A Coherence Analysis Model for English Essay Based on Sentence Semantic Graph

Guimin Huang, Hui Tang, Jiahao Wang, Xuan Ye

School of Computer Science and Information Security, Guilin University of Electronic Technology, China

sendhuang@126.com, gmhuang@guet.edu.cn

Abstract. Aiming at the problem that cannot be solved by the entity-based graph model, we propose a new discourse coherence quality analysis model (sentence semantic graph) by merging entity graph and semantic similarity graph of the text. Firstly, we use an improved coreference resolution module to process the coreference phenomenon in the text, and build a new type of entity graph with the processed results; secondly, we merge the entity graph and the semantic similarity based on the semantic space and get our sentence semantic graph lastly, we analyze the average outdegree of sentence semantic graph to indicate the degree of text coherence. Experiments show that this model is superior to the entity-based graph model and has a good effect on the automatic evaluation of English essay.

1. Introduction

In recent years, artificial intelligence and its subdivisions have developed rapidly, bringing tremendous changes to all areas of people's lives. In the field of English education, people are beginning to use computers to assist teaching. Among them, there is a huge demand for automatic correction of students' English essay, and the coherent quality of the text, as an important basis for manual scoring standards, is an important need for automatic correction systems. Therefore, how to quantify the coherence quality of a text has become a hot topic among researchers.

2. Related Work

In the research of the researchers, a very classic model for evaluating text coherence is the entity grid model, which is based on the centering theory proposed by Barzilay & Lapata[1,2]. It extracts the common entities of adjacent sentences in the text, and counts the frequency of grammatical role conversion of the same entities in different sentences to measure the coherence of the text. But it belongs to a supervised model, which is affected by data sparsity and domain dependence. And it can only analyze the coherence between adjacent sentences, which belongs to local coherence. In order to avoid these shortcomings of the entity grid model, Jiwei Li[3] proposed a neural network model for coherent tasks based on distributed sentence representation. This method uses a recurrent or recurrent neural network to automatically learn the syntactic symbol representation of sentences. A three-layer neural network based on sentence sliding window is constructed to analyze the text coherence. This model structure avoids the need for feature engineering, but cannot capture the long-distance dependence of sentences. Lin et al.[4] proposed a textual relationship model on the basis of the theory of rhetorical structure. The method specifically focuses on the discourse relation transitions between adjacent sentences and models them in a discourse role matrix. Guinaudeau and Strube[5] proposed an entity-based unsupervised graph model. In this model, it allows us to associate non-adjacent sentences
in the text to reflect global consistency. The model converted the text to a two-layer graph structure of sentences and entities, and then projects the bipartite graph in three different ways and the degree of coherence of the text is expressed by analyzing the average degree of the sentence graph.

However, in our actual writing, the coherence between sentences is not all formed by describing the same entity or an obvious rhetorical relationship. There are also many deep semantically coherent sentences, such as Fig 1.

![Fig.1 An example of sentences](image)

From Fig 1, we can see that S1 and S2 are semantically coherent, but there is no same entity or obvious rhetorical relationship between them, so the model mentioned before cannot accurately analyze them. In order to solve this kind of problem, Gotama et al.[6] proposed another graph model, which maps sentences in the text into the semantic space, expresses sentences as sentence vectors, and calculates the semantic similarity between sentences separately. According to the similarity between sentences, it uses three different methods to construct the text into a semantic similarity graph, and represents the text coherence through the features of the graph. Although the model can analyze this type of sentences, it analyzes them only based on the semantic similarity between the sentences. This is not comprehensive. It ignores the importance of entity words in sentences. In allusion to the deficiencies of the above model existing, we design and implement a sentence semantic graph model by merged entity graph and semantic similarity graph,

3. Sentence Semantic Graph Model

In our model, we use an improved coreference resolution module to deal with the coreference phenomenon in the text, and then construct an entity graph and a semantic similarity graph separately. By merging these two graph models, we get our Sentence semantic graph model. Finally, we analyze the graph features of sentence semantic graph to analyze text coherence.

3.1. Entity Graph

In Guinaudeau and Strube[5]’s approach, text is converted to a graph of sentences and entities and then projected as an entity graph. In this paper, we use a coreference module to resolve the coreference phenomenon of the sentences in the text, and then construct our entity graph through the obtained entity chain, which is distinguished from Guinaudeau and Strube[5]’s method.

