Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2172580, 10 pages
https://doi.org/10.1155/2022/2172580

Research Article

Numerical Analysis and Optimization of English Reading Corpus for Feature Extraction

Wu Juan, Xiong Wei, and Liao Hongyi

Jiangxi University of Engineering, Xinyu 338029, China

Correspondence should be addressed to Liao Hongyi; b20160904213@stu.ccsu.edu.cn

Received 7 June 2022; Revised 10 August 2022; Accepted 18 August 2022; Published 6 September 2022

Academic Editor: Jun Ye

In order to solve the problem of comparative analysis ability of English reading corpus, meet the needs of the construction of language feature analysis system of English reading corpus, make up for the shortcomings of English reading corpus, and improve the wide application of English, the construction method of English reading corpus is proposed. For a long time, English has been an important communication tool in terms of national politics and final economic communication. Especially with the in-depth development of economic globalization, English plays a vital role in international communication. Therefore, the study of the semantics and features of English texts can provide important information materials for English learning and communication in nonnative English speaking countries. Taking this as the starting point, this paper focuses on the numerical analysis and optimization of English reading corpus based on feature extraction and puts forward a language feature analysis system of English reading corpus. At the same time, feature extraction and text classification are used to improve the comparative analysis ability of English reading corpus. The research shows that the feature extraction oriented English reading corpus can solve the problem of English reading learning and improve the ability of reading and analyzing English articles.

1. Introduction

Language itself is an important tool to assist life, and corpus is a special language material and language material based on language. Corpus is extracted from things actually applied in real life. The extraction of the text content is mainly used to specify the regular rules according to the text content, match, search, and extract the target content, and export it to a new file. Its content is generally composed of written and oral language, while corpus is an important method of modern linguistic research formed on the basis of language and corpus and after a large number of collection, sorting and processing of corpus according to special rules. With the support of corpus, people can obtain language application rules or grammatical phenomena more regularly and conveniently. English corpus is a more widely used corpus, which plays a vital role in linguistic research and English Curriculum Teaching [1]. Especially with the development of English popularization, the existence of English corpus provides strong convenience for people to analyze and learn English. In order to ensure that the search of English corpus is more convenient and accurate, we must deeply analyze and process the relevant corpus by word segmentation, sentence segmentation, and classification. Put the corpus in one database to meet all requirements. Therefore, this paper proposes a language feature analysis system of English reading corpus based on Java web and improves the comparative analysis ability of English reading corpus with the help of feature extraction and text classification [2]. The structure of the web application is shown in Figure 1.

2. Literature Review

Ahsan et al. said that the research on Chinese text classification started in the early 1980s [3]. It has generally experienced three stages: feasibility study, auxiliary classification system, and automatic classification system. Rahman et al. believe that the research on the classification of Chinese texts is also based on the classification of English texts and

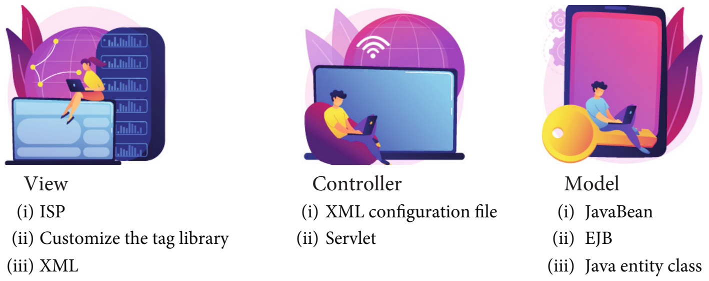

Figure 1: Structure diagram of web.

combined with the characteristics of Chinese texts to adjust the method and realize its application [4]. Liu et al. said that the main research units include dozens of colleges and universities, including the Institute of Computing of the Chinese Academy of Sciences and Tsinghua University, and have made a series of corresponding research results and experimental systems [5].

Massei said that in recent years, special education or representative education has also attracted the attention of many scholars and has been widely used in the practice of natural language and culture, exchange rate changes, and achieved positive results [6]. In 2003, Cheng and Jin first applied the distributed representation of words (also known as word representation or word embedding) to the statistical language model [7]. Zhang et al. proposed a hierarchical log bilinear model in 2008 [8]. Xu first introduced the calculation method of word embedding proposed by them in 2008. In 2011, it was combined with convolutional architecture to develop a Senna system for sharing word representation among multiple natural language processing tasks such as language model construction, named entity recognition, semantic analysis, and syntactic analysis, and achieved the best effect at present at a speed exceeding the traditional methods [9]. Xu and He extended the skip gram (continuous skip gram) model of continuous word bag and opened source word2vec, a tool based on deep learning, which provides an effective implementation of continuous word bag and skip gram architecture for calculating word vector [10]. In addition, some scholars have also done some research on the comparison of these word vector methods. For example, Joseph Turian compared the effects of several word vectors in 2010. American scholars have also done some research work on cross domain learning classification tasks, such as Xavier Glorot's cross domain emotion classification based on deep learning in 2011 and Bengio's deep learning research on feature representation for transfer learning in 2012.

