Research Article

Rihong Tang*

Application and research of English composition tangent model based on unsupervised semantic space

https://doi.org/10.1515/jisys-2023-0148
received September 01, 2023; accepted February 20, 2024

Abstract: Nowadays, major enterprises and schools vigorously promote the combination of information technology and subject teaching, among which automatic grading technology is more widely used. In order to improve the efficiency of English composition correction, the study proposes an unsupervised semantic space model for English composition tangent, using a Hierarchical Topic Tree Hybrid Semantic Space to achieve topic representation and clustering in English composition; adopts a feature dimensionality reduction method to select a set of semantic features to complete the optimization of the feature semantic space; and combines the tangent analysis algorithm to achieve intelligent scoring of English composition. The experimental data show that the accuracy and *F*-value of the English composition tangent analysis method based on the semantic space are significantly improved, and the Pearson correlation coefficient between the unsupervised semantic space English composition tangent model and the teacher's manual grading is 0.8936. The results show that the unsupervised semantic space English composition tangent model has a higher accuracy rate, is more applicable, and can efficiently complete the English composition grading task: essay review task.

Keywords: semantic space, tangential analysis, English composition, hierarchical topic tree, character filtering

1 Introduction

As the popularity of English has increased, the ability to teach English has gradually improved and students' English standards have also improved significantly [1]. Writing occupies an important place in the English learning process, reflecting students' ability to use language and express themselves in writing. Students' English composition reflects the extent to which they have mastered the basics of vocabulary and grammar, as well as their overall mastery of sentence and paragraph structure and the logic of the text [2,3]. English composition writing is commonly found in subject examinations, Level 4 and 6 examinations and major English competitions, and is a comprehensive reflection of students' English writing ability. Many experts and scholars agree that writing is an important way of assessing students' ability to use language. However, there are many problems with the current traditional English teaching model, such as the overwhelming number of students resulting in a heavy task of reviewing essays, and the difficult task of teachers resulting in untimely feedback on review information, all of which can lead to a failure to improve students' writing skills quickly and efficiently [4]. To address these problems, researchers have proposed an automated English grading system that combines technologies such as natural language processing and machine learning to reduce teachers' physical and mental workload [5,6]. However, this field still faces some challenges, such as how to accurately evaluate students' writing abilities and how to handle complex language structures and

*** Corresponding author: Rihong Tang**, Department of Basic Education, Weihai Ocean Vocational College, Weihai, 264300, China, e-mail: tangrh_trh@outlook.com

semantics. Therefore, this study designed a semantic space English composition topic analysis model based on unsupervised methods. The main innovation lies in the use of relational triples as carriers for topic clustering and distributed representation, and proposed topic analysis algorithms, topic coherence algorithms, and topic viewpoint algorithms for multi-dimensional topic analysis of English composition content. Through experimental verification, this model has high accuracy and application value in scoring the quality of relevance to the topic. The main contribution of the research is to propose a method for constructing a hierarchical topic tree (HTT) mixed semantic space and to extend topic relevance analysis from shallow semantic analysis to potential topic semantic analysis, improving the accuracy and granularity of topic relevance semantic analysis.

2 Related works

2.1 Semantic space analysis

Semantic space is an important concept in the study of communication effectiveness, and a prerequisite for communication to be realized is a common semantic space between the transmitter and the receiver. Xiao et al. addressed the problem of low accuracy of scene migration by proposing a simulated realistic decision model based on feature semantic space, combining environment representation, policy optimization, and intelligent decision modules to shorten the difference between real and virtual scenes. It is verified that the method has good stability in practical applications [7]. To optimize the semantic space of text, Kherwa et al. adopted the three-level weight model to distribute the weight of terms, documents, and corpora, so as to achieve the generation of semantic similarity and context clustering [8]. Yu’s research group addressed the convolutional neural network neural-style conversion problem, proposed a multi-scale style conversion algorithm based on deep semantic matching, combined with spatial segmentation and contextual illumination information to construct a deep semantic space, and used the loss function of nearest neighbour search to optimize the effect of deep-style migration, and the experimental results show that the algorithm synthesizes a more reasonable spatial structure image [9]. To optimize the conversion method of semantic space, Yu proposed a semantic space-based and automatic bibliographic classification algorithm based on conversion, combining text preprocessing and word vectors to achieve automatic classification of bibliographic semantic vectors. Experimental data show that the accuracy of this algorithm is higher than that of traditional classification algorithms [10]. Orhan et al. proposed an embedding method for learning word vectors through weighted semantic relations, finding the best weight for them by semantic relations and adjusting the Euclidean distance to obtain word vectors of synonymous sets. Experimental results show that the method is able to find word-level semantic similarities and weights [11].

2.2 Automatic scoring technology

With the development of the times and the progress of technology, major enterprises and schools vigorously promote the combination of information technology and subject teaching, forming the concept of Internet + education, among which the application of automatic scoring technology is more common. Zhao used English automatic scoring algorithm and English sentence feature scoring algorithm to achieve intelligent online scoring and elegant sentence extraction of English composition, and the experimental results show that the algorithm can reduce the workload of English and the experimental results show that the algorithm can reduce the workload of scoring English essays [12]. Wang et al. proposed an improved P-means-based automatic scoring algorithm for Chinese fill-in-the-blank questions, combining semantic lexicon matching and semantic similarity calculation to build an automatic scoring framework, using the improved P-means model to generate standard answers and sentence vectors and calculate semantic similarity, and the experimental data show that the highest accuracy rate of the algorithm is 94.3% [13]. Yuan's research group proposed an automatic essay scoring system based on a linear regression machine learning algorithm, combining linguistic features in a multiple regression approach to complete essay evaluation and model performance analysis [14]. Xia et al. proposed an automatic essay scoring model based on a neural network structure, which includes a long- and short-term memory layer and an attention mechanism layer; using pre-trained generated word vectors to calculate the experimental data showed that the quadratic weighted kappa coefficients of this model outperformed the bidirectional long- and short-term memory model [15]. Saihanqiqige proposed a multi-model fusion algorithm based on word vectors in order to improve the accuracy of English proficiency assessment, combining the text representation method of word vector clustering and the word vector space model to complete the automatic scoring of English composition [16].

In summary, many researchers have designed many algorithms and models for semantic space and automatic scoring, but the accuracy of these algorithms and models has yet to be improved. Therefore, the study proposes an unsupervised semantic space-based English composition tangent model, which is expected to improve the accuracy of automatic English composition scoring and reduce teachers' workload.

