Information and Knowledge Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/infkm.2023.040105
ISSN 2523-5893 Vol. 4 Num. 1

Automatic Mining Method for Heterogeneity Features of Prose-Chinese Translation Corpus Based on Artificial Intelligence

Yanhua Ma*

Zhejiang Yuexiu University, Shaoxing, Zhejiang, 312000, China
mora0914@163.com
**Corresponding author*

Keywords: Chinese Prose Translation Corpus, Artificial Intelligence, Heterogeneity Features, Automatic Mining

Abstract: With the diversification of culture and the universality of language, prose as an important literary material has also attracted more scholars' attention. At the same time, due to the current integration and development of science and technology and culture, corpus, as a large-scale electronic text library, is of great significance to the study of relevant language theories. However, after studying the heterogeneity characteristics of prose Chinese translation corpus, it was found that there were still some problems in the current automatic mining methods of heterogeneity characteristics. In order to solve this problem, this paper proposed a new method based on artificial intelligence (AI) to automatically mine the heterogeneity features of prose Chinese translation corpus. In order to verify the effectiveness of this method, this paper also conducted an empirical study. The research results showed that the method in this paper could increase the weight coefficients of heterogeneity features from dataset 1 to dataset 6 in the corpus by 57, 34, 28, 36, 16, 13 respectively, and effectively reduce the offset of dataset nodes and increase the mining amount of data node access, thus improving the effectiveness and practicability of the automatic mining method. In addition, the research of prose meaning corpus could also enrich the research content of corpus, and broaden the research scope of corpus, so as to promote its better development.

1. Introduction

Due to the gradual expansion of the scope of the current corpus research, the corpus of prose Chinese translation has also received more development opportunities, making it attract more attention. As a literary style juxtaposed with poetry, fiction and drama, prose does not need to be rigidly bound to rhythm, plot and characters, and its works are flexible, true and natural. It has a long history and a wide variety, especially modern prose. However, there are still major problems in the study of prose translation into Chinese. Since the 1990s, the development of corpus linguistics has received more and more attention and has strong vitality. However, for the interpretation and translation of the literature of various countries, most of them refer to the traditional literary theories

and relevant Chinese translation norms, and tend to translate prose works into Chinese under a certain theoretical framework. The research paradigm of corpus provides a way to obtain the relevant features of the text from the quantitative level. Therefore, it is of great significance to explore the heterogeneity characteristics in the corpus with the help of computer technology for the study of the explicit and implicit meaning of prose works and the improvement of their Chinese translation norms. The method proposed in this paper combined with artificial intelligence can effectively reduce the offset of corpus data nodes and improve the effectiveness of automatic mining methods through deep learning algorithm. In addition, this paper hopes that this method can provide more theoretical basis for the study of heterogeneity characteristics, and enrich the research content of prose Chinese translation corpus, so as to promote the better development of language application.

As for the heterogeneity of corpora, many scholars have studied it and said that most corpora have heterogeneity. Lee Thomas R studied corpus and criticism and stated that because of the heterogeneity in the corpus, some critics believe that it is wrong to use corpus linguistic tools for problem evaluation, but he believes that it is possible to evaluate and justify the problem by constructing a corpus [1]. He suggests a new construction method for building reference corpus because the heterogeneity in the corpus makes it unfavorable for text extraction and segmentation, and he shows that this method has better segmentation efficiency and reduces the negative effects of corpus heterogeneity [2]. Luo Jinru studied the Chinese-English translation corpus, and discussed the standardization and simplification of the language model in the Chinese-English translation corpus of WeChat translation. He analyzed the existing problems and said that the Chinese-English translation corpus was heterogeneous [3]. Davies Mark, who studied coronavirus corpora, discussed the creation and use of coronavirus corpora and stated that coronavirus corpora can be somewhat heterogeneous, so users need to regularly look at the frequency of words and phrases over time to see how particular topics are perceived over time [4]. These scholars' research on the heterogeneity of corpus can enrich their theoretical content, but there are also some deficiencies.

However, some scholars have discussed this issue from the perspective of corpus heterogeneity mining and put forward different views. Li Yawen studied the discovery of heterogeneous potential topics in semantic text mining. He developed a new method for discovering heterogeneous potential topics, which seamlessly integrated topic modeling and word embedding to discover heterogeneous potential topics. By combining the parameter server architecture with the new private sampling algorithm, heterogeneous potential topic discovery could be effectively trained to protect the underlying data privacy. Experiments showed that this method was superior to the current existing technology [5]. Prabu P studied sentiment comment analysis based on corpus heterogeneous feature mining, and he constructed an autoencoder convolutional neural network for sentiment analysis based on corpus sentiment heterogeneous feature mining, and he stated that the method can better mine the heterogeneous features in the corpus by autoencoder with convolutional neural network for classification [6]. These scholars can provide some theoretical support for the research of corpus heterogeneity mining. However, because the corpora studied are different, there is little mention of automatic mining and artificial intelligence for the heterogeneity of the corpus of prose Chinese translation, which makes the research conclusions of little reference value. It can be seen that the application of artificial intelligence to the automatic mining of heterogeneous features in the corpus of prose Chinese translation is a relatively new field, which needs more research.

