Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 1, pp. 327–339, DOI 10.12694/scpe.v25i1.2298

# CONSTRUCTION OF SEMANTIC COHERENCE DIAGNOSIS MODEL OF ENGLISH TEXT BASED ON SENTENCE SEMANTIC MAP

PENG GUO*

**Abstract.** The current English composition automatic correction system rarely involves coherence quality analysis of compositions. Therefore, a semantic coherence diagnosis model for English texts was constructed based on sentence semantic maps, and its effectiveness was verified through experiments. Experimental results show that sub Graphical model 10, 12 and 13 exceed 200 on the first two test texts and 1 on the last two test texts. However, the differences between these three subgraphs on different test texts are not significant, with differences below 30 and 0.3. In addition, when extracting incoherent sentences, the F1 value reaches the optimal value at a threshold of 0.34, which is 87.54%. When the threshold is fixed at 0.34, the accuracy of extracting non coherent sentences also increases with the number of articles, reaching a maximum of 88.43%. At the same time, there was no significant difference in accuracy, recall, and F1 values among different English composition numbers, maintaining between 83% and 89%. The Pearson coefficient calculated in the actual comparison with the teacher's manual composition score is 0.6025, indicating a strong correlation between the two, indicating that the diagnostic results are reliable. Overall, the diagnostic model constructed in the study has strong accuracy and effectiveness, and is practical in the diagnosis of semantic coherence in actual English texts.

**Key words:** Sentence semantic map; English text; Preprocessing; Diagnostic model

**1. Introduction.** The coherence of the text is highly subjective, and in the language field, it is regarded as the fluency of the text in the text [1]. The coherence of the article is an important characteristic of good text, and it is also an important criterion for measuring the readability of the article. In a coherent work, the arrangement of sentences and words in it is not random, but has certain logic and consistency [2]. The development of artificial intelligence has prompted the emergence of many English text correction systems, which not only reduces the pressure on teachers to correct, but also improves the writing ability of students. Based on this, scholars at home and abroad have conducted in-depth research on it. Hu L et al. constructed a model of English text grammar error correction by using neural network, so as to effectively realize the error correction of English grammar [3]. Gondaliya Y et al. achieved English grammar and spelling check by using rules, thus effectively improving the coherence of sentences [4]. By building a spelling error correction system, Chaabi Y et al. realized the error correction of English vocabulary, thereby improving the quality of the text [5]. Currently, there is very little research on automatic grading of English compositions to design the indicator of coherence quality, and the performance of the actual core construction system still needs to be improved. Therefore, a diagnostic model for semantic coherence in English texts was constructed based on sentence semantic maps, with the aim of achieving effective analysis of the coherence quality of English texts and accurately evaluating the overall quality of English function.

**2. Related Work.** With the rapid development of artificial intelligence and its related fields, computers can gradually independently evaluate the comprehensive quality of English texts [6]. However, the coherence quality is seldom included in the evaluation by computer, which leads to the unreasonable scoring results [7]. Based on this, scholars at home and abroad have conducted in-depth research on it. Srivastava K et al. ensured the semantic coherence between different parts of the article by using some sentences of the article to test the semantic coherence [8]. Yang X et al. ensured the semantic coherence of the original sentence by using context-aware word replacement [9]. Aminovna ensures the coherence of text writing by proposing the most effective methods in writing and emphasizing the importance of coherence in academics [10]. Saleh M et al. provided help for the coherence of people's text creation by analyzing the means of grammatical cohesion [11]. Dassanayake N

*College of Finance and Management, Guangzhou Institute of Technology, Guangzhou, 510630, China (wayne007008@163.com)

Table 2.1: Summary and Discussion of Previous Research

| **Author** | **Research questions** | **Research results** |
|---|---|---|
| Srivastava K et al. | The issue of semantic coherence in different parts of the article | Strengthening semantic coherence through the use of partial sentences in test articles |
| Yang X et al. | Semantic Incoherence of Original Sentences | Ensure semantic coherence of the original sentence by using context aware vocabulary substitution |
| Aminovna B D et al. | Discontinuity in text writing | Emphasizing the importance of coherence in academia, ensuring coherence in text writing |
| Saleh M et al. | Discontinuity in text creation | The analysis of grammatical cohesive devices provideassistance for the coherence of text creation |
| Dassanayake N et al. | Lexical Incoherence in Chinese Translation of Sri Lankan Language | Utilizing relevant texts translated by Sri Lankan learners to ensure lexical coherence in Sri Lankan language translation into Chinese |
| Akmilia P M et al. | Discontinuity of text | Utilizing cohesive devices to ensure text coherence |
| Abdusalomovna K Y et al. | Semantic Incoherence in Discourse | Analyzed the relationship between cohesion and coherence to provide assistance in ensuring semantic coherence in discourse |
| Zhang K et al. | Discontinuity in semantic themes | The Topic model of short text is constructed by using two related knowledge to ensure the coherence of semantic topics |
| Linnik A et al. | Discontinuity in overall semantics | And utilizing combination rating operations such as information content to ensure overall semantic coherence |
| Rebuffel C et al. | Discontinuity and fluency of text | Propose a multi branch decoder to enhance text coherence and fluency |
| Gaur M et al. | The problem of weak semantic logicality | The use of knowledge graphs to generate information ensures the logic of semantics |
| Ruzikulovich A T et al. | The Semantic Incoherence of Poetry | The analysis of the functional semantics of imperative structures ensures semantic coherence in poetry |

ensures the lexical coherence of Sri Lankan language translation into Chinese by using relevant texts translated by Sri Lankan learners [12]. Akmilia PM effectively enhances persuasiveness at research conferences by using cohesive devices to ensure the coherence of the text [13]. Abdusalomovna provides help in ensuring semantic coherence in discourse by analyzing the relationship between cohesion and coherence [14].