In our regular English writing, the most commonly used method is to use pronouns to maintain the continuity of the sentence. While in natural language processing, this kind of coreference phenomenon will bring us a lot of difficulties in analysis of the semantic information of the text. In order to analyze the text correctly, we need to eliminate these reference phenomena. In this paper, we have improved the coreference resolution method proposed by Soon et al.[7]. At first, we use Stanford NLP toolkit to preprocess the text for word segmentation clauses, part-of-speech tagging, syntactic tree analysis, and named entity recognition and so on. Next, we extract the markers in the text by using the preprocessing results, which contain coreference information and entity words syntax information and so on. Fig 2 shows the main entity words extracted from a coherent text. In order to make it be more suitable for students' English essay, we enriched Soon's feature set with some new characteristics like the wordnet similarity which get from WordNet2.1. By adding more feature sets, we can extract more common reference information. Last, we trained the module using a corpus containing abundant reference phenomena, and obtained a coreference analysis module.
S1: [Yesterday], [my whole family] went to [the lake] for [a picnic].
S2: [The scenery] around [the lake] is very beautiful.
S3: [The water] sparkled in [a gentle breeze], and [fish] jumped out of [the water] now and then.
S4: In [this charming scenery], [our whole family] had [a wonderful day].

Fig 2. A text as an example

Table 1. The coreference chains of the text in fig 2

| Coreference Entity Chains | Mentions (Mention ID) |
|---------------------------|-----------------------|
| Chain                     |                       |
| 1                         | yesterday (1)         |
| 2                         | my whole family (2); our whole family (12) |
| 3                         | the lake (3); the lake (6) |
| 4                         | a picnic (4)          |
| 5                         | the scenery (5); this charming scenery (11) |
| 6                         | the water (7); the water (10) |
| 7                         | a gentle breeze (8)   |
| 8                         | fish (9)              |
| 9                         | a wonderful day (13)  |

By using the coreference resolution module, we can form coreference entity chains with coreferential relations in the text. Table 1 shows the coreference entity chains of fig 2. We encode a text into a graph \( G (V, E) \) with those coreference entity chains, where \( V \) is a set of vertices and \( E \) is a set of edges in the graph. Each vertex represents a sentence, and each edge represents the entity relationship between them. When the two entities in the entity chain come from different sentences, we establish an edge between these two sentences. If multiple entity words of the entity chain appear in a sentence, the one with the highest importance is selected as the syntactic role of the entity word chain in the sentence. We set different weights for the syntactic role of each entity word (subject (S): \( \lambda_1 \), object (O): \( \lambda_2 \), other (X): \( \lambda_3 \), \( 1 > \lambda_1 > \lambda_2 > \lambda_3 > 0 \)). For the role of syntactic nothing (-), we do not take it into consideration. The formula for calculating the weight of an edge is determined by the grammatical role and distance of the entity in the sentence.

\[
weight_{\text{entity}}(v_i, v_j) = \frac{\sum_{e \in E_{ij}} (ew(e,i) \times ew(e,j))}{|j - i|}
\]  

(1)

Where \( ew(e,i), ew(e,j) \) is the weight of the grammatical role of entities in sentences \( j \) and \( i \), and \( E_{ij} \) is the collection of entities jointly owned by sentences \( i \) and \( j \). The text of fig 2 can be formed into an entity graph in the fig 3(a). We can see that s2 and s3 are coherent, but the entity graph cannot analyze it. We should find more semantic information of the two sentences.

Fig 3 Three different sentence graphs of the text in fig 2
3.2. Semantic Similarity Graph

Text coherence actually belongs to the semantic category. When the coherence cannot be analyzed from the perspective of sentence surface semantic information (such as entity repetition, coreference), we must obtain deeper sentence semantic information. And the neural network-based word embedding method can help us solve this problem.

In our Semantic Similarity Graph, we adopt the pre-trained Glove word embedding model of Pennington et al. (2014)[8], and express the sentence in the form of vector. A sentence is composed of multiple words. We map each word into a vector space and represent it with a 300-dimensional vector. By averaging the word vectors which are composed of sentences, we can encode the sentence into a vector form. The sentence vector $S$ is described as

$$ S = \frac{1}{m} \sum_{i=1}^{m} W_i $$

Where $m$ is the number of words in the sentence $S$, $W_i$ is a vector representation of the word. In the semantic space, the more similar sentences are, the closer they are. Therefore, we use the ratio of the cosine similarity between two sentence vectors to the sentence distance to express their semantic similarity.

$$ \text{similarity}(s_i, s_j) = \frac{\cos(s_i, s_j)}{|j - i|} $$

Where $\cos(s_i, s_j)$is the cosine of the sentence vector $s_i$ and $s_j$. We construct a semantic similarity graph by comparing the semantic similarity between sentences. When creating an edge from the vertices of a sentence, we set a threshold $\emptyset$. When the similarity between two sentences is greater than the threshold $\emptyset$, we create an edge between the two sentences. In this way, we can represent the text in Figure 1 as a semantic similarity graph, as shown in Figure 3(b).