All research on the linguistic corpus in the United States is independently developed by relevant universities. Since the linguistic corpus generally studies English and English is mostly the mother tongue of developed countries, the research is also relatively developed. However, there are some disadvantages in some research software. The number of software is insufficient, and the application of Computer Science in phrasal analysis of linguistic corpus is not paid enough attention. Most software development time is too early, and there is no update and secondary development, which makes the function disconnected. At the same time, there is no connection between software and software, and the function integration is very poor. In the analysis software wordsmith 50, you can make a vocabulary list of the article you are studying. In the process of making, you can count the frequency of a word in the article, and you can choose the arrangement of frequency or the arrangement of words. After the production, you can view the statistical results of frequency. Its function is to study the types of words in the corpus, determine the common lexical chunks in the corpus, and compare the frequency of specific words in different texts. Then, query and count the frequency of one or some words or phrases in the specified text, mark their positions, restore the sentences of 8 words on the left and 8 words on the right, and form moves at the same time. At the same time, we can compare multiple word lists and analyze the subject words. The analysis method is the subalgorithm to determine the subject words (subject words refer to those words whose frequency is significantly higher or lower than that of the corresponding words in the reference corpus). This software lacks the location map of displaying words, the vocabulary list of making multiword chunks, and the analysis after segmentation in paragraphs. The processing speed is relatively slow, the interface is not friendly enough, and the installation is troublesome. In the research of linguistic corpus, the use of computer technology to support the research of this discipline is a short board and cannot be used in the mobile terminal, resulting in the fixed research location. If you change the environment, you need to reinstall the software and configure the computer environment, which is very complicated, this is because there are scattered software, no integration, and uneven functions. In the research process, we should do a lot of preliminary work for the research projects and functions, understand the functions of each software, analyze its functions, and carry out research planning in combination with the research methods of linguistic corpus at this stage, which occupies a lot of research time, and the methods used are old means. Although the research methods of linguistic corpus have experienced a certain development time, the scalability of existing software is still insufficient.

3. Method

3.1. Overview of Java Web Technology. Java web is composed of JSP and Servlet. The so-called website development is actually the combination of JSP and Servlet. In JSP and

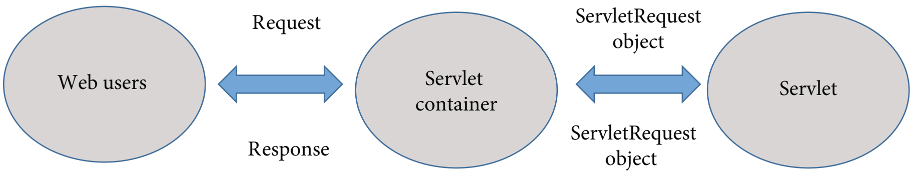

Figure 2: Servlet request process.

Table 1: Request response mode.

Name	Explain
HttpServletRequest	The Servlet container contains the HTTP request information in the HttpServletRequest object, and the Servlet component reads the user's request data from the request object. In addition, HttpServletRequest can store the shared data within the request range.
HttpServletResponse	The user generates HTTP response results.
Httpsession	The Servlet container creates an instance of Httpsession for each HTTP reply. Httpsession can store the shared data in the session range
ServletContext	The Servlet container creates a ServletContext instance for each web application, which can store the shared data of the application.

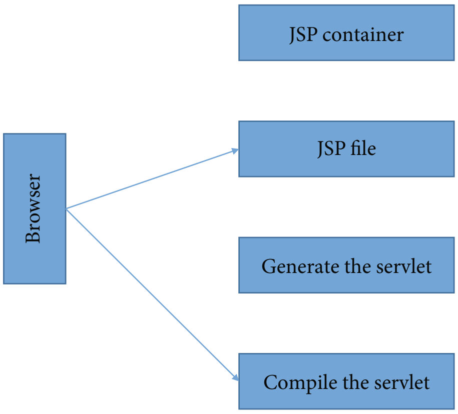

Figure 3: Analysis of JSP by container.

Servlet, JSP is responsible for realizing the function of front-end, while Servlet is responsible for realizing the function of background processing server. In the process of web project development, regardless of the choice of any framework structure, we need to rely on the mutual echo and combination of these two technologies to complete the development of web applications [11]. Servlet is responsible for processing technology and runs in the container, but its real running process is to perform its technology processing function when it is passively loaded in the container. For web applications, the working mode of Servlet is mainly to respond to the request of the client. Whenever the client initiates a request, the server will respond to the initiated request and push the results processed by the business layer to the client. There are two situations that will cause connection interruption. One is that the client request times out, and the other is that the browsing is disconnected. In the same Servlet container, each object corresponds to the corresponding request one by one. Then, the Servlet will fill the processed response results into the relevant attributes of the corresponding object Servlet response. The return process is in the charge of the container where it is located, and the content that the user can understand will be parsed by the browser and displayed in a corresponding way [12], as shown in Figure 2.

The following describes some of these classes, which are mainly used to deal with the corresponding methods of shared data and requests, as shown in Table 1.

