Journal of Computers Vol. 33 No. 1, February 2022, pp. 151-164
doi: 10.53106/199115992022023301014

An Analysis Model of Potential Topics in English Essays Based on Semantic Space

Guimin Huang, Xiaowei Zhang*

Guangxi Key Laboratory of Image and Graphic Intelligent Processing,
School of Computer Science and Information Security,
Guilin University of Electronic Technology Guilin 541004, China
sendhuang@126.com, zhang_xiaowei_2021@163.com
Received 18 May 2021; Revised 18 August 2021; Accepted 8 September 2021

Abstract. With the reform of English examination in China in recent years, the automatic evaluation of English essays as subjective questions has always been the focus and difficulty of research. The existing automatic essay evaluation system (AEE) has obtained good feedback on the vocabulary, syntactic and other features of English essays, but there is still a problem of low accuracy in the analysis of potential topic in English essay. In order to solve this problem, this paper takes relational triples as the carrier to analyze the potential topics of English essays. By constructing the hierarchical topic trees hybrid semantic spaces to carry out topic clustering, distributed representation of topic relational triples and topic set extension in English essays. Then, based on the improved on-topic analysis algorithm in this paper, the paper analyzes the topic of English essay in multiple dimensions to obtain more abundant potential on-topic semantic information. The experiment results show that the proposed model can reduce the noise caused by non-topic words effectively, and improve the fine-grained topic semantic space in English essays, and the proposed model has better performance than the current methods of on-topic analysis in English esssays.

Keywords: English essays, topic analysis, semantic spaces, hLDA, WP-nCRP

1 Introduction

In recent years, with the rapid development of artificial intelligence and natural language processing technology, the teaching field is becoming more and more technical. The automatic essay evaluation system (AEE) uses natural language processing, machine learning and other related technologies to achieve intelligent scoring and correction of English essays [1]. AEE can not only reduce the burden and pressure brought by a large number of English essays to English teachers, but also objectively and scientifically score English essays, increase the credibility and efficiency of the correction and feedback of essays, and provide more perfect results for English writing learners [2]. For second language learners, essay content is the most important indicator to evaluate essays. The results of relevant experimental studies show that the three characteristic variables of the article's topicality, topic coherence and opinion tendency account for 56% of the differences in the comprehensive quality of Chinese students' English essay content [3]. Therefore, on-topic analysis of essays plays an important role in the accuracy and robustness of automatic essays grading.

At present, AEE is poor in terms of composition content quality, structure fluency, for example, it focuses too much on the form and display semantic features of the essay while neglecting important potential semantic features such as topic distribution and essay coherence [4]. This increases the probability that students will mislead the machine by writing deceptive articles with incoherent topics and unfluent ideas [5]. For example, researchers have successively launched Bingguo English intelligent essay evaluation system and Kuju correction website and other automatic scoring systems. These systems are mainly based on superficial text features such as spelling and grammar, but lack the ability to analyze the deep sentence structure and content, which makes AEE prone to misjudge high scores [6].

The core problem of on-topic essay detection is to judge whether the topic of the essay deviates from the target prompt. Previous studies mainly focused on the improvement of semantic similarity between articles and target prompt. For example, Li X [7] et al. implemented an unsupervised off-topic essay detection system. The model use the semantic difference between the similarities of the essay with the target prompt and that of with the reference prompts to on-topic score calculation, which is used to better distinguish the on-topic essays and the off-topic essays. However, when the target prompt is short, then the semantic information between the essay and the target prompt will be less adequate. Meng C [8] et al. proposed the LDA coupling space model, and extracted the topic words of text to be approved by using the LDA and expressed the topic words with Word2vec. The experi-

* Corresponding Author

ment showed that the analysis accuracy of subjects with high divergence was effectively improved, but the model has the problem that the topic semantic space is not fine-grained enough. In fact, we find that by improving the hierarchical topic tree model, the relational triples are introduced to replace the unary phrases in HLDA, and the triples are given higher weight to show the topic features, so as to obtain more topic information. then, the original nested Chinese restaurant process (nCRP) is improved to a nested Chinese restaurant process based on phrase adjacent distance, which can effectively improve the accuracy of topic clustering.

Therefore, aiming at the existing problems of low accuracy in the topic analysis of English essays, this paper proposes an unsupervised topic analysis model based on semantic space. The improved hierarchical topic tree model is combined with distributed vector, and the knowledge base. And combined with the improved on-topic analysis algorithm in this paper, a detailed on-topic analysis is carried out on the content and structure of English essays, extract the on-topic sentences of English essays. Experimental results show that the proposed model effectively improves the accuracy of English essays topic clustering and enriches the topic semantic space of distributed vector representation.

The main contributions of this paper are as follows:

(1) Based on the nested Chinese restaurant process algorithm of phrase distance, a hierarchical topic tree model of relational triples was constructed, and the model clustering topic generation relational triples distributed vector representation.

(2) By matching the semantic concepts in the knowledge base, the optimal candidate topic set is analyzed to expand the topic semantic space. A large number of corpora are used to train the model to obtain better distributed vector representation and clustering.

(3) Based on the on-topic analysis algorithm proposed in this paper, the relationship triples in English essays are used as the carrier to analyze the potential topic features of English essays from the sentence level to the text level, and extract the relevant sentences. Finally, several experiments are carried out on different English essays test sets. The experimental results show that the proposed model has higher accuracy and practical value than the current unsupervised approach.

The rest of this paper follows. The second part introduces the relevant work of the topic analysis of English essays. In the third part, the hierarchical topic trees hybrid semantic spaces and on-topic analysis algorithm is introduced in detail. In the fourth part, the experimental results are compared and analyzed. The fifth part summarizes the work of this paper and looks forward to the next step.

2 Related Work

In recent years, topic analysis of English essays has been widely concerned by the industry and academia. Writing to the point means to express the content of ideas within the scope prescribed by the topic. At present, the topic analysis methods of English essays at home and abroad are mainly divided into supervised and unsupervised methods.

Supervised English essays on-topic analysis methods need to refer to model essays or manually annotated data, and are not suitable for correcting a large number of English essays.For example, Liu L [9] et al. proposed to use various corpora to train the word vector model, and combined with WordNet to narrow the semantic space distance of the essays.Chen Z [10] et al. proposed the method of essay divergence to set up the dynamic threshold of similarity between the essays to be approved and the model essays, and tested the on-topic degree of the essay by building a correlation regression model.