This paper made a deep research on the automatic mining method of the heterogeneity features of the corpus of prose Chinese translation based on artificial intelligence. After research, it was found that this method was effective. It could improve the defects of traditional automatic mining methods, and enhance the effectiveness and practical value of automatic mining methods, so as to

better complete the autonomous mining of heterogeneous features in the corpus, which has certain practical significance.

2. Theoretical Research on Automatic Mining of Heterogeneous Features of Prose Chinese Translation Corpus

2.1. Overview of Corpus Heterogeneity in Prose Translation

2.1.1. Corpus and Heterogeneous Meaning

Corpus is a combination of bilingual corpora formed by two different languages. It is an electronic text library formed by scientific selection and processing. It can rely on software or other tools to match the paragraphs to be translated, and provide help for translation according to the matching degree between corpora [7]. The formation of corpus is shown in Figure 1.

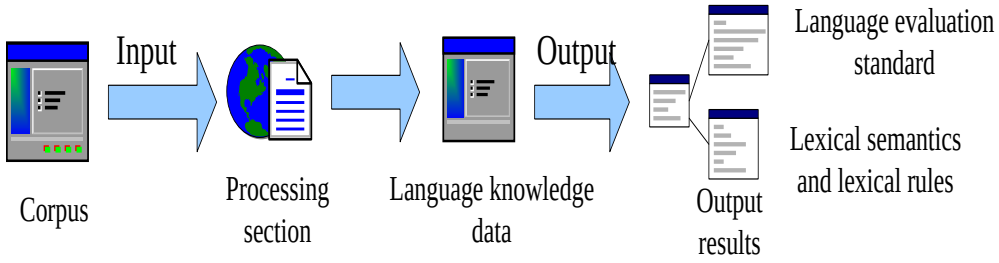

Figure 1: Formation of corpus.

Heterogeneity is a biological concept [8]. However, when heterogeneity is applied at the level of data analysis, it represents the characteristic difference of all individuals in a certain group [9]. In general, the higher the degree of heterogeneity, the more dispersed the distribution of corresponding personal characteristics. With the continuous enrichment of the Chinese translation corpus, the content of prose translation into Chinese gradually changes.

2.1.2. Corpus Translation

Corpus translatology is a research method that, under the guidance of linguistics and translation theory, aiming at a large number of bilingual actual corpus, uses probability and statistical methods, as well as intralingual and interlingual contrast, to describe and explain translation phenomena diachronically or synchronically, and finally discusses its essence [10].

The combination of descriptive translation studies and corpus linguistics is an important basis for the study of metacorpus translation. Compared with the prescriptive translation theory of evaluating the merits of a translation, descriptive translation research has its own characteristics. This characteristic determines that there are certain differences between corpus translatology and other disciplines in translatology. In addition, the focus of corpus translation studies is not on the analysis and statistics of corpus data, but on the study of translation phenomena and behaviors, as well as translation products, translation processes and translation functions [11]. The discussion of corpus translation studies not only inherits the characteristics of descriptive translation studies, but also inherits its basic ideas and research methods, and does not exclude the basic paradigm of linguistics. Therefore, corpus translatology is not only an empirical and descriptive study, but also a qualitative and quantitative study, even a basic or applied study.

Generally speaking, there are two main categories of translation corpus, namely parallel corpus and reference corpus [12]. According to their translation methods, they can be divided into translation, interpretation and multimodal corpus. According to the time span, it can be divided into

synchronic and diachronic corpora. According to the collection method, it can be classified into static and dynamic corpora. The study of corpus translation can be carried out from two aspects: description and interpretation.

2.1.3. Chinese Prose Translation Norms

In descriptive translation studies, translation norms are an extremely important category. As translation is a social behavior, translation norms are also an internal norm of a society, which reflects the constraints of the common values of a society on its activities. As a translation criterion, prose Chinese translation criterion is the choice of rules and customs for prose Chinese translation in a certain historical period, certain social relations and cultural conditions. At a specific historical stage, the emergence of the norms of prose translation requires that the Chinese translation of prose should have corresponding regularity and tendency. Only by analyzing and studying a large number of Chinese translated essays can the specific regularity of Chinese translation norms be found. If the above laws are thoroughly studied, the norms of Chinese translation can be reconstructed.

In recent years, the translation academia has paid more and more attention to the translation of prose into Chinese, and the research perspective and content have also been expanding. It is also increasingly common to use the free software analysis kit to promote the translation of literary corpus into Chinese [13]. However, there are still many problems in the translation of prose in different languages. After sorting out the research on the Chinese translation of modern and contemporary prose, it can be seen that the early translation research level is low, and the research methods and perspectives are single. The research on the reception of the Chinese version is basically blank.