In addition, Zhang et al. constructed a short text topic model by using two-item related knowledge to ensure the coherence of semantic topics [15]. Linnik et al. compared the differences between aphasia and non-aphasia, and used combined rating operations such as information content to ensure the coherence of the overall semantics [16]. Rebuffel et al. proposed a multi-branch decoder to effectively train word-level labels, thereby improving text coherence and fluency [17]. Gaur M et al. used dynamic meta-information retrieval to analyze the coherence of contextual sentences, and used knowledge graphs to generate information to ensure semantic logic [18]. Ruzikulovich analyzes the functional semantics of imperative structures to ensure that the poetic language is both poetic and semantically coherent [19]. Judging from the research of scholars at home and abroad, most scholars do not start from the overall coherence of the text to distinguish different degrees of sentence coherence. Therefore, the research on the construction of an English text semantic coherence diagnosis model based on sentence semantic maps is innovative. It can not only effectively improve the current lack of coherence quality assessment in English composition evaluation, but also provide assistance for students' English composition grades. At the same time, it also lays the foundation for the development of an English composition evaluation system while improving the overall evaluation quality of the role of English. In addition, a summary of specific research literature is shown in Table 2.1.

**3. Construction of a semantic coherence diagnosis model of English text based on sentence semantic graph.**

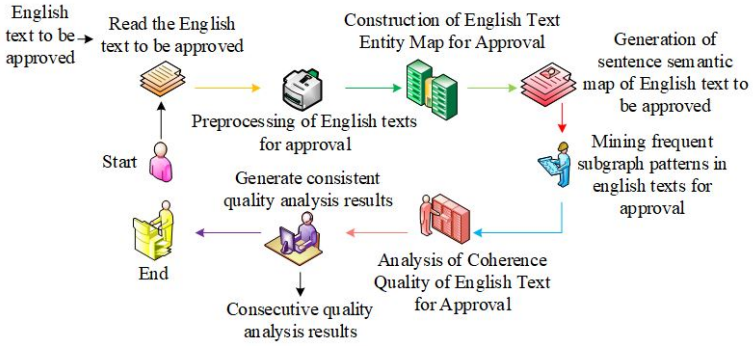



Fig. 3.1: The Model Structure of English Text Semantic Coherence Diagnosis

**3.1. Analysis of semantic coherence diagnosis model structure and semantic space correlation algorithm.** To realize the effective diagnosis of the semantic coherence quality of English composition texts, the research constructs a semantic coherence diagnosis model of English text based on the sentence semantic graph, and evaluates it through experiments. In English text scoring systems, semantic coherence is rarely involved as an evaluation metric. Semantic coherence refers to paying attention to the combination and cohesion of sentences in written writing. To achieve discourse coherence, attention should be paid to not only topic, word order, language use (cohesion, echo), but also context and sentence pattern. coordination [20]. Semantic coherence theory is a widely accepted theory in coherence analysis. In this theory, there are two very important concepts, namely macrostructure and connection. The macro structure represents the semantic relationship between the parts of the whole text, while the connection is the semantic relationship between the various elements in the idiom.

Research on the application of semantic coherence theory, Entitative graph model, related algorithms of vector semantic space and related knowledge of sub graph matching in the construction of semantic coherence diagnosis model of English text. Among them, the Entitative graph model is fundamentally an improvement on the grid model of test questions. It uses the form of sentence graph to make the connection between sentences more three-dimensional, and can analyze the semantic relationship within the text from a macro perspective. The related algorithms in Vector space are mainly used to map the text to the semantic space of the vector, and express the semantic similarity between text units by using the association between vectors. Therefore, the model structure of the semantic coherence diagnosis of English text constructed by the research is shown in Figure 3.1.

From Figure 3.1, for the input English text, the research first uses the preprocessing module of this article to perform a series of preprocessing, such as word segmentation, part of speech tagging, dependency analysis, etc; Secondly, through the results of preprocessing, entity words in English text are identified, and the boundaries of entity word phrases are determined based on the form of syntactic trees. On this basis, the co referential phenomenon between entity words is resolved. Then, the entity words are marked with grammatical roles, and the Entitative graph structure of English text is established by combining relevant information. Then, under the guidance of semantic coherence theory, the semantic similarity information between sentences is combined with Entitative graph to generate the studied sentence semantic map model. Then, the improved matching technology is used to mine the frequent sub Graphical model in the sentence semantic map to capture the coherence patterns in the text, and the coherence features in the sentence semantic map are extracted according to the coherence quality analysis module built by the research to conduct the corresponding coherence quality analysis, finally forming the coherence quality analysis results of the English text to be approved required by the research. In addition, in natural language processing, in order to obtain a large amount of lexical information and semantic information, it is necessary to perform a series of preprocessing on the text. The researched preprocessing module consists of three parts, whose contents are shown in Figure 3.2.

From Figure 3.2, the model preprocessing module constructed by the research includes text segmentation

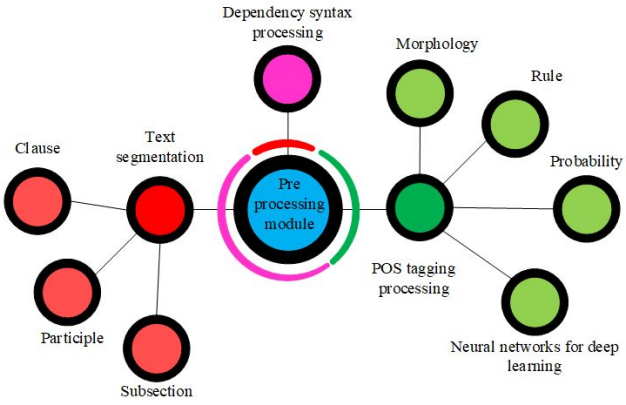



Fig. 3.2: Specific Contents of Text Preprocessing Module

processing, part-of-speech tagging processing and dependency syntax analysis processing. The text segmentation process includes segmentation, sentence segmentation and word segmentation. On paragraphs, since two consecutive line breaks appear between each paragraph, this feature is needed to divide the text. In clauses, the period is the only sign of the end of the sentence, and in English, the period can not only be used as the sign of the end of the sentence, but also can be used as the abbreviation of the first letter and surname. Therefore, the segmentation of Chinese should also be considered on a case-by-case basis. In word segmentation, there is a space between each word in English, so it can be used for word segmentation. In order to segment English text effectively and accurately, this paper studies the segmentation processing of English text using regular expressions. The part-of-speech tagging process includes part-of-speech tagging methods using lexical, rules, probability, and deep learning neural networks. In addition, dependency parsing is an important basic technique that can directly analyze words in text sentences. Dependency syntax refers to the semantic dependencies between words. On this basis, it is centered on the tree-like node in the root of the tree, and the rest are used as modifiers.