Among the typical unsupervised topic analysis methods, Yang Z [11] et al. proposed to combine the on-topic detection model of neural network, use the combination of neural network features and text dominant features as the representation form of essays to be approved, and constructed a neural network structure containing three convolutional layers and the maximum polling layer. However, the model lacks topic level information, which has a certain impact on the accuracy of topic clustering. Huang G [12] et al. proposed an off-topic essay detection model by calculating the similarity value between the essay and the essay prompt in a hybrid semantic space. This method relies on noun phrases extracted from essays and essay prompt, and lacks the analysis of potential topics in English essays. Chang Y [13] et al. proposed a topic detection based on semantic framework (SFTD) to simulate this process in human perception. This method can effectively use syntactic structure, semantic association and context to detect the topic of a document. In order to explore the combination of topic modeling and bidirectional LSTM, Wang g [14] et al. proposed a new probabilistic topic model GPU-LDA-LSTM, which uses LSTM network to improve context consistency in parameter reasoning. However, the model lacks the concept of topic level, which reduces the accuracy of topic clustering.

Furthermore, in order to avoid the interference of non-topic words and increase the analysis of the relevance of context topic, the researchers proposed to add the distributed representation method to the topic model. Chung S [15] et al. proposed to combine PTE heterogeneous network with distributed representation, in which the construction of the network uses the probability graph representation of the topic model. Li C [16] et al. used distributed word representations to improve the sampling process of the topic model and enhance model performance in terms of effectiveness and efficiency. Qu Q [17] et al. proposed an off-topic detection algorithm combining LDA and word2vec. The algorithm realizes the intelligent processing of off-topic detection. The input of the distributed semantic space based on neural network is the word and context of the document. However, the sparse words and polysemous words in the document lead to defects in the expression of the semantic space. Knowledge base contains rich semantic relations and structured knowledge, which can better improve the quality of word vector representation. The semantic features of integrating external semantic knowledge base have been proposed. Speer R [18] et al. proposed that ConceptNet5.5 combines specific knowledge and external resources with word embedding model to improve the expression of word vectors in common sense semantics and other recessive semantics. Shalaby W [19] et al. proposed to use the combination of word vector embedding model and probability concept map to represent the concept knowledge of two large knowledge bases (Wikipedia and Probase) with different structures. Derby S [20] et al. proposed that Feature2vec combines attribute knowledge with word vector and represents it in semantic space, which can better predict attribute conceptual features.

This paper proposes a hierarchical topic tree hybrid semantic space for English composition. By improving the existing hierarchical topic tree model and integrating the distributed representation and knowledge base, we can obtain better distributed vector representation and topic clustering. In addition, We improve the existing on-topic analysis algorithm from sentence level to discourse level to analyze the potential topic features of English compositions and extract the topic related sentences. The experimental results show that the model can not only meet the needs of teachers and students for topic analysis of English essay, but also improve the reliability and validity of the writing system.

3 Methodology

The hierarchical topic trees hybrid semantic spaces is composed of three parts: the relational triple hierarchical topic tree model, the distributed vector representation of topic relational triples, and the extension of topic semantic space based on Knowledge Base. Each part is described in detail below:

3.1 Relational Triple Hierarchical Topic Tree Model

In this paper, we propose to use relational triples instead of unary phrases in the hLDA(hierarchical Latent Dirichlet Allocation) [21], and give higher weight to the nouns in the triples to show the topic features, so as to obtain more topic information.Secondly, since adjacent phrases in English essays are more likely to belong to the same topic, Therefore, the nested Chinese restaurant process (nCRP), which divides and randomly distributes the essays according to the probability size of the word previously belongs to, is improved to the nested word proximity Chinese restaurant process(WP-nCRP).The relational triple hierarchical topic tree model is shown in Fig.1.

We define θ_m to represent the probability distribution of textual - relational triple topics, ϕ_k to represent the probability distribution of relational triple topics-relational triple. α is θ_m a hyperparameter subject to dirichlet distribution. β is ϕ_k a hyperparameter subject to dirichlet distribution. L Represents the level of hierarchical topic tree structure based on the adjacent distance of phrases, γ is the super parameter controlling the probability to create a new path, T represents the infinite set of l-level paths created from the process of the nested word proximity Chinese restaurant process. Potential topic $Z_{m.n}$ represents the topic allocation of relational triples, $W_{m,n}$ represents the relational triples of topic extraction, M represents the total number of essays, N represents the total number of relational triples, ∞ represents the uncertain number of topic numbers, and the relational triad graph is represented in the form of (S,R,O).

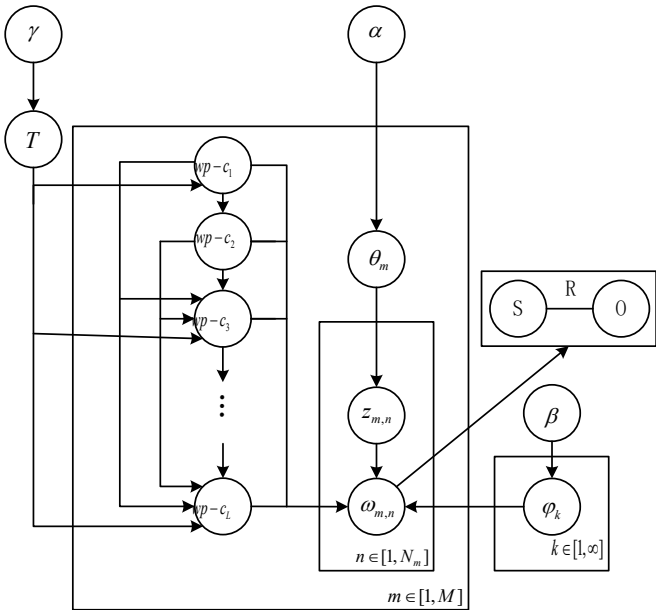

Fig. 1. Relational triple hierarchical topic tree model

3.1.1 Nested Word Proximity Chinese Restaurant Process (WP-nCRP)

The definition of the traditional Chinese restaurant process can be simplified into Equation (1), where it is assumed that words 1 to i-1 ($Z_{1:i-1}$) have occupied K tables:

$$p(z_i = k \mid z_{1:(i-1)}, \alpha) \propto \begin{cases} n_k, & k \le K \\ \alpha, & k = K+1 \end{cases} . \tag{1}$$

Where z_i is the random distribution of the i-th word table, n_k is the number of words assigned to the k-th table, and α is the given parameter. It can be seen from the formula that the specific partition probability of a word is determined by the number of words in the first k tables.

