integrated to realize human-computer interaction in this way [1]. At present, natural language processing technology has been widely used in the Chinese market, and has made remarkable achievements in the application of machine algorithm, social public opinion information acquisition, text resource classification and processing, language and character recognition, etc. Machine reading comprehension is one of the main application directions of natural language processing technology. The main purpose of its application is to accurately identify the key
words in the text and locate the core of the text content through the application of the technology, so as to solve the problems related to the text correctly [2]. In the early traditional research methods in this field, most of the text information extraction and recognition methods adopt machine modeling and manual extraction. However, based on the development of intelligent technology in China in recent years, the amount of information and data in machine reading comprehension has increased dramatically, and large-scale data sets are frequently used in the application. This trend belongs to the difficulty proposed in the research of this technology. Therefore, based on the significant advantages of natural language processing technology, this paper will design a keyword extraction method for machine reading comprehension, to realize the effective extraction of key text information for explosive information, and ensure the efficient processing of large volume data sets.

2. Key words extraction method for machine reading comprehension based on natural language processing

2.1. Machine reading and comprehension of Chinese text based on natural language preprocessing

Considering that the information in the machine reading comprehension text may come from the web page, the text information related to machine reading comprehension can be in pdf format; Caj format; Word format, etc. Text information content presents a trend of diversified data, so before text information extraction, it is necessary to combine with natural language processing technology to denoise text information.

In the process of processing, considering that different news pages involve more knowledge fields, we can start from the target level of different pages. According to the natural language content, multiple value information such as target topic information, content information and annotated information are obtained [3]. According to certain extraction principles, the content of information is filtered, the conditions of processing are positioned, and the information is structurally adjusted. Make the information present a natural form, and keep the original format of the text information after adjustment into the machine reading comprehension Chinese database.

In the research of this aspect, it is found that different websites and page templates are different, but most of the pages will contain keywords annotated by manual. In the process of identifying Chinese text, a single document content analysis (including extraction of text body, classification of text content and resources, analysis of text structure, expression of text semantics, etc.) can be carried out by referring to the NLP system of natural language algorithm. On this basis, the short cascade string parsing (including the replacement of the same semantic vocabulary, the unification of semantic structure, semantic error correction, etc.) is carried out, and the key character recognition (including login recognition, attribute recognition, semantic recognition, synonymy recognition, etc.) is carried out continuously in the term level of natural sentences. Finally, the pre-processing of natural language text information is sorted out (including the sharing of resource data, the invocation of knowledge base resources, and the in-depth analysis of text dictionaries, etc.). In short, according to the above mentioned content, the preprocessing of machine reading comprehension of Chinese text can be completed.

2.2. Design SKEA keyword extraction algorithm

After completing the processing and preprocessing of the data set, this chapter puts forward the way of designing the SKEA keyword extraction algorithm, and plans the keyword screening and extraction process by combining the application of the algorithm. This algorithm is designed according to the Naive Bayes model, in which semantic resources are invoked to improve machine reading comprehension. In the process of improvement, a semantic resource extraction method is proposed [4]. According to the model design, the SKEA keyword extraction algorithm can be shown in Figure 1 below.
In the process of the original machine-reading comprehension document set, it is necessary to merge the text terms with the same meaning [5]. The expression of SF is defined as the word frequency of the same meaning. After improvement, the expression of word frequency merging processing is expressed as:

$$SF = \frac{N_{synonym}}{N}$$  \hspace{1cm} (1)

In Formula (1), $N_{synonym}$ represents the combined amount of text vocabulary. On the basis of the above calculation, word frequency inversion algorithm is needed. Use UDF to represent as the reciprocal of frequency, then IDF can be expressed by the following calculation formula:

$$IDF_i = \log \left( \frac{|D|}{|\{ j : l_i \in d_j \}|} \right)$$  \hspace{1cm} (2)

In formula (2), $|D|$ represents the total number of documents in the Chinese resource library for machine reading and comprehension; $j$ represents the number of documents; $t$ represents an extraction period (file location information); $i$ represents the number of files; $d$ represents the Chinese resource library Existing resource storage capacity. Through the above calculation formula, the deletion processing of the invalid information in the phrase or sentence can be realized. That is, in the case of inversion, useless information may appear as "0". On this basis, the location information of the document resources is located to obtain a collection of word segmentation [6]. The set can be expressed as $(w_1, w_2, ..., w_k)$, where the corresponding position of $w_i$ in the document can be expressed as the following calculation formula:

$$p_i = \frac{\sum_{j=1}^k l_j}{L}$$  \hspace{1cm} (3)

In formula (3), $l$ is the length corresponding to $w_i$, and $L$ is the total length of the text content to be extracted. On the basis of ensuring that the information existing in the existing resources is all value information, the extracted information is compared with the word segmentation database, and the brief vocabulary is processed in dynamic rules, and combined with the level of the terminal subject, the information in the resource library is identified Annotated vocabulary. And after finishing its annotation processing, extract the annotation information from it [7]. The called information is normalized and imported into a Chinese database to complete the design of a method for extracting keywords for machine reading comprehension based on natural language processing.
3. Comparative Experiment

Based on the above discussion, this paper completes the theoretical design of keyword extraction method for machine reading comprehension based on natural language processing. In order to further verify the effect of this method in practical application, this paper introduces the extraction method proposed above and the traditional extraction method into the same machine reading comprehension environment at the same time, so as to complete the following comparative experiments:

In order to ensure the reliability of the experimental results, the same experimental environment was set when the two extraction methods were applied, as shown in Fig. 2.