JSP is similar to the principles of PHP/ASP and net/ASP. It is a tool for developing web applications based on HTML technology. The development languages used in the implementation of the server are different. There are two cases. The first is a single file. JSP will contain the file itself and the code content of the server and HTML. The second is multiple files. At this time, JSP will separate and save the code content of service order and HTML. When JSP recognizes Java code, it will add JSP tags before Java code to facilitate code reading [13], as shown in Figure 3.

3.2. Spring Overview. Spring's goal is to turn complexity into simplicity and make the coding and design of Java programs lively and interesting. This is one of the most important reasons why most Java programmers trace this open source framework [14]. Spring integrates modular programming, code writing, and testing, simplifies development steps, and provides convenience for Java programmers. According to the output function, the whole framework has five important output modules, including more than 1400 classes, as shown in Figure 4.

3.3. Bootstrap Overview. Bootstrap is an open source front-end development framework. Creating bootstrap is a framework for web front-end development used to build responsive websites [15]. Because bootstrap inherits the front-end development library of less (the latest version already contains the source code of SASS), many common CSS and JavaScript collections can be found in bootstrap, which is convenient for developers to call at any time. Bootstrap

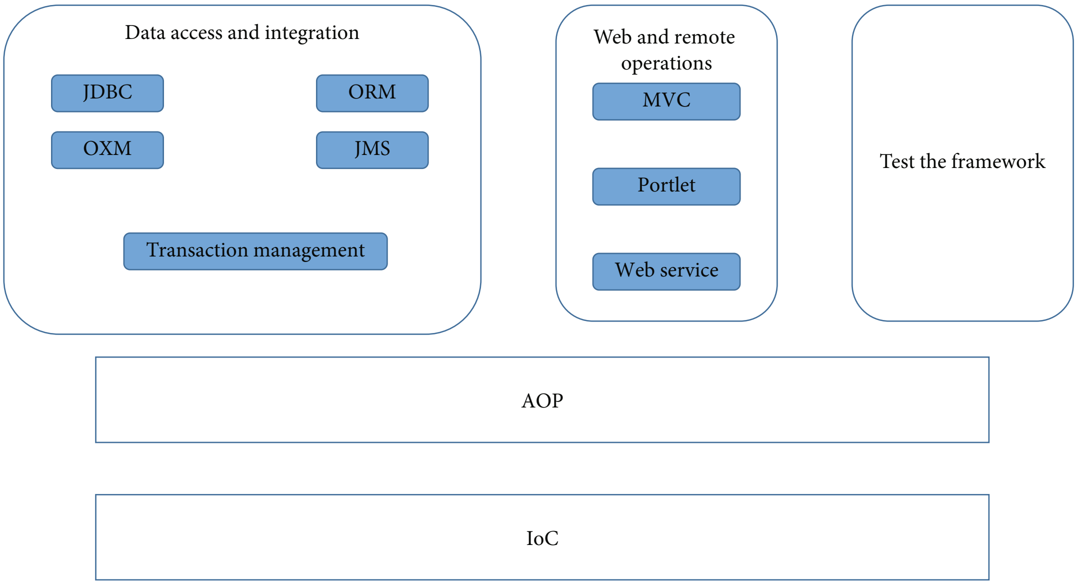

Figure 4: Spring framework.

has basic layout styles, JavaScript plug-ins, customized CSS style sheets, CSS components, grid system, and other functions [16].

(1) Basic layout style

Many layout styles can be found in bootstrap, such as tables, buttons, and forms. These styles can be applied to any HTML element. You only need to prestore the relevant CSS classes.

(2) JavaScript plug-in

Based on Jquery1.10 +, bootstrap has 12 JavaScript plug-ins. These plug-ins can provide rich user experience and website extension functions. When used in combination with the CSS group, these functions can achieve the effect of designing corresponding pages [17].

4. Results and Analysis

The system takes browser/server (B/S) as the basic architecture of the system, which is divided into three-tier structure of database, server, and user [18]. All the articles studied are stored in the English reading corpus in the form of text files. As the research corpus, they are retrieved by the server and used by users. Users can easily obtain the English reading corpus by accessing the web page through the browser. Pretrained models have been shown to generalize well to a given domain or into different modes. They show strong small-sample learning behavior and good learning ability. All business logic and functions are realized on the server side. The physical server is the analysis system server, while the logical server is the business server, database server, and web server. A computer software that manages resources and services users is usually divided into file servers, database servers, and application servers. The database (corpus) is connected with the database server. The user logs in through authentication and uses relevant functions. The user sends a service request to the server. The server accesses the query corpus through the connection with the corpus, completes the calculation of the required analysis function on the server, and then displays the analysis results to the user on the web page. Roles are divided into three categories: tourists, users, and administrators [19]. The relationships among the three types of roles are as follows:

Tourists use specific functions, which are described in the requirements analysis of the authentication module and will not be repeated here. Tourists log in as users through identity authentication and use the business functions in the server. From the administrator login portal, visitors can log in as administrators through authentication, add, delete, modify, and check the user data in the database server, and maintain the authority of the system [20].