In this paper, we propose the nested word proximity Chinese restaurant process. The probability between phrases is determined by the phrase distance in the semantic space. The definition formula is as follows:

$$p(wp - c_i = j \mid D, \alpha) \propto \begin{cases} f(d_{ij}), & i \ne j \\ \alpha, & i = j \end{cases} \tag{2}$$

In Formula (2), wp-c_i represents the assignment of the i topic relation triplet, d_{ij} is the adjacent distance between the i-th topic relation triplet and the j-th topic relation triplet. D is the set of adjacent distance values of all topic relations in English essays, and f is the attenuation function. From the definition, we can see that the current topic relationship triplet assignment does not depend on the probability assignment with other topic relationship triples, but only depends on the adjacent distance between topic relationship triples, and the range of j includes all the relationship triples in the whole essays. d_{ij} is the semantic distance. We use the cosine similarity algorithm to calculate the relative distance of topic relation triples in space. In the formula, we introduce the attenuation function to adjust the relationship between the distance of the relational triplet and the random distribution in the partition. Based on the nested word proximity Chinese restaurant process, the Gipps sampling pseudo code Algorithm 1 is as follows:

Algorithm 1. Based on WP-nCRP, the Gipps sampling pseudo code algorithm

INPUT: N: the number of all topic related triples in English essays, T: total number of samples, α: Dirichlet parameter, b: the number of discarded samples after sampling

OUTPUT: $wp - c$: distribution of topic words after sampling

BEGIN

$wp-c_1^{(0)}=1, wp-c_2^{(0)}=1,...,wp-c_N^{(0)}=0$

For (int t=1; t<T+b)

For (int n=0; n<N+1)

$wp-c_n^{(t)} \sim P(wp-c_n|d_{1,n}^{(t)}, d_{2,n}^{(t)}, \ldots, d_{i-1,n}^{(t)}, d_{i+1,n}^{(t-1)}, \ldots, d_{N,n}^{(t-1)}, \alpha)$

To calculate (the n-th topic word assignment of adjacent distance);

$wp\text{-}c_N^{(t)}=(d_{1,N}^{(t)}, d_{2,N}^{(t)}, \ldots, d_{N-1,N}^{(t)})$

To generate (the N-th topic word assignment after sampling);

To return $(wp-c^{(1+b)}, wp-c^{(2+b)}, ..., wp-c^{(T+b)})$: the last T samples after discard burn-in;

END

3.1.2 Gipps Sampling

In this paper, we will use the Gipps sampling algorithm to sample the topics of relational triples hierarchical topic tree model. The algorithm can be used to train the parameters of the hierarchical topic tree model of relational triples to evaluate the potential topic $Z_{m,n}$ and sampling path $C_{m,l}$ of relational triples. In the Gipps sampling algorithm, we iteratively obtain the conditional distribution of each potential topic variable according to the given potential topic variable and observed topic value:

$$p(c_{m,l}^{(new)}|c_{-m,l}, w, \eta) \propto p(c_{m,l}^{(new)}|D, \alpha) p(w|z(c_{-m,l} \cup c_{m,l}^{(new)})_{m,n}, H_0) \quad . \tag{3}$$

In formula (3), $c_{m,l}^{(new)}$ is the new sampling path. $Z(c)_{m,n}$ is a potential topic relation triplet. η={D, α, f, H_0} is the set of model super parameters, including the set of adjacent distance values between all topic relations triples in English essays D. The super parameters of attenuation functions f and θ_m obey the Dirichlet distribution α. And the basic distribution H_0 obey the Dir(α) distribution. W is the topic relation triplet of observation. P($c_{m,l}^{(new)}$ | D, α) is the prior distribution of WP – nCRP generated from formula (2). p($w|z(c\_{m,l} \cup c_{m,l}^{(new)})_{m,n}$ is the possibility of observing the subject value in a given partition. If the content of English essays changes, the random distribution will also change accordingly. L is the number of topics, $z^l(c)_{m,n}$ is the topic assignment of the l-th potential topic, and the likelihood term of the potential topic relation triplet assignment $z(c)_{m,n}$ is:

$$p(w|z(c)_{m,n}, H_0) = \prod_{l=1}^{L} p(w_{z^l(c)_{m,n}}|H_0) \quad . \tag{4}$$

The Gipps sampling type of WP-nCRP is divided into three parts:

$$p(c_{m,l}^{(new)} \mid c_{-m,l}, w, \eta) \propto \begin{cases} \alpha & if\ c_{m,l}^{(new)} = c_{m,l}^{(new)} \\ f(d_{ij}) & if\ c_{m,l}^{(new)} = \text{j and does not belong to the two topic at the same time} \\ f(d_{ij}) \dfrac{p(w_{z^l(c_{-m,l}) \cup z^l(c_{-m,l})} \mid H_0)}{P(W_{z^l(c_{-m,l})} \mid H_0) p(w_{z^l(c_{-m,l})} \mid H_0)} & if\ c_{m,l}^{(new)} = \text{j and both belong to the topic K and L} \end{cases} . \tag{5}$$

In formula (5), if $c_{m,l}^{(new)} = c_{m,l}^{(new)}$, then the sampling path $c_{m,l}^{(new)}$ is the sampling path of its own node, and the likelihood function will not change; If $c_{m,l}^{(new)} = j$, the sampling path $c_{m,l}^{(new)}$ of may be the same as that of the other node j, but because the topic belongs to two different topics, the partition will not change; If $c_{m,l}^{(new)} = j$ and the topic belongs to the divided topic k and topic l, then $c_{m,l}^{(new)}$ sampling path will create a new topic to connect the two topics. To ensure the correct clustering of topics:

$$p(\mathbf{w}_{z^l(\mathbf{c})_{m,n}} \mid H_0) = p(w_{z^l(\mathbf{c})_1} \mid H_0) \prod_{i \in z^l(\mathbf{c})_{m,n}} 1(w_i = w_{z^l(\mathbf{c})_1}) \ . \tag{6}$$

In formula (6), $z^l(\mathbf{c})_1$ is the l-th topic word in the first topic. Therefore, when WP - nCRP extracts a relationship triplet topic $\omega_{m,n}^{(new)}$ from the potential relationship triplet topic set $\mathbf{W}$, the conditional distribution is as follows:

$$p(\omega_{m,n}^{(new)} \mid \mathbf{w}, D, H_0, \alpha) = \sum_{c_{new}} p(c_{new} \mid D, \alpha) \sum_{\mathbf{c}} p(\omega_{m,n}^{(new)} \mid c_{new}, \mathbf{c}, \mathbf{w}, H_0) p(\mathbf{c} \mid \mathbf{w}, D, \alpha, H_0) \ . \tag{7}$$

