Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9853917, 1 page
https://doi.org/10.1155/2023/9853917

Retraction

Retracted: The Construction of College English Smart Classroom Based on Artificial Intelligence and Big Data

Wireless Communications and Mobile Computing

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Sun, D. Li, and M. Sun, "The Construction of College English Smart Classroom Based on Artificial Intelligence and Big Data," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3775803, 11 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3775803, 11 pages
https://doi.org/10.1155/2022/3775803

Research Article

The Construction of College English Smart Classroom Based on Artificial Intelligence and Big Data

Shan Sun,[1] Dapeng Li,[2] and Meng Sun[3]

[1]*School of Foreign Languages, College of Humanities & Information Changchun University of Technology, Changchun, 130000 Jilin, China*
[2]*School of Foreign Languages, Changchun University of Finance and Economics, Changchun, 130000 Jilin, China*
[3]*Department of Foreign Language Studies, Changchun Humanities and Sciences College, Changchun, 130000 Jilin, China*

Correspondence should be addressed to Meng Sun; sunmeng@ccrw.edu.cn

Received 24 May 2022; Revised 20 July 2022; Accepted 12 August 2022; Published 30 August 2022

Academic Editor: Chia-Huei Wu

In today's era, the English level of college students is very important. Different English classes can cultivate different English abilities. Smarter classroom is a concept put forward in the new century. This paper proposes the construction of a smarter classroom for college English with artificial intelligence and big data and proposes a deep neural network text semantic matching original model (OM), which uses different lexical information to match different English digital information. Combined with the K-means clustering method, different lexical semantic information is matched. After the comparison of experiments, the combination of the research theory and the algorithm in this paper is effective and has good use value.

1. Introduction

With the rapid development of economy and science and technology, artificial intelligence technology has been widely used in many fields of society, bringing great changes to the development of economy, society, education, medical treatment, and so on. In the era of educational informatization, using intelligent technology to build smart classroom has become a hot topic in educational reform research. In order to comply with the iterative renewal of information technology and the development and reform trend of educational informatization 2.0, a new smart classroom teaching mode came into being with the needs of the new era. In particular, the construction of college English smart classroom is conducive to the improvement of college students' English ability.

In today's era, the English level of college students is very important. Different English classes can cultivate different English abilities, such as listening, reading, and writing. Good curriculum design can cultivate college students' ability more pertinently. Different scholars have put forward different ideas and methods for the construction of college students' English wisdom classroom based on artificial intelligence. Smart classroom is a concept put forward since the new century. It is performed on the basis of the original ordinary classroom and becomes an intelligent classroom form focusing on technology. Especially in the past decade, thanks to the support of the Internet, Internet of things, big data, and other technologies, this form of classroom teaching has developed rapidly and played a great role during the epidemic. The "suspension of classes without suspension" ensures the students' learning progress and learning efficiency.

College English, a language discipline, requires teachers not to be complacent. They should open their horizons and fully combine the online and offline teaching modes to open up a new situation for their own teaching [1]. Smart classroom refers to an intelligent and efficient classroom based on constructivist learning theory and built by using new generation information technologies such as big data, cloud computing, the Internet of things, and mobile Internet to realize the whole process of application before, during, and after class [2]. Teachers can push a large number of curriculum-related resources on the platform to fully

mobilize students' curiosity and enthusiasm for the content to be explained. In classroom teaching, teachers can also use the network platform to understand the learning status of each student [3]. There are many problems in college English vocabulary teaching, such as the goal of vocabulary teaching is not clear; vocabulary teaching means are single; and lack of vocabulary evaluation system. The traditional teaching model is difficult to build a reasonable and complete vocabulary evaluation system, because it is limited by the source of resources, operation means, and so on [4].

In view of the above problems, this paper proposes the construction of college English intelligent classroom based on artificial intelligence and big data. For the artificial intelligence method, this paper proposes the original text semantic matching model (OM) of deep neural network, which uses different word semantic information for different English digital information. Combined with K-means clustering method, different lexical semantic information is matched. In the construction of college English vocabulary teaching wisdom classroom, we should warm up before class, push information, vocabulary test and enlightenment to teachers' teaching, cooperation in class, interaction between teachers and students, interaction between students and students, secondary detection in class, feedback after class, etc. After experimental comparison, the combination of theory and algorithm in this paper is effective and has good application value.

The main contributions of this paper are as follows:

(1) This paper uses the method of deep learning for text semantic matching, which is the main algorithm of this system and an important combination of deep learning in smart classroom

(2) K-means clustering method is used to match different lexical semantic information. It can quickly identify different English categories of information

(3) This paper designs a perfect wisdom curriculum system, considering many factors, including different classes of teachers and students

2. Related Work

At present, there are mainly two kinds of knowledge base question answering methods on English datasets. The first is semantic parsing method, which directly recognizes entities, entity relationships, and entity combinations from question sentences by compiling rule base, auxiliary dictionary, artificial reasoning, machine learning, and deep learning. Wang et al. [5] used the sequence annotation model to identify the entities in the question, used the sequence to sequence model to predict the relationship sequence in the question, and used the answer verification mechanism and cyclic training method to improve the performance of the model, which has reached an advanced level in the English multirelationship question dataset web question. Hu et al. [6] proposed a framework of state transition, designed four state transition actions and constraints, combined with multichannel convolutional neural network and other methods, and reached the most advanced level in the English complex problem dataset complex question. The method based on semantic analysis usually uses classification model to predict the relationship, which faces the problem of unregistered relationship, that is, the relationship that does not appear in the training set is difficult to be predicted. Chinese data usually contains more than thousands of relationships. When the number of relationships is very large, the effect of semantic analysis method is often not very good, which makes the semantic analysis method applied to Chinese knowledge base question and answer. Yu et al. [7] proposed a method to enhance relationship matching, which uses two-layer bi-LSTM for multilevel matching with candidate relationships and uses relationship matching to reorder entity link results, which has achieved the most advanced level in English multirelationship problem dataset. At present, the question answering method of knowledge base in Chinese domain is mainly improved based on information retrieval and vector modeling. Lai et al. [8] used convolutional neural network to identify semantic features in questions and determined the results through the matching degree of answers and questions; Dai et al. [9] proposed a method, which first carries out named entity recognition, then carries out attribute mapping through two-way LSTM [10] based on attention mechanism, and finally selects the answer from the knowledge base based on the results of the first two steps; Chen et al. [11] proposed a relationship extraction method integrating artificial rules to improve the accuracy of relationship recognition.

