Hindawi
Mobile Information Systems
Volume 2022, Article ID 1019508, 10 pages
https://doi.org/10.1155/2022/1019508

Research Article

Similarity Measurement and Classification of English Characters Based on Language Features

Linna Miao,[1] **Zhixin Fang**,[1] **and Junping Zhang**[2]

[1]*School of Foreign Language, Henan University of Chinese Medicine, Zhengzhou 450046, China*
[2]*School of International Studies, Zhengzhou University, Zhengzhou 45001, China*

Correspondence should be addressed to Linna Miao; miaolinna@hactcm.edu.cn

Received 10 June 2022; Revised 4 July 2022; Accepted 12 July 2022; Published 24 August 2022

Academic Editor: Le Sun

English is now widely used in the world as an international language. As a symbol of the development of human civilization, English characters provide an important medium and tool for mankind. In the current information age, the vocabulary of English words is more quantitative, and it is almost everywhere. Under the background of the multiquantification of English words and the quantification of the relationship between words, the similarity measurement analysis and calculation of English words and the classification of vocabulary measurement calculation are carried out by integrating the characteristics of language. The experimental results are as follows: (1) the development situation of English words is analyzed, the research direction of the experiment is determined, the concept of English character features is proposed, and the similarity calculation method is selected according to different features, in order to simplify the complex and difficult-to-understand word meaning relationship between English words; (2) the text features are extracted through the similarity feature selection of language and text. The extraction of features indirectly affects the effectiveness of classification. The similarity word embedding vector is used to map English words into the vector for analysis and comparison, calculate the distance between the similarity numerical variables between English words and their similarity coefficient, measure the distance between them, and evaluate the similarity between them, including the angle cosine method and correlation coefficient method which are the two main methods for calculating the similarity coefficient.