An important part of the system is the English language analysis function. There are three problems to be solved in the realization of this part. The first is to carry out secondary development and integration by analyzing the past linguistic corpus analysis software and drawing lessons from the functions of Antconc, BFSU_Colloctor1.0, WordSmith5.0, Kf Nfram, and Claws4 software. The second aspect is to develop the analysis function after code assignment combined with the code assignment function of a university. The third aspect is to realize its research and analysis function. English reading corpus is a collection of English language materials stored on a computer to study how English is used, according to different research and analysis schemes (comparative study between corpus and corpus in corpus, comparative study between upload research factors and corpus in corpus, and comparative study between upload factors and upload factors). The system functions are divided into general functions, characteristic functions, and special functions according to the use frequency and research order. Common functions include English reading related data analysis and vocabulary search.

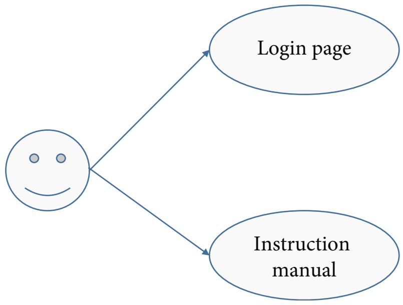

Figure 5: Use case diagram of tourist login role.

Corpus is the basic resource for corpus linguistics research and the main resource for empirical language research methods. It can be applied to lexicography, language teaching, traditional language research, and statistical or example-based research in natural language processing. Language feature analysis is inseparable from the choice of vocabulary or lexical chunks. There are simple and accurate search and complex regular expression search in the demand. Search is the basis of research and is used frequently [21].

4.1. Authentication Module

4.1.1. The Role of Tourists. The system is not a free and open source project, so set the identity of tourists. Tourists can only browse the user manual on the login page and consult the use methods and main functions of the system. If they are in a nonlogin state, they cannot use the functions of the system and can only have a better understanding. If you need to use the system, you need to contact the administrator to obtain the user name and password to use [22], as shown in Figure 5.

4.1.2. User Role. Visitors check the user name and password in the database through the user name and password to log in. After logging in, they become users, that is, they can use all the functions in the system. If you want to change your password or account name, or forget your user name or password, you need to contact the administrator to change or find your user name and password. When you add security questions when setting up your local account, you can answer security questions to log in again.

4.1.3. Administrator Role. The administrator is only responsible for adding, deleting, modifying, and querying users. The administrator is responsible for all permissions in the system. If tourists apply to use the system, contact the administrator. The administrator creates user account and password by default or according to the user's requirements [23]. If users want to change or query their user name or password, they also need to contact the administrator for operation, but the administrator cannot use the system and exists only as user management [24], as shown in Figure 6.

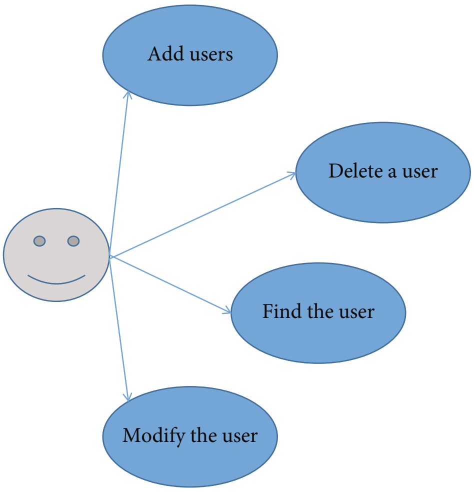

Figure 6: Use case diagram of administrator login role.

4.2. General Function Module

4.2.1. Vocabulary Search and Data Analysis. For a word in an article, match a selected phrase, an uploaded phrase, or search to find it correctly. As required, finding a word in each sentence is the easiest task. The corpus contains the language material that appears in the actual use of a language, so it should not usually be counted as a corpus. By understanding the requirements, he discovered that the task was to search not just single words, but also word chunks, with many variations. Real research is aimed at improving and preparing future projects [25].

In the process of density analysis, in order to make the searched words or lexical chunks more intuitive, the function of black-and-white location map needs to be introduced. Black and white location map is an incidental function combined with density analysis. It is made up of a single point called pixels (picture elements). These points can be differently arranged and stained to form the pattern. Count and display the position of a single word or lexical chunk in the article or corpus, and show it with a black bar chart on a white background, which is equivalent to hot spot analysis. It intuitively displays the specific position of the text search word, so as to facilitate the researchers to identify the distribution of words or lexical chunks in the article or corpus [26].

4.2.2. MI Value. Mutual information value (MI value for short) is a common method used to calculate the collocation strength of words. Mutual information indicates whether the two variables X and Y are related to each other, and the relationship between them. The MI value is calculated in bits. The use of MI in linguistics is different from its common methods in finance, especially its numerical range [27]. In Information Science, the MI value is in the range of 0~1, while the MI value used between linguistic words is only separated by 0 bits. The larger the value, the greater the mutual encounter and attraction between words. Mutual information is actually a broader special case of relative independence. If the variables are not independent, then we can judge whether they are "close" to being mutually independent by examining the Kullback-Leibler divergence between the product of the joint probability distribution and the edge probability distribution. Specifically, the MI value calculates the frequency of one word in the corpus and can provide information about the probability of another word [28].