3 Designing a model for English composition tangents in unsupervised semantic space

3.1 Unsupervised semantic space and tangent analysis algorithm design

The unsupervised semantic space, namely, the Hierarchical Topic Tree Hybrid Semantic Space (HTTHSS), is proposed for implementing topic representation and clustering in English composition. The hierarchical tree topic tree hybrid space is mainly composed of a ternary hierarchical theme plants (HTP) model, a distributed vector group of topic relationships, and a topic semantic space based on a knowledge base [17]. The monads in the HTP model are replaced by relational triads to achieve a thematic semantic representation of sentence semantics and structural components, and Figure 1 illustrates the triad hierarchical tree topic model.

In Figure 1, the hyperparameters controlling the probability of a new path are γ, the set of infinite paths in the restaurant process nested under the adjacent distance of the phrase is T, the assignment of the ith topic relational triad is $wp - c_i$, the probability distribution of the topics of the text-relational triad is θ_m, the hyperparameters of θ_m obeying the Dirichlet distribution are α, the potential topics are $z_{m,n}$, the relational triad from which the text is extracted is $\omega_{m,n}$, M represents the quantity of texts, N represents the quantity of relational triads, and the quantity of uncertainties in the number of topics is ∞, and the form (S, R, O) is used to represent the relational triad. The formula for defining the restaurant process in nesting under the improved word group adjacency distance was studied as shown in the following equation:

$$p(wp - c_i = j|D, \alpha) \propto \begin{cases} f(d_{ij}), & i \neq j \\ \alpha, & i = j, \end{cases} \tag{1}$$

where d_{ij} denotes the i and j topic relation triads' adjacency distances, the set of all topic relation triads in English composition is D, and the decay function is f. The current allocation of topic-relational triples is only affected by the adjacency distance between topic triples, and the decay function is introduced to regulate the relationship between distance and random distribution [18]. The study uses the Gipsy sampling algorithm to implement topic sampling in the relational triad hierarchical tree model, and the distribution conditions for each potential topic variable are shown in the following equation:

$$p(c_{m,l}^{\text{new}}|c_{-m,l}, w, \eta) \propto p(c_{m,l}^{\text{new}}|D, \alpha)p(w|z(c_{-m,l} \cup c_{m,l}^{\text{new}})_{m,n}, H_0), \tag{2}$$

where the new sampling path is $c_{m,l}^{\text{new}}$, the relational triad of potential topics is $z(c)_{m,n}$, the set of model hyperparameters is $\eta = \{D, \alpha, f, H_0\}$, the underlying distribution obeying the Dirichlet distribution is H_0, the

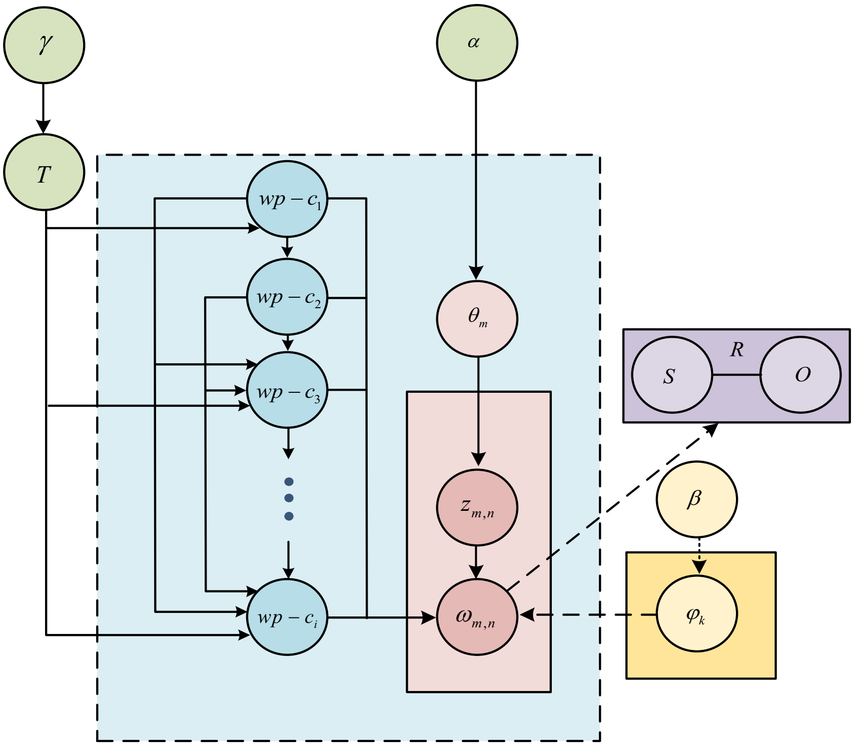


Figure 1: Relational triple hierarchical tree theme model.

relational triad of observed topics is w, the prior distribution is $p(c_{m,l}^{\text{new}}|D, \alpha)$, the probability of observed topic values is $p(w|z(c_{-m,l} \cup c_{m,l}^{\text{new}})_{m,n})$, and the number of topics is l. The relational triad hierarchical tree topic model converts the text semantic space from high-dimensional to low-dimensional, where each word in the relational triad (S, R, O) needs to be represented as a vector, whose N-dimensional distribution vector and calculation formula are shown in the following equation:

$$\begin{cases}\begin{cases}\mathrm{vec}(S) = [s_1, s_2, \cdots, s_n]\\ \mathrm{vec}(R) = [r_1, r_2, \cdots, r_n]\\ \mathrm{vec}(O) = [o_1, o_2, \cdots, o_n]\end{cases}\\ \mathrm{vec}(R - \mathrm{triad}) = \lambda_1\mathrm{vec}(S) + \lambda_2\mathrm{vec}(R) + \lambda_3\mathrm{vec}(O),\end{cases} \tag{3}$$

where the subject, relation, and object hyperparameters are λ_1, λ_2, and λ_3, respectively, and the sum of the three is 1. In this research, the set of relational triples is $\{T_1, T_2, \ldots, T_L\}$, the corresponding candidate triples are $K = \{K_1, K_2, \ldots, K_M\}$, and a new candidate topic set $K^{(\text{new})} = \{K_1, K_2, \ldots, K_N\}$ is iteratively generated. This study embeds distributed vectors and trains a HTP model with three parameters: topic smoothing, the predicted number of clusters of thematic relational triples, and the smoothed distribution of relational triples. The richness of semantics makes it possible to express the same meaning with different words, and the same word can correspond to different semantics. To solve the problems of lexical mismatch and linguistic ambiguity, the study uses feature dimensionality reduction to select the set of semantic features and complete the optimization of the feature semantic space. The feature selection algorithm is an important part of the feature reduction method, extracting the set of features that best reflect the model category and improving the efficiency of text classification. The study applies the local dimensionality reduction and global dimensionality reduction methods to the semantic space construction, and the mutual information algorithm is used in the tangent model to achieve topic coherence and viewpoint feature selection, and the mutual information value PMI is calculated as shown in the following equation:

$$\mathrm{PMI}(\omega_i) = \sum_{j}^{N-1} \log \frac{P(\omega_i, \omega_j)}{P(\omega_i)P(\omega_j)}, \tag{4}$$

where the list of the first N words in the topic list is ω, the first words in the topic list is $i\omega_i$, the first j words in the topic list is ω_j, the probability that the word ω_i and the word ω_j occur together is $P(\omega_i, \omega_j)$, the probability that the word occurs is $\omega_i P(\omega_i)$, and the probability that the word ω_j occurs is $P(\omega_j)$. The study used the distributed vector of relational triads as the premise of the tangent analysis and analysed the semantic similarity between sentences and composition topic semantics, sentences and composition paragraph topic semantics, paragraph and topic topic topic semantics, and full text and topic topic topic semantics, and the flow of the tangent analysis algorithm is shown in Figure 2.

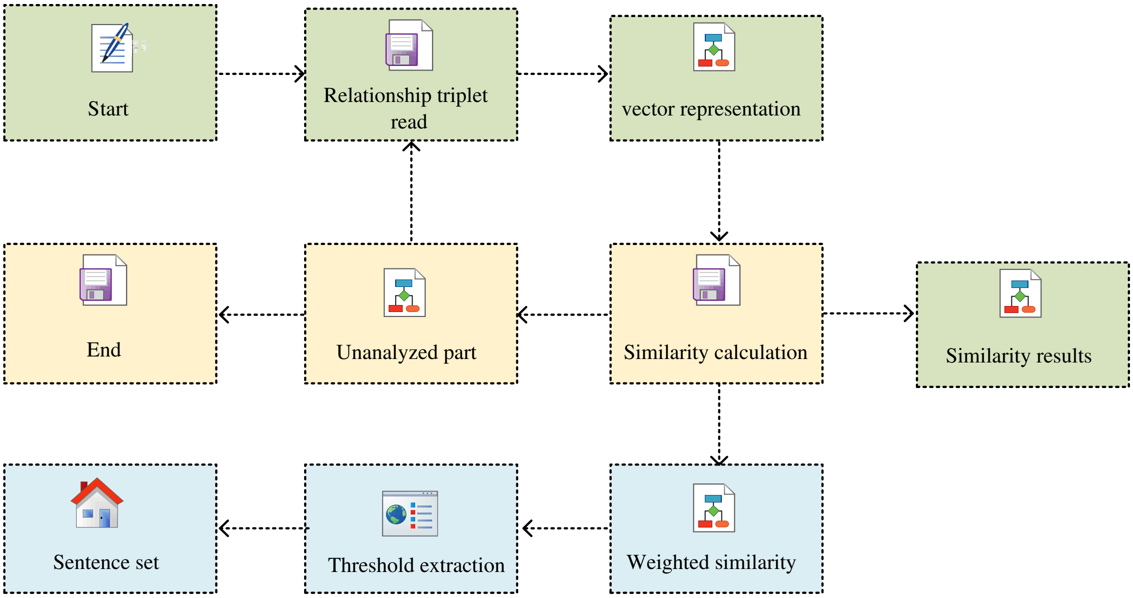


Figure 2: Process of relevance analysis algorithm.

The HTTHSS represents the English composition topics as a topic relation triplet distributed vector. Let the relation triplet contain i topic relations, and each topic distribution triplet distributed vector is shown in the following equation:

$$T_{(S_i,R_i,O_i)} = [\lambda_1 s_{1,i} + \lambda_2 r_{1,i} + \lambda_3 o_{1,i}, \lambda_1 s_{2,i} + \lambda_2 r_{2,i} + \lambda_3 o_{2,i}, \ldots, \lambda_1 s_{n,i} + \lambda_2 r_{n,i} + \lambda_3 o_{n,i}], \tag{5}$$

where the hyperparameters are λ_1, λ_2, and λ_3, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$, the distributed vector of English composition topics T_{title}, S denotes the sentence topic relations triplet's distributed vector, the distributed vector of paragraph and full text topic relations triplet P, and the distributed vector of C for the full text of English composition are shown in the following equation:

$$\begin{cases} T_{\mathrm{title}} = T_{(S_1,R_1,O_1)} + T_{(S_2,R_2,O_2)} + \ldots + T_{(S_i,R_i,O_i)} \\ S = T_{(S_1,R_1,O_1)} + T_{(S_2,R_2,O_2)} + \ldots + T_{(S_j,R_j,O_j)} \\ P = T_{(S_1,R_1,O_1)} + T_{(S_2,R_2,O_2)} + \ldots + T_{(S_k,R_k,O_k)} \\ C = T_{(S_1,R_1,O_1)} + T_{(S_2,R_2,O_2)} + \ldots + T_{(S_m,R_m,O_m)}, \end{cases} \tag{6}$$

where i, j, k, and m represent the dimensions of the four relational triadic distributed vectors, and then, the semantic similarity between English composition sentences and composition topics, sentences and composition paragraphs, paragraphs and topics, and full text and topics is calculated as shown in the following equation:

$$\begin{cases} \cos\theta_{S-T} = \dfrac{\sum_{i=1}^{n}(s_i - \mu_s) \times (t_i - \mu_t)}{\sqrt{\sum_{i=1}^{n}(s_i - \mu_s)^2} \times \sqrt{\sum_{i=1}^{n}(t_i - \mu_t)^2}} \\ \cos\theta_{S-P} = \dfrac{\sum_{i=1}^{n}(s_i - \mu_s) \times (p_i - \mu_p)}{\sqrt{\sum_{i=1}^{n}(s_i - \mu_s)^2} \times \sqrt{\sum_{i=1}^{n}(p_i - \mu_p)^2}} \\ \cos\theta_{P-T} = \dfrac{\sum_{i=1}^{n}(p_i - \mu_p) \times (t_i - \mu_t)}{\sqrt{\sum_{i=1}^{n}(p_i - \mu_p)^2} \times \sqrt{\sum_{i=1}^{n}(t_i - \mu_t)^2}} \\ \cos\theta_{C-T} = \dfrac{\sum_{i=1}^{n}(c_i - \mu_c) \times (t_i - \mu_t)}{\sqrt{\sum_{i=1}^{n}(c_i - \mu_c)^2} \times \sqrt{\sum_{i=1}^{n}(t_i - \mu_t)^2}}, \end{cases} \tag{7}$$