After studying this, this paper has found that the current prose translation should be improved in the following three aspects. First of all, it is necessary to have different research perspectives and make full use of the advantages of different perspectives. Secondly, more research objects are needed to broaden the research field. Finally, there is a need for diversified research methods to promote the scientific development of research. At present, the research of Chinese translated prose is mainly carried out from a qualitative perspective, most of which are based on corpus-based empirical critical translation as the main research method, and lack of objective and quantitative analysis of the text [14]. This paper believes that the study of prose translation into Chinese should start from many aspects to reverse the current single research method.

2.1.4. Construction of Corpus

At present, the corpus of translation research can be divided into two categories: parallel corpus and reference corpus. The former is a comparison between the original text and the translated text, and the latter is a comparison between the translated text and the non-translated text. Parallel corpora help to provide theoretical basis for the translation norms used in the target language. The important role of reference corpus is to identify the difference between translation behavior and original text. The main components of the corpus are shown in Figure 2.

The research methods used in this paper include the basic concepts and tools of corpus linguistics and stylistic statistics. In short, corpus linguistics is an important topic in the study of corpus. Corpus linguistics is a study of language and linguistic phenomena in corpus. The research methods of stylistic statistics or quantitative stylistics are based on the extensive use of computer technology and the practical needs of researchers for quantitative research of style characteristics. With the method of stylistic statistics, the author's writing style and stylistic characteristics can be quantitatively analyzed, which is closely related to literary research and literary criticism. The application of corpus linguistics and stylistic statistics in linguistics and literature mainly includes

criminal investigation linguistics, author identification, literary text analysis, language style quantification, etc. The research of this discipline applied to Chinese translation mainly lies in the fields of translation generality, translation norms, translator's style, etc.

This paper establishes a parallel corpus and a reference corpus for prose translation into Chinese, and in some ways draws on some data from the modern Chinese balanced corpus constructed by other scholars. The parallel corpus constructed in this paper includes 180 English prose and corresponding translated works, and the reference corpus includes 180 modern prose. These works are some famous works, and most of them are from famous writers. Their translations should well reflect the current situation of prose translation into Chinese. There are 249857 words in the original English text and 485359 words in the translation into Chinese. The selected articles are all modern prose works. The reference corpus has a total of 581230 words, and the proportion of Chinese in the two corpora is basically the same. The modern Chinese balanced corpus selects the corpus from 1989 to 1993, covering different types of subjects and texts. It is almost in the same period as the original prose and translation included in this article, and can be used as the reference corpus of this article. After building the corpus, researchers can analyze different language features through Word Smith tool to reflect different translation norms.

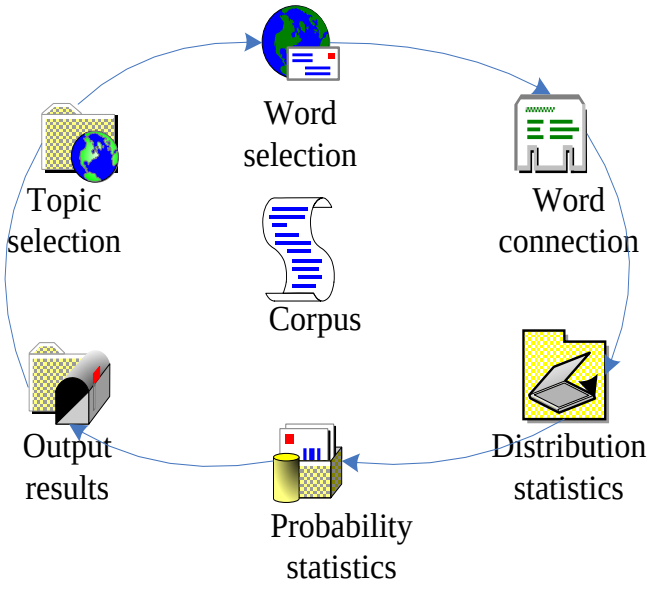

Figure 2: Main components of corpus

2.2. Corpus Heterogeneity of Prose Chinese Translation

Most Chinese translation corpora have many search conditions. In order to achieve the heterogeneous feature selection of the Chinese translation corpus, the size of the solution space selected should be $2^{|A|}$. The available height of space in the corpus can be replaced by a complete binary tree of $|A|$. The spatial tree structure formed by the corpus of prose Chinese translation is shown in Figure 3.