For example, a and B are two random words divided by body, the total potential of body is w, and S is the interval. The actual observed frequencies of their bodies are $F(a)$ and $F(b)$, and the resulting frequency is $F(a, b)$. The MI value is calculated as shown in the following equation:

$$I(a, b) = \log_2 \bullet \frac{P(a, b)}{P(a) \bullet P(b) \bullet 2S} = \log_2 \bullet \frac{F(a, b) \bullet W}{F(a) \bullet F(b) \bullet 2S}. \quad (1)$$

If a body's potential is w-word determinant, $F(a)$ is a frequency test for multiword sequences or lexical-driven models, $F(b)$ is a collocation frequency test for multiple temporal or lexical-driven models a, and $F(a, b))$ is the frequency of the sum of the two parts of speech, and the MI value can be calculated according to the following formula:

$$I(a, b) = \log_2 \bullet \frac{W \bullet F(a, b)}{F(a) \bullet F(b)}. \quad (2)$$

4.2.3. Z Value. If the unit total capacitance is w and the nominal frequency of the integer in the body is C_1, then the average frequency of the integer in each function is calculated as C_1/W . If the collocation span is limited to s, the frequency of the collocation with each node word is $C_1 \cdot (2S + 1)/W$. $2S$ refers to the span position set on the left and right sides of the node word, and 1 is the word position occupied by the node word. However, this form can contain block words and similar phrases, so the word position in the design will not be 1. However, when determining the probability of occurrence of a word with probability frequency N, the theoretical probability P must be calculated according to the following equation:

$$P = \frac{C_1(2S + 1)}{s} \bullet \frac{N}{s}. \quad (3)$$

The desired coupling frequency of the coupling term can be obtained by giving the theoretical P value for the coupling produced by the potential W. Then, the combined frequency of the combination item and the instruction item is shown in the following equation:

$$SD = \sqrt{(2S + 1) \bullet N \bullet \left(1 - \frac{C_1}{s}\right) \bullet \frac{C_1}{s}}. \quad (4)$$

Next, calculate the difference of the collocation distribution in the text, as shown in the following equation:

$$E = \frac{C_1(2S + 1) \bullet N}{W}. \quad (5)$$

Divide the difference between the actual frequency C_2 and the expected frequency e of the node words of the collocation by the standard deviation to obtain the Z value. The value of Z value can be used to judge the strength of word collocation. The Z value of word collocation needs to be at the level of 0.01 to be significant, and the Z value must be equal to or greater than 2.576. By setting a critical value of 2.576, researchers can obtain significant word collocations and filter out accidental collocations that have no effect on node words, as shown in the following formula:

$$Z = \frac{C_2 - E}{SD}. \quad (6)$$

4.2.4. Log Likelihood Function Value. In linguistic English reading corpus, the value of log likelihood function is mainly used to determine the collocation relationship value of two words in the left and right span of node words, display the collocation rate in the left and right span, and quickly locate the search for effective collocation words near node words. When A is set as the vocabulary of corpus A, B is defined as the vocabulary of corpus B, c is the frequency of words in corpus A, and d is the frequency of words in corpus B, the algorithm is shown in the following formula:

$$\text{log} - \text{likelihood} = 2c \bullet \log_e \frac{c}{A \bullet c + b/A + B} + 2d \bullet \log_e \frac{d}{B \bullet c + d/A + B}. \quad (7)$$

4.2.5. T Value. T value is a relative position quantity of the most common straight-line conversion standard score, which is used to represent the relative position of an individual in its group. In linguistics, it is to describe the weight of vocabulary in its research factor compared with that of another vocabulary in the same research factor. The basic principle is that the original score of an individual has several standard deviations above or below the average, which is the Z value. The time sharing obtained by expanding the Z score is the T value, as shown in the following formula:

$$T - \text{Score} = \frac{N \bullet F(n, c) - FN \bullet F(c)}{N \bullet \sqrt{F(n, c)}}. \quad (8)$$

4.2.6. K-SVD Algorithm. An over complete dictionary is used to represent any known signal. The number of atoms in the dictionary is K. After sorting all atoms according to the column vector, it is set as the following formula; then, the signal y can be approximately expressed as the following formulas:

$$D \in R^{n \times k}, \quad (9)$$

$$\{d_j\}_{j=1}^{K}, \quad (10)$$

$$y \approx Dx. \quad (11)$$

In the above formula, the sparse representation coefficient of signal y is $X \in R^k$. The solution process of signal sparse representation is transformed into the following optimization problem, as shown in the following formula:

$$\hat{x} = \arg \min_x \|x\|_0. \quad (12)$$

The constraint conditions of this formula are shown in the following formula:

$$\|y - Dx\|_2 \leq \varepsilon. \quad (13)$$

In the above formula, the zero norm of x is $\|x\|_0$, which refers to the number of nonzero elements in vector x, and the smaller positive value is ε. In order to obtain the optimal dictionary D and the most sparse coefficient matrix X, the problem of solving the known signal $Y = \{y_1, y_2, \cdots, y_N\}$ is transformed into the optimization problem described by the following expression, as shown in the following formula:

$$\left(\widehat{D}, \widehat{X}\right) = \arg \min_{x} \|Y - DX\|_F^2. \tag{14}$$

The constraint conditions of this formula are shown in the following formula:

$$\|x_i\|_0 \leq T_0, \forall i = 1, 2, \cdots, N. \tag{15}$$

In the above formula, the column vector of coefficient matrix X is x_i, the F norm of matrix A is defined as $\|A\|_F$, and the sparsity is defined by t0.