where the vector dimension is, the sentence distributed vector mean is, the topic distributed vector mean is $n\mu_s\mu_t$, the paragraph distributed vector mean is μ_p, and the full-text distributed vector mean is μ_c. δ_1 represents the semantic likeness between the sentence and the text. δ_2 denotes the semantic likeness between the sentence and the essay. The final sentence tangent similarity was calculated as shown in the following equation:

$$\cos\theta_{\text{in-topic}} = \delta_1 \cos\theta_{S-T} + \delta_2 \cos\theta_{S-P}, \tag{8}$$

where the sentence cut semantic similarity is $\cos\theta_{\text{in-topic}}$, the cut semantic similarity is ranked, and the cut sentence extraction threshold is set.

3.2 Design of English composition tangent model in unsupervised semantic space

The English composition cut model in unsupervised semantic space generates cut analysis results through English composition pre-processing, semantic space establishment, English composition cut analysis, and English composition cut quality analysis, and the processing flow of the English composition cut model in unsupervised semantic space as presented in the figure in the articles [19].

The basic link of text analysis is pre-processing, which has an important role in the semantic analysis later on. Using natural language processing tools, special character filtering, segmentation, sentence division, word division, and information extraction are completed [19]. The special characters and Chinese characters that appear in the English writing process can affect the segmentation, sentence, and word division of English compositions. The study establishes a special character set for English composition writing, combines regular expressions to filter special characters, and then slices the filtered English compositions in a global to partial manner to obtain sliced paragraphs, sentences, and words. The deactivated word set is then used to delete words in the sentences that do not affect the semantics, and then, the words in the composition are converted to their initial state for subsequent processing operations.

Words with the same grammatical properties belong to the same lexical property. There are 10 types of English lexical properties, and 2 special lexical properties are transitive and intransitive. The lexical property of each word needs to be labelled to facilitate the subsequent building of a dependency syntactic tree [20]. The study uses a recurrent dependency neural network lexical annotator, which obtains lexical labels through symmetric inference, relying on the existence of correlations between each node in the neural network and neighbouring nodes, and obtains conditional similarity maximization of node training data through local model training, to prepare for English text processing and analysis. The study injects a large number of rare word set features into the lexical annotator to achieve correct annotation of unknown words.

Information extraction is a text processing technique that extracts factual information such as entities, relations, and events of specified types from natural language text and forms structured data output. The relational triad form contains subjects, relations, and objects. The relationships between words in a sentence are realized through dependency syntactic analysis, and the dependency syntactic tree expression is shown in the following equation:

$$f = \{(s, r, o) : 0 \le s \le m, 1 \le o \le m, r \in C\}, \tag{9}$$

where the textual dependency arc is (s, r, o), the dependency type of the core word and the modifier is r, and the set of dependency types is C. There is a wide variety of dependencies between words, and the results of the dependency syntactic analysis are shown in Figure 3.

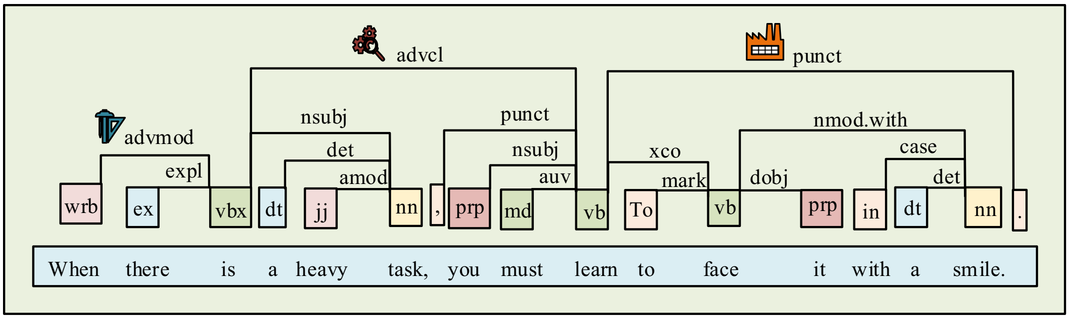


Figure 3: Dependency parsing results.

The study comprehensively analyses the quality of English composition tangency by calculating the semantics of English composition sentences and composition topics, the semantics of sentences and composition paragraphs, the semantics of paragraphs and topic topics, and the semantic similarity of full text and topic topics, and analyses the quality of composition topics and opinion expressions by combining the topic coherence algorithm and the topic opinion algorithm. The formula for calculating the cut score of English composition is shown in the following equation:

$$G_{\text{inTopic}} = \gamma_1 \frac{\sum_{i=1}^{N} \cos \theta_{S-T}}{N} + \gamma_2 \frac{\sum_{i=1}^{N} \cos \theta_{S-P}}{N} + \gamma_3 \frac{\sum_{j=1}^{M} \cos \theta_{P-T}}{N} + \gamma_4 \cos \theta_{C-T}, \tag{10}$$

where the hyperparameter is γ and $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$, the score of English composition tangential analysis is G_{inTopic}, the sum of the semantic similarity of sentences and paragraphs in N and $\sum_{i=1}^{N} \cos \theta_{S-T}$ and $\sum_{i=1}^{N} \cos \theta_{S-P}$, respectively, the sum of the semantic similarity of paragraphs and topics in M and $\sum_{j=1}^{M} \cos \theta_{P-T}$, and the semantic similarity of full English composition and topics in $\cos \theta_{C-T}$. The formula for calculating the score of topic coherence is shown in the following equation:

$$G_{\text{TopicCoherence}} = \varepsilon_1 \frac{\sum_{i=1}^{N} \text{TPMI}(s_{\text{center}})}{N} + \varepsilon_2 \frac{\sum_{i=1}^{2M-2} \text{TPMI}(p_{\text{center}-p})}{2M-2} + \varepsilon_3 \frac{\sum_{i=1}^{M} \text{TPMI}(p_{\text{center}-c})}{M}, \tag{11}$$