1. Introduction

Whispering is a natural way of speaking. Although its perceptual ability is reduced, it still contains information about expectations (i.e., comprehensibility) and the identity and gender of the speaker. However, considering the acoustic differences between whispered speech and normal voiced speech, speech applications trained on the latter but tested with the former showed unacceptable performance levels. In the automatic speaker verification task, previous studies have shown that I) traditional features (e.g., the Mel frequency cepstrum coefficient, MFCC) cannot transmit enough speaker discrimination clues in two utterance efforts and II) multiconditional training often reduces the performance of normal speech while improving whisper performance. In this paper, we aim to solve these two shortcomings by proposing three innovative features, which can provide reliable results for normal speech and whispered speech when fused at the score level. Overall, the relative improvement rates of the whisper group and the normal group were 66% and 63%, respectively [1]. Although the correctness of the characteristic measurement largely blames it on the changing external environment so far, little attention has been paid to the consequences of this fact in pattern recognition tasks. In this paper, we explicitly consider the uncertainty of feature measurement and illustrate how to improve diversified classification rules and research methods to make up for the impact of uncertainty. This experimental method can be effectively used in various multistyle scenes, and the feature vectors evolved from different sences are merged. For the development of this kind of operation, if we can estimate the uncertainty of noise generated by each vector characteristic flow, this kind of development will achieve high efficiency and adapt to various pattern fusion rules. The study further shows that under some assumptions, the multimodal fusion method depending on the flow weight can be naturally generated from our scheme; this relationship can help us give a helpful view on the use of uncertainty compensation methods. This reveals how to apply those views to audio-visual intelligent induction. In this similar event, an emerging technology is developed and proposed, which works when using the framework in feature extraction and variability evaluation of human perception and also studies how to effectively calculate the enhanced audio features and their uncertainty estimation. The effectiveness of our multimodal integration method in the audio-visual database is proved [2]. It is a very difficult and meaningful challenge to identify and classify complex people's actions and behaviors from animated videos. The article uses Indian sign language (ISL) video to explore this kind of problem. The feature of discretization through the scale and translation of basic wavelet, the new segmentation algorithm, is proposed. The fusion functions form a two-dimensional point cloud to show some characteristics in uninterrupted animation playback. The most perfect feature extraction of symbols in animation playback is carried out on each single classifier to check the feasibility of the designed feature extraction framework. In the experiment, we can see that the feature proportion of some binary patterns can better represent the value of symbol recognition data than other most advanced features. The specific reason is that the designed feature model combines the overall features and some features. The obtained and classified characteristics are transmitted to the network database remotely, and they correspond to their own nominal words. The accuracy and correctness of the recognition mark are tested. Through the largest training example, an artificial neural network classifier with a recognition rate of 92.79% is obtained, which is much higher than the existing artificial neural network classifiers on sign language and ISL data sets with other features [3]. This paper mainly studies the temporal retrieval of activities in videos through sentence queries. Given the sentence query describing the activity, time moment retrieval aims to locate the time period that can best describe the text query in the video. This is a common and challenging task because it requires understanding both the video and the language. Existing studies mainly use coarse frame level features as visual representation, blurring specific details in the video (for example, the required object "girl," "Cup," and action "dumping,") which may provide key clues for the time required for positioning. In this paper, we propose a new spatial and linguistic time tensor fusion (SLTF) method to solve these problems [4]. This study investigates the generation and perception of English vowels by Korean English learners in two English learning sessions about one year apart. A preliminary experiment shows that Korean adults use two different Korean vowels to classify some comparative English vowels, while others show classification overlap, which means that it is difficult for Korean English learners to distinguish these vowels. In two subsequent experiments, NK adults and children living in North America for different periods of time (3 years vs. 5 years; 4 groups, 18 years in each group) were compared with age-matched native English speakers. In Experiment 2, NK children identified English vowels more accurately than NK adults but more accurately than NE children. In Experiment 3, a picture naming task was used to extract images containing/, i.e., εД/English words. Some vowels produced by NK children are easier to hear than those produced by NK adults. Acoustic analysis shows that the vowel contrast of NK children is significantly higher than that of NK adults [5]. A size and color invariant character recognition system based on the feedforward neural network is proposed. Our feedforward network has two layers. One is the input layer, and the other is the output layer. The whole recognition process is divided into four basic steps: preprocessing, standardization, network establishment, and recognition. Preprocessing includes digitization, noise removal, and boundary detection. After boundary detection, the input character matrix is normalized to a 12×8 matrix for size invariant recognition and input it into the network composed of 96 input and 36 output neurons. Then, we use the proposed training algorithm to train the network in a supervised way and establish the network by adjusting the weight. Finally, we test our network by averaging more than 20 samples per character. By considering the similarity measure between classes, we give 99.99% accuracy for numbers (0~9), 98% accuracy for letters (a~z), and more than 94% accuracy for alphanumeric characters [6]. Using the perceptual assimilation model (PAM) of best (1995), we studied the dictation and observation ability of Cantonese tone, as well as touch, smell, hearing, and visual perception, including Thai and English [7]. This paper identifies six social science research methods that help to describe the social and cultural significance of nanotechnology: web-based questionnaire survey, episode experiment, network link analysis, recommendation system, quantitative content analysis, and qualitative text analysis. Data from a range of sources are used to illustrate how these methods describe the knowledge content and institutional structure of the emerging nanotechnology culture. These methods will make it possible to test hypotheses in the future. For example, nanotechnology has two competing definitions, namely science and technology and science fiction, which affect public cognition through different ways and directions [8]. In the study of biomedical field, the identification and standardization of medical case literature is an important step of biomedical text extraction. In addition, a gene symbol recognition system is also described to obtain special text content from biomedical materials and standardize its content. The composition of this gene symbol recognition system includes gene symbol recognition, gene text content mapping, gene text standardization, and text content filtering. Gene symbol recognition is a process based on fund symbol matching and monitoring. It uses a large number of labeling methods to achieve the recognition of gene symbols. In the gene text content mapping stage, the data set connection is established in the system context around the principles of accurate matching and priority matching [9]. If we lack relevant problem specific knowledge, we can use the cross validation method to select the classification method empirically. We test this idea here to illustrate the meaning of cross validation to solve and not solve the selection problem. As experience shows, cross validation may bring higher average performance than the application of any single classification strategy and can also reduce the risk of poor performance. On the other hand, compared with simpler strategies, cross validation is more or less a bias. The correct application of cross validation ultimately depends on previous knowledge. In fact, cross validation may be seen as a way to apply some information about the applicability of alternative classification strategies [10]. A new intelligent fault diagnosis method of rotating machinery based on wavelet packet transform (WPT), empirical mode decomposition (EMD), dimensionless parameters, distance estimation technology, and radial basis function (RBF) network is proposed. The experimental results show that the method combining WPT, EMD, distance evaluation technology, and RBF network can accurately extract fault information and select sensitive features so as to correctly diagnose different fault types of bearings. This method is applied to the fault diagnosis of slight rub impact in heavy oil catalytic cracking units. The actual results show that this method can be effectively applied to the fault diagnosis of rotating machinery [11]. Decision tree classification provides a fast and effective data set classification method. There are many algorithms to optimize the structure of decision tree, although these methods are vulnerable to the changes of the training data set. This method is tested with two different data sets, and the results are equivalent to or better than other classification methods. The last discussion demonstrates the utility of decision trees relative to algorithms or other alternative methods (such as neural networks), especially when considering a large number of variables [12]. An objective classification method of weather situation in Europe and northeast Atlantic is established. The mean air pressure of each mser40 in winter and the mean air pressure of each mser40 in winter are calculated, respectively. Then, according to the original concept of Hess and Brezovsky, by using the mode correlation of these composite fields, the daily directory of the target GWL is constructed, and some filtering methods are used to detach the instantaneous feature vector, which can help to keep the GWL task at least more than four days. The essential difference between the fact and the original GWL system is found. The reason is that the original system is mainly concentrated in Central Europe and has a certain subjectivity, while the reality system treats power more in terms of spatial standards. The data transformation fluctuation of most air flows in Central Europe usually comes from GWL series, which is used to calculate the law of anticyclone change, reanalyze the change of anticyclone fluctuation in Central Europe during this period, and predict the development situation. [13]. In this paper, a fault classification method based on neural network and orthogonal least squares (OLS) learning process is adopted to identify various relevant voltage and current modes. This paper also compares the RBF neural network with the BP neural network. The results show that the RBF method can calculate all kinds of faults quickly and accurately. The simulation results also show that this method can be used as an effective tool for high-speed relay protection [14]. A fully automatic multiscale fuzzy c-means classification method is proposed. We use diffusion filters to process MR images and construct multiscale image sequences. On the scale from coarse to fine, the multiscale fuzzy c-means classification method is adopted. The final function of the old implicit averaging method is basically modified and its classification is diversified, in which the coarse scale supervises the classification of the next fine scale. Due to its multiscale diffusion filtering scheme, this means has high stability for noisy and weak contrast animation images. By comparing and improving the new design method with the old method, the synthetic images with different contrasts and the McGill brain magnetic resonance image database were verified. Our MsFCM method is always superior to the traditional FCM and MFCM methods. The actual ground verification shows that the MsFCM method achieves an overlap rate of more than 90%. The availability of this method is proved in the actual animation image. It is proved that the diversified average classification methods are correct and stable for all kinds of animation images. It can become a tool for animated images and other application scenes [15].