3.2 Topic Relation Triple Distributed Vector Representation

After the hierarchical topic tree model of relational triples is adopted, the relational triples will form topic clustering. This section proposes the integration of Word2Vec [22], a word embedding model based on neural network, with relational triples hierarchical topic tree model. Word2vec is one of the commonly used word embedding models proposed by Google in 2013. It belongs to the shallow neural network Model, including two kinds of neural network structures, CBOW and skip-gram Model, and the efficient training methods of these two kinds of structures, including the negative sampling method and the hierarchical classification method. Due to the rich sentence types in English composition, this paper will use skip-gram model to train word vectors, and combine with the negative sampling training method to improve the computational efficiency and accuracy of word embedding model, and carry out distributed semantic representation of essays.

We use the distributed vector representation of the word embedding model to represent each word in the relation triplet (S, R, O) as a vector, and give different weights to the vector by giving the subject, relation and object components. Then, the weight of subject, relation and object is added to get the distributed vector representation of topic relation triplet. The detailed steps of the distributed vector representation method for triples of topic relation proposed in this paper are as follows:

Words in topic relation triples are represented as n-dimensional distributed word vectors by word embedding model:

$$\begin{cases} vec(S) = [s_1, s_2, ..., s_n] \\ vec(\mathrm{R}) = [r_1, r_2, ..., r_n] \\ vec(O) = [o_1, o_2, ..., o_n] \end{cases} . \tag{8}$$

The calculation formula of distributed vector representation of topic relation triples is as follows:

$$vec(R-triad) = \lambda_1 vec(S) + \lambda_2 vec(\mathrm{R}) + \lambda_3 vec(O) \ . \tag{9}$$

In formula (9), λ_1 represents the super parameter of the subject component S in the relational triplet, $\lambda_2$represents the super parameter of the relational component R in the relational triplet, and λ_3 represents the super parameter of the object component O in the relational triplet, where $\lambda_1+\lambda_2+\lambda_3=1$. In the training process of the model, because the subject and object components in English composition sentences are more likely to affect the topic, the weight of them will also be increased in this paper. Through the training samples, the super parameters are optimized to make the distributed representation of relation triples the best.

Under the same topic, the modified cosine similarity algorithm is used to calculate the semantic similarity of two topic relationship triples under the same topic, The calculation formula is as follows:

$$\mathbf{X}_{(S_i,\mathrm{R}_i,O_i)} = [\lambda_1 \mathrm{s}_{1,i} + \lambda_2 \mathrm{r}_{1,i} + \lambda_3 o_{1,i}, \lambda_1 \mathrm{s}_{2,i} + \lambda_2 \mathrm{r}_{2,i} + \lambda_3 o_{2,i}, ..., \lambda_1 \mathrm{s}_{n,i} + \lambda_2 \mathrm{r}_{n,i} + \lambda_3 o_{n,i}] . \tag{10}$$

$$\mathbf{Y}_{(S_j,\mathrm{R}_j,O_j)} = [\lambda_1 \mathrm{s}_{1,j} + \lambda_2 \mathrm{r}_{1,j} + \lambda_3 o_{1,j}, \lambda_1 \mathrm{s}_{2,j} + \lambda_2 \mathrm{r}_{2,j} + \lambda_3 o_{2,j}, ..., \lambda_1 \mathrm{s}_{n,j} + \lambda_2 \mathrm{r}_{n,j} + \lambda_3 o_{n,j}] . \tag{11}$$

$$\cos\theta=\frac{\sum_{i=1}^{n}((\lambda_1 s_{n,i}+\lambda_2 r_{n,i}+\lambda_3 o_{n,i})-\mu_x)\times((\lambda_1 s_{n,j}+\lambda_2 r_{n,j}+\lambda_3 o_{n,j})-\mu_y)}{\sqrt{\sum_{i=1}^{n}((\lambda_1 s_{n,i}+\lambda_2 r_{n,i}+\lambda_3 o_{n,i})-\mu_x)^2}\times\sqrt{\sum_{i=1}^{n}((\lambda_1 s_{n,j}+\lambda_2 r_{n,j}+\lambda_3 o_{n,j})-\mu_y)^2}}. \quad (12)$$

In formula (12), μ_x is the mean value of distributed vector X of topic relation triplet, and μ_y is the mean value of distributed vector **Y** of topic relation triplet. In this paper, the top L topic relation triplet of semantic similarity under each topic is selected as the optimal topic relation triplet.

3.3 Topic Semantic Space Extension Based on Knowledge Base

In order to further enhance the semantic representation of topic relation triples in distributed vector space and extend the semantic relationship of hierarchical topic tree mixed semantic space, this paper will use semantic vocabulary (PPDB2.0) [23] and common sense knowledge (ConceptNet5.5) [18] base to expand the topic semantic space.

In this paper, the subject word distributed vector and object word distributed vector of the above optimal topic relationship triples are extracted, and semantic similarity query matching is carried out in semantic vocabulary database and common sense knowledge database. The top ten types of relationship triples with similarity weight are selected as candidate topic relationship triples. We express the candidate relation triples under each topic as topic relation triples distributed vectors. Then, the sequential iteration method is used to extend the topic relation triples in the hybrid semantic space of hierarchical topic tree. The expansion process of topic semantic space is shown in Fig. 2:

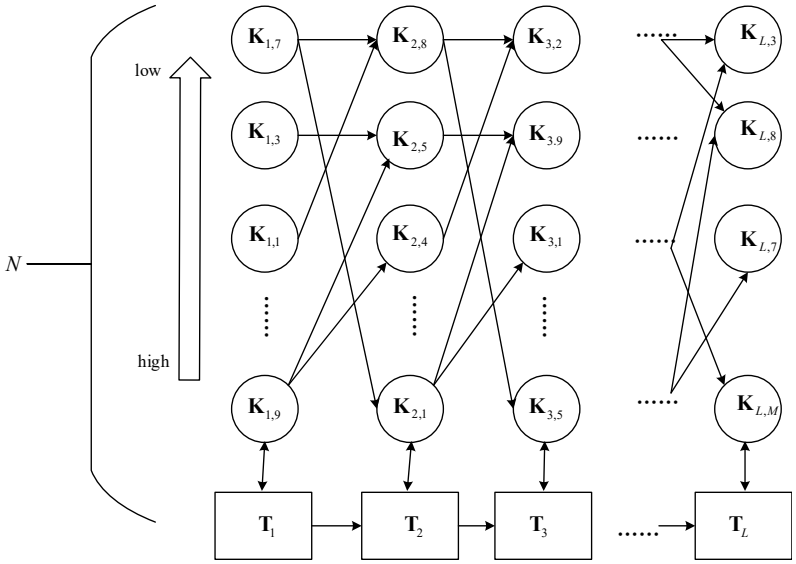

Fig. 2. Flow chart of semantic space expansion of knowledge base topics

Firstly, N relational triples with the best semantics are selected according to T_1 and stored in the new candidate topic set. Then, N relation triples with the best semantics are selected according to T1T2 and stored in the new candidate topic set. We update the ranking of relation triples in the new candidate topic set by calculating the semantic similarity of relation triples, and further generate the optimal N relation triples. Iterate in turn until all the original L relationship triples under the topic are selected and the optimal relationship triples are selected. The total number of iterations of the whole process is L, and the new candidate topic set generated by the last iteration $K^{(new)}=\{K_1,K_2,...,K_N\}$. The optimal relation triplet generated before creating a new candidate topic set $K^{(old)}=\{K_1^{'},K_2^{'},...,K_M^{'}\}$, Semantic relevancy is $\{r_1^{'},r_2^{'},...,r_M^{'}\}$, $K^{(new)'}=\{T_1^{'},T_2^{'},...,T_l^{'},K_1^{'},K_2^{'},...,K_M^{'}\}$.

3.4 On-Topic Analysis Algorithm

In this paper, the hierarchical topic trees hybrid semantic spaces generated in the mixed semantic space of hierarchical topic tree is used as the basis of English essays topic analysis. By calculating the topic semantic similarity between sentence and topic, between sentence and paragraph, between paragraph and topic, and between full text and topic, the paper analyzes English essays from the sentence level Topic semantic similarity from paragraph level to full text. The calculation formula is as follows:

$$G_{inTopic} = \gamma_1 \frac{\sum_{i=1}^{N} \cos\theta_{S-T}}{N} + \gamma_2 \frac{\sum_{i=1}^{N} \cos\theta_{S-P}}{N} + \gamma_3 \frac{\sum_{j=1}^{M} \cos\theta_{P-T}}{M} + \gamma_4 \cos\theta_{C-T} \quad . \tag{13}$$

In the formula(13), $G_{inTopic}$ is the score of English essays topic analysis, $\Sigma_{i=1}^{N} cos\theta_{S-T}$ is the sum of semantic similarity between N sentences and topics in the English essays to be approved, $\Sigma_{i=1}^{N} cos\theta_{S-P}$ is the sum of semantic similarity between N sentences and paragraphs in the English essays to be approved, $\Sigma_{i=1}^{N} cos\theta_{P-T}$ is the sum of semantic similarity between M paragraphs and topics in the English essays to be approved, $cos\theta_{S-T}$ is the topic semantic similarity between the full text and the topic in the English essays to be approved, γ is the super parameter, and $\gamma_1+\gamma_2+\gamma_3+\gamma_4=1$.

By calculating the thematic semantic similarity of the triple distributed vectors of the topic relationship between sentence and topic and sentence and paragraph, and assigning the different weights δ_1, δ_2 of the thematic semantic similarity between sentence and topic and sentence and paragraph, we get the relevant semantic similarity of the final sentence, and extract the relevant English essays sentences:

$$\cos\theta_{in-topic} = \delta_1 \cos\theta_{S-T} + \delta_2 \cos\theta_{S-P} = \delta_1 \cdot \frac{\sum_{i=1}^{n} (s_i - \mu_s) \times (t_i - \mu_t)}{\sqrt{\sum_{i=1}^{n} (s_i - \mu_s)^2} \times \sqrt{\sum_{i=1}^{n} (t_i - \mu_t)^2}} + \delta_2 \cdot \frac{\sum_{i=1}^{n} (s_i - \mu_s) \times (p_i - \mu_p)}{\sqrt{\sum_{i=1}^{n} (s_i - \mu_s)^2} \times \sqrt{\sum_{i=1}^{n} (p_i - \mu_p)^2}} \quad . \tag{14}$$

We will use formula (14) to calculate the semantic similarity of all sentences in the English essays to be approved and sort them. We will set a threshold to extract the sentences in the essays paragraph.

4 Experiment

4.1 Data Set

The training sets used in this paper include the Chinese Learner Corpus (CLEC), the International Corpus of Asian English Learners (ICNALE)and the Wikipedia corpus. The test set is mainly composed of two parts: on-topic test set and off-topic test set. This article selects five topics of Chinese students' English essay from CLEC, two topics of Chinese students' English essay from ICNALE, and three topics from Kaggle, the foreign English essay competition data set. A total of 10 themed student English essays are used as a test set for the topical analysis of the English essays of this article. In the test set of each subject, a fixed number of English essays on other topics are added as off-topic essays, as shown in Table 1. Shown. Based on the principle of topic differentiation, this paper selects a large number of topic category composition samples in the test set, and will use the above test set to comprehensively verify the effectiveness and accuracy of the analysis method of essays topic in this model.

Table 1. Data sources of test set

The topic of essays	Source	On-topic	Off-topic
My Future	CLEC	1594	406
Practice Makes Perfect	CLEC	1592	408

Getting to Know the World Outside the Campus	CLEC	1315	685
How to make good use of college life	CLEC	1322	678
Chinese Traditional Festival	CLEC	1039	961
Whether it is important for college students to have a part time job	ICCNALE	1018	982
Whether smoking should be completely banned at all the restaurant in the country	ICCNALE	1136	864
Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, should be removed from the shelves?	Kaggle	1600	400
Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.	Kaggle	1526	474
Describe the mood created by the author in the memoir. Support your answer with relevant information from the memoir.	Kaggle	1605	395

As shown in Table 1, the test set of this experiment consists of 14000 Chinese students' English essays and 6000 foreign students' English essays selected from different corpora under 10 topics, with a total of 20000 students' English essays, including 13610 on-topic English essays and 6390 off-topic English essays.