It is worth noting that in the English semantic coherence diagnosis model, text analysis occupies a large proportion. In natural language research, it is usually necessary to understand the semantics and internal connections of each component of the text, and then analyze it as a whole [21]. The study uses a group of embedding models used to generate word vectors (Word to vector, Word2vec). The focus of this model is to learn the weights of the neural network to obtain a distributed lexical representation. It includes two neural network models, namely skip mode and continuous bag of words (CBOW). Among them, the skip word mode mainly predicts the usage of the current word, while CBOW predicts the current word through the context-related vocabulary [22]. Therefore, the research uses CBOW for word vector training, and combines it with negative value sampling, so as to improve the expression efficiency and accuracy of English text, and achieve distributed semantic representation related to English text.

When calculating the output of the relevant hidden layer in the CBOW model, usually in the case of using the weight matrix of the hidden layer and the average vector, the average value of the input context word is calculated. The relevant output formula of the hidden layer is shown in equation 3.1.

$$h = \frac{1}{A} W^T (x_1 + x_2 + \ldots + x_A) = \frac{1}{A} (v_{\omega 1} + v_{\omega 2} + \ldots + v_{\omega A})^T \tag{3.1}$$

In equation 3.1, $h$ represents the output value of the hidden layer; $A$ represents the number of words in the context; $W$ represents the weighting matrix; $x_A$ represents the input vocabulary; $v$ represents the number of dimensions of the vocabulary vector; $\omega_A$ represents the words of the context $T$ ; There is a different weighting matrix in the process from the hidden layer to the output layer $W'$. According to the weights in the weighting

matrix, the calculation formula of the word score in the vocabulary is shown in equation 3.2.

$$p_j = v_{\omega j}^{'} h \tag{3.2}$$

In equation 3.2, it $p_j$ represents the score value of each word; $v_{\omega j}^{'}$ represents the degree of the $j$-th column dimension in different weighted matrices $W'$ . On this basis, the study uses the correlation log-linear classification model to obtain the posterior distribution of English words, and outputs the multinomial distribution. The relevant calculation formula is shown in equation 3.3.

$$O\left(\omega_j \mid \omega_I\right) = y_j = \frac{\exp\left(p_j\right)}{\sum_{j'=1}^{v} \exp\left(p_{j'}\right)} = \frac{\exp\left(v_{\omega_j}^T v_{\omega_I}\right)}{\sum_{j'=1}^{v} \exp\left(v'^T_{\omega_{j'}} v_{\omega_I}\right)} \tag{3.3}$$

In equation 3.3, $O\left(\omega_j \mid \omega_I\right)$ represents the multinomial distribution value; $y_j$ represents the output value of the th unit of the $v_\omega$ output layer ; $j$ represents $W$ the row of the $v'_\omega$ weighting matrix; represents $W'$ the column of the weighting matrix. For the input of context words, the purpose of research and training is to maximize the probability in equation 3.3, and according to the weight of the input context words = words, the conditional probability of the actual output words is obtained, and the final loss function is calculated. The formula is shown in equation 3.4.

$$\begin{aligned} E &= -\log\left(\omega_O \mid \omega_I, 1, \ldots, \omega_I, A\right) \\ &= -p_{j*} + \log \sum_{j'=1}^{v} \exp\left(p_{j'}\right) \\ &= -\left(v'\right)_{\omega_O}^T h + \log \sum_{j'=1}^{v} \exp\left(\left(v'\right)_{\omega_O}^T h\right) \end{aligned} \tag{3.4}$$

In equation 3.4, it $E$ represents the loss function; $j^*$ it represents the relevant index of the actual output word of the output layer. After the corresponding loss function is obtained, the correction formula for the weights of the hidden layer and the output layer can be derived. Among them, for the reciprocal of the net input of the first unit of $E$ the output layer $j$, the relevant calculation formula is shown in equation 3.5.

$$\frac{\partial E}{\partial r_j} y_j - t_j := e_j \tag{3.5}$$

In equation 3.5, $r_j$ represents the net input value; $t_j$ represents $y_j$ the label; $e_j$ represents the reciprocal value. Similarly, the formula for calculating the reciprocal $E$ of the weighting matrix from the hidden layer to the output layer is shown in $W'$ equation 3.6.

$$\frac{\partial E}{\partial \omega'_{ij}} = \frac{\partial E}{\partial p_j} \frac{\partial p_j}{\partial \omega'_{ij}} = e_j h_i \tag{3.6}$$

In equation 3.6, it represents the $h_i$th unit in the hidden layer. $i$ Combining equation 3.5 and equation 3.6, the modified formula of the weight from the hidden layer to the output layer is obtained as shown in equation 3.7.

$$v_{\omega j}^{'new} = v_{\omega j}^{'old} - \eta e_j h \quad for\ j = 1, 2, \ldots, v \tag{3.7}$$

In equation 3.7, it $\eta$ represents the learning rate; it $v'_{\omega,j}$ represents the relevant output vector of the word . $\omega_j$ Similarly, the correction formula of the weights from the output layer to the hidden layer is shown in equation 3.8.

$$v_{\omega_{I,a}}^{(new)} = v_{\omega_{I,a}}^{(old)} - \frac{1}{A} \eta E H^T \quad for \quad a = 1, 2, \ldots, A, \ldots \tag{3.8}$$

