Hindawi
Mobile Information Systems
Volume 2022, Article ID 2652622, 10 pages
https://doi.org/10.1155/2022/2652622

*Research Article*

# Automatic Recognition and Extraction of English Verb Types Based on Index Line Clustering

**Hui Zhao**,[1,2] **Kexin Jin**,[3] **and Jing Wang**[3]

[1]*Office of Development and Planning, Chengdu Vocational & Technical College of Industry, Chengdu 610000, Sichuan, China*
[2]*School of Rail Transit, Chengdu Vocational & Technical College of Industry, Chengdu 610000, Sichuan, China*
[3]*Office of Educational Administration, Chengdu Vocational & Technical College of Industry, Chengdu 610000, Sichuan, China*

Correspondence should be addressed to Hui Zhao; zhaohui@cdivtc.edu.cn

Received 17 May 2022; Accepted 4 July 2022; Published 20 July 2022

Academic Editor: Abid Yahya

Copyright © 2022 Hui Zhao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Languages are not uniform and certain words are used differently by speakers of different languages more or less often, or with distinct meanings. In both linguistics and natural language processing (NLP) problems, the classification that groups together verbs and a collection of similar syntactic and semantic features are of great interest. In the modern era of science and technology, NLP technology is developing rapidly. However, the interpretation of index lines still needs to be realized manually. This method takes a long time, especially after entering the era of big data, the number of corpora has increased rapidly and it is normal to have a corpus with hundreds of millions of words. The quantity of text generated every day is increasing intensely and the word index based on search words is as high as tens of thousands of lines, so it is very difficult to analyze index lines manually. Automatic lexical knowledge acquisition is essential for a variety of NLP activities. Particularly knowledge about verbs is critical, which are the major source of relationship information in a sentence. Due to this issue, this study attempts to automatically identify and extract English verbs by index line clustering. Each index behavior can be regarded as microtext automatic clustering to realize the automatic identification and extraction of English verb forms. This study first focuses on the clustering index algorithm including the C-means clustering algorithm and fuzzy C-means clustering algorithm, then describes in detail the automatic recognition and extraction process of English verbs based on index line clustering, and creates a verification set and completes the index line clustering of English verbs. Finally, the effect of index line algorithm is analyzed from two aspects: automatic recognition of English verb types and recall rate. At the same time, the verbs are selected to analyze their types and judge the probability of each type. The experimental results show that the average recognition rate of English verbs in the manual classification is 91.01%, and the average accuracy of automatic recognition and extraction of English verb patterns based on index row clustering is 95.99%.

## 1. Introduction

Verbs in every language may be classified into semantic classes that share similar aspects of meaning. The semantics of a verb at least partially controls its syntactic behavior and is widely known in linguistics [1]. Verbs can be semantically categorized based on their syntactic alternation behavior in subcategory frames and their argument selection preferences inside those frames. Automatic lexical knowledge acquisition is essential for a variety of NLP activities. Knowledge of verbs, which are the main source of relational information in a sentence, is particularly critical [2]. NLP is a branch of AI that allows machines to interpret human speech. NLP integrates linguistic and information science to study the principles and structure of the language and build expert systems capable of reading, analyzing, and extracting information from the text and voice (using machine learning (ML) and NLP algorithms). In NLP, automated identification and correction of grammar, spelling, word order, and punctuation errors discovered in English text written by non-native language learners are essential [3]. Due to a large amount of language vocabulary, the complexity of grammatical rules, the ambiguity of semantics, and the ambiguity of speech are the only ways to enhance language processing and recognition and employ computer programming technology to examine and investigate these challenges [4].

There are currently three techniques for correcting grammatical errors: the first is a rule-based method, which formulates particular error correction rules to fix certain sorts of errors, which depends on the quality of the rules and it can only change certain types of errors. The second method is based on statistics, and it extracts text characteristics from associated information in words, models, and language and then selects applicable statistical models to fix text problems. The third method is a depth-based strategy that uses word vector representation to build a deep neural network that corrects text mistakes from beginning to end, and it is independent of error kind [5]. Semantic verb classes generalize semantic qualities across verbs, capturing substantial quantities of verb meaning without describing the distinctive specifics for each verb. The classes belong to a general semantic level, and the idiosyncratic lexical semantic properties of a verb are either included to the class description or left unspecified [6].

With the development of globalization, the relationship between China and other countries has become increasingly close. Our country gives great importance to English education for students. There are a large number of English verb patterns in the process of English teaching and teachers cannot list them all in the process of teaching. It is difficult and inefficient to interpret English verb patterns manually, which makes it impossible for students to master more English verb contents [7]. To solve this problem, this study uses the index row clustering method as a mini-text for each index behavior to complete the clustering operation. Indexed row clustering allows you to quickly process large datasets and tag English verb patterns based on a list of existing styles that match text strings in the corpus [8]. The corpus-based research is one of the most fundamental transformations in modern linguistics. A corpus is a systematic collection of real texts saved electronically that may be utilized to uncover language information [9]. The feasibility and necessity of automatic recognition of English verb patterns are strong such as the patterns can be accurately described, can reduce the confusion of language users, and can help them to use English verb patterns more accurately [10, 11].