4.2 Preprocessing and Parameter Setting

We perform some preprocessing on the data set. (1) We use regular expression matching method to filter special characters for batch English essays, so as to prepare for the next step of segmentation.(2) The English slicer with higher accuracy is used to segment the English essay in a global to local way, and the stop word set is used to get rid of stop words in the English essays, and then the word stemming processing is carried out to restore the words in the essay to the initial state. (3) A part of speech tagging algorithm based on cyclic dependent neural network is used to tag words. (4) The essays is represented as a structured, computer-readable representation, usually in the form of a relational triplet, including the subject, the relation, and the object of the sentence. The English Splitter, Part-of Speech Labeller and Relational Triple Information Extractor used in the model in this paper are StandfordCorenlp [24], an open source natural language processing package from Stanford University with better processing effect.

The training set described in 4.1 of this paper optimizes the three main parameters of the hierarchical topic tree model: alpha, gamma and eta. Through the training set optimization, the optimal parameters are alpha = 10.0, gamma = 1.0 and eta= 0.1, and use the above training set to train word2vec. The word embedding training model used in this paper is based on negative sampling skip-gram model, and the dimension of the generated distributed semantic space word vector is set to 300.

4.3 Results Analysis and Discussion

4.3.1 Overall Evaluation and Discussion of The Model

In order to verify the performance of the proposed method in different corpus test sets, we choose different topic test sets from different corpora for experiments. The experimental results are shown in Table 2:

Table 2. The experimental results of topic analysis method based on semantic space under 10 topics

The topic of essays	Accuracy	Recall	F1
My Future	93.91%	87.14%	90.40%
Practice Makes Perfect	94.66%	86.81%	90.56%

Getting to Know the World Outside the Campus	93.40%	88.29%	90.77%
How to make good use of college life	94.46%	88.96%	91.62%
Chinese Traditional Festival	94.62%	88.07%	91.23%
Whether it is important for college students to have a part time job	92.17%	87.13%	89.58%
Whether smoking should be completely banned at all the restaurant in the country	94.63%	86.88%	90.59%
Write a persuasive essay to a newspaper reflecting your vies on censorship in libraries. Do you believe that certain materials, should be removed from the shelves?	89.86%	86.44%	88.12%
Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion.	90.32%	86.83%	88.54%
Describe the mood created by the author in the memoir. Support your answer with relevant information from the memoir.	90.83%	86.42%	88.57%

As can be seen from Table 2, the method is more effective in the test set of English essays with different topics of different corpora. The average F1 of the method under 10 topics is 90.00%. From the test results of English essay topic test sets with different lengths, the effect of this paper's topic analysis is basically not affected by the length of the topic. The reason is that this paper expands the topic set and enriches the semantic information of the topic. The accuracy of the CLEC test set and the two topics in the ICNALE test set is slightly higher than that of the foreign English essay data sets. There may be two reasons. One is that this paper mainly uses Chinese students' English essays as the training set, the other part may be that the topics of foreign English essays are relatively open, and the topic semantic information is relatively scattered, so it is not easy to cluster the topics. Therefore, there is a small difference in the experimental results of the above test set between Chinese students' English essays and foreign students' English essays.

In order to verify the analysis effect of the hierarchical topic trees hybrid semantic spaces, comparative experiments are carried out. This paper uses the topic semantic space which combines word2vec with topic hierarchical tree, the topic semantic space which combines word2vec with the improved relational triplet hierarchical topic tree model, the topic semantic space which combines word2vec with topic hierarchical tree and knowledge base, and the mixed semantic space which combines relational triplet hierarchical topic tree model with word2vec and knowledge base extension .A contrastive experiment was conducted to analyze the relevance of English essays. The test set uses the above 20000 English texts, and the experimental results are shown in Fig. 3.

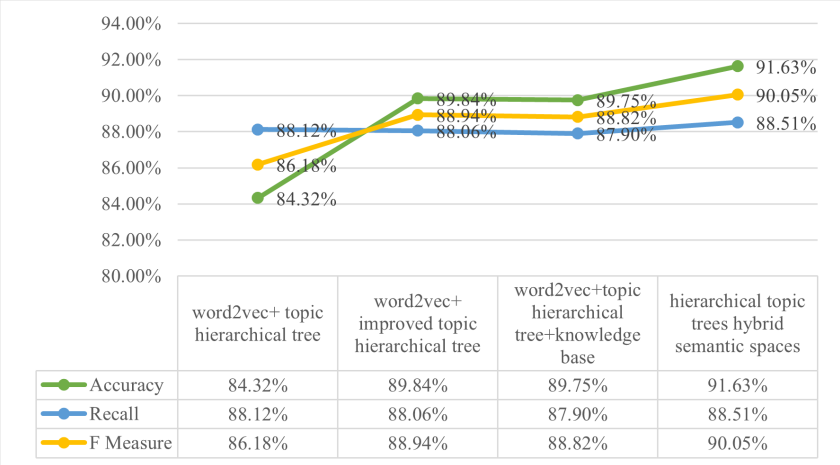

Fig. 3. Experimental comparison diagram of topic analysis method

From the broken line graph of the experimental results, compared with the model before the improvement, the hierarchical topic trees hybrid semantic spaces model constructed in this paper keeps the recall rate stable, the accuracy rate is 91.63%, F1 is 90.05%, and the experimental effect is good. Therefore, adding topic model into word2vec distributed semantic space can effectively reduce the noise interference caused by non topic words. At the same time, the expansion of topic set of knowledge base can obviously improve the fine-grained of topic semantic space in English essay.

Then, we make a comparative experiment with the current typical unsupervised topic analysis method. At present, the typical unsupervised topic analysis methods include:in foreign research, Rei and Cummins [25] et al. use word embedding model word2vec to generate word distributed vector representation, and combine word weight with TF-IDF feature weight to represent sentence vector to realize English text topic analysis, which is referred to as "word embedding distributed vector representation" in the table. In domestic research, Li X [26] et al. used LDA topic model to extract the core topic words set, and used distributed word vector to represent the expansion of the core topic words of text extraction, combined with the similarity between text and topic to get the accuracy of English essay topic, which is referred to as "LDA + words embedded in distributed vector representation" in the table. In this paper, the unsupervised semantic space based English essay topic analysis method is referred to as " method of this paper ". In this paper, 20000 English essays in the test set are used to test the above three methods, and the experimental results are shown in Fig. 4.