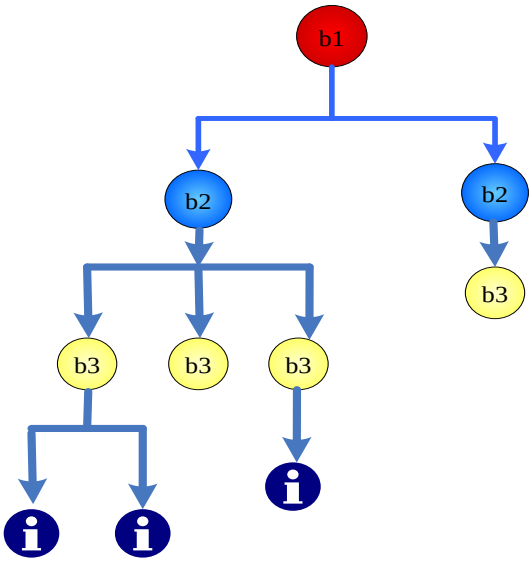

Figure 3: Spatial tree structure of prose Chinese translation corpus

As can be seen from the spatial tree structure in Figure 3, the heterogeneity feature can be divided into three feature choices, namely, cost-sensitive, minimum test cost and constraint feature. First of all, it is considered that cost-sensitive corpus B is an independent state, and $B \subseteq A$. At this time, A is a Chinese translation corpus, then the independent Chinese translation corpus process of this sensitive corpus can be described as follows:

$$A(B) = \sum_{b \in B} a(\{b\}) \tag{1}$$

Among them: b - Sensitive corpus collection.

In practical applications, the same type of sensitive corpus has the same heterogeneity. In order to avoid repeated selection of heterogeneous characteristics, it is assumed that a common test cost of $ha(y)$ is met by a test cost, and the test cost can meet $0 < ha(y) < \min a(b)$. Therefore, the processing of heterogeneous screening of sensitive corpus can be expressed as:

$$a(B \cup \{b\}) = \begin{cases} a^*(B) + a(b) - ha(y) \\ a^*(B) + a(b) \end{cases} \tag{2}$$

Among them: y - Sensitive vocabulary groups.

After the feature selection of the sensitive corpus is completed, the relevant reduction combination of the sensitive corpus is used as the minimum test cost set. Then, the feature selection method of the minimum test cost set is expressed as:

$$a(T) = \min\left\{a\left(T'\right) \middle| T' \in Red(Z)\right\} \tag{3}$$

Among them: Z - Relative reduction of sensitive corpus;

T' - Test set;

T - Feature set;

a - Test cost function.

Due to the restriction of the Chinese translation corpus, after selecting the heterogeneity of the constraint features, the selected sub-reduction can be regarded as the selected process. The processing process can be expressed as follows:

$$W(Z, a_n) = \{D \subseteq A | a(D) \leq a_n\} \tag{4}$$

Among them: a_n - Upper limit of constraint feature test cost;

D - Heterogeneity feature output.

After analyzing the heterogeneity features, it is necessary to evaluate the impact potential of the heterogeneity feature nodes and realize automatic mining based on the impact potential of different nodes. Based on the research of previous scholars, this paper proposes an automatic mining method of heterogeneous features of prose Chinese translation corpus based on artificial intelligence.

2.3 Artificial Intelligence

2.3.1. Overview of Artificial Intelligence

Artificial intelligence is a kind of technological science. It uses computer systems to learn and imitate human intelligence, so as to achieve behaviors similar to human intelligence [15]. It is a way based on big data technology, through training specific goals, to generate a logical thinking ability and obtain a higher probability of occurrence. In this process, it uses an adaptive computer algorithm to produce a behavior and performance similar to that of human beings. However, AI is different from human intelligence. With the current level of science and technology, there is still a long way to go to make science and technology have the same intelligence as human beings. It has to be admitted that the computer is indeed superior to the human brain in some aspects. In terms of computing power, computers have surpassed human brains. The operation process of AI is shown in Figure 4.

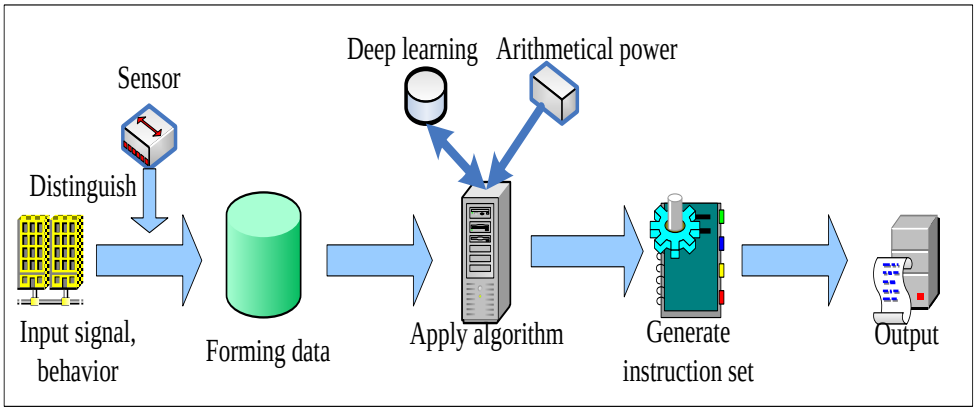

Figure 4: AI operation process diagram

Data, algorithms and computing power are the three pillars of the rapid development of AI. The algorithm can be said to be a kind of judgment logic, and the judgment of logic depends on its data characteristics. The source of data features is a large number of data that are similar to human cognition. Then, through operation, its specific tag is extracted and analyzed. Until familiar data is encountered, correct instruction operation can be performed. However, there are great requirements for computing power. The strength of computing power directly affects the amount of data that the machine can process. It can be said that computing power is the core of AI. Deep learning is a self-improvement algorithm, which sends instructions after recognizing its own characteristics. The machine can transmit the required behavior and information according to the command set issued.