In order to meet the information needs of users, the online Q&A community with both social and Q&A functions came into being under the background of social network. Wang et al. [12] took the health information data released by the users of "home for the elderly" as the research object, identified keywords and topics based on the cooccurrence network, and analyzed the needs of users in the online community; Zhang [13] took tu-niu as an example, extracted text keywords by using TF-IDF and text rank, and built a cooccurrence network through Gephi, so as to master users' tourism information needs; Rezgui et al. [14] conducted research on user service demand aggregation based on the user comments of doctor clove and innovatively proposed a service demand aggregation method based on canopy K-means and MMR on the basis of Word2Vec word vector expression; Ning et al. [15] used latent semantic index model and MapReduce distributed text clustering technology and took the tumor section of medical network as an example to mine user information needs.

Expand English vocabulary teaching methods and teaching contents according to the context. For the traditional teaching mode of English vocabulary in senior high school, most teachers only tell students the meaning of vocabulary. Classroom teaching is usually read by teachers, and then ask students to read the vocabulary independently. Finally, teachers are explaining the Chinese interpretation of vocabulary for students. This English vocabulary teaching model will reduce students' learning efficiency [16, 17]. In order to strengthen the intuitiveness and vividness of English vocabulary teaching in senior high school and attract

students' attention, English teachers can build teaching scenes for students with the help of multimedia equipment or simple strokes and use diversified teaching modes to mobilize students' interest in learning [18, 19].

3. The Method

The research on the construction of college English smart classrooms based on artificial intelligence and big data is an important research direction at present. For the artificial intelligence method, a deep neural network text semantic matching original model (OM) is proposed, which uses different lexical information to match different English digital information. Combined with the K-means clustering method, different lexical semantic information is matched.

3.1. Text Semantic Matching Original Model (OM) for Deep Neural Networks. On the basis of the existing deep text semantic matching model, the self-supervised learning model is used to extract the interactive information of sequence conversion between sentence pairs, and the multitask learning method is used to dynamically participate in the extracted interactive information in the deep text semantic matching model train. The framework of this paper is divided into two parts: the original model (OM) and the self-supervised model (SSM) (Figure 1). The overall framework adopts the hard parameter sharing of multitask learning to build the connection between the two parts of the model.

Learn to get the feature interaction vector Vector_E of two sentences. Vector_E is calculated as follows:

(1) Given two sentence sets $A = \{a1, a2, \cdots, \text{an}\}$ and $B = \{b1, b2, \cdots, bn\}$, each with n sentences, and form a dataset of n sentence pairs. Denote the A sentence in the i-th sentence pair as $a_i = \{\omega_1^{a_i}, \omega_2^{a_i}, \cdots, \omega_m^{a_i}\}$, and $\omega_x^{a_i}(x \in [1, m])$ represents the x-th character/word of sentence a_i (characters for Chinese text, words for English text). m represents MaxLen in Figure 1, that is, the maximum length of the sentence sequence. Likewise, there is $b_i = \{\omega_1^{b_i}, \omega_2^{b_i}, \cdots, \omega_m^{b_i}\}, \omega_x^{b_i}(x \in [1, m])$

(2) The two sentences a_i and b_i of sentence pair are obtained through the embedding layer of TSSM to obtain the embedded representation, that is, the matrices $\text{Embed\_}a_i \in R^{m\times\text{Dim}}$ and $\text{Embed\_}b_i \in R^{m\times\text{Dim}}$; Dim represents the dimension of the embedding layer, which is set to 300 in the experiment

(3) Input the embedding representation of the two sentences of sentence pair i into TSSM to get Vector_E_i

$$\text{Vector\_}E_i = \text{TSMM}_i(\text{Embed\_}a_i, \text{Embed\_}b_i). \quad (1)$$

(4) Input Vector_E_i into the fully connected layer with the Sigmoid function as the activation function, and get the similarity score Sim_i of the two sentences

$$\text{Sim}_i = \text{Sigmoid}_i(W_o\text{Vector\_}E_i + b_o). \quad (2)$$

Among them, W_o and b_o are the parameters that can be learned and updated.

Denote the label of the sentence pair as $L = \{y1, y2, \cdots, yn\}$; $yi(i \in [1, n])$ represents the label of the i-th sentence pair, and use binary cross-entropy as the loss function:

$$\text{LOSS}_\text{OM} = -(L \cdot \log(\text{Sim}_i) + (1 - L) \cdot \log(1 - \text{Sim}_i)). \quad (3)$$

The self-supervised model (SSM) designed in this paper uses sequence generation to extract the interaction information of sentence pair vector matrix mutual generation and uses the interaction information to assist the task of text semantic matching. The pretext task of SSM is the mutual sequence generation of sentence pairs. The specific algorithm is as follows.