In equation 3.8, it represents $v_{\omega_{I,a}}$ the input vector of the th word in the $RH$ input context; $a$ it is a vector of dimension N, which represents the sum of the output vectors of all words in a vocabulary. If the weight is calculated according to its expected error, the definition expression of each component is shown in equation 3.9.

$$EH_i = \frac{\partial E}{\partial h_i} = \sum_{j=1}^{v} \frac{\partial E}{\partial p_j} \frac{\partial p_j}{\partial h_i} = \sum_{j=1}^{v} e_j \omega'_{ij} \tag{3.9}$$

On this basis, in order to improve the efficiency of the Word2 vec embedding model, a training method using negative sampling is used to optimize the output vector update of the model. When sampling, a better probability distribution is selected, and it is called the noise distribution. By simplifying the corresponding training objective, high-quality word embeddings can be produced. The relevant calculation formula is shown in equation 3.10.

$$E = -\log \sigma\left(v_{w_k}^T h\right) - \sum_{w_j \in W_{neg}} \log \sigma\left(-v_{w_k}^T h\right) \tag{3.10}$$

In equation 3.10, $w_k$ represents the output word; $v_{w_j}$ represents $W_k$ the output vector of the word; represents $W_{neg}$ the set of negative samples of the noise distribution. In order to obtain the correction formula of the relevant words under the condition of negative sampling, the net input sum of the output unit needs to be derived, and $E$ the formula is shown in equation 3.11.

$$\frac{\partial E}{\partial v_{w_j}'^T h} = \begin{cases} \sigma(v_{w_j}'^T h - 1), & \text{if } w_j = w_k \\ \sigma(v_{w_j}'^T h), & \text{if } w_j \in W_{\text{neg}} \end{cases} = \sigma(v_{w_j}'^T h) - t_j \tag{3.11}$$

In equation 3.11, $W_j$ when it is a positive sample, the $t$ value is 1; $w_j$ when it is not a positive sample, the $t$ value is 0. At this point, the relative derivative of the output vector of the word $w_j$ can be obtained. $E$ The calculation formula is shown in equation 3.12.

$$\frac{\partial E}{\partial v'_{w_j}} = \frac{\partial E}{\partial v_{w_j}'^T h} \frac{\partial v'^T{}_{w_j} h}{\partial v'_{w_j}} = \left(\sigma(v_{w_j}'^T h) - t_j\right) h \tag{3.12}$$

According to equation 3.12, the correction formula of the output vector can be obtained. The relevant calculation formula is shown in equation 3.13.

$$v_{w_j}'^{(new)} = v_{w_j}'^{(old)} - \eta(\sigma(v_{w_j}'^T h) - t_j)h \ldots \tag{3.13}$$

Therefore, the relevant iterative calculation of the embedding model only needs to $w_j$ update the words belonging to the words in the vocabulary accordingly, so as to improve the calculation efficiency. In order to pass the prediction error to the hidden layer, take the $E$ derivative of the output sum of the hidden layer. The calculation formula is shown in equation 3.14.

$$\frac{\partial E}{\partial h} \sum_{w_j \in \{w_k\} \cup W_{neg}} \frac{\partial E}{\partial v_{w_j}'^T h} \frac{\partial v_{w_j}'^T h}{\partial h} = \sum_{w_j \in \{w_k\} \cup W_{neg}} \left(\sigma\left(v_{w_j}'^T h\right) - t_j\right) v'_{w_j} := EH \tag{3.14}$$

Output vector can be updated accordingly by substituting the $EH$ value calculated by equation 3.14) into equation 3.8. In addition, in order to verify the validity of the diagnostic model, the study introduces the Pearson correlation coefficient, and the calculation formula is shown in equation 3.15.

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{3.15}$$

In equation 3.15, $\rho_{X,Y}$ represents the Pearson coefficient; $n$ represents the number of samples; $X$ and $Y$ represents the value of a certain sample in the sample; $\overline{X}$ and $\overline{Y}$ represents the mean value of the sample. $\rho_{X,Y}$ A value greater than zero indicates a positive correlation between the two samples, and a value smaller than zero indicates a negative correlation. And the larger the absolute value, the stronger the correlation between samples.

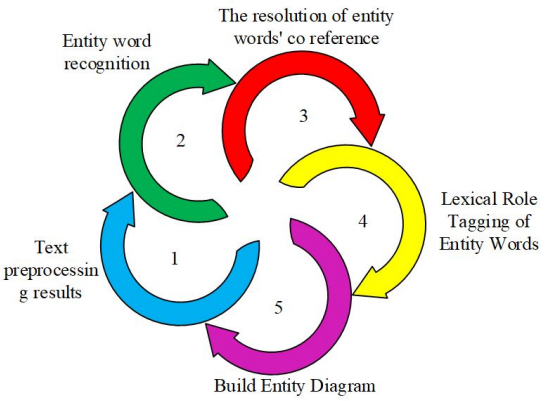



Fig. 3.3: Construction Process of Entity Diagram

**3.2. Research on entity graph and sentence semantic graph based on model preprocessing results.** After the English text preprocessing results are obtained, the relevant entity words in the text can be identified accordingly, and the co-referential phenomenon between the entity words can be eliminated at the same time. And combine the above information to build a text entity graphic model. The corresponding process of entity graph construction is shown in Figure 3.3.

From Figure 3.3, in the construction process, the first step is to preprocess the text and recognize entity words based on its results. In English text, entity words are usually used as morphological attributes of nouns or pronouns, so they can be identified according to this feature. Since most of the entity words in the English language are nouns or pronouns, the accuracy of part-of-speech tags is critical when identifying entity words. By marking the obtained parts of speech, nouns and pronouns can be extracted from the text. However, in some English languages, words such as numbers and symbols are sometimes marked as nouns. However, these words not only fail to explain the coherence of the article, but also act as a distraction. Therefore, after extracting all the entity words, in order to reduce the noise, the research also filters the entity words. In English texts, the research only focuses on the subject, predicate and existence of three grammatical roles played by entity words for the time being. Therefore, when marking the function of the entity word, it is only necessary to determine whether the entity word in the sentence is the subject or object, and the rest are marked as "existence". In the entity graph model constructed by the research, the process of the grammar role labeling module is shown in Figure 3.4.