The instruction set is a function that expresses the result, and its output expression and content must be understood by human beings. The instruction set is closely related to the development objectives of intelligent products. For example, the instruction set of the voice assistant is a previously set conversation. When it receives a command, it automatically extracts the corresponding reply from the corpus. The process of deep learning is a process of constantly expanding the corpus. With more and more knowledge learned, it becomes more and more

intelligent, and the conclusions drawn become more and more accurate. Therefore, this paper evaluates the impact potential of prose Chinese translation corpus nodes and excavates their heterogeneity characteristics by combining the deep learning algorithm under artificial intelligence.

2.3.2. Impact Potential and Automatic Mining of Corpus Nodes of AI

The above heterogeneous features are unified to form a formal set $a(l)$, because in different heterogeneous feature sets, they can be processed into node sets. The node processing expression can be as follows:

$$Q_m = f_m + \sum_{l=\gamma(m)} f_l\left(1 - a(l)\right) \tag{5}$$

Among them: $\gamma(m)$ - Adjacent node set of node m;

f_m - Aggregation coefficient of node m;

f_l - Aggregation coefficient of node l.

By using the nodes obtained by the above methods, the processing time of different nodes can be regarded as an attribute that affects the potential. It is expressed as:

$$W = \sum_{l=\gamma(m)} f_l \left(1 - \frac{2V_l}{f_l(f_l 1)-}\right) \tag{6}$$

Among them: V_l - Node complexity parameter.

When controlling the processing time, it is assumed that the influence of node m on node l is d_{ml}, the influence between nodes can be expressed by the following formula:

$$d_{ml} = \frac{1}{\sum_{y\in\gamma(l)}|Y|} \tag{7}$$

Among them: $\gamma(l)$ - The direct adjacent node set of node l;

Y - Node degree coefficient.

Due to the influence of nodes, there are nodes with different distances between nodes. In order to solve the unclear hidden danger caused by the different distances from the nodes in this part, this paper introduces the aggregation coefficient into the impact potential evaluation function and modifies it. The improved impact evaluation function can be expressed by the formula:

$$D_{ml} = \frac{Q_m}{Q_m + Q_l} * \left(1 - A(m)\right) \tag{8}$$

Among them: Q_m - Impact potential of node m;

Q_l - Impact potential of node l;

$A(m)$ - Aggregation coefficient of node m.

According to the size of influencing factors, the marginal effect of nodes is calculated, and its expression is as follows:

$$n(s|Z_m) = \frac{\omega(Z_m \cup \{s\})}{\omega(Z_m)} \tag{9}$$

Among them: $\omega(Z_m)$ - The scope of activating the translation corpus node;

s - Existing active node set.

The influence potential of the corpus nodes is evaluated by the above deep learning algorithm under artificial intelligence. The influence potential of the nodes is regarded as the heterogeneity mining feature, thus completing the automatic mining of the feature. This method can well realize the automatic mining of the heterogeneity features of prose meaning corpus, and maintain the stability of the variation of the error energy of the heterogeneity features.

3. Automatic Mining of Heterogeneous Features of Prose Chinese Translation Corpus

Based on the theoretical analysis of AI, this paper proposed a method to optimize the automatic mining of heterogeneous features, and theoretically analyzed the method. In order to verify the practical utility of this method, this paper also needed to conduct empirical research.

3.1. Methods for Automatic Mining of Heterogeneous Features of Prose Chinese Translation Corpus

In order to verify the feasibility of the automatic mining method of heterogeneity features in prose Chinese translation corpus based on artificial intelligence, this paper compared it with the traditional methods. The differences between this paper and the traditional automatic mining methods in terms of the amount of data node's out and in, the weight coefficient of heterogeneity features, and the node offset were compared, so as to draw relevant conclusions.

The experimental software equipment in this paper used Windows 2019 as the operating system, and the hardware equipment included the central processing unit (CPU). Its model is i5-11400, and the main frequency is 2.60GHz (gigahertz). The number of cores is 6, and the graphics card is 64GB (Gigabyte). The hard disk capacity is 1TB (Terabyte). The text index and the corresponding English index in the Chinese translation corpus were taken as the target of data processing, and the English discrepancy in the corpus was taken as the measurement standard. The actual effects of the two automatic mining methods were compared. The basic information about the selected dataset is shown in Table 1.