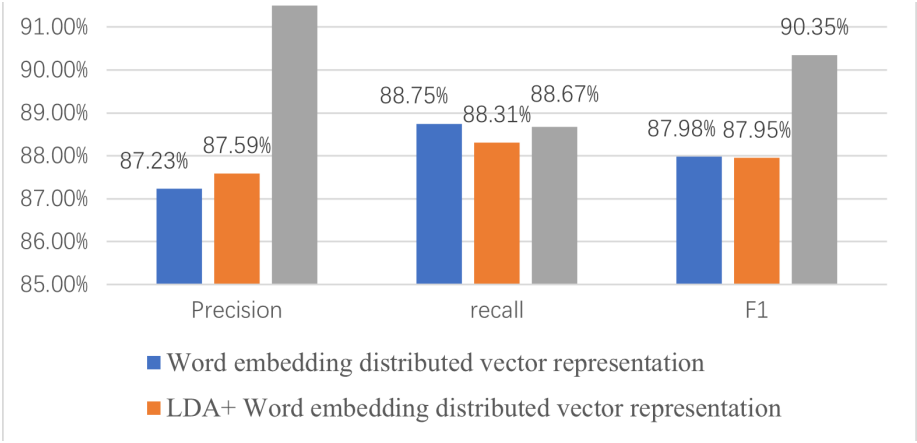

Fig. 4. Comparison experiment of unsupervised analysis method

From the experimental results in the table, the "word embedding distributed vector representation" method uses the word embedding model word2vec and the algorithm of inverse document word frequency to extract the key features of English essay to a certain extent, but the method is based on the idea of statistics and relies heavily on the training set. The "LDA+ Word embedding distributed vector representation" further reduces the interference caused by the non topic words in the topic semantic space by extracting the topic words in the essay, but the topic words clustering method based on LDA ignores the hierarchical relationship between the topic words. In this paper, the topic relation triplet is used as the carrier, and the hierarchical topic trees hybrid semantic spaces is further expanded through the knowledge base. Under the condition of stable recall rate, the accuracy of the model is improved by 3.91% compared with the LDA + Word embedding distributed vector representation. Experiments show that the model has higher accuracy and practical value than the current English essay topic analysis method.

4.3.2 Experiment and Analysis of On-topic Analysis Algorithm

This paper selects two writing topics from ICNALE: "what it is important for college students to have a part time job" and "what smoking should be completely banded at all the restaurant in the country", Under each topic, 500 English essays are selected as the test set of the topic analysis algorithm experiment. According to statistics, a total of 15930 sentences are included in 1000 English essays. The English teachers of our research group manually mark 1-5 relevant sentences for each article, and a total of 6832 sentences are relevant sentences. Firstly, this paper uses the on-topic analysis algorithm to test the sentences in the 1000 English essays, sets different thresholds of on-topic, and calculates the corresponding accuracy rate, recall rate and F1. The experimental results are shown in Table 3. From the experimental results in the table, we can see that when the threshold value of the on-topic

sentence is set to 0.60, the F1 of the algorithm is 86.45%, and the effect of extracting the sentence is ideal. If the extraction threshold is set too high or too low, it will cause the problem of high accuracy, low recall or low accuracy, high recall, which can not achieve the experimental effect of the topic sentences to be extracted in this paper.

Table 3. Comparison experiment of unsupervised analysis method

Topic sentence extraction threshold	Precision	recall	F1
0.5	65.65%	89.95%	75.90%
0.52	70.00%	87.24%	77.67%
0.54	75.26%	88.13%	81.19%
0.56	80.60%	87.57%	83.94%
0.58	84.51%	85.69%	85.10%
0.6	**89.12%**	**83.93%**	**86.45%**
0.62	88.23%	82.71%	85.38%
0.63	88.00%	80.95%	84.33%
0.64	86.67%	70.67%	77.86%
0.66	88.62%	65.18%	75.11%

Therefore, in order to verify the accuracy of on-topic analysis algorithm to extract the on-topic sentences, the threshold of on-topic sentence extraction is set to 0.60, and the experimental results of extracting on-topic sentences under different essay quantity are shown in Table 3.

Table 4. Experimental results of different essay numbers of relevant sentences set to 0.6

Number of essays	Precision	recall	F1
20	87.46%	82.42%	84.87%
40	87.69%	83.32%	85.45%
60	88.00%	83.44%	85.66%
80	88.17%	82.91%	85.46%
100	87.93%	83.27%	85.54%
200	88.54%	82.90%	85.63%
400	88.86%	83.28%	85.98%
600	88.92%	83.17%	85.95%
800	89.08%	83.02%	85.94%
1000	89.12%	83.93%	86.45%

20, 40, 60, 80, 100, 200, 400, 600, 800 and 1000 English essays were randomly selected from the test set. A total of 10 groups of on-topic sentences were extracted. 1-2 on-topic sentences were extracted from each paragraph of each article, and 1-5 on-topic sentences were extracted from each article, The extracted sentences are compared with the artificially marked sentences to verify the correctness of the extraction. The experimental results are shown in Table 4. With the gradual increase of the number of English essays, the accuracy of the on-topic related sentences extracted by the on-topic analysis algorithm is also improved, and the recall rate ends to be stable. After the above 10 groups of experiments, the average F1 of the on-topic related sentences extracted is 86.45%, with a high accuracy. Therefore, it shows that the weighted on-topic analysis algorithm of sentence and topic and sentence and paragraph can effectively improve the extraction of on-topic sentences.

4.3.3 Comparison of Model and Teacher Rating

In order to verify the effectiveness of this model in the actual English essay correction system, we invited a number of professional English teachers of our research group to analyze the composition topic “Whether it is important for college students to have a part time job” selected from ICNALE. The results of the experiment are shown in Fig. 5.