where the thematic coherence score of the English composition is $G_{\text{TopicCoherence}}$, the hyperparameter is ε and $\varepsilon_1 + \varepsilon_2 + \varepsilon_3 = 1$, the sum of the thematic coherence of the sentences of N and the paragraphs to which they belong is $\sum_{i=1}^{N} \text{TPMI}(s_{\text{center}})$, the sum of the thematic coherence of the paragraphs of M and the context, and the sum of the thematic coherence of the whole text is $\sum_{i=1}^{2M-2} \text{TPMI}(p_{\text{center}-p})$ and $\sum_{i=1}^{M} \text{TPMI}(p_{\text{center}-c})$, respectively. The formula for calculating the thematic opinion score of the English composition is shown in the following equation:

$$G_{\text{TopicSen}} = \eta_1 \frac{\sum_{j=1}^{M^2-M} \cos \theta_{\text{Senti}-p_1 p_2}}{M^2 - M} + \eta_2 \frac{\sum_{j=1}^{M} \cos \theta_{\text{Senti-pc}}}{M}, \tag{12}$$

where the topic perspective score of English composition is G_{TopicSen}, the hyperparameter is η and $\eta_1 + \eta_2 = 1$, and the sum of the semantic correlation of affective tendency between the paragraphs of English composition M and the paragraphs, and the sum of the semantic correlation of affective tendency between English composition and the whole text are $\sum_{j=1}^{M^2-M} \cos \theta_{\text{Senti}-p_1 p_2}$ and $\sum_{j=1}^{M} \cos \theta_{\text{Senti}-pc}$, respectively. The English essay quality score was obtained by weighting the scores of English essay tangency, coherence, and thematic viewpoints as shown in the following equation:

$$G = (\rho_1 G_{\text{inTopic}} + \rho_2 G_{\text{TopicCoherence}} + \rho_3 G_{\text{TopicSen}}) \times 100, \tag{13}$$

where the English composition cut quality score is G, the cut quality score hyperparameter is ρ, and $\rho_1 + \rho_2 + \rho_3 = 1$. The generation process of an English essay topic cutting model in an unsupervised semantic space is a complex process that involves multiple steps, including data preprocessing, feature extraction, model training, and prediction. In the data preprocessing stage, it is necessary to clean and standardize the original data to ensure the quality and consistency of the data. Next, in the feature extraction stage, meaningful feature vectors need to be extracted from the original data, which should be able to capture important information in the text data. Then, during the model training phase, machine learning algorithms can be used to classify feature vectors. Finally, in the prediction phase, the trained model is applied to new text data to predict the category or label of the text.

4 Experimentation and analysis of English composition tangent model in unsupervised semantic space

4.1 Experimental test sets and assessment criteria

The relational triad HTP model's parameters were optimized by the Wikipedia corpus, The International Corpus Network of Asian Learners of English (ICNALE), and the Chinese learner English corpus. The HTP model's parameters were optimized by setting the topic smoothing distribution to 10, the predictive clustering parameter of the topic-relational triad to 1, and the relational triad smoothing distribution to 0.1. The test set data are shown in Figure 4.

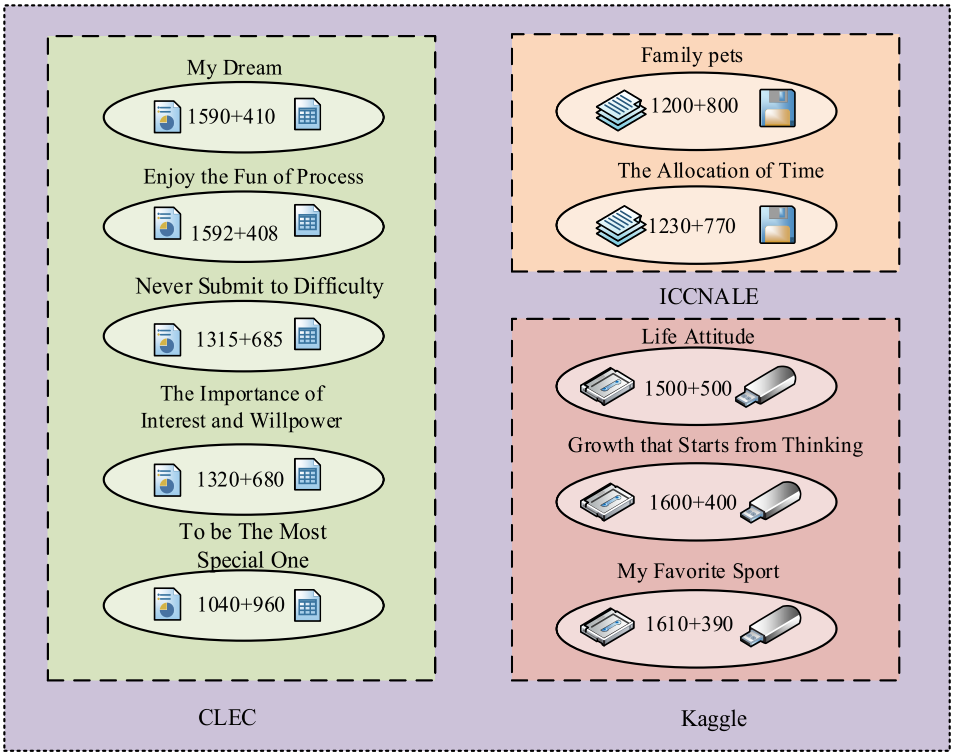


Figure 4: Test set data.

A corresponding number of run-on essays were added to each topic in the test set. The sentence test set was a random selection of 16,010 sentences from 1,000 essays under two essay topics from ICNALE, including 6,930 tangential sentences and including 5,200 thematically incoherent sentences. The number of samples with positive model and manual assessment is A, the number of samples with positive model but inconsistent manual assessment is B, the number of samples with negative model but inconsistent manual assessment is C, and the number of samples with negative model and manual assessment is D. The accuracy rate of model is P, the higher the value the better the model measurement effect; the recall rate is R, the higher the value the higher the model checking rate; the comprehensive evaluation index is F, the higher the value the better the model classification effect. The experiment referred to the scoring standard of English 4/6 level essay to develop the evaluation criteria of English essay tangency, as shown in Table 1.