The main contributions in this research process are as follows: (1) detailed description of the clustering index algorithm used in this study including the C-means clustering algorithm and fuzzy C-means clustering algorithm, which are the basis of this study. (2) Focus on the automatic recognition and extraction of English verb patterns based on index row clustering, create a verification set of English verbs, establish index row clustering, introduce the algorithm process of extracting patterns, and complete the extraction of English verb patterns according to this process.

The rest of the article is structured as follows: Section 2 illustrates the related work, Section 3 demonstrates the materials and methods, and Section 4 represents the results and discussion. The research study is finally concluded in Section 5.

## 2. Related Work

Nowadays, the study of English patterns has become a major topic due to the rapid development of English corpus linguistics [12]. The term "corpus" is derived from the Latin word "corpus," which means "body." It could be used to refer to any written or spoken information. But in current linguistics, this phrase refers to vast collections of texts that provide a sample of a given variety or usage of languages and are supplied in machine-readable format [13]. Arai et al. [14] summarized and analyzed thousands of English forms such as English nouns, English verbs, and adjectives and sorted out the concept, recognition standard, type relationship, and significance of English forms. S. Liu and W. Liu developed a classification-based fundamental model for English grammar error correction, assessed the classification and translation models for English grammar error correction, and presented an English grammar error correction algorithm based on the classification model [4]. Megariani et al. proposed that the components of phrase word bits include nouns, verbs, adjectives, and complement components. The word bits of the phrase are called patterns [15]. Balakrishnan et al. [16] used the Bank of English corpus to summarize and analyze the valence patterns of 274 nouns, 511 verbs, and 544 adjectives in English and pointed out that the element code similarity in English coding method and pattern grammar is high. However, English patterns should describe all objects exhaustively and abstract summarizes all patterns of various parts of speech.

A unique approach for automated thesaurus creation has been presented by Bourigault and Jacquemin It is based on the employment of two tools in combination: (1) a word extraction tool that extracts term candidates from tagged corpora using a shallow grammar of noun phrases and (2) a term clustering tool that clusters syntactic variations (insertions) [17]. Newman et al. proposed a word segmentation model based on the Dirichlet process (DP), in which multiword segments are either retrieved from a cache or newly generated [18]. It is an unsupervised approach for detecting index terms in a document collection as well as key terms for a single document. Wu et al. [19] used the chain parser method combined with the type grammar method to automatically analyze learners' language type errors. Naismith et al. proposed that at least 100 million word-level corpora should be used to better find English patterns. The content of corpus inventory is small, and a large number of language phenomena can be committed [20]. Smith et al. summarized and analyzed the interpretation of index lines based on the extended meaning unit method, forming a complete system analysis concept [21]. Chen first tried to automatically identify English verb patterns and match the strings with the text in the corpus based on the current list of existing patterns, so as to complete the verb pattern marking [22].

Zhou et al. proposed that English linguistics is a complex subject, unifying the specification and integration of various types of corpora, strengthening English linguistics, and expanding the translation range and oral application range of English corpora [23]. Ağçam refers to the academic oral corpus of the University of Michigan, compares the differences in the expression patterns of cognitive position markers in Chinese and foreign academic English with three types of cognitive position markers such as adverbs, verbs,

and adjectives, analyzes the factors that the total frequency of cognitive position markers is higher than that of native speakers in the process of academic oral communication, and analyzes the disadvantages of fuzzy adverbs [24]. Shei et al. start from corpus driven, organically combine valence grammar and type grammar to analyze the valence system, better retain the part of speech and words in type grammar, and strengthen the sentence function of valence grammar [25]. Xiao compared the given corpus in the native English corpus with that in China's spoken English corpus. The purpose is to find out the lexical pattern characteristics of the virtualized verb give in different spoken styles that students can summarize the lexical pattern errors of the given virtualized verb [26].

*2.1. Natural Language Processing (NLP).* NLP is a subfield of AI that aims to teach computers how to read the text and spoken words in the same way that people do [27]. NLP integrates statistical, ML, and DL algorithms with computational linguistics rule-based modeling of human language. When these technologies are used combined, computers can analyze human language in the form of text or speech data and "understand" its whole meaning, including the goal and emotion of the speaker or writer [28]. Natural language processing (NLP) is used to enable computer programmers that translate text from one language to another, respond to spoken requests, and quickly summarize massive amounts of text in real time. It allows machines to analyze and understand human language, so that they may execute repetitive jobs automatically [29]. Machine translation, summarization, ticket categorization, and word check are some examples of NLP. The NLP main concept is shown in Figure 1.

*2.2. Corpus Linguistics.* The Latin term "corpus," which means "body," is the source of the term "corpus." It can refer to any type of written or spoken information. However, in current linguistics, this word refers to massive collections of texts that are given in computer understandable format and constitute a sample of a certain variety or usage of languages [30]. Corpus linguistics is the scientific study of language based on enormous collections of "real-world" language usage stored in corpora, which are computerized databases dedicated to linguistic research. Corpus-based research is another name for it. Corpus linguistics is viewed as a research tool or technique by some linguists, but it is viewed as a separate science or theory by others. Corpus linguistics is the study and construction of a collection of spoken and written texts as a piece of data for understanding the nature, organization, and usage of languages. This method frequently provides a quantitative layer to language descriptions by including statistics on the probability of linguistic items or processes happening in various situations [4]. Corpora are available in a range of sizes and styles; however, the majority are now digital, with specially designed computer applications to aid study. Annotations of grammatical groups and functions are common in corpora. Figure 2 shows the three core areas of the corpus.