3.1.1. Input Design of SM. The two sentences a_i and b_i of sentence pair i are trained separately by the Skip-gram algorithm, and the Word2Vec [20] vector representation is obtained, namely $W_{a_i} \in R^{m\times\text{Dim}}$, $W_{b_i} \in R^{m\times\text{Dim}}$, m represents the length of the sentence sequence, Dim represents the vector dimension, and the matrix $\text{W2V\_ABi} \in \text{R2m} \times \text{Dim}$ is obtained by splicing:

$$W2V\_AB_i = \begin{bmatrix} W_{a_i} \\ W_{b_i} \end{bmatrix}. \quad (4)$$

3.1.2. Output Design of SSM. Concatenate the Word2Vec vector representations of the two sentences b_i and a_i of sentence pair i to obtain the matrix $\text{W2V\_BAi} \in \text{R2m} \times \text{Dim}$:

$$W2V\_BA_i = \begin{bmatrix} W_{b_i} \\ W_{a_i} \end{bmatrix}. \quad (5)$$

The SSM input takes the matrix $\text{W2V\_ABi} \in \text{R}^{2m\times\text{Dim}}$ of the sentence pair AB as input. The label of the SSM framework is $\text{W2V\_BAi} \in \text{R}^{2m\times\text{Dim}}$. SSM does not change the sequence length in the training process, so that the output of each input vector corresponds to the output vector one by one. The training mode of W2V_AB_i generating W2V_BA_i can make the interaction information extracted by the self-supervised model not only contain the contextual semantic information of the two sentences, but also contain the information of sequence transformation.

3.1.3. Feature Extraction of Convolution Layer. A one-dimensional convolution layer (Conv1D) of C layer was used to construct multi-CNN, and n-tuple features of W2V_AB_i were extracted, and these features were combined to form a matrix containing n-tuple features, denoted as $N_g \in R^{2\text{m}\times256\text{C}}$:

$$U_k = \text{Conv1}D_k^{k+1}(W2V\_AB_i), k \in [1, C], \quad (6)$$

$$N_g = [U_1, U_2, \cdots, U_C], \quad (7)$$

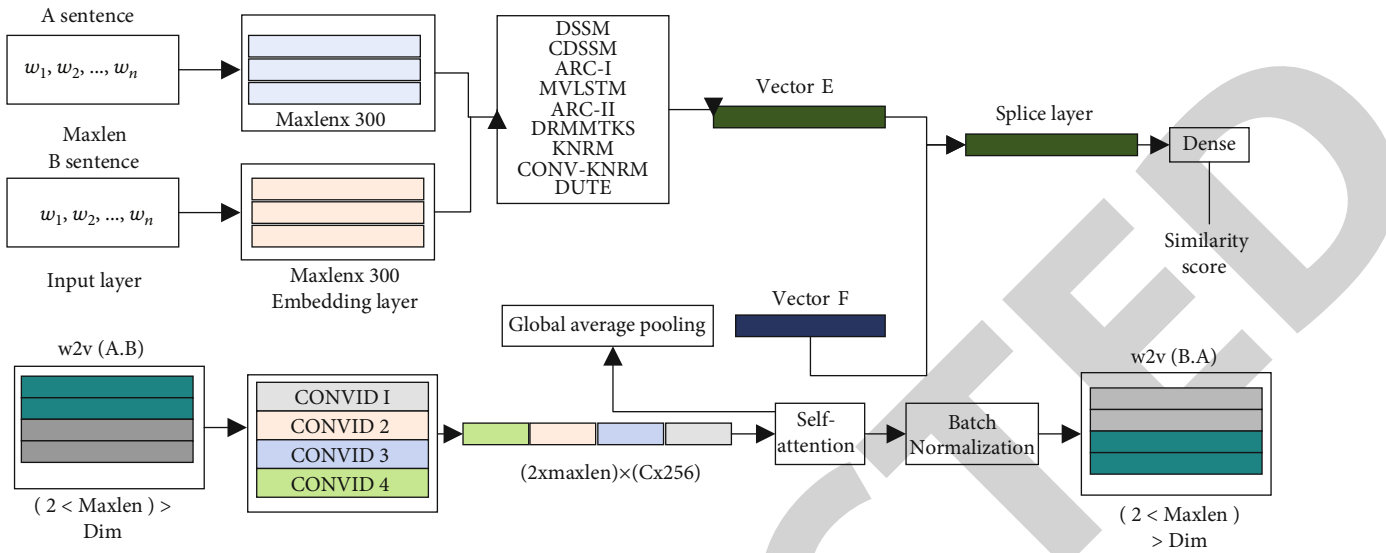


Figure 1: Model structure.

where $\text{Conv1}D_k^{k+1}$ stands for Conv1D at layer K and the width of convolution kernel is $K+1$, and U_k stands for the output of Conv1D at layer K.

As for the setting of multi-CNN level, for Chinese text, considering the existence of multiword words, C is set to 4, and the convolution kernel size is 2, 3, 4, and 5, respectively, so that the binary, ternary, quad, and quintuple features of the word vector matrix can be extracted simultaneously. For the English text, C is set to 3, and the convolution kernel size is 2, 3, and 4, respectively, to extract the binary, ternary, and quaternary features of the word vector matrix at the same time.

3.1.4. Sequence Feature Extraction and Model Output. The output of multilayer convolutional network in step 3 is taken as the input of self-attention to extract the sequence features of N-tuples, and each node of self-attention output contains the information of the whole sequence. After standardizing the output of the attention mechanism, time distributed fully connected network with Softmax as the activation function is used to obtain the output of SSM, denoted as $W2V\_\widehat{BA}_i$:

$$\text{BN} = \text{Batch\_Normalization}\left(\text{Self\_Attention}\left(N_g\right)\right), \quad (8)$$

$$W2V\_\widehat{BA}_i = \text{Soft max}_i\left(W_s \cdot BN + b_s\right) \quad (9)$$

Among them, W_s and B_s are parameters that can be learned and updated.