4.3. Special Functional Requirements. Concordance is a research method provided according to the needs of linguistic English reading corpus language feature analysis. Concordance retrieval is mainly used to study the frequency of random chunks in articles. Each discipline has different key points and keywords, and there are also differences in the mode of language expression. When it comes to professional knowledge, the composition of lexical chunks is different, resulting in different collocation of technical language. Consistency retrieval is mainly to analyze the above situation.

The function of thesaurus production is to serve the following functions of thesaurus production. Starting from the analysis of linguistic English reading features, the list of words can roughly know the main thrust of the article and predict the research direction of lexical chunks through the number of hits.

Keywords are the function of directly referring to the context and meaning words screened from the research factors, which is the top priority. Compare according to the theme words. Generally, the comparison is made between two theme words. The theme words of an article cannot be analyzed only according to the frequency of words or lexical chunks. The key function is the Keyness coefficient. The value of K determines the importance of words, and screen the high-frequency words of each article. Suppose A is the frequency of the same word or lexical chunk in corpus A, B the frequency of the same word or chunk in corpus B, C the vocabulary of corpus A, and D the vocabulary of corpus B. The algorithm is shown in the following formula:

$$\text{Keyness} = (A + B + C + D) \bullet \frac{[\|A * D - C * B\| - (A + B + C + D)/2]^2}{(A + C) \bullet (B + D) \bullet (A + B) \bullet (D + C)}. \tag{16}$$

The specific information of the model building platform is shown in Table 2.

Table 2: Statistics of development platform information.

Name	Content
Development language	Java
Development environment	Eclipse .JDK
Operating system	Uhuntu 14.04
Main tools	HanLP, Weka

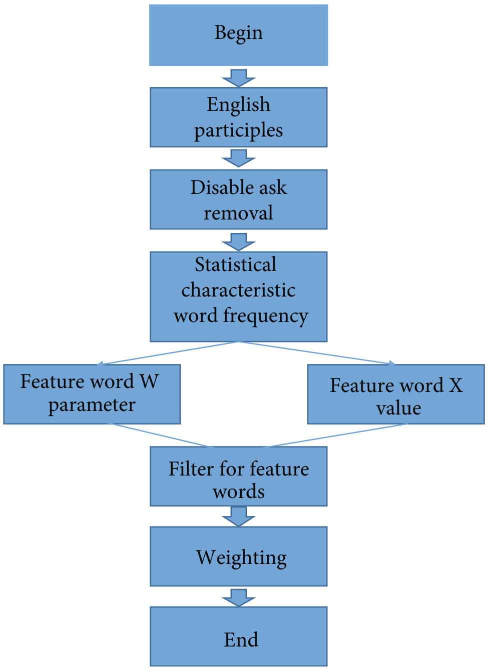

Figure 7: Schematic diagram of model operation flow.

As a Java toolkit in the field of natural language processing, HanLP has many advantages, such as efficient processing speed, real-time updating corpus, and user-defined processing mode. Moreover, the tool, including dictionary, is completely open source, with fast word segmentation speed and small memory occupation; Weka is one of the open source data mining platform tools, which is mainly used for data processing operations such as regression, classification, and acquisition of association rules.

The English reading corpus preprocessing module consists of three functions: word segmentation, removing stop words, and dependency parsing. After processing the input initial text data, two sets of output data are obtained, namely, the set of text feature words without stop words and the set of dependent word pairs. The four functions of effective dependent word pair acquisition, feature word frequency statistics, network edge weight determination, and directed network architecture are used to form a text network construction module for further processing the output data of the preprocessing module. The output data of this module is text network data and feature word frequency statistics. The feature extraction module is composed of three functions: inverse document frequency, feature word w parameter solution, and feature word extraction. The dimension-reduced feature space output data is obtained by calculating the output data of the previous module. The function of calculating the weight of feature words is the feature weighting module, which is used to assign values to the feature items in the feature vector space.