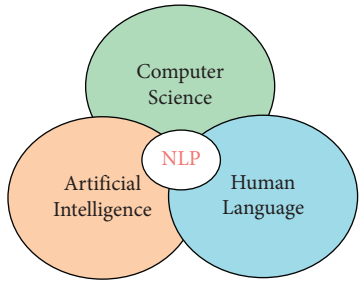


Figure 1: NLP analysis process.

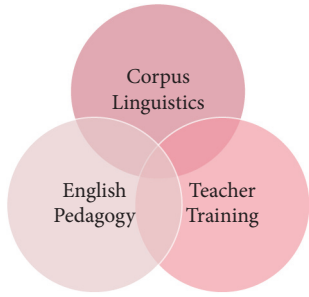


Figure 2: Three core areas of corpus.

## 3. Materials and Methods

*3.1. Clustering-Based Indexing Algorithm.* The C-means clustering algorithm can complete the processing of large datasets and has a fast iteration speed. The disadvantage of C-means clustering algorithm is that the number of clusters must be set ahead of time, and the clustering effect is related to the order of initial events and the clustering process, which does not conform to the basic characteristics of image database. The fuzzy C-mean algorithm uses pseudo-random numbers to generate initial class centers, which results in a lack of stability in the clustering effect. In this study, an improved fuzzy C-means clustering algorithm is used to retrieve English verb patterns. This algorithm effectively handles the clustering problem of selecting initial values as well as the functions of splitting, deleting, merging, and inserting. English verb patterns data can be clustered and the retrieval time will not increase linearly with the increase of English verb patterns in the corpus, which greatly improves the retrieval efficiency.

The basic idea of clustering algorithm is that if "Q" represents the number of images in the image library, "K" represents the number of clusters, and "N" represents the dimension of the eigenvectors, it makes English corpus N-dimensional eigenvectors. The main components of clustering technology are data representation model/text, computing similarity, clustering model, and clustering algorithm. The basic process of categorizing English verb patterns into text or using them in documents is as follows:

(1) *Text Feature Representation.* It includes the suffix tree model and vector space model.

(2) *Text Feature Dimension Reduction.* There is mainly a conceptual index, implicit semantic analysis, non-

negative matrix decomposition, and random projection.

(3) *Compute Text Similarity*. It includes Manhattan distance, Euclidean distance, Minkowski distance, cosine distance, weighted Euclidean distance, and correlation coefficient.

(4) *Assess the Quality of Clustering*. It includes measurement characteristics, direction, purity, and overall similarity.

(5) *Text Clustering*. It includes the K-means algorithm and its improved algorithm, PDDP algorithm, BK-means algorithm, competitive learning clustering, and hierarchical clustering. Figure 3 shows the clustering process.

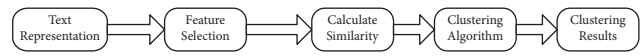


Figure 3: Cluster analysis process.

*3.1.1. C-Mean Clustering Algorithm.* C-means clustering algorithm needs continuous iteration and adjustment of "K" clustering centroids. The basic principle of this method is to minimize the distance within the class that is the comprehensive distance length between each kind of sample and the centroid, adjust K clustering centroids $C_k$, and then allocate samples to the nearest centroid category [31]. The process of C-means clustering algorithm is as follows:

(1) *Clarify the Initial Clustering*. Assuming "K" is the number of initial clusters, "K" feature vectors in the corpus are randomly selected as the initial clustering center and can be measured by the following formula:

$$C_n = X_n, \quad n = 1, 2, \ldots, N. \tag{1}$$

(2) *Allocate Samples*. Classify different samples to the nearest centroid category, which may be calculated using the following formula:

$$D(X_m, C_k) = \min\left(D\left(X_q, C_k\right),\ q = 1, 2, 3, ..., Q\right) X_m \in \text{clust}[k]. \tag{2}$$

(3) *Update Cluster Center*. Select the centroid of different class members as a new clustering center, and redistribute the samples based on the center until the clustering center of each sample remains stable. The following formula is used to describe it:

$$C_k = \frac{1}{\text{count}(k)} \sum_{\text{clust}[k] \in q} X_q, q = 1, 2, \ldots, Q, k = 1, 2, \ldots, K. \tag{3}$$

In the above formula, clust$[k]$ represents the category number.

*3.1.2. Fuzzy C-Means (FCM) Clustering Algorithm.* The fuzzy C-means clustering algorithm is widely used in pattern recognition and image processing. Its essence is to iterate and optimize the objective function to divide the set, which is used to represent the degree that different pixels belong to the corresponding type. The corresponding C-means clustering algorithm divides all pixels into a unique category [32]. The following is the clustering process:

(1) *Initial Fuzzy Weight*. The value range of the initial fuzzy weight is (0-1) and can be described using the following equations:

$$w_{qk} = \frac{w_{qk}}{\sum_{k=1}^{K} w_{qk}}, \quad W_{qk} = \frac{w_{qk}}{\sum_{k=1}^{Q} w_{qk}}. \tag{4}$$

(2) *Fuzzy Weight and Sample Center*. The following equation is used to calculate the cluster centers:

$$C_k = \sum_{q=1}^{Q} W_{qk} * X_q. \tag{5}$$

The fuzzy weight is calculated by the following formula:

$$w_{qk} = \frac{\left(1/D_{qk}^2\right)^{1/p-1}}{\sum_{r=1}^{K} \left(1/D_{qk}^2\right)^{1/p-1}}. \tag{6}$$

The Euclidean distance between $X_q$ and $X_k$ is represented by $D_{qk} = D(X_q, X_k)$.