Cosine similarity considers the angle between vectors and is applicable to judge the similarity between the generated vector and the real vector. In contrast, MSE (mean square error) and MAE (mean absolute error) take more account of the distance between the predicted value and the true value, and do not consider similarity. SSM takes cosine similarity as the loss function:

$$\text{Loss}_{\text{SSM}} = -\text{Cosine}_i\left(W2V\_BA_i, W2V\_\widehat{BA}_i\right) \quad (10)$$

In multitasking learning (OM + SSM), the multitask learning framework proposed in this paper firstly needs to provide the interaction information of text exchange generation acquired in self-supervised learning process to the downstream core task (i.e., the original model).Specifically, this paper sums and averages normalized interaction information (BN) extracted by SSM through pooling layer to obtain vector Vector_F:

$$\text{Vector\_}F_i = \text{GlobalAveragePooling}_i(\text{BN}). \quad (11)$$

Then, after stitching the original model Vector_E_i and interactive information Vector_F_i, input the full connection layer with Sigmoid function as the activation function to obtain the similarity score Sim_Score_i:

$$\text{Sim\_Score}_i = \text{Sigmoid}\left(Wm[\text{Vector\_}E_i, \text{Vector\_}F_i] + b_m\right). \quad (12)$$

W_m and B_m are learnable and updatable parameters. In the training process, the overall loss function of multitasking learning is

$$\text{Loss}_{\text{ML}} = \text{Loss}_{\text{OM}} + \lambda\text{Loss}_{\text{SSM}}. \quad (13)$$

$\lambda \in (0, 1)$ is the weight coefficient of the loss function of the self-supervised model. In this experiment, the value of λ is 0.5.

Table 1 lists the SSM parameters. The number of self-attention neurons in each layer of multi-CNN is fixed at 256, and the activation function is ReLu.

This paper designs decomposition model and multitask model combining decomposition model and SSM. Self-attention model (SA) refers to inputting Vector_F_i, the text interaction information acquired in the process of self-supervised learning, into the full connection layer with Sigmoid function as the activation function to get the similarity

Table 1: Neural network parameters.

Dataset	Sequence length	Batch Size	OM embedding	SSM input	SSM output	Multi-CNN
MSRP	34	64	(34,300)	(34,300)	(68,300)	2,3,4
CCKS18-T3	40	64	(40,300)	(40,100)	(80,300)	2,3,4,5
TCA120	20	64	(20,300)	(20,300)	(40,300)	2,3,4,5
GAIIC21-T3	37	64	(37,300)	(37,100)	(74,300)	2,3,4,5
GAIIC21-T3M	30	64	(30,300)	(30,100)	(60,300)	2,3,4,5

score of sentence pairs. This model evaluates whether the text interaction information Vector_F_i learned in pretext task in self-supervised learning can be used independently for text similarity calculation. Based on this, this paper also designed a multitask SA + SSM model; that is, the loss function of the SA decomposition model and the loss function of the SSM model are weighted to sum as the similarity of sentence pairs predicted by the multitask SA + SSM model, and the weighted way is the same as the loss function of OM + SSM multitask learning; λ is 0.5 [21–23].

3.2. The K-Means Clustering Method Matches Different Lexical Semantic Information. The central idea of K-means determine the constant K in advance, the constant K means the final number of cluster categories; first randomly select the initial point as the centroid and calculate the similarity between each sample and the centroid (here is the Euclidean distance),, assign the sample points to the most similar class, then recalculate the centroid of each class (that is, the class center), repeat this process until the centroid does not change, and finally determine the class to which each sample belongs and the centroid of each class. Since the similarity between all samples and each centroid is calculated each time, the convergence speed of the K-means algorithm is relatively slow on large-scale datasets.

The biggest difference between the clustering algorithm and the classification algorithm is that the clustering algorithm is an unsupervised learning algorithm, while the classification algorithm belongs to the supervised learning algorithm, and the classification is to know the result. In the clustering algorithm, the samples are divided into different categories according to the similarity between the samples. For different similarity calculation methods, different clustering results will be obtained. The commonly used similarity calculation method is the Euclidean distance method.

For each point, calculate the center point that is closest to all center points, and then classify this point into the cluster represented by this center point. After one iteration, recalculate the center point for each cluster class, and then refind the center point closest to itself for each point. This cycle is repeated until the cluster class of the two iterations before and after does not change. [24, 25]

In K-means, first define a class, class K-means; since the implementation of this algorithm needs to read and store external data, a container vector is defined at a time, in which the data type is the structure *st_point*, which contains three-dimensional point coordinates and a char type the ID of the class to which it belongs. Next is the function declaration. The specific flow chart is shown in Figure 2:

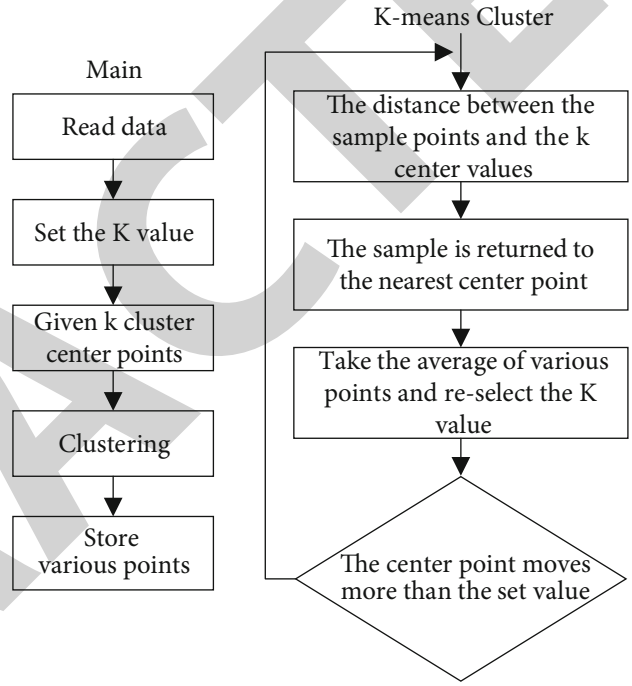


Figure 2: Basic program structure and corresponding functions.