Table 1: Evaluation criteria for the degree of relevance in English compositions

Score	Evaluation criterion
90–100	Content to the point; good thematic coherence Express clearly and fluently
80–90	Content to the point; good thematic coherence; express opinions clearly Minor language errors
70–80	Content is basically relevant to the topic; theme is generally coherent Inadequate expression of thematic viewpoints
60–70	Content is basically relevant to the topic; poor thematic coherence and expression of thematic viewpoints
0–60	Content is not well organized; poor thematic coherence; chaotic expression of thematic viewpoints

The experiment reflects the relevance of the two by evaluating the degree of tangibility of the English composition through the model and calculating the correlation index Pearson's correlation coefficient by combining the results of the teachers' evaluation. The Pearson correlation coefficient is $r_{x,z}$, and its value ranges from [−1,1], with a higher value indicating a stronger correlation between the two samples. The i sample in the teacher manual assessment sample set is xx_i, the i sample in the model assessment sample set z is z_i, and the sum of the sample trees in the sets x and z is n.

4.2 Analysis of model experimental data

The experimental test set consisted of 22,000 English essays under ten topics in the corpus, of which 15,000 were tangential essays as well as 7,000 run-on essays. Table 2 indicates the experimental results of the research that proposed a semantic space-based approach to the analysis of tangential English essays.

Table 2: Experimental results of relevance analysis method

Composition title	*P* (%)	*R* (%)	*F* (%)
My dream	94.01	87.21	90.48
Enjoy the fun of process	94.86	86.97	90.74
Never submit to difficulty	93.42	88.31	90.79
The importance of interest and willpower	94.56	89.01	91.70
To be the most special one	94.89	88.78	91.73
Family pets	92.18	88.12	90.10
The allocation of time	94.78	87.03	90.74
Life attitude	89.96	87.54	88.73
Growth that starts from thinking	90.36	89.82	90.09
My favourite sport	90.93	87.56	89.21

Analysis of the experimental data in Table 2 shows that the highest value of accuracy of the semantic space-based English composition tangent analysis algorithm is 94.86%, the average accuracy is 93.00%, the highest value of recall is 89.82%, the average recall is 88.04%, the highest *F*-value is 91.73%, and the average *F*-value is 90.43%. Due to the study's extended topic set, the effectiveness of the tangential analysis method was stable under different lengths of English composition topics. The experiments were conducted to verify the effectiveness of the analysis of the hybrid semantic space of HTP (marked as D) by Word2Vec + topic hierarchical tree semantic space (marked as A), Word2Vec + improved topic hierarchical tree semantic space (marked as B), and Word2Vec + topic hierarchical tree + knowledge base semantic space (marked as C) as comparisons, and the experimental results are shown in Figure 5.

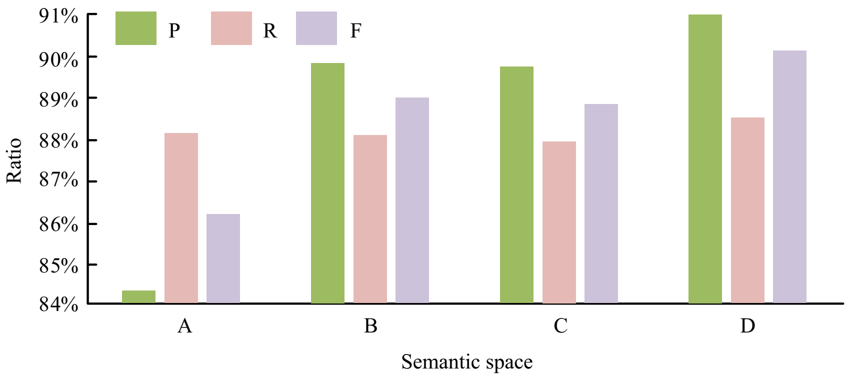


Figure 5: Experimental results of semantic space relevance analysis.

As can be seen in Figure 5, the recall rates of the four semantic spaces for cut-topic analysis do not differ much, all remaining around 88%, but the recall rate under the HTTHSS is slightly higher at 88.53%. The accuracy rate of English tangent analysis in the HTTHSS was 91.65% with an *F*-value of 90.08%, indicating that the experiment was effective. The experiments compared the WEDVRM (marked as A), the LDA +

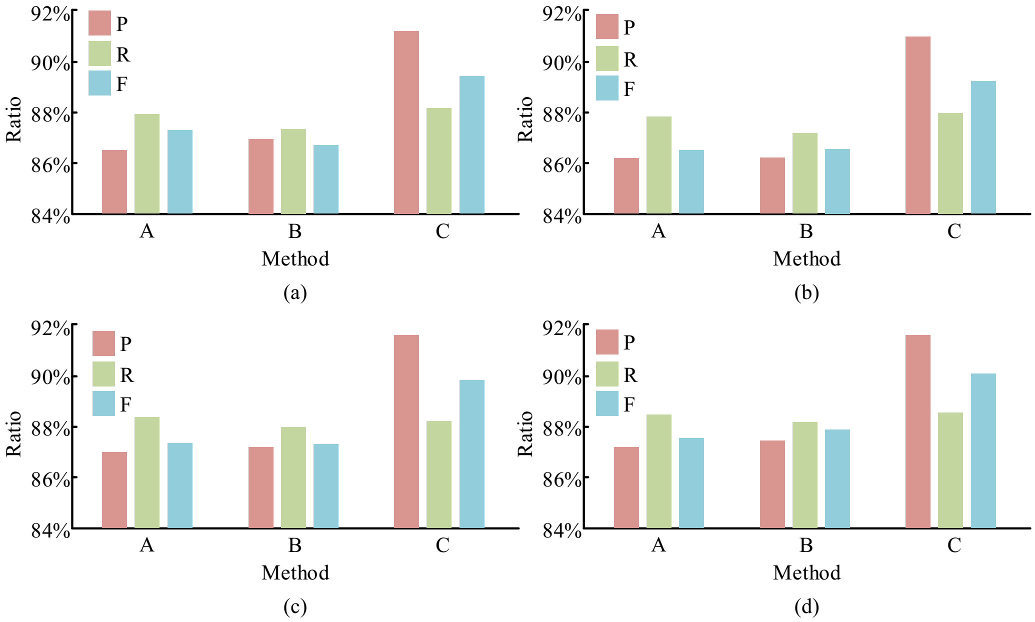
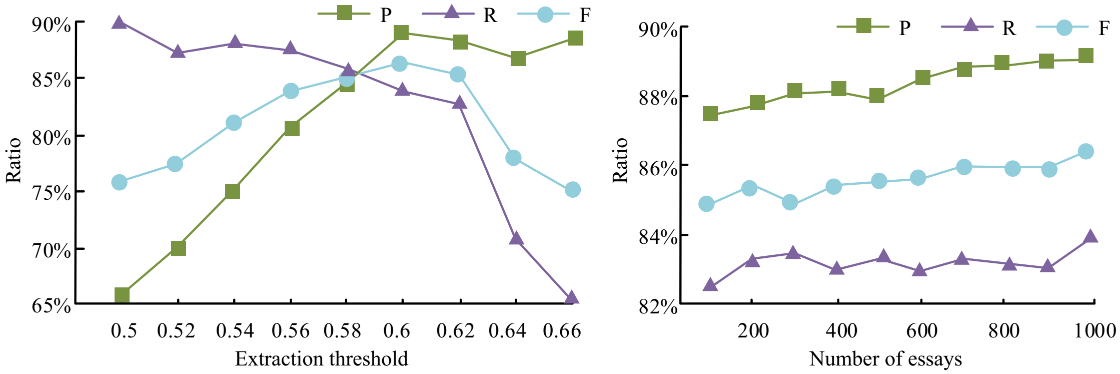


Figure 6: Comparison of experimental results of different relevance analysis methods. (a) 5,000 essays, (b) 10,000 essays, (c) 15,000 essays, and (d) 20,000 essays.