(3) *Allocate Samples to the Category Closest to the Centroid*. The algorithm is consistent with the above clustering algorithm. After continuous cycle consistency, we can obtain a stable clustering center and fuzzy weight:

*3.2. Automatic Recognition and Extraction of English Verb Patterns Based on Index Line Clustering.* In this study, the automatic recognition and extraction of English verb forms are carried out based on index line clustering. The research period is realized by the following five processes:

(I) *Prepare Data*. Analyze the index line of core words in the coded corpus. In order to ensure that the meaning of English sentences is not changed, the whole line of sentences is analyzed.

(II) *Summarize the Language Features on the List of English Verb Forms*. The eigenvalues are constructed based on the summarized linguistic features of English verb forms, which mainly include adjacent word combination, word item, semantic category, part of speech, and grammatical category labels.

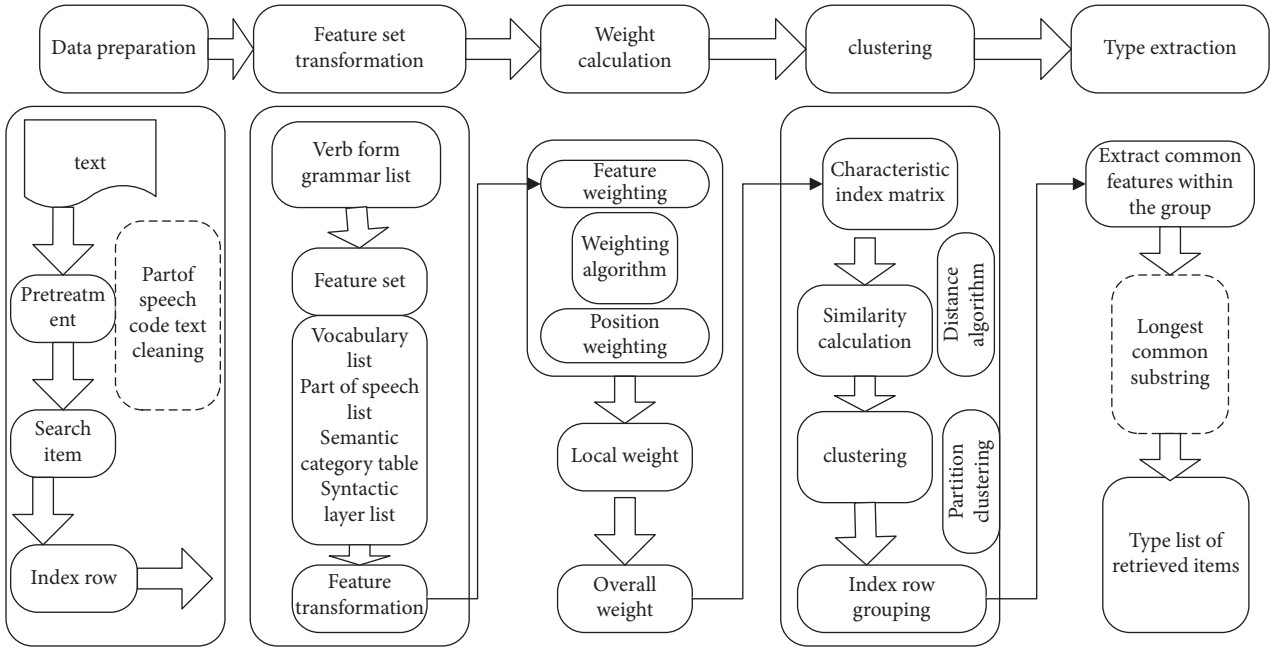


Figure 4: Automatic clustering and pattern extraction of index rows.

(III) *Transformational Language Features*. The language information encoded in the index line is transformed into the pattern coding on the feature set.

(IV) *Calculate the Similarity between Index Rows and Cluster Automatically*. In this process, we need to calculate the feature weight of English verb form, calculate the feature position, generate the feature index line matrix, calculate the similarity, and automatically cluster the index lines.

(V) Automatic recognition and extraction of English verb types by clustering index lines are shown in Figure 4.

*3.2.1. Create Validation Set.* When establishing the validation set, the external validity of the model should be evaluated, including recall rate and accuracy rate. In this study, the verification set is a list of specific verb types based on manual recognition and the corresponding index line which are the examples of each type. The node data and the node word information of the verification set are listed in Table 1.

*3.2.2. Index Row Clustering.* Index line clustering is to gather feature item index lines with a strong similarity in a group, which can reduce the difference of feature items in the group and maximize the difference between groups. The K-means algorithm is efficient and simple and it is the most used partition clustering algorithm at present. In this study, the K-means algorithm is used to achieve the optimal clustering results. During the application of this algorithm, the problem of k-value selection and similarity measurement should be taken into account.