The public functions of different functions are specifically given in K-means. As shown in Figure 1, the functions are relatively refined, which is convenient for later application expansion. The more specific clustering function is cluster, which is strictly based on the basic principle of K-means. Similarity is the simplest Euclidean distance, and the end of iteration is judged by whether the deviation of the two central values is greater than the given Dist_near_zero value [26].

The flow of the K-means algorithm is as follows:

(1) Select the number of clusters k (when the K-means algorithm passes hyperparameters, you only need to set the maximum *K* value)

(2) Generate k clusters arbitrarily, and then determine the cluster centers, or directly generate k centers

(3) For each point, determine its cluster center point

(4) Determine whether the categories of the sample points before and after the clustering are the same. If they are the same, the algorithm terminates; otherwise, go to step 5

(5) For the sample points in each category, calculate the center point of these sample points as the new center point of the class, and continue step 2

3.3. Construction of Wisdom Classroom in College English Vocabulary Teaching

3.3.1. Warm-up before Class. Preclass vocabulary warm-up activities focus on helping students solve the basic content of the words taught, such as word pronunciation, meaning recognition, synonym and synonym analysis, part of speech, and grammatical function, which lays a good foundation for teachers to extend the explanation of vocabulary in class and guide students to use it in practice. The warm-up before class is mainly divided into three parts.

Information push. Before class, teachers can use Dingding platform, WeChat group, email, or other network platforms to push videos, audio, pictures, memory methods, and master degree of word explanation to students, so as to help students to clarify the teaching objectives and learning points of each word. Students can also send some new ways of understanding and reciting words to the group and discuss with teachers and other students.

Vocabulary test. Teachers should test students' vocabulary mastery according to the teaching objectives of the pushed vocabulary and set reasonable, effective, and targeted topics. For example, read the word topic, and ask the students to upload the voice file or the word dictation topic; choose words to fill in the blanks, focusing on the students' knowledge of parts of speech, grammar, and word meaning. Cloze question: Distinguish synonyms from synonyms. Through this part of the test, students can further clarify the mastery of the learned words and the key and difficult points, laying a good foundation for classroom practice.

Enlightenment to teacher teaching. Teachers can master first-hand information through online evaluation of students' vocabularies and then summarize which vocabularies students have mastered well. They can directly do extended exercises in class and which vocabularies are not mastered well and to what extent. Further explanation and reinforcement should be made in class. At the same time, we keep updating the vocabulary explanation resources to find better and more appropriate ways to explain vocabulary, so as to gradually systematize and perfect vocabulary teaching.

3.3.2. Collaboration in the Class. The classroom is mainly the cooperation and interaction between teachers and students as well as between students and students, mainly engaged in the output of vocabulary, to help students dig and understand the essence of vocabulary from a deeper level and stand on a higher level to feel and experience vocabulary, on the basis of the warm-up before class, a higher level, including teacher-student interaction, student-student interaction, and classroom secondary testing.

Table 2: Dataset statistics.

Question type	Training set	Validation set	Test set
Single entity single relationship	1159	476	484
Single entity multiple relationship	682	156	160
Multientity	356	133	121
Total	2297	765	765

3.3.3. After-Class Feedback. After-class feedback mainly refers to teachers' further tracking of students' after-class learning and helping them to review. Based on the second vocabulary test results of each student, teachers can classify them as excellent, medium, and inferior. According to different grades, the way, degree, and content of teachers' tracking will be very different. Since after-class feedback is the final stage for students to learn a certain unit's vocabulary, teachers should try to be specific and give the most targeted guidance for each student.

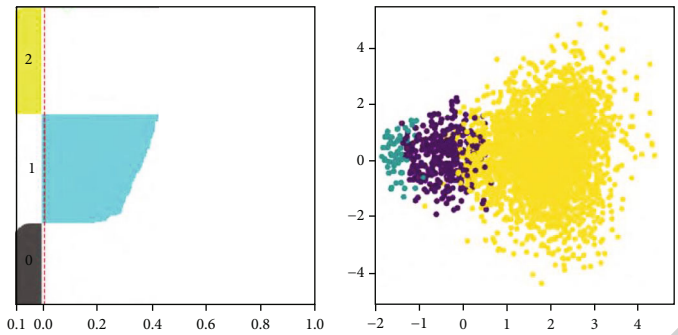
Top students already have the highest level of knowledge and understanding of the vocabulary taught and can use the word flexibly and thoroughly internalize it as their own knowledge. Middle school students' grasp of vocabulary mostly stays at the level of pronunciation, part of speech, and meaning, but they are not able to use words to make sentences, carry out conversations, or write articles, that is, there are problems in the application of vocabulary. The inferior students belong to the students with very weak foundation and have great problems in the basic skills of words. Teachers should pay more attention to this kind of students and make more efforts.

4. Experiment Analysis

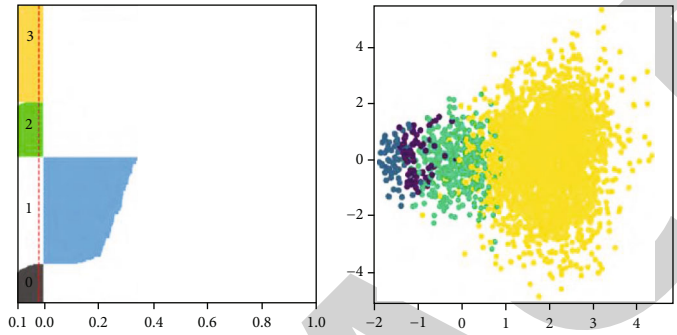
Our method combines neural network and dictionary classification model. The specific methods are as follows. In auxiliary dictionary construction, in the process of this method, multiple dictionaries are required for word segmentation and word frequency calculation, including entity link dictionary, word segmentation dictionary, word frequency dictionary, and attribute dictionary; entity recognition and attribute value recognition include entity recognition and attribute value recognition. The attribute value contained in the problem is less standardized, which may be a long word sequence, or it may not be able to directly correspond to the knowledge base entity. Some entities will be ignored only through the word segmentation dictionary; entity link and filtering and calculate some features for each entity; candidate query path generation and text matching; and entity splicing and answer retrieval.