WEDVRM (marked as B) with the research proposed semantic space-based English composition tangent analysis algorithm (marked as C) for 5,000, 10,000, 15,000, and 20,000 compositions in the test set, respectively, and the experimental results are shown in Figure 6.

As can be seen from Figure 6, the three indicators corresponding to the three algorithms increased as the number of compositions increased, indicating that the more English compositions were analysed, the more stable the performance of the tangential analysis algorithm became. The accuracy and *F*-value of the semantic space-based English composition tangent analysis algorithm were significantly higher than those of the Word Embedding Distributed Vector Representation Method (WEDVRM) and the LDA + WEDVRM, with an accuracy and *F*-value of 91.65 and 90.12%, respectively, when 20,000 English compositions were analysed, increasing the accuracy and *F*-value of the tangent analysis algorithm by 5 and 3%, respectively. The recall rates for the three algorithms did not differ significantly, all remaining at around 88%. The tangent analysis model analysed 16,010 sentences (which included 6,930 marked tangent sentences) from 1,000 essays, with different tangent thresholds set in the tangent experiments, and the experimental results are shown in Figure 7.



(a) (a)

Figure 7: Experimental results of topic analysis models under different extraction thresholds for topic specific sentences. (a) Results of extracting relevant sentences and (b) the threshold is 0.6 for extracting results.

Figure 7(a) shows the experimental results of the tangent analysis model under different tangent sentence extraction thresholds. Analysis of the experimental data shows that the highest *F*-value of the tangent analysis algorithm is 86.67% when the threshold is 0.6, and the tangent sentence extraction effect is more satisfactory at this time. When the extraction threshold was low, the recall rate of tangent extraction was higher but the accuracy rate was lower; when the extraction threshold was high, the accuracy rate of tangent extraction was higher but the recall rate was lower, both of which could not complete the task of tangent extraction stably. Figure 7(b) shows the experimental results of the model's tangential sentence extraction of 1,000 articles when the tangential extraction threshold is set to 0.6. Analysis of the experimental data shows that when the quantity of English essays increases, the precision of the tangential analysis algorithm also increases, with the highest accuracy rate being 89.12% and the recall rate being basically stable, with the mean value of F being 86.67%. The model analysed 16,010 sentences (which included 5,200 marked thematically incoherent sentences) from 1,000 essays, and the experiments were also set to different thresholds, and the results are shown in Figure 8.

Analysis of the experimental data in Figure 8 shows that the highest *F*-value of 87.68% was achieved when the extraction threshold for thematically incoherent sentences was 0.32. When the quantity of English compositions increases, the precision of the tangential analysis algorithm also increases, with a maximum accuracy of 90.16% and a decrease in the recall rate. To verify the practical application of the English composition tangent model in unsupervised semantic space, the experiments used the professional English teachers' ratings as a comparison and scored 1,000 English compositions tangentially, respectively, and the experimental results are shown in Figure 9.

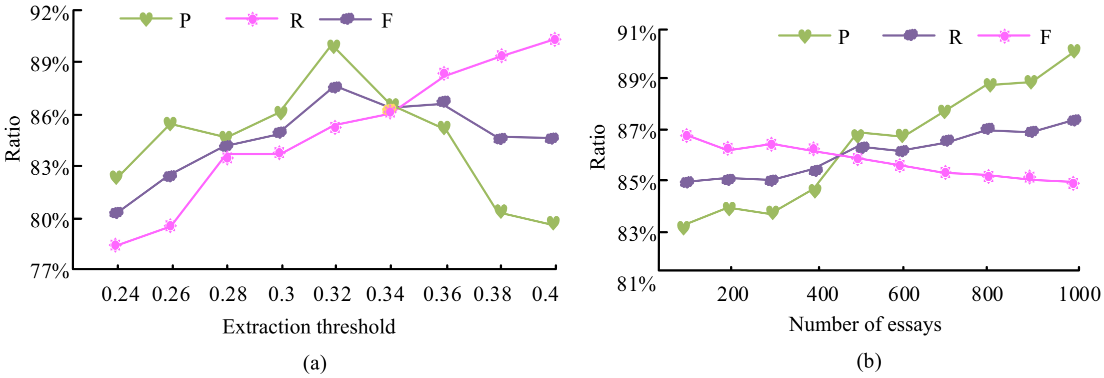


Figure 8: Analysing the results of disconnected sentences on the topic. (a) Incoherent sentence extraction result and (b) extracting incoherent sentences with a threshold of 0.32.

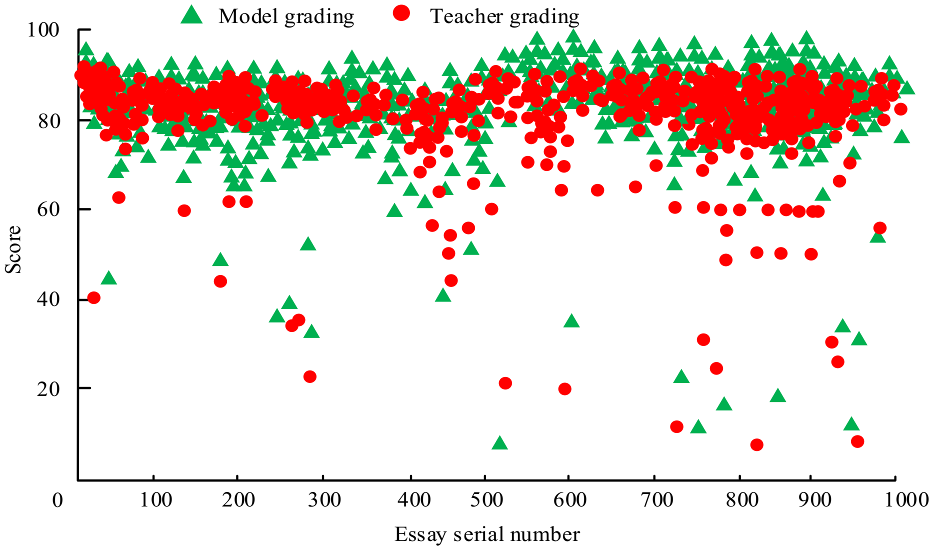


Figure 9: Comparison of model scoring and teacher scoring results.