2. Similarity Measurement of English Characters Based on Language Features

In the field of language learning and recognition, features are important research objects. In the process of language similarity calculation and classification, based on the analysis, recognition, and text inspection of different research objects, in essence, they can be regarded as extracting and classifying the features of research objects and calculating the similarity between two feature vectors through a measurement criterion. Therefore, the selection of features has a far-reaching impact on the results of calculating similarity.

2.1. Features. At present, feature extraction is in the primary stage in horizon perception. In the process of learning in the field of science, the most important thing is analytical theory. One of the important viewpoints in this theory is that horizon perception is a process extending from a local feature of something to a global feature, which makes it clear that the local feature is perceived at the first time. However, as for the principle of global priority theory, the global feature is regarded as the first perceived object, followed by the local feature. What is feature extraction? The so-called feature extraction is a method of transforming the original space into the space to be calculated through a certain mapping relationship. The initial feature is the first feature of the extracted object. If the dimension of the object to be calculated is high, it will produce too high time complexity in the calculation. Therefore, in general, try to map the high-dimensional space vector to the low-dimensional space. This method is helpful to complete the analysis and extraction of the features of the research object, and different features can complement each other. Therefore, in theory, the accuracy of multiobject combined feature extraction is higher than that of single-object feature extraction. Therefore, in the feature extraction of similarity measurement, it is best to extract and measure the features of multiobject combinations and then select some obvious features for linear or nonlinear combinations.

2.1.1. Statistical Characteristics

(1) Conversion Coefficient Method. The idea of the transformation coefficient method is to calculate the number of the whole global characteristic variables. It is to carry out different transformations on the model and take the results of different transformations as a feature. The transformation coefficient methods often used in the process of statistical features include KL transformation, Fourier transformation, Hough transformation, and so on. The conversion coefficient method takes each pixel in the graph as each unit. Therefore, when using the conversion coefficient method, it will also produce the problems of difficult calculation and resource consumption. Therefore, in practical application, some special correction methods will be adopted to reduce the difficulty of calculation.

(2) Contour Feature. The edge contour of English text form rich features. Although the features cannot be displayed inside the text and are not obvious, its edge contour can still reflect some rich features. Because this feature starts from the edge, it can be used as the classification of general features to a certain extent.

(3) Pixel Density Characteristics. Due to the wide variety of English characters, the pixel distribution represented by different kinds of English characters is very different. The coarse pixel density characteristics can be obtained by dividing the text image horizontally or vertically and calculating the effective number of pixels in each area. For some English text pictures, the difference in their own structure is not very obvious. Although the pixel density obtained by different division methods is different, the characters they actually represent are very similar. Therefore, the pixel density feature can be used to classify the English character features. The advantage of the pixel density feature is that it can prevent the influence of external things, and a small amount of information will not seriously affect the actual results. However, due to the diversity of text types, the features formed by different English words take a long time. Therefore, for different English text types, the feature extraction method needs to be improved.

2.2. Similarity Measure. Similarity reflects the degree of relationship between different objects or different features. The similarity is an important index to indicate whether the model samples are similar. It is usually represented by a value between 0 and 1. The vector similarity can be divided into the vector similarity and system similarity. Different research objects correspond to different similarities. The calculation methods of similarity measure mainly include the distance calculation method and function method. The two methods have their own differences. The accuracy of the results obtained by the distance calculation method is smaller, while the results calculated by the function method are more accurate, especially when studying the similarity between vectors.

3. English Text Similarity Measurement Algorithm Based on Language Features

3.1. Similarity Feature Selection. In the process of English text similarity measurement and classification, feature extraction is the most important content. The quality of feature selection directly affects the efficiency of similarity classification, so this paper uses the chi square test to extract features. What is chi square test? Chi square test is to score and sort the features of the research object after feature extraction so as to select the top features as the extraction result set.

Chi square test formula is

$$x^2 = \sum \frac{(A - T)^2}{T}. \tag{1}$$

3.2. Similarity Word Embedding Vector. Word embedding vector is a process of mapping a word into a measurement space. The computer itself cannot directly extract the features of English text, so it is necessary to convert English text into a spatial vector. Nowadays, the most important text space vector models are the skip-gram model and CBOW model. This paper selects the former as the training text vocabulary vector.

The skip-gram model obtains a weight model from the input layer to the output layer through simulation training in a certain scale corpus according to the probability of the characteristics of n words before and after text center vocabulary prediction. The probability of maximizing text obtained by the model is

$$\operatorname{argmax} \prod_{w_{ij}}^{D} \left[\coprod_{c \in C_{ij}} P\left(c \middle| w_{ij}, \theta\right) \right]. \tag{2}$$

A support vector machine algorithm is essentially a supervised classification algorithm. It can be divided into linear separable and linear nonseparable. It has achieved good results in the process of classification training.