The data statistics of the dataset are shown in Table 2.

4.1. Data Clustering. After data preprocessing and word vector conversion, the experiment enters the clustering process. The value range of the silhouette coefficient is generally [-1, 1], and the larger the value, the farther the distance between the cluster and other clusters, and the more compact the



(a) The number of clusters is 3; the silhouette coefficient value is 0.008



(b) The number of clusters is 4; the silhouette coefficient value is -0.035

Figure 3: The contour coefficient values and scatter plots of GMM with different clustering numbers.

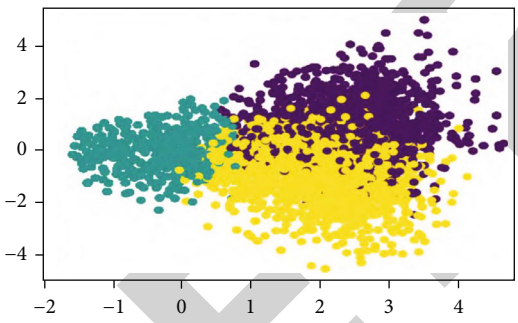


Figure 4: K-means clustering scatter diagram.

Table 3: Comparison results of evaluation indexes of different algorithms.

Algorithm type	DB index	CH index	Silhouette factor
G-K-means	4.578480326	1746.4285063	0.24638803
Birch	4.849882165	1387.9746962	0.24825625
DBSCAN	460.6305243	38.682026442	0.36329085
Hierarchical clustering	6.537886241	1381.5631847	0.21776102

distance within the cluster. In addition, too much or too little data in the same cluster will have a certain impact on the persuasiveness and representativeness of the clustering results. Therefore, the determination of the K value needs to comprehensively consider the silhouette coefficient value and the distribution of the data in the scatter plot. In short, on the basis of the maximum possible value of the silhouette coefficient, the more uniform the distribution of cluster data, the more reasonable the K value at this time. In order to obtain the most appropriate cluster assignment results, this paper conducts a control experiment of different cluster assignments. The specific experimental results are shown in Figure 3.

Observing the experimental results, we can see that when the number of clusters is 2 and 3, the value of the silhouette coefficient is relatively high, and the more the number of clusters, the smaller the value of the silhouette coefficient. The generated scatterplot shows that when $K = 2$, the distribution of the two clusters is obviously uneven, and when $K = 3$, the data distribution of the three clusters is obviously relatively uniform. Therefore, through the comparative experiment, it can be seen that the more suitable number of clusters for the data collected in this paper is 3. The center point at this time is used as the cluster center point of the experiment, and the value of parameter K is 3 to perform the final K-means clustering. The clustering effect of the sample data is shown in Figure 4.

In order to further verify the effectiveness of the user demand aggregation method G-K-means proposed in this paper, a comparative experiment is specially set up. Under the condition that other conditions remain unchanged, three traditional clustering algorithms of birch algorithm, hierarchical clustering algorithm, and DBSCAN algorithm are randomly selected, and the specified number of clusters is

Table 4: Comparing model performances (%).

	MODEL	MSRP	CCKS18-T3	TCAI20	GAIIC21-T3	GAIIC21-T3M
Representation-based OM model and OM SSM multitask model	ARC-I	75.51	69.53	74.88	85.34	93.08
	ARC-I + SSM	77.64	71.40	77.04	88.69	94.35
	DSSM	80.24	73.22	68.72	76.86	81.78
	DSSM+SSM	80.64	76.93	83.19	89.29	91.96
	CDSSM	79.75	70.47	70.88	79.83	69.51
	CDSSM+SSM	80.75	76.67	84.03	89.42	91.56
	ARC-II	77.41	71.23	75.21	80.53	83.87
	ARC-II + SSM	78.85	71.32	76.37	85.37	90.18
Interaction-based OM model and hybrid OM model and OM SSM multitasking model	DRMMTKS	78.81	75.73	87.02	68.51	67.03
	DRMMTKS +SSM	81.09	77.74	88.02	88.11	91.16
	K-NRM	78.13	74.01	73.88	67.01	63.10
	K-NRM + SSM	79.77	77.25	82.70	89.75	93.04
	CONV-KNRM	77.73	76.42	74.54	83.61	81.95
	CONVKNRM+SSM	80.17	78.03	86.86	87.31	90.28
	MVLSTM	76.04	75.17	79.67	81.18	83.63
	MVLSTM+SSM	79.87	77.43	83.36	88.43	93.30
	DUET	76.69	74.78	83.53	82.34	85.39
	DUET+SSM	78.56	76.25	85.69	89.29	93.37
Decomposition model and SA + SSM multitasking	Self-attention(SA)	80.50	75.57	81.03	89.75	93.13
	SA + SSM	80.76	75.79	82.69	89.88	94.12

3. The comparison and analysis of the experimental results are carried out in the two dimensions of graph and evaluation index. From the evaluation index values of different algorithms in Table 3, it can be seen that the DB index value of the G-K-means algorithm is the smallest, the CH index value is the largest, and the contour coefficient is large, indicating that compared with the other three algorithms, the G-K-means algorithm is more in line with the "principle of determining the optimal clustering quality." However, the birch algorithm, hierarchical clustering algorithm, and DBSCAN algorithm all show that their clustering effect is poor, and the cluster distribution is not ideal, especially in the DBSCAN algorithm. It can be seen from the scatter diagram that the algorithm hardly divides the dataset into clusters, which is obviously not suitable for the clustering of question texts in the online Q&A community.