Table 1: Validation set node word information.

| Verb | Type category quantity | Index row |
|---|---|---|
| Admit | 6 | 1065 |
| Agree | 5 | 210 |
| Argue | 7 | 285 |
| Claim | 6 | 1162 |
| Lead | 7 | 830 |
| Tell | 19 | 1420 |

This study analyzes the automatic recognition and extraction of English verb types based on index line clustering. First, the feature weighted index line features are used to form the feature index line matrix and then the Euclidean distance algorithm is used to calculate the similarity between two index lines to determine the index line grouping. The clustering algorithm is an unsupervised machine learning. There are a large number of factors that affect the number of groups. To analyze the internal structure of the dataset to clarify the number of groups, this study uses the internal validity evaluation index of clustering to get the $K$ value and adjusts three basic parameters to achieve the purpose of overall optimal clustering. First of all, the number of index row classification groups ($K$ value) is set and an interval (usually 5–30) is delimited. The other two parameters during the clustering period are adjusted, which include the initial centroid iteration times and index times, so as to obtain the optimum clustering results. Finally, based on the internal validity evaluation index of clustering, the "$K$" value curve and the sum of squares of residuals are used to help select the "$K$" value, and then the clustering index is carried out after the "$K$" value is defined.

*3.2.3. Pattern Extraction Algorithm.* English verb patterns have three quantitative features. In this study, we use the

following three features to establish a model when extracting English verb patterns:

(1) *Type Typicality*. A verb form has a wide range of applications. By quantifying the typicality of form, it is explained by using the proportion formed by this form on forms with equal length. In the calculation process, the ratio of the medium-sized option probability in the dataset to the sum of the type probability with equal length is used, that is, gravity ($C_x$). The following equation is used to describe it:

$$\text{Gravity}(C_x) = \frac{P(C_x)P}{\sum_{i=1}^{n} P(C_i)}. \quad (7)$$

In the above formula, ($C_x$) represents the type candidate, $P$ ($C_x$) represents the probability of ($C_x$) in the dataset, and $P$ ($C_i$) represents the sum of the probabilities of all equal length type candidates in the dataset.

(1.1) *Equal Length Pattern*. The equal length pattern is the element in the pattern component sequence in the pattern candidate, and the number distribution on the left and right sides of the verb is the length of a pattern candidate. The following are some examples of sentence analysis.

(a) *She Would Lie Her Way Out of Trouble*. PRP lie way out of is the type candidate in this sentence, the type element on the left is one, and the three elements are way, out, and of. The element on the left side of the verb is one and there are three elements on the right side. The left 1 and right 3 forms form the length of the form, which is represented by 1v3. If the length of two type candidates is the same, then it means that the type is the same, which is explained by the following example sentences.

(b) *He Argued His Way Out of Tough Situations*. PRP argument way out of is a candidate for this sentence type. According to the above definition, the length of this sentence type is 1v3, while PRP argument way out of and PRP lie way out of have the same length.

(2) *Viscosity*. It is used to indicate whether the affinity and selectivity between the elements in the form are strong or not. When studying the extraction of word collocation, viscosity is used as the feature of word collocation, so as to judge the collocation of words. The similarity between English type features and word collocation features is high, so in this study, we use a similar way to judge the nature of type elements. When calculating the extraction of English verb forms, mutual information can be set as a measurement method, as shown in the following equation:

$$MI(X) = \log\left(\frac{P(X)}{P(x_1)P(x_2)\ldots P(x_i)\ldots P(x_n)}\right). \quad (8)$$

Table 2: Comparative contingency table of interverb forms.

| | $Y$ = appear | $Y$ = not present | |
|---|---|---|---|
| $X$ = verb $A$ | $O_{11} = a$ | $O_{21} = b$ | $R_1 = a + b$ |
| $X$ = verb $B$ | $O_{12} = c$ | $O_{22} = d$ | $R_2 = c + d$ |
| | $C_1 = a + c$ | $O_2 = b + d$ | $N$ |

Table 3: Expected value contingency table of interverb type comparison.

| | $Y$ = appear | $Y$ = not present |
|---|---|---|
| $X$ = verb A | $E_{11} = (a + b)(a + c)/N$ | $E_{21} = (a + b)(b + d)/N$ |
| $X$ = verb B | $E_{12} = (c + d)(a + c)/N$ | $E_{22} = (c + d)(b + d)/N$ |

The above formula $x = x_1, x_2, x_3, \ldots x_n$ is a type candidate, and a type element is represented by $x_i$.

Finally, the differences in the use of forms are found in verbs. By comparing multiple verb forms, we can judge whether other verbs have the same type and whether they are consistent in application frequency. In this study, we will compare the application of interverb forms. Aiming at the difference modeling, this research work uses the comparison of the same types of multiple different verbs to judge the differences between the types in the same type and selects the proportion of other verbs with obvious differences in the same type as an index to judge the differences of the verbs. Here, the chi-square test is used to compare whether the difference of the same form between the two verbs is significant. The following formula is used to calculate the chi-square test:

$$I(C_x) = x^2 = \sum \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}. \quad (9)$$

In the above formula, $E_{i,j}$ represents the expected value, $Q_{i,j}$ represents the observed value, and $C_x$ represents the type options. The following tables are the two contingency tables, in which Table 2 lists comparative contingency of interverb forms and Table 3 lists the expected value contingency of interverb type comparison.