The support vector machine can map the research object data from low-dimensional space to high-dimensional space and select kernel function for a solution. The mathematical expression is

$$\begin{aligned} &\max \sum_{i=1}^{n} a_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j K\left(x_i, x_j\right) \\ &\text{s.t.} \sum_{i=1}^{n} a_i y_i = 0 \\ &0 \leq a_i \leq C, \\ &i = 1, 2, \ldots, n. \end{aligned} \tag{3}$$

In formula (3), $K(x_i, x_j)$ represents the kernel function, and the final classification function is

$$f(x) = \operatorname{sign}\left\{\sum_{i=1}^{n} a_i y_i K\left(x_i, x_j\right) + b\right\}. \tag{4}$$

According to the Bayesian formula

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P\left(B_j\right)P\left(A|B_j\right)}. \tag{5}$$

It can be concluded that

$$P(C_i|X) = \frac{P(C_i)P(X_k|C_i)}{P(X)}. \tag{6}$$

(When x condition is independent)
In formula (6)

$$P(C_i) = \frac{N_{\mathrm{ci}}}{d}. \tag{7}$$

Calculate in set C

$$C(C_i) = P(C_i)\prod_{k=1}^{n} P(X_k|C_i). \tag{8}$$

Final classification result is

$$C_{\max} = \arg\max P(C_i)\prod_{k=1}^{n} P(X_k|C_i). \tag{9}$$

The composition of random forest is composed of a variety of decision trees. Compared with the composition of a single decision tree, it avoids getting consistent assumptions and making the assumptions too strict. The method of increasing the amount of data and testing the sample set is usually used to evaluate the performance of the classifier. While solving the classification problem, each decision tree in the forest judges the simulated training samples in turn, which are selected by most decision trees as the final result.

$$\{h_1(X, \theta_{w1}), h_2(X, \theta_{w2}), \ldots, h_m(X, \theta_{wm})\}. \tag{10}$$

The marginal function of random forest is

$$m_g(X, Y) = av_k I\left(h_k(x) = y\right) - \max_{j\neq y} av_k I\left(h_k(x) = j\right). \tag{11}$$

3.3. Distance between Similarity Numerical Variables.

If the attributes of decision variables are continuous or discontinuous, how to measure the similarity or distance between variables?

3.3.1. European Distance

$$d_{i.j} = \sqrt{\sum_{k=1}^{n}\left(X_{ik} - X_{jk}\right)^2}, \tag{12}$$

$d_{i.j}$ refers to the overall distance in the n-dimensional space, that is, the dissimilarity. The larger $d_{i.j}$ means the farther the distance, that is, the more obvious the dissimilarity is. On the contrary, the smaller $d_{i.j}$ means the more obvious the similarity between the whole. X_{ik} means the i-dimensional coordinate of the first point, and X_{jk} means the second two-dimensional coordinate of the second point.

3.3.2. Manhattan Distance

$$d_{\mathrm{i}.j} = \sum_{k=1}^{n}\left|X_{ik} - X_{jk}\right|. \tag{13}$$

3.4. Similarity Coefficient.

Order $O = \{x_1, x_2, \ldots, x_n\}$. All numerical sets of simulation research objects are set to $(x_1, x_2, \ldots, x_n)$. The range of values for each simulation study was set to $x_i, x_j \in O$, wherer_{ij} is the similarity coefficient of x_i and x_j; the specific conditions are as follows:

(1) $\mathrm{r}_{ij} = 1 \Leftrightarrow x_i = x_j$

(2) $\forall x_i, x_j, r_{ij} \in [0, 1]$

(3) $\forall x_i, x_j, r_{ij} = r_{ji}$

The following methods are commonly used to measure and calculate the similarity coefficient:

3.4.1. Quantity Product Method

$$r_{ij} = \begin{cases} 1, & i = j, \\ \displaystyle\sum_{i=1}^{m} X_i X_{ik} X_{jk}, & i \neq j, \end{cases} \tag{14}$$

where M is a positive number, satisfying $M \geq (\sum_{k=1}^{m} X_{ik}X_{jk}), i \neq j$.

3.4.2. Included Angle Cosine

$$r_{ij} = \frac{\left|\sum_{k=1}^{m} X_{ik}X_{jk}\right|}{\sqrt{\left(\sum_{k=1}^{m} X_{ik}^2\right)\left(\sum_{k=1}^{m} X_{jk}^2\right)}}. \tag{15}$$

A vector is a directed line segment in a multidimensional space. If two vectors have the same direction, their included angle is 0. Therefore, the cosine value can be used to express the similarity of two vectors. When two vectors are orthogonal, r_{ij} =0 indicates that the vectors are completely different.

3.4.3. Correlation Coefficient Method

$$r_{ij} = \frac{\sum_{n=1}^{m}\left(X_{ik} - X_i\right)\left(X_{jk} - X_j\right)}{\sqrt{\left(\sum_{k=1}^{m}\left(X_{ik} - X_i\right)^2\right)}\sqrt{\left(\sum_{k=1}^{m}\left(X_{jk} - X_j\right)^2\right)}}. \tag{16}$$

Among them, $X_i = (\sum_{k=1}^{m} X_{ik}/m)$,$X_j = (\sum_{k=1}^{m} X_{jk}/m)$, and the numerical range of r_{ij} is in $[-1, 1]$. When the result is 0, it indicates that there is no correlation between the whole;

when the result is 1, it indicates that the whole is positively correlated; when the result is −1, it indicates that there is a negative correlation as a whole.