4.2. Text Semantic Matching Original Model (OM) Combined with Deep Learning. The two research questions proposed in this paper are discussed through the experimental results in Table 4. F1-score, accuracy, and AUC are all within the range of 0 to 1. RQ1 is discussed first. From the experimental results of the representation-based model in Table 4, it can be seen that after adding SSM, the ARC-I model is improved by 2.8%, 2.7%, 2.9%, 3.9%, and 1.4% in the five datasets, respectively; the DSSM model is improved by 0.5%, 5.1%, 21.1%, 16.2%, and 12.4%; CDSSM improvements are 1.3%, 8.8%, 18.6%, 12.0%, and 31.7%, respectively. It can be seen that the performance of the representation-based model on the five datasets has been improved after adding SSM. The interaction information extracted by self-supervised learning can make up for the shortcomings of these models.

By comparing the experimental results of decomposition model SA and original model OM on five datasets, it is found that the SA model is better than all OM models on three of the datasets. This indicates that the self-supervised auxiliary task designed in this paper can learn effective text interaction information that can be used for text similarity calculation. At the same time, the multitask model SA + SSM combined with SA also achieves optimal results on a dataset (GaiIC21-T3). The multitask model combined with self-supervised learning (OM + SSM) proposed in this paper achieves the best results on the other four datasets, indicating that it is effective for downstream tasks.

Table 5 shows the improvement of the performance of each dataset after combining the self-supervised model. It can be seen that for the MSRP dataset, the improvement effect of the 9 models is not obvious, with an average improvement of 2.44%. The sentences of this dataset are extracted from multiple news websites, and each sentence is from a different news article, which well eliminates the possible semantic similarity between sentences, and may also lead to less commonalities between sentences and themes complex, which shows that SSM is less robust to the mutual generation of sentence pairs in response to different topics. For the CCKS18-T3 dataset, the improvement effect of all models is not good enough, with an average improvement of 3.42%. The dataset comes from We-Bank intelligent customer service question matching, and its core is the intent matching between sentence pairs, while SSM lacks the extraction and representation of sentence intent features,

Table 5: Self-supervised model improvement by dataset (%).

Dataset	ARC-ISSM	DSSM+SSM	CDSSM+SSM	ARC-IISSM	DRMMTKS+SSM	KNRM+SSM	CONV+SSM	MVL+SSM	DUET+SSM	AVG
MSRP	2.8	0.5	1.3	1.9	2.9	2.1	3.1	5.0	2.4	2.44
CCK-T3	2.7	5.1	8.8	0.1	2.7	4.4	2.1	3.0	2.0	3.43
TCA120	2.9	21.1	18.6	1.5	1.1	11.9	16.5	4.6	2.6	8.98
GAIIC21-T3	3.9	16.2	12.0	6.0	28.6	33.9	4.4	8.9	8.4	13.59
GAIIC21-T3M	1.4	12.4	31.7	7.5	36.0	47.4	10.2	11.6	9.3	18.61

Table 6: Result of entity link on test set.

Feature	Reserved quantity	Recall@n/%
Question entity + entity feature	All (avg 12.6)	95.3
Question entity + entity feature	1	80.5
Question entity + entity feature	3	92.6
Question entity + entity feature	5	93.4
Question entity + entity feature	10	94.7
Solid features only	3	88.4
Question-only entity features	3	70.3

Table 7: Knowledge base Q&A results on the test set.

Search method	Number of negative cases	*F*-score
Text only match results	3	68.7
+ solid stitching	3	70.5
+ entity splicing and literal matching	3	75.6
+ entity splicing and literal matching	1	74.1
+ entity splicing and literal matching	5	72.1
+ entity splicing and literal matching	10	66.7

which shows that the self-supervised model based on sequence generation lacks deep semantic feature extraction ability. For the other three datasets, the model improvement brought by adding self-supervised learning is more obvious. TCAI20 is the judgment of similar sentences of the new crown epidemic, and GAIIC21-T3(M) is the artificial intelligence assistant dialogue short text matching. The sentences in these datasets have similar topics, and SSM can extract high-quality interactive information.

4.3. English Corpus Information Matching. For the entity linking link, ablation experiments are performed on the test set for 5 kinds of features, and the recall rates of retaining different numbers of candidate entities are recorded. The experimental results are shown in Table 6, where Recall@n/% represents the recall rate of all question annotated entities while retaining the first n candidate entities.

The results show that (1) the selected question entity features and knowledge base entity features have a great influence on the accuracy of entity linking; (2) from the experimental results, only keeping the top 5 candidate entities can reach nearly all the number of entities, while choosing to keep only the top five entities can also reduce training time and data noise.

On the test set, the F values of different numbers of negative examples and different retrieval schemes in the text matching process are calculated. This paper compares the performance of three schemes: (1) directly select the query path with the highest similarity after text matching; (2) use bridging for all questions to obtain possible query paths for multientity situations. (3) Rematch the top 3 paths and multientity paths of text matching with the question in terms of overlapping words, and select the one that is literally the most similar as the final query path.

From the experimental results and analysis in Table 7, it can be concluded that in the text matching process, a suitable number of negative examples can obtain better learning text similarity, and three negative examples have the best effect on this task; entity splicing can consider multientity However, some errors will be introduced, that is, some problems that are actually single entities get query paths for multientity cases, and overlapping word count matching can effectively alleviate this problem.

Advantages of the model are as follows: (1) Using the pretraining model and the knowledge base word segmentation technology greatly improves the recognition accuracy of the subject words of the question; (2) use the text matching technology to match the question and the query path of the entity in the knowledge base to avoid the problem of unregistered relationships; and (3) use entity splicing to explore multientity and multirelationship problems. Model defects are as follows: (1) the entity linking technology based on machine learning is more dependent on the features of question entities and knowledge base entities; (2) too many candidate query paths are generated, which affects the efficiency of the model. Therefore, the author believes that in the future, deep learning technology can be used to link entities, reduce feature dependencies, and improve accuracy; add entity type and entity quantity information in questions to further improve the accuracy of multientity and multirelationship problems.