When evaluating the internal typicality of verb types, the harmonic average value is measured by the following formula:

$$C_x = \frac{2 * \text{Gravity}(C_x) * MI(C_x)}{\text{Gravity}(C_x) + MI(C_x)}. \quad (10)$$

In the above formula, CX represents the type candidate. The current retrieval system uses formula (10) to calculate the type ranking, which is regarded as the third quantitative feature of the type and can be described as the observed value. During the calculation of the above formula, the following examples are selected to introduce. A dataset that contains 100 sentences is taken to explain verbs. The verbs that have three different forms are explained, which include A: v wh clause, B: V about clause, and C: v for n. The three forms used are 10 times, 20 times, and 60 times, respectively. After calculation, the weight value of type "A" is 60/100 (60/100 + 20/100 + 10/100) = 0.67, and the weight result of type "B" is 20/100 (60/100 + 20/100 + 10/100) = 0.22.

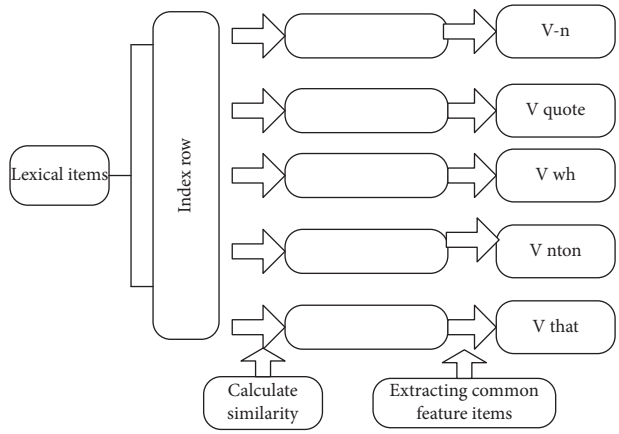


Figure 5: Index line clustering and type extraction process.

Table 4: Coincidence frequency, recall rate, and accuracy of manual and automatic type labels.

| Lexical items | Total frequency of index rows | First classification result | | | | Second classification result | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Frequency | Group | Accuracy (%) | Recall (%) | Frequency | Group | Accuracy (%) | Recall (%) |
| Admit | 1065 | 1000 | 7 | 94.2 | 100 | 1033 | 10 | 97.1 | 100 |
| Agree | 210 | 179 | 4 | 85.3 | 100 | 206 | 9 | 98.12 | 100 |
| Argue | 685 | 600 | 6 | 87.6 | 100 | 640 | 8 | 93.5 | 100 |
| Claim | 1161 | 1125 | 5 | 97.2 | 100 | 1135 | 21 | 97.8 | 100 |
| Lead | 830 | 772 | 6 | 92.9 | 100 | 789 | 11 | 95.1 | 100 |
| Tell | 1420 | 1268 | 18 | 89.3 | 100 | 1337 | 34 | 94.3 | 100 |
| Mean value | | | | 91.01 | 100 | | | 95.99% | 100% |

In this study, the list of English verb patterns is further analyzed to establish the set of elements required for the pattern, and then the language information on the index line is transformed into pattern elements. The index lines separated from the same search words are analyzed by the clustering method, so as to extract the common feature items on each group of index lines to achieve the purpose of automatic recognition and extraction of English verb patterns. The process of index row clustering and type extraction is shown in Figure 5.

## 4. Results and Discussion

*4.1. Automatic Recognition of English Verb Patterns and Analysis.* In this study, we study the automatic recognition and extraction of English verb types based on the index line clustering algorithm. In order to test the automatic recognition and extraction effect of this algorithm, the recall rate and accuracy rate of the model are selected as two main parameters. For the first time, the number of manual type label index lines on the two classification verification sets is equal to that of manual grouping, which is used to test the fit between the automatic classification results of the model and the manual classification results. In the second classification, according to the validity evaluation in clustering, after identifying the K value, it refers to the clustering analysis. Clustering is an unsupervised machine learning, in which the number of classifications is not clear. Therefore, the second classification verification set can test the model in the selection of K value.

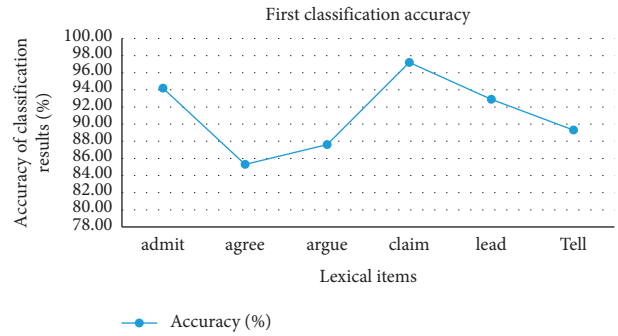


Figure 6: Accuracy of first classification.

The manual type and automatic type labels index lines frequency, recall rate, and accuracy, which are listed in Table 4.