3.4.4. Arithmetic Mean Minimum Method

$$r_{ij} = \frac{2 \times \sum_{k=1}^{m} \left(X_{ik} \wedge X_{jk} \right)}{\sum_{k=1}^{m} \left(X_{ik} + X_{jk} \right)}. \tag{17}$$

3.4.5. Exponential Similarity Method

$$r_{ij} = \frac{\sum_{k=1}^{m} \exp - \left(X_{ik} - X_{jk} \right)^2 / S_k^2}{m}. \tag{18}$$

3.4.6. Paste Progress. If the characteristics of X_i and X_j are unified so that X_{ik} and X_{jk} belong to [0,1] ($k = 1, 2, \ldots, m$), the similarity of X_i and X_j is defined as their pasting progress. Distance paste progress

$$r_{ij} = 1 - c\left(d\left(X_i - X_j \right) \right)^a, \tag{19}$$

C and a are appropriate selection parameters, and their values can be any value, but their selected values should meet the $0 \le r_{ij} \le 1$ inequality, and $d(X_i - X_j)$ represents the distance between them.

$d(X_i - X_j)$ is a certain distance, which can be taken as Minkovsky distance

$$d\left(X_i - X_j \right) = \left(\sum_{k=1}^{m} \left| X_{ik} - X_{jk} \right|^p \right)^{1/p}. \tag{20}$$

4. Experimental Analysis of Similarity Measurement and Classification of English Characters Based on Language Features

4.1. Comparative Analysis of Similarity Algorithm Efficiency. The cosine similarity algorithm, keyword similarity algorithm, word meaning similarity algorithm, common subsequence similarity algorithm, and the experimental algorithm are used to analyze and calculate the similarity measurement of the simulation research sample data.

Method1: cosine similarity algorithm, method2: keyword similarity algorithm, method3: word meaning similarity algorithm, method4: common subsequence similarity algorithm, and method 5: experimental algorithm in this paper.

As shown in Table 1, which represents the average similarity value of the five methods in the state of 1 under different numbers of data. Since the similarity between data vocabulary pairs is tested, if the similarity state between vocabulary pairs is 1, that is, the similarity value between the vocabulary pairs is also very high.

The average value of similarity in this paper is higher than other algorithms and remains around 0.84, and the difference between the maximum value and the minimum value is no more than 0.01, which shows that this algorithm has good results in the calculation of similarity and the stability of the algorithm. Among them, the average similarity of the cosine similarity algorithm, keyword similarity algorithm, and common subsequence similarity algorithm is also high, while the average similarity of the word meaning similarity algorithm remains low.

As shown in Figure 1, it shows the comparative analysis of the accuracy and efficiency of the five algorithms under the condition of anonuniform similarity threshold.

As shown in Figure 2, the recall rates of the five algorithms are compared and analyzed in the case of inconsistent similarity thresholds.

As shown in Figure 3, it shows the comparison and analysis of F result values of five algorithms under the condition of a nonuniform similarity threshold. The harmonic average calculated by each algorithm is basically the same as the calculated recall rate. Since the growth rate of each algorithm P is lower than the reduction rate of R, the result of R has an obvious impact on the result of F value, and the change curve of F value is close to that of R.

4.2. Experimental Analysis of Similarity Calculation. By collecting and analyzing the usage of English vocabulary resources participating in the system comparison, some English vocabulary pairs that cannot be calculated are selected from the English vocabulary data set test, and the final ten pairs of words are tested.

As shown in Table 2, the calculation results of English vocabulary similarity are shown. It can be seen that the calculation results of S_1 column are lower than those of other columns. The reason for this phenomenon is that the high similarity of English vocabulary selects the system design method and notes the common characteristics of a large number of English vocabulary, which may include the influence of external interference factors, resulting in the low similarity of English vocabulary vector. The jump of S_2 column value is too high. The reason for this phenomenon may be that the selection and design of highly similar English words in the Baidu library do not match the artificial idea in some way.

The calculation results of English word similarity in S_1 are generally low in value, which is mainly due to the design method of the high similarity English word autonomous selection system based on the database, considering many English word features and the influence of some other interference factors, resulting in the low similarity of high-dimensional vectors of English word features.

As shown in Figure 4, it shows the selection efficiency of the corresponding similar English vocabulary selection system design when the number of data for selecting English vocabulary is 200, 400, and 600. If α corresponds to 1, the selection efficiency of a similar English vocabulary selection system design is 30%, 32%, and 45%, respectively. If α corresponds to 3, the selection efficiency of a similar English vocabulary selection system design is 40%, 44%, and 60%, respectively. If α corresponds to 5, the selection efficiency of a similar English vocabulary selection system design is 55%,

Table 1: Average similarity of vocabulary pairs with status 1 under different data numbers of different algorithms.

Number of data	500	1000	1500	2000	2500	3000	3500	4000	4500	5801
method1	0.717	0.742	0.728	0.731	0.730	0.729	0.731	0.731	0.730	0.730
method2	0.713	0.718	0.721	0.725	0.724	0.722	0.724	0.724	0.723	0.723
method3	0.373	0.378	0.376	0.378	0.378	0.377	0.376	0.380	0.378	0.380
method4	0.649	0.652	0.658	0.661	0.663	0.662	0.664	0.664	0.663	0.664
method5	0.841	0.837	0.839	0.841	0.845	0.844	0.846	0.846	0.846	0.846

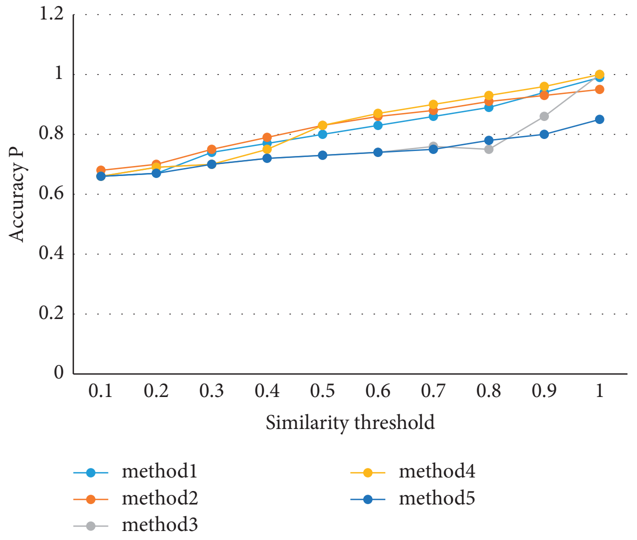

Figure 1: Comparison of accuracy of five algorithms under different similarity values.