5. Conclusion

In the era of educational informatization, the use of intelligent technology to build smart classrooms has become a hotspot in educational reform research. In particular, the construction of college English smart classroom is conducive to the improvement of college students' English ability. In view of the difficulty of English learning and the complex goals, this paper proposes the construction of a smart classroom for college English with artificial intelligence and big data. For the artificial intelligence method, a deep neural network text semantic matching original model (OM) is proposed, aiming at different English digital information, using different lexical information for matching. Combined with the K-means clustering method, different lexical semantic information is matched. The experimental results show that the method proposed in this paper has a good effect in the smart classroom.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there is no conflict of interest with any financial organizations regarding the material reported in this manuscript.

References

[1] R. Wrdhugh, *An Introduction to Sociolinguists*, Blackwell Publishers Ltd, Oxford, 1998.

[2] D. Simon, "Towards a new understanding of codeswitching in the foreign language classroom," *Trends in Linguistics Studies and Monographs*, vol. 126, pp. 311–342, 2001.

[3] M. Turnbull, "There is a role for the L1 in second and foreign language teaching, but…," *The Canadian Modern Language Review*, vol. 57, no. 4, pp. 531–540, 2001.

[4] S. Krashen and T. Terrell, *The Natural Approach*, Prentice Hall, New Jersey, 1983.

[5] Y. Wang, R. Zhang, C. Xu, and Y. Mao, "The APVA-TURBO approach to question answering in knowledge base," in *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1998–2009, Santa Fe, New Mexico, USA, 2018.

[6] S. Hu, L. Zou, and X. Zhang, "A state-transition framework to answer complex questions over knowledge base," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2098–2108, Brussels, Belgium, 2018.

[7] M. Yu, W. Yin, K. S. Hasan, C. D. Santos, B. Xiang, and B. Zhou, "Improved neural relation detection for knowledge base question answering," 2017, http://arXiv:1704.06194.

[8] Y. Lai, Y. Jia, Y. Lin, Y. Feng, and D. Zhao, "A Chinese question answering system for single-relation factoid questions," in *National CCF Conference on Natural Language Processing and Chinese Computing*, pp. 124–135, Cham, 2018.

[9] Z. Dai, L. Li, and W. Xu, "Cfo: conditional focused neural question answering with large-scale knowledge bases," 2016, http://arXiv.1606.01994.

[10] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, vol. 2012, pp. 37–45, 2012.

[11] H. C. Chen, Z. Y. Chen, S. Y. Huang, L. W. Ku, Y. S. Chiu, and W. J. Yang, "Relation extraction in knowledge base question answering: from general-domain to the catering industry," in *International Conference on HCI in Business, Government, and Organizations*, pp. 26–41, Cham, 2018.

[12] J. Wang, L. Wang, J. Xu, and Y. Peng, "Information needs mining of COVID-19 in Chinese online health communities," *Big Data Research*, vol. 24, article 100193, 2021.

[13] W. Zhang, "Text mining applied in evolution of Q & A platforms users' information demand on tourism in COVID-19," in *Normalization//2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA)*, pp. 29–40, Changsha, China, 2021.

[14] Y. Rezgui, S. Boddy, M. Wetherill, and G. Cooper, "Past, present and future of information and knowledge sharing in the construction industry: Towards semantic service-based e-construction?," *Computer-Aided Design*, vol. 43, no. 5, pp. 502–515, 2011.

[15] C. Ning, J. Xu, H. Gao, X. Yang, and T. Wang, "Sports information needs in Chinese Online Q&A Community: topic mining based on BERT," *Applied Sciences*, vol. 12, no. 9, p. 4784, 2022.

[16] L. Mengze, "Research on the teaching mode of "teacher-student collaborative development" in high school English teaching," *New Curriculum Research*, vol. 35, pp. 89-90, 2020.

[17] H. Y. Zhang, "A study on the effectiveness strategy of teaching objectives design based on core literacy in high school English class hours," *Campus English*, vol. 37, pp. 223-224, 2020.

[18] L. Biying, "Application of flipped classroom teaching model based on micro-lesson in high school English grammar teaching," *Campus English*, vol. 38, pp. 162-163, 2020.

[19] W. Bingyue, "The reform of English teaching model in high school under the background of "Internet +"," *Anhui Education and Research*, vol. 4, pp. 80-81, 2020.

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, 2019.

[21] S. Palanisamy, B. Thangaraju, O. I. Khalaf, Y. Alotaibi, S. Alghamdi, and F. Alassery, "A novel approach of design and analysis of a hexagonal fractal antenna array (HFAA) for next-generation wireless communication," *Energies*, vol. 14, no. 19, p. 6204, 2021.

[22] S. N. Alsubari, S. N. Deshmukh, A. A. Alqarni et al., "Data analytics for the identification of fake reviews using supervised learning," *CMC-Computers, Materials & Continua*, vol. 70, no. 2, pp. 3189–3204, 2022.

[23] L. Qingfeng, L. Chenxuan, and W. Yanan, "Integrating external dictionary knowledge in conference scenarios the field of personalized machine translation method," *Journal of Chinese Informatics*, vol. 33, no. 10, pp. 31–37, 2019.

[24] R. Ali, M. H. Siddiqi, and S. Lee, "Rough set-based approaches for discretization: a compact review," *Artificial Intelligence Review*, vol. 44, no. 2, pp. 235–263, 2015.

[25] D. Wu, Y. Lei, M. He, C. Zhang, and L. Ji, "Deep reinforcement learning-based path control and optimization for unmanned ships," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 7135043, 8 pages, 2022.

[26] F. M. Abd Algalil and S. P. Zambare, "Effects of temperature on the development of Calliphorid fly of forensic importance Chrysomya megacephala (Fabricius, 1794)," *Journal of Applied Research*, vol. 5, no. 2, pp. 767–769, 2015.