In Figure 9, the results of the model scoring and teacher scoring are mainly concentrated between 80 and 90 points. Analysis of the experimental data shows that the mean value of the English composition tangent model scoring 1,000 English compositions under unsupervised semantic space is 83.56 points and the mean value of teacher scoring is 82.36 points, and the difference between the two scores as well as the Pearson correlation coefficient are 1.2 points and 0.8936 points, respectively, with a strong correlation grade. The experiment shows that the English composition tangent model under unsupervised semantic space has high credibility and practicality.

5 Conclusion

In the form of globalization, English occupies a major position as a common language, and various English examinations and competitions usually involve the writing of English essays, and the marking of essays is also a big project. To improve the efficiency of English composition correction, the study proposes an unsupervised semantic space-based English composition tangent model, combining a relational triad hierarchical tree topic model with a tangent analysis algorithm to achieve intelligent scoring of English compositions. The experimental data show that the accuracy and *F*-value of the semantic space-based English composition tangent analysis method are significantly higher than those of the WEDVRM and the LDA + WEDVRM, and the precision and *F*-value are 91.65 and 90.12%, respectively, when 20,000 English compositions are analysed: 86.67%, which is a good result for tangential sentence extraction. The highest *F*-value of 87.68% was obtained for a threshold of 0.32 for the extraction of thematically incoherent sentences. When the quantity of English compositions increases, the precision of the tangential analysis algorithm also increases, with a maximum accuracy of 90.16%. The mean score of the English composition tangent model in unsupervised semantic space for 1,000 English compositions was 83.56, and the mean score of the teacher's score was 82.36, and the difference between the two scores and the Pearson correlation coefficient were 1.2 and 0.8936, respectively. The results show that the English composition tangent model in unsupervised semantic space is more stable and applicable, and can accurately and efficiently complete the English composition criticism task. The study is a tangential analysis of the topics, sentences, and paragraphs of English composition, and the tangential analysis can be further improved subsequently by combining features such as chapter structure.

Funding information: The research was supported by National level, A Study on the construction and promotion of Jiaodong Culture into English Quality Courses in Higher Vocational Colleges (No. WYW2022A03).

Author contributions: Rihong Tang: methodology, conceptualization, Writing – Original Draft, Writing – Review & Editing.

Conflict of interest: The author reports there are no competing interests to declare.

Data availability statement: The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

References

[1] Cowen AS, Keltner D. Semantic space theory: A computational approach to emotion. Trends Cognit Sci. 2021;25(2):124–36.

[2] Sato N, Matsumoto R, Shimotake A, Matsuhashi M, Otani M, Kikuchi T, et al. Frequency-dependent cortical interactions during semantic processing: an electrocorticogram cross- spectrum analysis using a semantic space model. Cereb Cortex. 2021;31(9):4329–39.

[3] Huang GM, Zhang XW. An analysis model of potential topics in English essays based on semantic space. J Comput. 2022;33(1):151–64.

[4] Neumeyer L, Franco H, Digalakis V, Weintraub M. Automatic scoring of pronunciation quality. Speech Commun. 2000;30(2–3):83–93.

[5] Nimrah S, Saifullah S. Context-free word importance scores for attacking neural networks. J Comput Cognit Eng. 2022;1(4):187–92.

[6] Waziri TA, Yakasai BM. Assessment of some proposed replacement models involving moderate fix-up. J Comput Cognit Eng. 2023;2(1):28–37.

[7] Xiao W, Luo X, Xie S. Feature semantic space-based sim2real decision model. Appl Intell. 2023;53(3):4890–906.

[8] Kherwa P, Bansal P. Three level weight for latent semantic analysis: an efficient approach to find enhanced semantic themes. Int J Knowl Learn. 2023;16(1):56–72.

[9] Yu J, Jin L, Chen J, Xiao Y, Tian Z, Lan X. Deep semantic space guided multi-scale neural style transfer. Multimed Tools Appl. 2022;81(3):3915–38.

[10] Yu HF. Bibliographic automatic classification algorithm based on semantic space transformation. Multimed Tools Appl. 2020;79(13):9283–97.

[11] Orhan U, Tulu CN. A novel embedding approach to learn word vectors by weighting semantic relations: semspace. Expert Syst Appl. 2021;180:115146–53.

[12] Zhao Y. Research and design of automatic scoring algorithm for english composition based on machine learning. Sci Program. 2021;3429463–72.

[13] Wang H, Zhao Y, Lin H, Zuo X. Automatic scoring of Chinese fill-in-the-blank questions based on improved P-means. J Intell Fuzzy Syst. 2021;40(3):5473–82.

[14] Yuan Z. Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm. J Intell & Fuzzy Syst. 2021;40(2):2069–81.

[15] Xia L, Luo D, Liu J, Guan M, Zhang Z, Gong A. Attention-based two-layer long short-term memory model for automatic essay scoring. J Shenzhen Univ Sci Eng. 2021;37(6):559–66.

[16] Saihanqiqige HE. Application research of english scoring based on TF-IDF clustering algorithm. IOP Conf Ser: Mater Sci Eng. 2020;750(1):12215–301.

[17] Lewis M, Marsden D, Sadrzadeh M. Semantic spaces at the intersection of NLP, physics, and cognitive science. FLAP. 2020;7(5):677–82.

[18] Shi L, Du J, Liang M, Kuo F. Dynamic topic modeling via self-aggregation for short text streams. Peer-to-Peer Netw Appl. 2019;12(1):1403–17.

[19] Kou F, Du J, Lin Z, Liang M, Li H, Shi L, et al. A semantic modeling method for social network short text based on spatial and temporal characteristics. J Comput Sci. 2018;28(1):281–93.

[20] Shi L, Song G, Cheng G, Liu X. A user-based aggregation topic model for understanding user's preference and intention in social network. Neurocomputing. 2020;413(1):1–13.