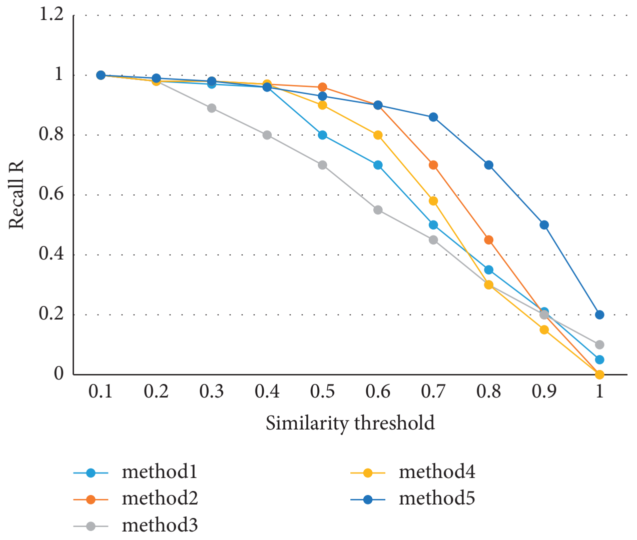

Figure 2: Comparison of recall rates of five algorithms under different similarity values.

63%, and 80%, respectively. Through comparative analysis, the recognition rate and weight of stable English lexical features can be obtained as α. In the interval [1,5], the selection efficiency will be the highest.

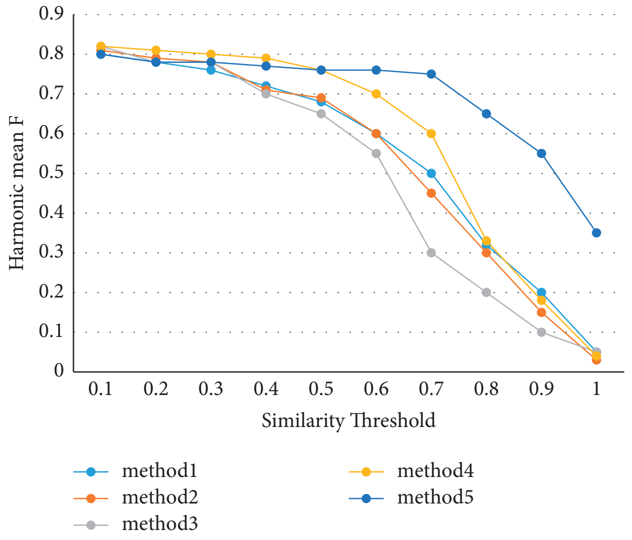

Figure 3: Comparison of F values of five algorithms under different similarity thresholds.

Table 2: Calculation results of English vocabulary similarity.

ID	W_1	W_2	S	S_1	S_2	S_3
1	Automobile	Car	0.923	1.032	0.998	0.935
2	Jewellery	Glass	0.856	7.698	0.789	0.864
3	Noon	Noon	0.015	0.203	0.036	0.017
4	Forest	Woodland	0.816	0.786	0.839	0.822
5	Phone	Telephone	0.805	0.963	0.823	0.811
6	Chair	Stool	0.236	0.354	0.254	0.240
7	Rope	Line	0.369	0.478	0.372	0.359
8	Worry	Worried	0.359	0.423	0.397	0.361
9	Hospital	Clinic	0.413	0.512	0.438	0.419
10	Reflection	Consider	0.716	0.836	0.725	0.712

4.3. *CD_Sim Test and Analysis of Methods.*

To verify the CD_, the calculation results of the SIM method show accuracy and time efficiency in practical application. Four types of data are randomly selected from English vocabulary as research simulation samples. Through the keyword extraction of the experimental results, the similarity measurement results are tested by cluster analysis and classification methods.

4.3.1. *Cluster Analysis.*

The results of similarity measure calculation indirectly affect the accuracy of the English vocabulary clustering algorithm. In addition, in the simulation sample, the accuracy of the clustering algorithm can in

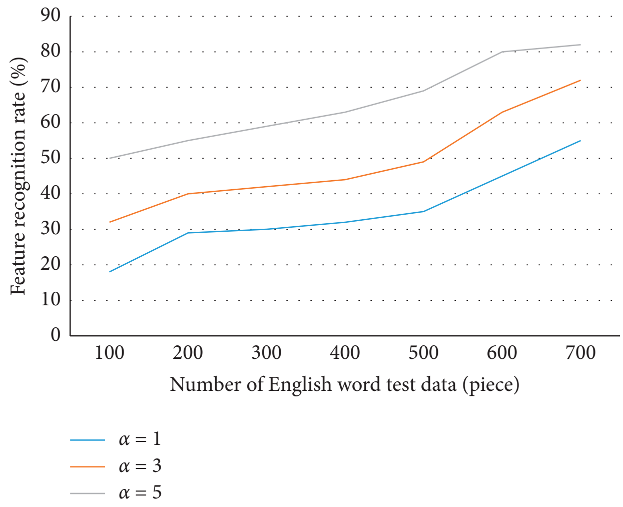

Figure 4: Selection efficiency of high similarity English vocabulary selection system design.