The specific operation process of English reading corpus word segmentation feature extraction model is shown in Figure 7. The operation contents of each stage are described as follows:

(1) *Text Preprocessing Stage*. Use the open source language processing tool HanLP to perform word segmentation and dependency parsing on the input text data, and remove the stop words contained in the feature word set and dependency word pair set

(2) *Text Network Construction Stage*. After counting the word frequency of the feature word set, the feature word is used to construct the text network node to obtain the effective dependent word pair set, complete the construction of the text network edge and its weight, and obtain the weighted directed text network

(3) *Feature Extraction Stage*. The w parameter value of the feature word node is solved by the improved K-SVD algorithm, and the value of the corresponding feature word is solved according to the word frequency information of the feature word. After comparing the weight, the feature word screening is completed to realize feature extraction

(4) *Feature Weighting Stage*. Weigh the feature words after feature extraction. The operation method is the same as the feature extraction method. The weighted value of the feature words is the weight of the solved feature words

In order to verify the performance of the constructed model, a comparative verification experiment is designed. The experimental data selects English news reading materials in recent ten years as the corpus. The number of news articles is about 511065, and the number of sentences is about 21412825. The training data is randomly extracted from the news content. The overall iterative training of the English reading corpus takes about 3 days, and each training time is about 2 seconds.

(1) Comparative analysis of word segmentation feature output results

Under the same experimental data, the comparison experiment of word segmentation feature output accuracy is carried out by using design model and model a and model B. The comparison results of word segmentation feature output accuracy of the three models are shown in Figure 8.

As can be seen from Figure 8, for feature extraction, the model has a high accuracy of word segmentation feature output in the increasing experimental time, which shows that the design model has a better extraction effect and can better master the semantics and grammar of the original sentence. The output accuracy of word segmentation features of the two literature comparison methods shows a gradual downward trend. The above comparison results fully prove the good performance of the design model.

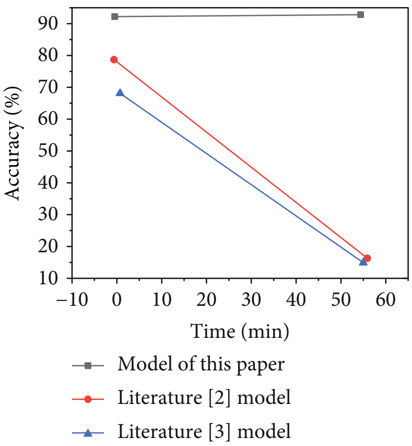

Figure 8: Comparison results of word segmentation feature output accuracy.

(2) Comparative analysis of accuracy and recall of output results

Using the model, the accuracy and recall of feature extraction are obtained. The accuracy and recall of the model have more significant advantages, and the feature extraction is more accurate. The reason is the introduction of K-SVD optimization algorithm, which considers the word length of feature words. After accurately screening word segmentation features and removing stop words, the model uses feature words to construct text network nodes, obtain an effective set of dependent word pairs, complete the construction of text network edges and their weights, obtain a weighted directed text network, achieve a relatively balanced accuracy and recall, and complete the effective extraction of word segmentation features.

5. Conclusion

In the word segmentation stage of English reading corpus, to remove redundant features and extract features conducive to classification, it is necessary to reduce the dimension of the feature vector space. Among them, one of the most commonly used and effective methods is feature extraction, which uses the most representative feature items of text category information to complete the component word task. Compared with the existing analysis system, there is only the comparison means between corpus and corpus. This system adds the function of corpus upload and comparison, which makes the system add the function of corpus and upload corpus and the function of comparative analysis between upload corpus and upload corpus, and expands the application scope of the system. The work of this paper will lay a solid foundation for the analysis of English language features of the linguistic corpus of a university and solve the embarrassing situation that linguistics has research means but does not realize methods to a certain extent. At the same time, through the innovation of function and research object, the scientific research of linguistics has turned a new page. It is proved that the construction of language feature analysis system of English reading corpus can effectively solve the problem of comparative analysis ability of English corpus, meet the needs of global economic and trade exchanges, make up for the lack of comparative analysis ability of English reading corpus, and improve the development of global economy and trade.

Data Availability

The data used to support the findings of this are available on request from the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Provincial Project for the 2019 Teaching Reform Research of Colleges and Universities in Jiangxi Province: analysis and research on the graded teaching of College English in private colleges (No.: jxjg-19-28-8).

References

[1] R. G. Hashish and M. Zeidouni, “Injection profiling through temperature warmback analysis under variable injection rate and variable injection temperature,” *Transport in Porous Media*, vol. 141, no. 1, pp. 107–149, 2022.

[2] S. Kushwah, S. Parekh, H. Mistry, M. Bhatt, and V. Joshi, “A methodological study of leaf spring by material comparison and Taguchi's DOE,” *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 16, no. 1, pp. 239–252, 2022.

[3] Z. Ahsan, H. Dankowicz, M. Li, and J. Sieber, “Methods of continuation and their implementation in the COCO software platform with application to delay differential equations,” *Nonlinear Dynamics*, vol. 107, no. 4, pp. 3181–3243, 2022.

[4] A. U. Rahman, M. Saeed, and F. Smarandache, “A theoretical and analytical approach to the conceptual framework of convexity cum concavity on fuzzy hypersoft sets with some generalized properties,” *Soft Computing*, vol. 26, no. 9, pp. 4123–4139, 2022.

[5] J. Liu, J. Li, Z. Wang, Y. Tian, and H. Wang, “Optimization of heating process for bearing rings in a vacuum furnace based on numerical analysis,” *ISIJ International*, vol. 61, no. 1, pp. 302–308, 2021.