From Figure 3.4, the process of the grammar role labeling module first reads the relevant syntactic analysis results of English discourses, and then classifies them. The specific operation steps first traverse the nodes in the dependency syntax tree of the statement. When traversing the entity word, look up the synonym of the entity word, and look up the entity word and its sibling nodes in the corresponding word tree library, and if it exists, it is set to mark it. Then look up all the dependencies of the current entity word, if it contains a noun subject relationship or a clause component subject relationship, mark the entity word as the subject. If there is a relationship with a direct object or an indirect object, mark the entity word as an object, if not, mark it as "existing", and save the marking result. Finally, the syntactic tree traversal is completed, and the result of the entity lexical role tagging of the English text is output.

After the sentence semantic map of the English text is generated, its features need to be analyzed, and then the entire coherence of the English text is analyzed. The English discourse has good coherence, and the discourse must have logic inside. This structure is reflected in the sentence semantic graph as the different connection modes between sentences, that is, the subgraph pattern. On this basis, the research first extracts features such as frequent subgraph frequencies, graph signatures and subgraph semantic values from the graph, and uses these feature values to further study the consistency of the text. On this basis, a coherent quality analysis can be performed. The specific analysis process is shown in Figure 3.5.

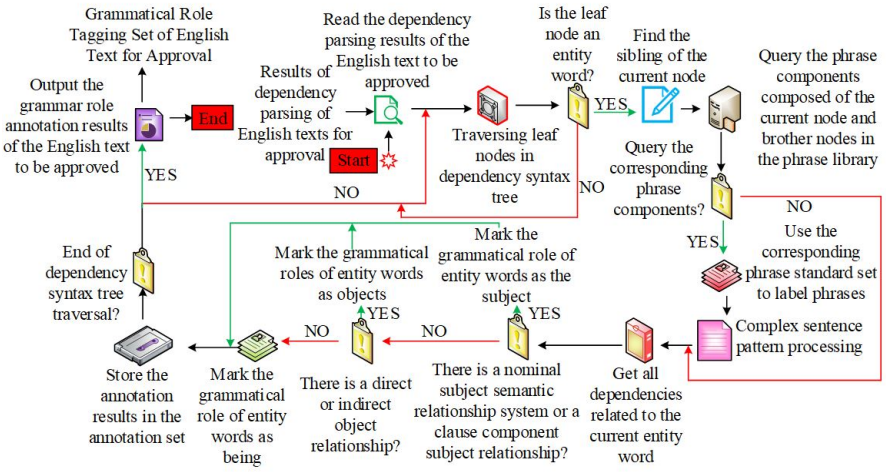



Fig. 3.4: Process of Syntax Role Annotation Module

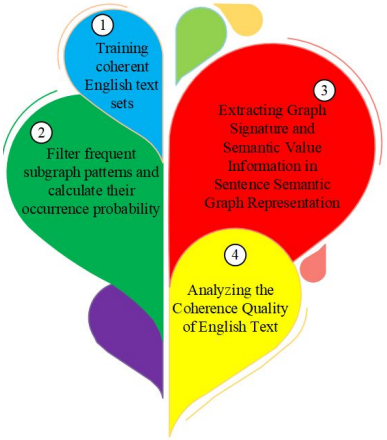
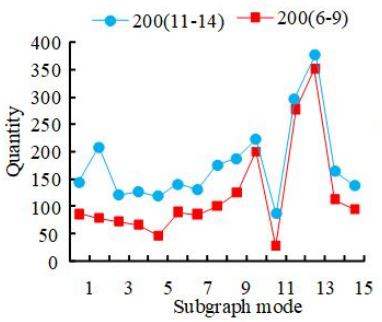



Fig. 3.5: Process of Coherent Quality Analysis

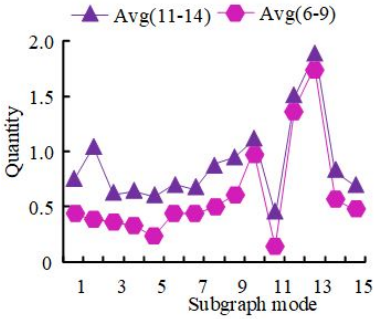
From Figure 3.5, the coherence quality analysis first uses a large number of English texts with good continuity as the training set, and counts the dot subgraphs of all three nodes and four nodes; Concentrate the subgraphs that appear frequently, and calculate the occurrence probability of each frequency subgraph to generate a frequency subgraph model and use it as the continuous subgraph distribution feature of the English text; then extract the graph features and the subgraphs in the English text. Semantic value information; finally, an algorithm is designed to analyze the consistency of English discourse by using the distribution characteristics of frequent subgraphs in the semantic graph of sentences. In addition, in the experimental setup of the English text semantic coherence diagnosis model, the research experimental environment is divided into hardware and software environments. The hardware environment includes model number 880@3.07GHz Intel processor with a memory setting of 16.00GB; The software environment is Microsoft's 64 bit Bitwise operation system. Eclipse development tools, Java programming language and Excel 2016 data analysis tools are selected. The experimental data set contains the International Corpus for Asian English Learners, which has 1.2 million words and a single text length of 200 400 words; The college English textbook corpus contains 930 English articles, all of which are from college English textbooks; The Chinese English Learner Corpus, which contains 1