Table 1: Basic information of data set

Data set	Number of test sets	Characteristic sequence length
1	421	287
2	752	900
3	835	957
4	300	214
5	625	819
6	700	501

3.2 Experimental Results of Two Methods

3.2.1. Comparison of Input and Output Results in Node Volume

Outgoing is the number of arcs that take the vertex as the arc tail and start at the arc tail. The degree of penetration refers to the number of arcs ending at the vertex with the vertex as the arc head. The automatic mining method can mine the access degree of nodes. The more the access degree is mined, the better the performance of this method is. The specific mining situation of the two methods for the access degree of the selected data nodes is shown in Figure 5.

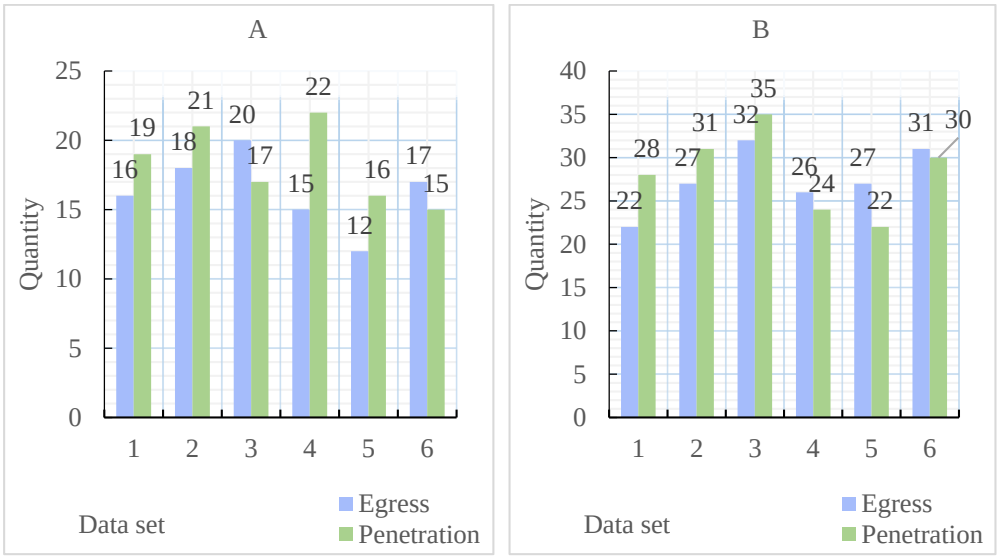

Figure 5 (A): The amount of data node access mining in traditional methods.
Figure 5 (B): The data node access mining volume of this method.

Figure 5: Data node access mining volume of two methods.

It can be seen from Figure 5 (A) and Figure 5 (B) that under the traditional method, the number of out-degrees of six data sets mining is 16, 18, 20, 15, 12, 17, and the number of in-degrees is 19, 21, 17, 22, 16, 15. However, the number of out-degree mining of the six data sets in this paper is 22, 27, 32, 26, 27, 31, and the number of in-degree mining is 28, 31, 35, 24, 22, 30. Compared with the traditional method, the method in this paper increases 6, 9, 12, 11, 15, 14 for the number of entries of 6 data sets, and 9, 10, 18, 2, 6, 15 for the number of entries. It can be seen that this method can better mine the egress and ingress of data set nodes. This means that the performance of this method is better, and the number of data nodes that can be mined is also more.