[6] S. Massei, “Some algorithms for maximum volume and cross approximation of symmetric semidefinite matrices,” *BIT Numerical Mathematics*, vol. 62, no. 1, pp. 195–220, 2022.

[7] J. Cheng and H. Jin, “An adaptive extreme learning machine based on an active learning method for structural reliability analysis,” *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 43, no. 12, pp. 1–19, 2021.

[8] G. Zhang, Y. Song, S. Liao, L. Qu, and Z. Li, “Uncertainty measurement for a three heterogeneous information system and its application in feature selection,” *Soft Computing*, vol. 26, no. 4, pp. 1711–1725, 2022.

[9] D. Xu, “Observability inequalities for Hermite Bi-cubic orthogonal spline collocation methods of 2-D integro-differential equations in the square domains,” *Applied Mathematics & Optimization*, vol. 84, no. 2, pp. 1341–1372, 2021.

[10] S. Xu and B. He, “A parallel splitting alm-based algorithm for separable convex programming,” *Computational Optimization and Applications*, vol. 80, no. 3, pp. 831–851, 2021.

[11] Y. Li, “Numerical analysis and optimization of feature extraction-oriented english reading corpus,” *Mathematical Problems in Engineering*, vol. 2022, Article ID 9883201, 13 pages, 2022.

[12] F. Mohammaddokht and J. Fathi, “An investigation of flipping an English reading course: focus on reading gains and anxiety,” *Education Research International*, vol. 2022, Article ID 2262983, 10 pages, 2022.

[13] L. Chang and G. O. Deák, “Adjacent and non-adjacent word contexts both predict age of acquisition of English words: a distributional corpus analysis of child-directed speech,” *Cognitive Science*, vol. 44, no. 11, p. e12899, 2020.

[14] X. He, “An English reading and learning system based on web,” *Scientific Programming*, vol. 2021, Article ID 7281269, 8 pages, 2021.

[15] G. Ling, “Corpus-driven resource recommendation algorithm for English online autonomous learning,” *Computational and Mathematical Methods in Medicine*, vol. 2022, Article ID 9369258, 10 pages, 2022.

[16] F. Ghanami, G. A. Hodtani, B. Vucetic, and M. Shirvanimoghaddam, “Performance analysis and optimization of NOMA with HARQ for short packet communications in massive IoT,” *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4736–4748, 2020.

[17] A. K. Sarnaghi, A. Rais, A. Kovryga, W. F. Gard, and J. W. G. V. D. Kuilen, “Yield optimization and surface image-based strength prediction of beech,” *European Journal of Wood and Wood Products*, vol. 78, no. 5, pp. 995–1006, 2020.

[18] Y. Ding and T. Wang, “Environmental affection-driven English tense analysis: a healthcare exercise-based corpus case study over public English environment,” *Journal of Environmental and Public Health*, vol. 2022, Article ID 9497554, 8 pages, 2022.

[19] Z. Sun, “Development of corpus linguistic using lexical teaching to improve English writing,” *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 4024149, 7 pages, 2022.

[20] G. Yufang, W. Wu, M. White, H. Aziz, and K. Liew, “Evaluation of college English textbooks based on computer-aided analysis corpus,” *Security and Communication Networks*, vol. 2022, Article ID 4648957, 7 pages, 2022.

[21] Q. Dai, “Construction of English and American literature corpus based on machine learning algorithm,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 9773452, 9 pages, 2022.

[22] S. Wang and X. Shi, “Research on correction method of spoken pronunciation accuracy of AI virtual English reading,” *Advances in Multimedia*, vol. 2021, Article ID 6783205, 12 pages, 2021.

[23] S. Huang, “Optimization and simulation of an English-assisted reading system based on wireless sensor networks,” *Journal of Sensors*, vol. 2022, Article ID 6823502, 11 pages, 2022.

[24] X. Yu and L. Zhang, “Effectiveness of interactive reading mode based on multisensor information fusion in English teaching,” *Mobile Information Systems*, vol. 2022, Article ID 7993728, 12 pages, 2022.

[25] R. Futrell, E. Gibson, H. J. Tily et al., “The natural stories corpus: a reading-time corpus of English texts containing rare syntactic constructions,” *Language Resources and Evaluation*, vol. 55, no. 1, pp. 63–77, 2021.

[26] L. Lowphansirikul, C. Polpanumas, A. T. Rutherford, and S. Nutanong, “A large English-Thai parallel corpus from the web and machine-generated text,” *Language Resources and Evaluation*, vol. 56, no. 2, pp. 477–499, 2022.

[27] G. Fatima, R. M. A. Nawab, M. S. Khan, and A. Saeed, “Developing a cross-lingual semantic word similarity corpus for English-Urdu language pair,” *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 2, pp. 1–16, 2021.

[28] L. Plug, R. Lennon, and E. Gold, “Articulation rates' intercorrelations and discriminating powers in an English speech corpus,” *Speech Communication*, vol. 132, pp. 40–54, 2021.