![Diagram of experimental environment](image)

The CPU of the server in Figure 2 uses Intel Core i5 processor, the size of the hard disk is 850GB, and the memory is 5GB. During the experiment, the whole operating system uses 128-bit, Windows 9.0, IntelliJidea 15.1.5 as the programming tool, and Apache POI as the open source library. The data of reading files used in the experiment is the Chinese database with keyword annotation that is not open in the industry, and all the data in a certain webpage is taken as the experimental data to construct the keyword extraction data set. The main reason for choosing the data in a certain webpage as the experimental data is that a large number of manually annotated keywords are provided in each article. Therefore, the specific topic in the article can be more accurately reflected, so as to meet the rationality requirements of the experiment in this paper. We set the time of extracting this page as February 2021, and contain a total of 500 articles. The proposed extraction method and the traditional extraction method were used to extract keywords from the experimental data set. Under the condition that all other data are the same, complete the experiment. The extraction accuracy was used as the evaluation index to verify the practical application effect of the two extraction methods. The calculation formula of accuracy is: \((\text{the number of correctly extracted keywords}/\text{the number of all keywords in the experimental data set}) \times 100\%\). According to the above formula, the experimental results were counted and the comparison chart of experimental results as shown in Fig. 3 was drawn.

![Comparison chart](image)
As can be seen from the results of the two curves in Figure 3, the keyword extraction accuracy of the proposed method is significantly higher than that of the traditional method. At the same time, the accuracy of the two methods in the initial keyword extraction is not much different, but with the increasing number of keywords, the accuracy of the traditional method of keyword extraction presents an obvious trend of decline. This indicates that the accuracy of keyword extraction is gradually decreasing. However, with the increase of the number of keywords, the extraction accuracy of the method in this paper does not show a downward trend, but has a certain upward trend. This experimental phenomenon is mainly due to the introduction of the understanding of the main structure and semantics of the document when designing the keyword extraction method, and the full consideration of the semantic similarity existing in Chinese words in the process of keyword extraction, which can achieve a better keyword extraction effect. Therefore, through comparative experiments, it is proved that the NLP-based machine reading comprehension keyword extraction method proposed in this paper can accurately extract more keywords in practical application to achieve higher accuracy of machine reading comprehension.

4. Conclusion
This paper designs a method of keyword extraction for machine reading comprehension based on natural language processing. By designing a comparative experiment, it is proved that the extraction accuracy of this method does not decrease with the increase of the number of keywords in the practical application, but has a certain upward trend. Compared with the traditional method, the extraction accuracy is higher. Therefore, it can be considered that the method designed in this paper is of certain value. We can try to apply the method in the field of linguistics, mathematics, computer science and so on in the later research. In this way, the construction and development of China's high-tech industry is truly put forward technical guidance and core technical support. Although the research in this paper may be slightly inadequate, it can also be improved in this aspect through later practice.

References
[1] Li Lu, He Lili. A Keyword Extraction Algorithm for Marketing Reviews Fusing Information Entropy and Multi-weight TF-IDF, Intelligent Computers and Applications. Vol. 09, pp. 69-72, October, 2020.
[2] Yu Bengong, Zhang Hongmei, Cao Yumeng. Improved TextRank keyword extraction method based on multivariate feature weighting. Digital Library Forum. Vol. 03, pp. 41-50, 2020.
[3] Zhu Yancheng, Cai Manchun, Lu Tianliang, et al. SVM based fusion multi-feature TextRank keyword extraction algorithm. Software Guide. Vol. 19, pp. 88-91, February, 2020.
[4] Zhan Fei, Zhu Yanhui, Liang Wentong, et al. Entity link method based on Bert and TextRank keyword extraction. Journal of Hunan University of Technology. Vol. 34, pp. 63-70, April, 2020.
[5] Yang Yanying, Jiang Ying. Design and Implementation of Micro-blog Personal Keyword Cloud Extraction Based on TextRank Algorithm. Computer Programming Skills and Maintenance. Vol. 02, pp. 14-16, 2019.
[6] Zhang Shaodi, Zheng Jiong, etc. Design and Implementation of Chinese Keyword Extraction System Based on Djangos. Computer Knowledge and Technology. Vol. 13, pp. 220-222, 2019.
[7] Li Wenfeng, Zong Jiajia, Tang Shancheng, et al. WeChat Tax Service Platform Based on Natural Language Processing -- A Case Study of Tax Service in Weinan City, Shaanxi Province. Modern Electronic Technology. Vol. 43, pp. 19-21, May, 2020.