Table 4.1: Relevant data set before experiment

| **Hardware environment** | | **Software environment** | | | | | |
|---|---|---|---|---|---|---|---|
| Memory | 16 GB | Operating system | Windows 10 64 bit operating system | Programming language | Java Java | | |
| **Corpus of model experiment data** | | | | | | | |
| - | Data _ | - | Data _ | - | Data _ | - | Data _ |
| ICNALE | 1000 articles | COLEN | 100 texts | CELC | 400 compositions | TECCL | 500 test sets |
| Grade (Points) | Criterion of comments | | | | | | |
| 0-7 | The whole passage is abrupt, the semantics between sentences are not necessarily connected, the language is fragmented, and the coherence is poor | | | | | | |
| 7-13 | The overall transition of the essay is poor, the semantic connection between sentences is not close enough, and the coherence is poor | | | | | | |
| 13-20 | The overall transition of the essay is more natural and smooth, the semantic connection between sentences is closer, and the coherence is better | | | | | | |
| 20-25 | The overall transition of the article is natural, smooth, and coherent | | | | | | |




(a) Number of frequent subsets in CELC




(b) Average number of frequent subsets in CELC

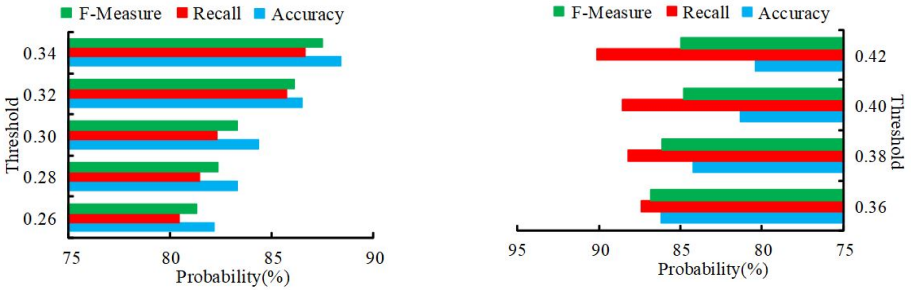
Fig. 4.1: Number of Frequent Subgraphs in Different Test Texts

million words, is a collection of essays written by college English majors and non major students; The Chinese Student English Composition Corpus, which contains 1817335 words, contains 10000 compositions written by Chinese middle school and college students for English reasons.

**4. Evaluation and analysis of the semantic coherence diagnostic model of English texts.** To verify the actual effect of the English text semantic coherence diagnosis model constructed by the research, the research analyzes it through three experiments, namely subgraph screening, incoherent sentence extraction, and sentence sorting. and compared it with the teacher's rating. Before the experiment, the research set up the experimental environment, experimental data and evaluation indicators of the diagnostic model, the contents of which are shown in Table 4.1.

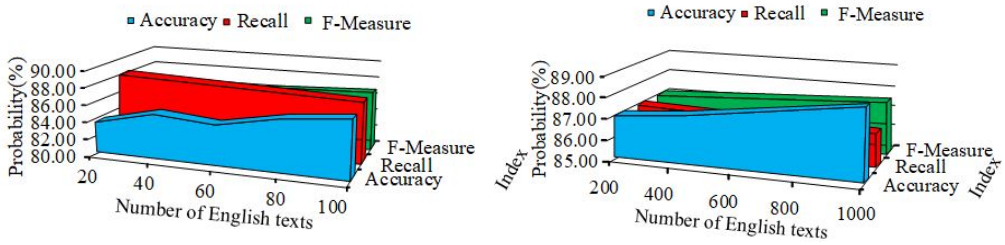
From Table 4.1, the model experimental data selected for the study are 4 known corpora, namely the International Corpus of Asian English Learners (ICNALE) and the Corpus of College English Textbooks. English Textbooks (COLEN), Chinese English Learner Corpus (CELC) and Chinese Students 'English Composition Corpus (TECCL). The study selected 1000 articles in ICNALE as a test set for incoherent sentence extraction, 100 English texts in COLEN as a test set for sentence ordering, 200 essays in CELC with scores of 11-14, and 6 -9 points 200 essays as a test set for frequent sub-atlas and 500 essays in TECCL for comparison experiments with teacher ratings. On this basis, the research first conducts an experimental analysis on the screening of subgraphs, and the results are shown in Figure 4.1.

From Figure 4.1, it can be found that the subgraph patterns 10, 12, and 13 all exceed 200 on the first two




(a) Probability% (b) Probability%

Fig. 4.2: Extraction of Discontinuous Sentences under Different Thresholds




(a) Results of Discontinuous Sentences Extraction in the First Five Different English Texts

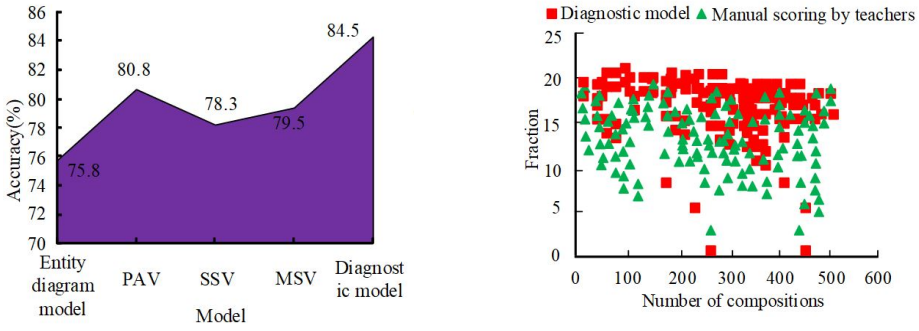
(b) Results of Discontinuous Sentence Extraction in the Last Five Different English Texts

Fig. 4.3: Extraction Results of Discontinuous Sentences in the Number of Non English Compositions

test texts, and exceed 1 on the last two test texts. However, the differences between these three subgraphs on different test texts are not large, and the gaps are all lower than 30 and 0.3. Taken together, most of the spectrograms show large differences in both types of test text, and can distinguish between coherent text and incoherent text very well. But not all subgraphs can capture the coherent information of the text well, such as subgraph patterns 10, 12, 13, etc. There is not much difference between the average number of occurrences of these subgraphs in the test text with high degree of continuity and the test text with discontinuity, and it cannot capture the coherence of the text very well. If it is placed in the frequent subgraph, it will affect the frequency distribution of the frequent subgraph. In addition, the study extracted incoherent sentences to evaluate the constructed coherent diagnosis model, and the experimental results are shown in Figure 4.2.