According to the data in Table 4, the average accuracy of automatic recognition of clustering type of six test English verb index lines in the first classification results on the index line is 91.01%, which indicates that the index effect is good. The accuracy of the six English verb cluster type recognition is in the range of 85.3% to 97.2%. According to the data, the common words in English verbs and words with low frequency are more suitable. Analyzing the results of the second classification, the average accuracy of the six English verb indexes is 95.99%, which is 4.98% higher than that of the first classification. After the second classification experiment, the number of index groups of each English verb is significantly higher than that of manual classification. By analyzing the automatic

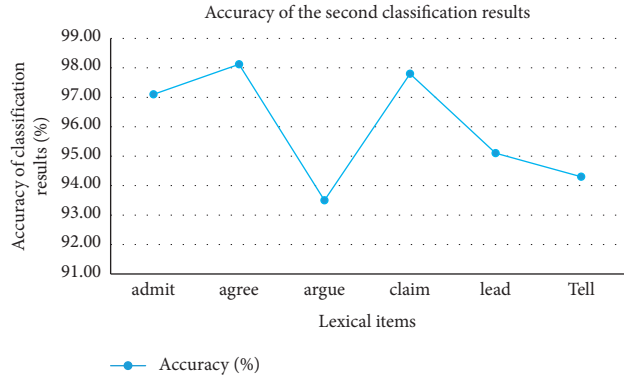


Figure 7: Accuracy of the second classification.

Table 5: Pattern extraction results of 42 English verbs.

| Verb | Type and quantity |
|---|---|
| Take | 284 |
| Do | 212 |
| Go | 192 |
| Make | 191 |
| Tell | 181 |
| Ask | 163 |
| Call | 133 |
| Bring | 129 |
| Know | 122 |
| Use | 119 |
| Spend | 115 |
| Rise | 107 |
| Put | 103 |
| Want | 64 |
| Join | 63 |
| Add | 62 |
| Need | 59 |
| Lose | 58 |
| Show | 58 |
| Meet | 58 |
| Hear | 57 |
| Feel | 56 |
| Send | 56 |
| Lead | 55 |
| Increase | 55 |
| Run | 52 |
| Follow | 43 |
| Set | 43 |
| Try | 42 |
| Involve | 42 |
| Speak | 40 |
| Decide | 39 |
| Play | 39 |
| Base | 39 |
| Build | 39 |
| Learn | 39 |
| Remain | 36 |
| End | 36 |
| Support | 35 |

recognition and extraction results of the second classification of English verb forms, it can be obtained that most of the subclasses of English verb forms can be recognized accurately.

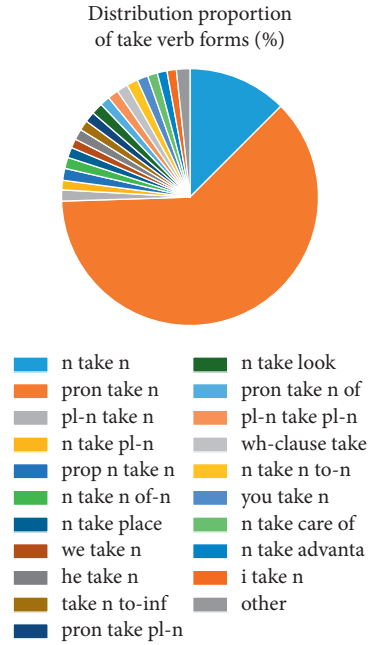


Figure 8: Pattern distribution of verbs.

Figure 6 shows the accuracy rate of the first classification and Figure 7 shows the accuracy rate of the second classification.

*4.2. Analysis of English Verb Pattern Extraction.* We select 42 English verbs for analysis and listed the extracted number of English verb forms (Table 5). In this study, take English verbs are selected as examples to describe the pattern extraction results. There are 20 different patterns of take, which are analyzed according to the occurrence probability of each pattern. The experimental results are shown in Figure 8.

## 5. Conclusion

With the rapid development of globalization, the importance of English is gradually increasing. There are a large number of verb patterns in the process of English learning and teaching. The inefficiency and difficulty of English verb patterns teaching make it almost impossible for students to learn English verb patterns in a better way. This study uses the index row clustering method to simulate the manual interpretation of indexed rows and computes the automatic classification of indexed rows based on similarity to complete the automatic recognition and extraction of English verb patterns. This study first focuses on the clustering index algorithm including the C-means clustering algorithm and fuzzy C-means clustering algorithm, then describes in detail the automatic recognition and extraction process of English verbs based on index line clustering, and creates a verification set and completes the index line clustering of English verbs. The use of the clustering-based index rows method has completely changed the traditional manual interpretation of automatic recognition and extraction of index rows.

The number of index rows and the differences between rows have a direct impact on the number of index rows in clustering groups. The experimental results show that the average accuracy of automatic recognition and extraction of English verb patterns based on index row clustering is 95.99%, which is 4.98% higher than the method of manual classification.

## Data Availability

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

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## References


[1] E. Joanis, "Automatic verb classification using a general feature space," Master's thesis, Department of Computer Science, University of Toronto, Toronto, Canada, 2002.

[2] S. Abney, "Partial parsing via finite-state cascades," *Natural Language Engineering*, vol. 2, no. 4, pp. 337–344, 1996.

[3] X. Xu, "Exploration of English composition diagnosis system based on rule matching," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 13, no. 7, p. 161, 2018.

[4] S. Liu and W. Liu, "English grammar error correction algorithm based on classification model," *Complexity*, vol. 2021, Article ID 6687337, 11 pages, 2021.

[5] M. M. Abdel Latif, "Sources of L2 writing apprehension: a study of Egyptian university students," *Journal of Research in Reading*, vol. 38, no. 2, pp. 194–212, 2015.

[6] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *Proceedings of the COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, PA, USA, August 1998.

[7] Ł. Groom and N. Groom, "Functionally-defined recurrent multi-word units in English-to-Polish translation," *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, vol. 35, no. 1, pp. 1–29, 2022.