Table 3: Calculation and test results based on the cluster test method.

Method		AP clustering		Spectral clustering		Kmeans clustering	
	NUM	Entropy	Purity	Entropy	Purity	Entropy	Purity
Mean clustering	14	0.96	0.74	0.74	0.47	1.84	0.41
Hierarchical clustering	9	2.13	0.28	0.28	0.41	2.22	0.24
SOM clustering	18	0.33	0.90	0.90	0.82	1.26	0.66
FCM clustering	18	0.60	0.85	1.60	0.50	1.68	0.51

Table 4: Time complexity of similarity measurement method.

Method	Mean clustering	Hierarchical clustering	SOM clustering	FCM clustering
Time/s	10266	1014682	5734	10.6

Table 5: Overall classification accuracy of different feature numbers.

Number of features	Total classification accuracy (%)	Spend time (s)
110	67.17	830
550	72.30	945
1100	76.62	1216
1600	77.78	1536
2100	79.30	1872
3300	80.70	2305
4100	81.98	2742
5000	81.84	3062
5500	82.60	3177
6800	82.53	3684
8300	82.61	3752

turn test the quality of similarity results. Commonly used clustering algorithms include the distance matrix-based clustering algorithm, AP clustering algorithm, and gradually developed spectral clustering algorithm. Both the distance-based clustering algorithm and the spectral clustering algorithm are suitable for a given number of data, with high time complexity and clustering accuracy. If the given data are unknown, the results calculated by the two algorithms will have a certain deviation. Cluster analysis is formed according to the similarity measurement analysis. The specific experimental results are shown in Table 3.

As shown in Table 3, the four similarity measurement methods are compared and analyzed, The clustering result obtained by the sim method is the best, but there are only four documents in the data simulation sample, while the number of SIM clustering methods has reached 18, which is obviously unreasonable. Through the analysis of experimental clustering data, the experimental results of SIM are

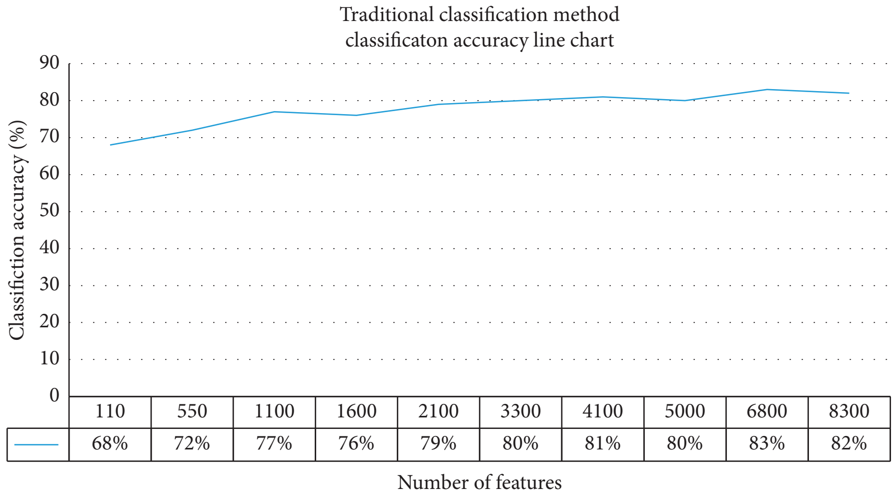

Figure 5: The broken line diagram of classification accuracy of traditional classification methods.

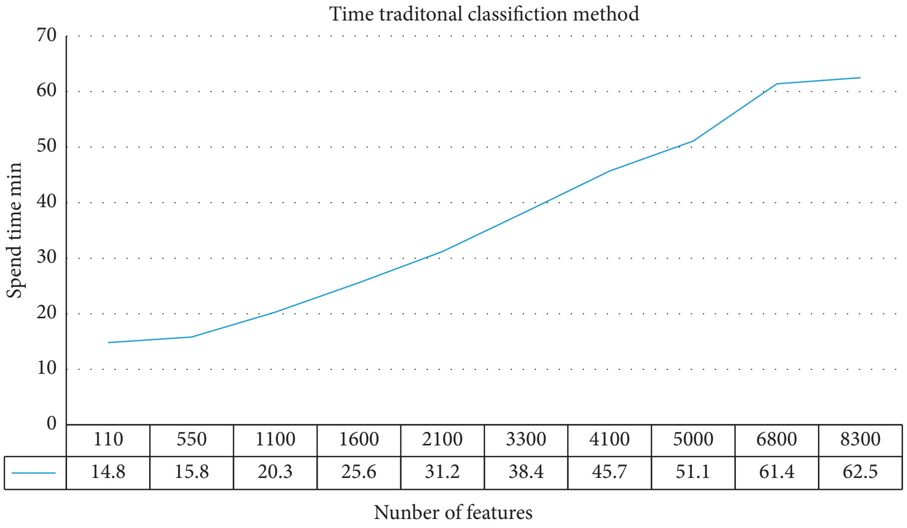

Figure 6: Test time of the traditional classification method.

better than CL_ SIM and ZWS_ SIM, the clustering entropy is the smallest, and the purity is the largest.