3.2.2. Comparison of Weight Coefficients of Heterogeneity Characteristics

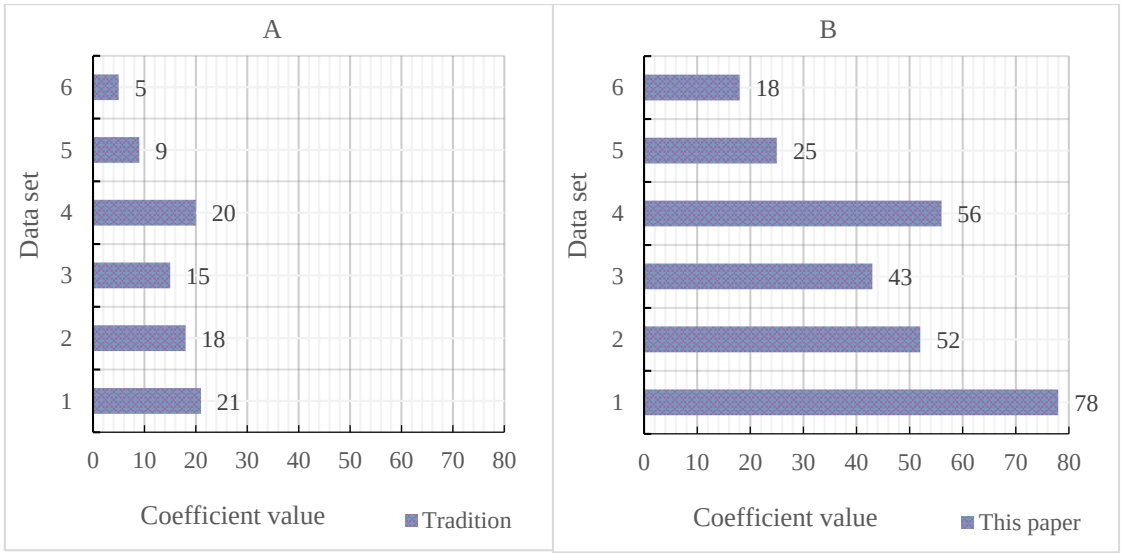

Figure 6 (A): The weight coefficient of the heterogeneity feature of the data set under the traditional method.
Figure 6 (B): The weight coefficient of the heterogeneity feature of the data set under this method.

Figure 6: Heterogeneity characteristic weight coefficient of data set under two methods

The weight coefficient is the importance of an indicator item in the indicator project system. It reflects the impact of the change of the indicator item on the evaluation results when other indicator items are unchanged. The weight coefficient of heterogeneity features is the most intuitive embodiment of testing automatic mining methods. The specific weight coefficient values of heterogeneity characteristics in six data sets under the two methods are shown in Figure 6.

It can be seen from Figure 6 (A) and Figure 6 (B) that the weight coefficients of the heterogeneity characteristics of the six data sets under this method are 78, 52, 43, 56, 25 and 18, respectively. The weight coefficients of the heterogeneity characteristics of the six data sets under the traditional method are 21, 18, 15, 20, 9 and 5 respectively. Compared with the traditional method, the weight coefficients of the heterogeneity characteristics of the six data sets in this method are improved by 57, 34, 28, 36, 16 and 13 respectively. It can be seen that the weight coefficient of the heterogeneity feature of the corpus is small due to the automatic mining under the traditional method. However, the weight coefficient of heterogeneity features under this method is relatively large, which is helpful to better analyze the different factors that affect the heterogeneity features in the mining process, so as to achieve a comprehensive mining of the heterogeneity features of prose Chinese-Chinese translation corpus.

3.2.3. Comparison of Node Offset Results

After analyzing the number of access degrees and the weight coefficient, the last thing to be verified is the offset. Offset refers to the distance from the real address of the memory unit to the segment bit position of the segment in which it is located. The smaller the offset, the higher the effectiveness and practicability of the method. The specific offset size of the six data set nodes is shown in Figure 7.

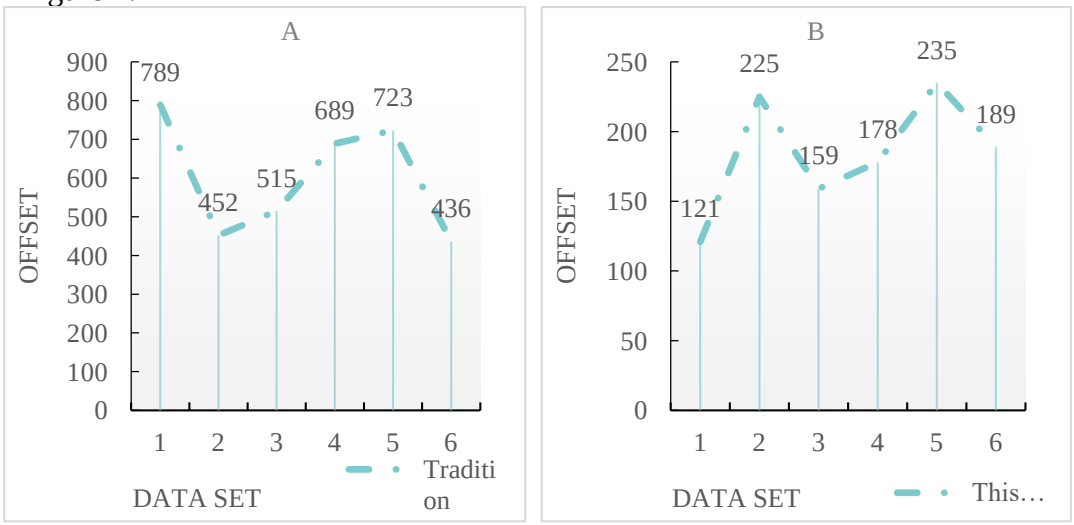

Figure 7 (A): Offset value of data set node under traditional method.
Figure 7 (B): Offset value of data set node under the method in this paper.

Figure 7: Comparison of data set node offset values under two methods

It can be seen from Figure 7 (A) and Figure 7 (B) that the offsets of the six data set nodes under the traditional method are 789, 452, 515, 689, 723 and 436 respectively, while the offsets of the six data set nodes under the method in this paper are 121, 225, 159, 178, 235 and 189 respectively. Compared with the traditional method, the node offset values of the six data sets in this method are reduced by 668, 227, 356, 511, 488 and 247 respectively. It can be seen that the data set node offset value based on the traditional automatic mining method is generally large, and the value is basically above 400. However, the data set node offset value of the automatic mining method based on

artificial intelligence is smaller, and the value is controlled between 100 and 250. Compared with the traditional method, the node offset of the dataset has been greatly reduced. This shows that the method in this paper can effectively reduce the offset of data set nodes, so that the offset of data set nodes is not too large, thus improving the effectiveness and practicability of the automatic mining method. This eventually promotes the excavation of the heterogeneity of the corpus of prose translation into Chinese, and make it more suitable for the current needs.