From Figure 4.2, the precision rate and recall rate are different under different thresholds, and the F1 value (F-Measure) value under the combination of the two is also different. Among them, the highest precision appears at 0.34, which is 88.43%, and the highest recall rate appears at 0.42, which is 90.15%. Taken together, the optimal value of the diagnostic model occurs when the threshold is 0.34, and the F1 value is 87.54%. Too high or too low a threshold will decrease the probability of both indicators, so the optimal threshold for diagnosing the model is 0.34. On this basis, in order to verify the performance of the diagnostic model in extracting incoherent sentences, the study sets the extraction threshold to 0.34. And randomly select a certain amount of essays from the test set, and divide them into ten groups for experiments according to the number. The results are shown in Figure 4.3.

From Figure 4.3, the precision rate, recall rate and F1 value are not very different under different numbers of English compositions, which are maintained between 83% and 89%. Among them, the highest precision




(a) Sentence Sorting Result Under Different Model

(b) Comparison results between diagnostic model and teacher's manual scoring

Fig. 4.4: Sentence Sorting Experiment and Actual Composition Scoring Results

rate is 88.43%, the highest recall rate is 88.20%, and the highest F1 value is 87.54%. On the whole, when the number of English writings is used for testing, the accuracy of the discontinuous sentences extracted by the model still maintains a good stability, and the extraction accuracy of non-coherent sentences also increases with the increase of the number of articles. Therefore, the diagnostic model shows better performance in the actual English incoherent sentence extraction experiment. Finally, the research conducted an experiment on English sentence ordering, and compared the diagnostic model score with the teacher's manual score, and the results are shown in Figure 4.4.

In Figure 4.4(a), the study introduces an entity graph model and three semantic similarity models. The semantic similarity models are a single similar vertex (Single S imilar Vertex, SSV), multiple similar vertices (Multiple S imilar Vertices)) and the preceding adjacent vertex (P receding Adjacent Vertex). As can be seen from the figure, the diagnostic model given by the study is significantly better than the other four models, and the sentence sorting accuracy rate reaches 84.50%. As can be seen from Figure 4.4(b), in the actual English composition scoring, the diagnostic model constructed by the research is generally consistent with the teacher's manual scoring, roughly in the range of 5-25 points. Although there is a big difference in the scores of English compositions, this is a subjective task and will be affected by many subjective factors. For example, there are many mistakes in students' compositions, and teachers' understanding and correction requirements are also different. So some gaps are understandable. In addition, the Pearson correlation coefficient between the calculated diagnostic model and the manual score was 0.6025, indicating a strong correlation, indicating high effectiveness. On the whole, the diagnostic model constructed by the research has better performance in English text recognition, and it also shows high accuracy in the actual composition correction and scoring, and has strong practicability.

**5. Discussion.** The coherence of a discourse is a major indicator of its readability. Excellent discourse does not have a random combination of sentences and vocabulary, but rather has certain logical and coherent rules. In English articles, the brain naturally searches for special ways of logic and coherence. If this way of logic and coherence can be found, then the content of the article can be understood. However, in some large English exams in China, candidates often use texts containing a large number of advanced vocabulary and complex sentence structures in English articles to please the examiners in order to pass. However, when expressing the theme and content of the article, its logic is very chaotic and incoherent. At the same time, current research on automatic grading of English compositions has rarely designed the indicator of coherence quality, and the performance of the actual kernel construction system still needs to be improved. Therefore, an English text semantic coherence diagnosis model was constructed on the basis of sentence semantic graph, and its accuracy and effectiveness were verified through experiments.

The experimental results show that sub Graphical model 10, 12 and 13 exceed 200 in the first two test

texts and 1 in the last two test texts; The accuracy and recall rates are different under different thresholds, and the F1 value (F-Measure) value under the combination of the two is also different. Among them, the highest accuracy occurs at 0.34, which is 88.43%, and the highest recall rate occurs at 0.42, which is 90.15%. This result is consistent with the results of Farkhodovna M S [23]. At the same time, there was no significant difference in accuracy, recall, and F1 values among different English composition numbers, maintaining between 83% and 89%. Among them, the highest accuracy rate is 88.43%, the highest recall rate is 88.20%, and the highest F1 value is 87.54%. This result is superior to Uru OB et al.'s 87.65% [24]. The diagnostic model provided by the study is significantly superior to the other four models, with a sentence sorting accuracy of 84.50%. This result has advantages compared to Keskin D et al [25].

Overall, the diagnostic model constructed in the study has better performance in English text recognition, and it also shows high accuracy in actual essay grading.

**6. Conclusion.** To effectively diagnose the semantic coherence quality of English texts, a diagnostic model is constructed by using the sentence semantic graph, and its performance is experimentally verified. The experimental results show that in the frequent subgraph screening experiment, the subgraph modes 10, 12 and 13 have little difference in different text tests, and the rest all show large differences. Therefore, in order to improve the performance of the diagnosis model, it is necessary to delete these Subgraph; in the incoherent sentence extraction experiment, the highest precision reached 88.43, the highest recall rate reached 90.15%, and the best threshold for the comprehensively calculated F1 value was 0.34, and the probability reached 87.54% at this time; the threshold was fixed at 0.34. In the experiment of extracting incoherent sentences, the precision rate and recall rate were 88.43% and 88.20% respectively under different numbers of English compositions; in the comparison experiment between sentence sorting and actual scoring, the diagnostic model had the highest accuracy rate of 84.50%, and in practice There is not much difference between the scores and the manual scores of teachers, showing high accuracy. In addition, the calculated Pearson correlation coefficient was 0.6025. On the whole, the diagnostic model constructed by the research shows good performance, high accuracy and high reliability in the actual English incoherent sentence extraction experiment and English composition correction. It is worth noting that although the research has introduced semantic information between sentences into the Entitative graph model, the semantic relationship between sentences cannot be well expressed, so its ability to represent the semantic information of the whole text has limitations, which needs further improvement in the future. At the same time, the frequency of nodes that the research mainly focuses on in coherence quality analysis is not comprehensive enough, so it is easy to solve the problem of sparse data. In the future, it is necessary to increase the focus to solve this problem.