4.3.2. Time Complexity Analysis. According to the experimental data in Table 4, among the four text similarity measurement methods, FCM clustering similarity measurement method based on statistics has higher time efficiency, while the SOM clustering similarity measurement method has lower time efficiency than the mean clustering.

4.4. Experimental Results and Analysis of Classification Methods. In traditional classification experiments, word types are usually divided into simulation training, and only nouns, verbs, and verbs with nominality are selected as feature selection objects.

When the threshold values of feature number are 110, 550, 1100, 1600, 2100, 3300, 4100, 5000, 5500, 6800, and 8300, respectively, the overall classification accuracy is obtained.

As shown in Table 5, the overall classification accuracy table represents the number of individuals with different characteristics.

As shown in Figure 5, when the number of features is sorted from small to large, the classification accuracy rate and features increase linearly. When the number of features reaches about 5000, the classification accuracy rate is basically stable.

As shown in Figure 6, when the number of features is sorted from small to large, the test time and time spent basically increase linearly.

5. Conclusion

Firstly, this paper defines the concept of features and introduces the methods of statistical features of English characters and the research direction and background of the subject. Then, it analyzes the current situation of language development. The diversification of word meaning relationships between words has become the primary task of language word meaning research, that is, how to choose the correct method and model to express the relationship between language words, which is the purpose of this paper. Then, it introduces the meaning of similarity measurement and the calculation algorithm of similarity measurement, mainly including the feature selection of similarity, the embedding amount of similarity words, the distance between similarity numerical variables, and the calculation of the similarity coefficient. Finally, the efficiency of the similarity algorithm is compared and analyzed, the similarity measurement of fused language features is calculated and analyzed, and the CD is tested_ according to the classification method, the experimental calculation and analysis are carried out, and the experimental results are analyzed.

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 regarding this work.

Acknowledgments

This work was sponsored in part by (1) Research on the Curriculum System of Level-based College English Teaching Model under the Mode of Enrolling Students in Large Categories: A Case Study of Henan University of Chinese Medicine, Education and Teaching Reform Research Project of Henan University of Chinese Medicine (2021JX97). (2) Research on Cultivating the Translation Competence of MTI Candidates from the Perspective of "Translating Henan" (2020ZDB91); (3) Research and Practice on the Cultivation Model of Translation Talents of the Central Plains Culture Driven by the Project of "Translating Henan" (2021ZZUJGLX020); (4) Research on the Transformation of MTI Translation Theory Course in the Context of "Translation Henan" (2020YB0004).

References

[1] Y. Jiang, H. G. Wang, and N. Xi, "Target object identification and location based on multi-sensor fusion," *International Journal of Automation and Smart Technology*, vol. 3, no. 1, pp. 57–65, 2013.

[2] G. Papandreou, A. Katsamanis, and V. Pitsikalis, "Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition," *Audio, speech, and language processing*, vol. 17, no. 3, pp. 423–435, 2008.

[3] S. Ravi, S. Maloji, V. V. K. Polurie, and K. K. Eepuri, "Sign language recognition with multi feature fusion and ANN classifier," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 26, no. 6, pp. 2872–2886, 2018.

[4] B. Jiang, X. Huang, C. Yang, and J. Yuan, "SLTFNet: a spatial and language-temporal tensor fusion network for video moment retrieval," *Information Processing & Management*, vol. 56, no. 6, Article ID 102104, 2019.

[5] K. Tsukada, D. Birdsong, E. Bialystok, M. Mack, H. Sung, and J. Flege, "A developmental study of English vowel production and perception by native Korean adults and children," *Journal of Phonetics*, vol. 33, no. 3, pp. 263–290, 2005.

[6] W. S. Bainbridge, "Sociocultural meanings of nanotechnology: research methodologies," *Journal of Nanoparticle Research*, vol. 6, no. 2/3, pp. 285–299, 2004.

[7] A. Reid, D. Burnham, B. Kasisopa et al., "Perceptual assimilation of lexical tone: the roles of language experience and visual information," *Attention, Perception, & Psychophysics*, vol. 77, no. 2, pp. 571–591, 2015.

[8] W. S. Bainbridge, "Sociocultural meanings of nanotechnology: research methodologies," *Journal of Nanoparticle Research*, vol. 6, no. 2/3, pp. 285–299, 2004.

[9] Y. Hu, Y. Li, H. Lin, Z. Yang, and L. Cheng, "Integrating various resources for gene name normalization," *PLoS One*, vol. 7, no. 9, Article ID e43558, 2012.

[10] C. Schaffer, "Selecting a classification method by cross-validation," *Machine Learning*, vol. 13, no. 1, pp. 135–143, 1993.

[11] Y. Lei, Z. He, and Y. Zi, "Application of an intelligent classification method to mechanical fault diagnosis," *Expert Systems with Applications*, vol. 36, no. 6, pp. 9941–9948, 2009.

[12] M. J. Aitkenhead, "A co-evolving decision tree classification method," *Expert Systems with Applications*, vol. 34, no. 1, pp. 18–25, 2008.

[13] P. M. James, "An objective classification method for hbg," *Theoretical and Applied Climatology*, vol. 88, no. 1-2, pp. 17–42, 2007.

[14] W. M. Lin, C. D. Yang, J. H. Lin, and M. Tsay, "a fault classification method by RBF neural network with OLS learning procedure," *IEEE Power Engineering Review*, vol. 21, no. 8, p. 60, 2001.

[15] H. Wang and B. Fei, "A modified fuzzy C-means classification method using a multiscale diffusion filtering scheme," *Medical Image Analysis*, vol. 13, no. 2, pp. 193–202, 2009.