To sum up, through the research on the automatic mining method of the heterogeneity features of prose Chinese translation corpus based on artificial intelligence, it is found that the proposed method is feasible. Compared with traditional methods, this method can increase the weight coefficients of heterogeneity features of the six selected data sets by 57, 34, 28, 36, 16 and 13, respectively. At the same time, it can also significantly reduce the offset value of the dataset node, and control its value between 100 and 250. In addition, this method can better mine the number of egress and ingress in data nodes. On the whole, this method has better performance and more data nodes are mined. Its effectiveness and practicability are also higher.

4. Conclusions

With the faster and better development of science and technology and the integration of cultural diversity, the study of corpus has gradually attracted more attention. The topic of this paper was the automatic mining method of heterogeneous features of prose Chinese translation corpus based on artificial intelligence. First of all, the research background of the article was briefly summarized. Then the advantages and disadvantages of previous scholars in the research of corpus heterogeneity were summarized and analyzed, and the relevant theoretical analysis of corpus heterogeneity and artificial intelligence was carried out. Finally, combined with relevant theoretical content, the idea of applying AI to the automatic mining of heterogeneous features in the corpus of prose Chinese translation was proposed. In order to verify this assumption, this paper also carried out a practical verification. It was found that through the deep learning algorithm under AI, this method could realize the automatic mining of the heterogeneity features of prose Chinese translation corpus. It has also solved the shortcomings of traditional automatic mining algorithms and improved the efficiency of automatic mining of heterogeneous features, which has laid a theoretical foundation for further research on automatic mining of heterogeneous features in the future. However, this paper also has some shortcomings. Due to the limitations of actual conditions, the samples selected in this paper are few and have certain limitations.

References

[1] Lee Thomas R., and Stephen C. Mouritsen. (2021) "The corpus and the critics." The University of Chicago Law Review 88.2: 275-366.

[2] Feng Haoda, Ineke Crezee, and Lynn Grant. (2018) "Form and meaning in collocations: a corpus-driven study on translation universals in Chinese-to-English business translation." Perspectives 26.5: 677-690.

[3] Luo Jinru, and Dechao Li. (2022) "Universals in machine translation? A corpus-based study of Chinese-English translations by WeChat Translate." International Journal of Corpus Linguistics 27.1: 31-58.

[4] Davies Mark. (2021) "The coronavirus corpus: Design, construction, and use." International journal of corpus linguistics 26.4: 583-598.

[5] Li Yawen. (2021) "Heterogeneous latent topic discovery for semantic text mining." IEEE Transactions on Knowledge and Data Engineering 35.1: 533-544.

[6] Prabu P., R. Sivakumar, and B. Ramamurthy. (2021) "Corpus based sentimenal movie review analysis using auto encoder convolutional neural network." Journal of Discrete Mathematical Sciences and Cryptography 24.8: 2323-2339.

[7] Yang Lu, and Averil Coxhead. (2022) "A corpus-based study of vocabulary in the new concept English textbook series." RELC Journal 53.3: 597-611.

[8] Bryan Christopher J., Elizabeth Tipton, and David S. Yeager. (2021) "Behavioural science is unlikely to change the world without a heterogeneity revolution." Nature human behaviour 5.8: 980-989.
[9] Zaki Rezgar. (2022) "Observed and unobserved heterogeneity in failure data analysis." Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 236.1: 194-207.
[10] De Sutter, Gert and Marie-Aude Lefer. (2020) "On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach." Perspectives 28.1: 1-23.
[11] Rebechi Rozane, and Stella Tagnin. (2020) "Brazilian cultural markers in translation: A model for a corpus-based glossary." Research in Corpus Linguistics 8.1: 65-85.
[12] Park Chanjun, and Heuiseok Lim. (2020) "A study on the performance improvement of machine translation using public Korean-English parallel corpus." Journal of Digital Convergence 18.6: 271-277.
[13] Sayogie Frans, and Moh Supardi. (2021) "Equivalence Levels of Literary Corpus Translation Using a Freeware Analysis Toolkit." Buletin Al-Turas 27.1: 55-70.
[14] Vosiljonov Azizbek. (2022) "Basic theoretical principles of corpus linguistics." Academicia Globe: Inderscience Research 3.2: 1-3.
[15] Keding Christoph. (2021) "Understanding the interplay of artificial intelligence and strategic management: four decades of research in review." Management Review Quarterly 71.1: 91-134.