REFERENCES

[1] Najafi, E., Valizadeh, A. & Darooneh, A. The Effect of Translation on Text Coherence: A Quantitative Study. *Journal Of Quantitative Linguistics*. **29**, 151-164 (2022)
[2] Akmilia, P., Faridi, A. & Sakhiyya, Z. The Use of Cohesive Devices in. (Conference to Achieve Texts Coherence. English Education Journal, 12(1): 66-74,2022)
[3] Hu, L., Tang, Y., Wu, X. & Zeng, J. Considering optimization of English grammar error correction based on neural network. *Neural Computing And Applications*. **34**, 3323-3335 (2022)
[4] Gondaliya, Y., Kalariya, P., Panchal, B. & Nayak, A. A Rule-based Grammar and Spell Checking. *SAMRIDDHI: A Journal Of Physical Sciences, Engineering And Technology*. **14** pp. 01 (2022)
[5] Chaabi, Y. & Allah, F. Amazigh spell checker using Damerau-Levenshtein algorithm and N-gram. *Journal Of King Saud University-Computer And Information Sciences*. **34**, 6116-6124 (2022)
[6] Li, X., Li, X., Chen, S., Ma, S. & Xie, F. Neural-based automatic scoring model for Chinese-English interpretation with a multi-indicator assessment. *Connection Science*. **34**, 1638-1653 (2022)
[7] Chen, J., Zhang, L. & Parr, J. Improving EFL students' text revision with the self-regulated strategy development (SRSD) model. *Metacognition And Learning*. **17**, 191-211 (2022)
[8] Srivastava, K., Dhanda, N. & Shrivastava, A. Optimization of Window Size for Calculating Semantic Coherence Within an Essay. *ADCAIJ: Advances In Distributed Computing And Artificial Intelligence Journal*. **11**, 147-158 (2022)
[9] Yang, X., Zhang, J., Chen, K., Zhang, W., Ma, Z., Wang, F. & Yu, N. . *Tracing Text Provenance Via Context-aware Lexical Substitution//Proceedings Of The AAAI Conference On Artificial Intelligence*. pp. 11613-11621 (2022)
[10] Aminovna, B. Importance of coherence and cohesion in writing. *Eurasian Research Bulletin*. **4**, 83-89 (2022)
[11] Saleh, M. & Bharati, D. The Use of Cohesive Devices in Descriptive Text by English Training Participants at PST". *English Education Journal*. **12**, 95-102 (2022)

[12] Dassanayake, N. Exploring Coherence among Sri Lankan CFL Learners in Chinese-English Translation: Decoding and Interpreting of Culture-loaded Content. *International Journal Of Language And Literary Studies*. **4**, 350-363 (2022)
[13] Akmilia, P. & Faridi, A. Sakhiyya. (Z. The Use of Cohesive Devices in,2022)
[14] Abdusalomovna, K. Theoretical background of using cohesion in discourse. *Web Of Scientist: International Scientific Research Journal*. **3**, 687-693 (2022)
[15] Zhang, K., Zhou, Y., Chen, Z. & Others (2022) Incorporating Biterm Correlation Knowledge into Topic Modeling for Short Texts. *The Computer Journal*. **65**, 537-553 (0)
[16] Linnik, A., Bastiaanse, R., Stede, M. & Khudyakova, M. Linguistic mechanisms of coherence in aphasic and non-aphasic discourse. *Aphasiology*. **36**, 123-146 (2022)
[17] Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R. & Gallinari, P. Controlling hallucinations at word level in data-to-text generation. *Data Mining And Knowledge Discovery*. **36**, 318-354 (2022)
[18] Gaur, M., Gunaratna, K., Srinivasan, V. & Jin, H. . *Iseeq: Information Seeking Question Generation Using Dynamic Meta-information Retrieval And Knowledge Graphs//Proceedings Of The AAAI Conference On Artificial Intelligence*. pp. 10672-10680 (2022)
[19] Ruzikulovich, A. Functional-semantic and linguo-poetic capabilities of imperative structures. *EPRA International Journal Of Multidisciplinary Research (IJMR)*. **8**, 219-221 (2022)
[20] Zhao, L. & Xu, J. A Study on the Translation Strategies of The Nine Songs from the Perspective of Cognitive Construal: A Comparative Analysis of the Yangs' and Waley's Versions. *International Journal Of Linguistics, Literature And Translation*. **5**, 56-62 (2022)
[21] Wang, Z. & Wang, J. The Grammatical and Semantic Functions of "with" Structure in Chinese-English Translation. *International Journal Of Linguistics, Literature And Translation*. **5**, 109-116 (2022)
[22] Feng, Y., Hu, C., Kamigaito, H. & Others (2022) A simple and effective usage of word clusters for CBOW model[J]. *Journal Of Natural Language Processing*. **29**, 785-806 (0)
[23] Farkhodovna, M. THE STRUCTURE OF PROVERBS AND PHRASEOLOGICAL UNITS IN ENGLISH. *IJTIMOIY FANLARDA INNOVASIYA ONLAYN ILMIY JURNALI*. **3**, 83-87 (2023)
[24] Uru, O., Sudirman, A. & Nugroho, A. Exploring cohesions in EFL academic writing: A state of the art on the study of cohesions. *Elsya: Journal Of English Language Studies*. **3**, 141-149 (2021)
[25] Keskin, D. & DEMİR, B. The role of theme and rheme in thematic progression patterns in English argumentative essays by Turkish University students. *Edu*. **10**, 64-82 (2021)

*Edited by:* Mudasir Mohd
*Special issue on:* Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
*Received:* May 16, 2023
*Accepted:* Nov 1, 2023