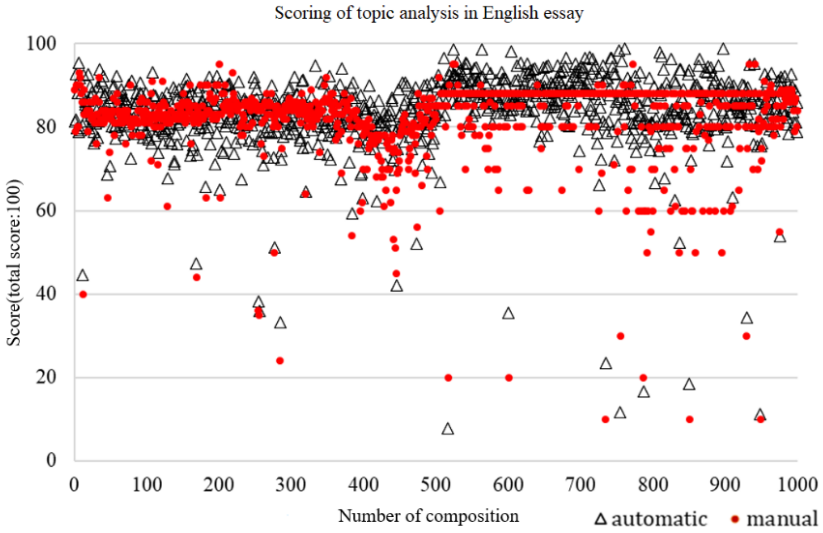

Fig. 5. Comparison chart of model and teacher rating

From the contrastive experiments of 1000 English essays, it can be seen that the score of this model is close to the teacher's score in most of the test sets. The average score of teachers is 81.73, and the average score of this model is 83.64. The average score difference between teachers' score and this model is 1.91. As the analysis of English on-topic belongs to subjective analysis and judgment, English teachers are more strict in correcting the analysis of English essay on-topic, and the score of this model is higher than that of teachers' score.

To sum up, this paper evaluates the contrastive experiments in different test sets, the comparison experiments between the improved and the original models, the comparison experiments with the existing typical unsupervised models, and the comparison experiments with the teachers' scores of the quality of English writing. The experiments show that the model based on Chinese students has a high accuracy.

5 Conclusion

In this paper, we use relational triples as the carrier to analyze the potential topics of English essays. By constructing a hierarchical topic tree hybrid semantic space, we cluster the topics of relational triples in English essays, express them in a distributed way, and expand the topic set. Based on the improved on-topic analysis algorithm in this paper, the on-topic sentences of English essays are extracted to achieve the score of the quality of English essays, which effectively improves the accuracy of topic clustering of relation triples in English essays, and enriches the topic semantic space of distributed vector representation.

This model has achieved better results in analyzing topics, but some details still need to be improved. Due to the limited training time, the problem of data sparsity may arise because the corpus does not contain words with less frequency in the text. Therefore, future research will consider using various types of corpus such as Google news corpus to further train the word embedding model to minimize data sparsity. At the same time, because word2vec belongs to the static distributed vector representation model, it cannot dynamically learn the context information in the text, so we can add dynamic adjustment in the word embedded distributed representation model, such as long-term and short-term memory neural network LSTM, and use a large number of corpora for training. In the future, we need to do further research to improve the text topic analysis model.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62066009)

References

[1] M. Liu, Design of Intelligent English Writing Self-evaluation Auxiliary System, Informatica 43(2) (2019).

[2] J. Zhou, Application investigation and technological development of intelligent product in college English teaching, Journal of Physics: Conference Series. IOP Publishing 1237(2) (2019).
[3] Q. Wen, A Study on the construct validity of “Composition Content” -- an attempt to use structural equation model software AMOS5, Foreign language study 3(2007) 66-71.
[4] K. Zupanc, B. Zoran, Automated essay evaluation with semantic analysis, Knowledge-Based Systems 120(2017) 118-132.
[5] M.A. Hussein, H. Hassan, M. Nassef, Automated language essay scoring systems: a literature review, Peer J Comput Sci 5(2019) e208.
[6] J. Wang, Study on validity of essay automatic grading system, Hainan university (2015).
[7] X. Li, Q. Wen, K. Pan, Unsupervised Off-Topic Essay Detection Based on Target and Reference Prompts, in: Proc. 2017 13th International Conference on Computational Intelligence and Security (CIS), IEEE, 2017.
[8] C. Meng, W. Song, L. Fu, Research on off-topic Writing Detection Method based on LDA coupled space Model, Computer Application Research 338(2019) 30-33.
[9] L. Liu, A Comparative Study of Different Text Similarity Measures for Identification of Off-topic Student Essays, Boletín Técnico 55(11) 2017.
[10]Z.P. Chen, W.L. Chen, Off-topic Essay Detection Based on Document Divergence, Journal of Chinese Information Processing 31(1) (2017) 23-30.
[11]Z. Yang, H. Liu, M. Chen, Off-Topic Text Detection Based on Neural Networks Combined with Text Features, in: Proc. 2018 14th International Conference on Computational Intelligence and Security (CIS). IEEE, 2018.
[12]G. Huang, J. Liu, C. Fan, T. Pan, Off-topic English Essay Detection Model Based on Hybrid Semantic Space for Automated English Essay Scoring System. in: Proc. MATEC Web of Conferences, 2018.
[13]Y.-C. Chang, Y.-L. Hsieh, C.-C. Chen, A semantic frame-based intelligent agent for topic detection, Soft Computing 21(2) (2017) 391-401.
[14]G. Wang, Z. Yang, H. Wang, A Semantic Enhanced Topic Model Based on Bi-directional LSTM Networks, Journal of Computers 30(6) (2019) 60-72.
[15]S.S. Chung, M. D’Arcy, Unsupervised Topic Model Based Text Network Construction for Learning Word Embeddings, in: Proc. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), IEEE, 2019.
[16]C. Li, H. Wang, Z. Zhang, Topic modeling for short texts with auxiliary word embeddings, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016.
[17]Q. Qu, R. Cui, J. Zhao, Off-topic Detection of English compositions based on LDA and word2vec, Computer Application Research 2(2019).
[18]R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proc. Thirty-first AAAI conference on artificial intelligence, 2017.
[19]W. Shalaby, W. Zadrozny, H. Jin, Beyond word embeddings: learning entity and concept representations from large scale knowledge bases, Information Retrieval Journal 22(6)(2019) 525-542.
[20]S. Derby, P. Miller, B. Devereux, Feature2Vec: Distributional semantic modelling of human property knowledge, arXiv preprint arXiv:1908.11439, (2019).
[21]D.M. Blei, T.L. Griffiths, M.I. Jordan, Hierarchical topic models and the nested Chinese restaurant process, NIPS, 2003.
[22]T. Mikolov, K. Chen, G. Corrado, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.
[23]E. Pavlick, P. Rastogi, J. Ganitkevitch, B. Van Durme, C. Callison-Burch, PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.
[24]G. Angeli, M.J.J. Premkumar, C.D. Manning, Leveraging linguistic structure for open domain information extraction, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015.
[25]M. Rei, R. Cummins, Sentence similarity measures for fine-grained estimation of topical relevance in learner essays, arXiv preprint arXiv:1606.03144, (2016).
[26]X. Li, Q. Wen. An unsupervised off-topic detection method based on local density, Journal of Chinese Information 31(6) (2017) 205-213.