[8] S. R. Kumaran, M. S. Othman, and L. M. Yusuf, "Hybrid of hierarchical and partitional clustering algorithm for gene expression data," *IOP Conference Series: Materials Science and Engineering*, vol. 864, Article ID 012071, 2020.

[9] J. Bai, D. Song, P. Bruza, J.-Y. Nie, and G. Cao, "Query expansion using term relationships in language models for information retrieval," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pp. 688–695, Bremen Germany, October 2005.

[10] A. Reghunadhan and R. Reghunadhan, "Applying deep neural networks for the automatic recognition of sign language words: a communication aid to deaf agriculturists," *Expert Systems with Applications*, vol. 185, Article ID 115601, 2021.

[11] A. Linarsih, D. Irwan, and M. I. R. Putra, "The interferences of Indonesian grammatical aspects into English: an evaluation on preservice English teachers' EFL learning," *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, vol. 5, no. 1, p. 69, 2020.

[12] J. Rose and H. Rose, "English language teaching and English-medium instruction," *Journal of English-Medium Instruction*, vol. 1, no. 1, pp. 85–104, 2022.

[13] S. T. Gries, "What is corpus linguistics?" *Language and linguistics compass*, vol. 3, no. 5, pp. 1225–1241, 2009.

[14] E. Arai and T. Arai, "The perception of English words with consonant clusters and vowel deletion by young normal-hearing listeners under noise," *Journal of the Acoustical Society of America*, vol. 148, no. 4, 2505 pages, 2020.

[15] Y. M. Megariani, N. A. Listyantari, and B. Bram, "Mispronunciations in graduate students' presentation projects," *Metathesis: Journal of English Language, Literature, and Teaching*, vol. 4, no. 1, p. 56, 2020.

[16] M. Balakrishnan and V. Balakrishnan, "An integrated semi-automated framework for domain-based polarity words extraction from an unannotated non-English corpus," *The Journal of Supercomputing*, vol. 76, no. 12, pp. 9772–9799, 2020.

[17] D. Bourigault and C. Jacquemin, "Term extraction+ term clustering: an integrated platform for computer-aided terminology," in *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 15–22, PA, USA, June 1999.

[18] D. Newman, N. Koilada, J. H. Lau, and T. Baldwin, "Bayesian text segmentation for index term identification and keyphrase extraction," *Proceedings of COLING*, vol. 2012, pp. 2077–2092, 2012.

[19] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," *Frontiers of Multimedia Research*, vol. 25, pp. 3–29, 2017.

[20] B. Naismith, N.-R. Han, and A. Juffs, "The university of pittsburgh English language institute corpus (PELIC)," *International Journal of Learner Corpus Research*, vol. 8, no. 1, pp. 121–138, 2022.

[21] J. P. Smith, S. Meerow, and B. Turner II, "Planning urban community gardens strategically through multicriteria decision analysis," *Urban Forestry and Urban Greening*, vol. 58, Article ID 126897, 2021.

[22] J. Chen, "A corpus-based analysis of although errors in Chinese EFL learners' written output," *Study in English Language Teaching*, vol. 5, no. 3, p. 429, 2017.

[23] Z. Xiaojing and Z. Xiaoqiong, *Migrant Ecologies: Zheng Xiaoqiong's Women Migrant Workers*, Lexington Books, DC, USA, 2021.

[24] R. Ağçam, "Author stance in academic writing: a corpus-based study on epistemic verbs," *Journal of Teaching English for Specific and Academic Purposes*, vol. 3, no. 1, p. 9, 2015.

[25] C. Shei, *The Routledge Handbook of Chinese Discourse Analysis*, Routledge London, London, UK, 2019.

[26] R. Xiao, "How different is translated Chinese from native Chinese?" *International Journal of Corpus Linguistics*, vol. 15, no. 1, pp. 5–35, 2010.

[27] J. Bouaziz, R. Mashiach, S. Cohen et al., "How artificial intelligence can improve our understanding of the genes associated with endometriosis: natural language processing of the PubMed Database," *BioMed Research International*, vol. 2018, Article ID 6217812, 7 pages, 2018.

[28] A. E. Thessen, H. Cui, and D. Mozzherin, "Applications of natural language processing in biodiversity science," *Advances in bioinformatics*, vol. 2012, Article ID 391574, 17 pages, 2012.

[29] X. Chen, R. Ding, K. Xu, S. Wang, T. Hao, and Y. Zhou, "A bibliometric review of natural language processing

empowered mobile computing," *Wireless Communications and Mobile Computing*, vol. 2018, Article ID 1827074, 21 pages, 2018.

[30] J. Zhang, "Data-driven learning teaching model of college English based on mega data analysis," *Scientific Programming*, vol. 2022, Article ID 3490594, 2022.

[31] A. S. Aziz, R. A. El-Khoribi, and S. A. Taie, "AFCM model to predict the learner style based on questionnaire and fuzzy C mean algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 99, 2021.

[32] V. N. Phu and V. T. N. Tran, "A CO-training model using a fuzzy C-means algorithm, a K-means algorithm and the sentiment lexiCOns-based multi-dimensional vectors of an otsuka COefficient for English sentiment classification," *Journal of Theoretical and Applied Information Technology*, vol. 96, 2018.