3.3. Sentence Semantic Graph

As mentioned above, we know that neither the entity graph model nor the semantic similarity graph model can analyze the coherent form of text very well, and the entity graph model cannot analyze the coherent sentences with high semantic similarity. And semantic similarity graphs cannot distinguish sentences containing the same entity well, so we merge the two graphs. The fusion of the two graphs is mainly the fusion of edges, so that the weight of edges contains both entity information and semantic information. The formula for calculating the weight of the edge of the fusion graph is as follows:

$$ \text{weight}(i, j) = \begin{cases} \lambda_1 \omega_e(i, j) + \lambda_2 \omega_e(i, j) & \omega_e(i, j) = 0 \\ \omega_e(i, j) & \omega_e(i, j) = 0 \end{cases} $$

where $\omega_e(i, j)$ is the edge weight of the semantic similarity graph and $\omega_e(i, j)$ is the edge weight of the entity graph, $\lambda_1$ and $\lambda_2$ are the parameter. In this way, we can represent the text as a sentence semantic graph. Fig 3(c) shows the sentence semantic graph of the text in fig 2.

From this sentence semantic graph representation, the coherence of a text $T$ can be measured by computing the average outdegree of the graph $G$. Because the average out-degree allows us to evaluate the extent to which a sentence is related to other sentences in the text in terms of semantic information.

$$ \text{Coherence}(T) = \frac{1}{m} \sum_{i=1}^{m} \text{OutDegree}(v_i) $$

Where $\text{OutDegree}(v_i)$ is the sum of the weights associated to edges $v_i$ and $m$ is the number of sentences in the text.

4. Experiments

In order to evaluate our model, we conducted a series of experiments on Chinese Learner English Corpus (CLEC)[10]. First, in order to evaluate our model's ability to distinguish coherent text, we conducted a sentence ranking experiment and applied our model. Compared with the entity graph
model and the semantic similarity graph model. Then, in order to evaluate the feasibility of applying our model to the automatic correction system, we conducted an incoherent sentence extraction experiment. The experimental results prove that our model works well in the automatic correction system.

4.1. Sentence Order
In the first part of the experiment, we selected 100 articles from the Chinese English Learner Corpus as the test set. We shuffle the sentence order of each article to generate 10 disordered articles, and then use the original article and one disordered article as a set of examples. There are 1000 test samples in total. We compared the three kinds of the entity graph models of Guinaudeau et al[5]. with the semantic similarity graph model and the sentence semantic graph model on the test set.

| Model                  | Acc(%) | F-measure |
|------------------------|--------|-----------|
| Entity Graph           |        |           |
| P_U                   | 72.80  | 72.80     |
| P_W                   | 73.50  | 73.50     |
| P_Acc                 | 73.80  | 73.80     |
| Semantic Similarity Graph |    |           |
|                       | 75.20  | 75.20     |
| Our Model             | 78.50  | 78.50     |

From the experimental results in Table 2, we can find that the performance of the entity graph model in the field of student essay is not so good. The main reason is that, on the one hand, each article in the original paper is shuffled into 20 rankings, and the accuracy will be higher; on the other hand, there is the difference between student essays and news reports. In contrast, our model has a certain improvement in accuracy and F-measure, which shows that our model is more able to distinguish student papers and has better results.

4.2. Extract Incoherent Sentence
We selected 200 student essays from the Chinese English learner corpus as the test set of our model. The articles in this corpus are all from the test essay of Chinese students. They are highly representative and conform to the field in which our model will be applied. We insert 5 incoherent sentences in each article, a total of 1000 sentences, and the insertion position of each sentence is random. The range of content of these 200 articles is wide, so the 5 sentences we selected are concise statement. We reproduced the entity graph model of Guinaudeau et al[5]. and compared it with the model in this article. The experimental results are shown in Figure 4.

As shown in the figure, in different composition articles, our sentence semantic graph model is better than the entity-based entity graph model, and the accuracy rate gains an improvement by about 8% on average, which indicates that the merger of sentence entity information and the semantic graph model of sentences with semantic information can more accurately identify the incoherent sentences in the composition, which is helpful for students' learning improvement.
5. Conclusion
In order to evaluate the coherence of text, this paper proposes a text coherence analysis model based on sentence semantic graph. In order to improve the performance of the model, on the one hand, we use an improved coreference analysis algorithm to deal with the phenomenon of reference in sentences. According to the processed results, we constructed an entity graph model. On the other hand, we merged the entity graph and the semantic similarity graph composed of the semantic similarity between sentences to obtain a sentence semantic graph model containing sentence entity information and semantic information. The coherence of the text is analyzed based on the average outdegree of sentence semantic graph. The experimental results show that compared with the entity-based entity graph model, our model has certain improvements in two evaluation tasks. Especially in the task of extracting incoherent sentences, the accuracy of our model can reach 93%, which means that our model is applied to the composition evaluation system, which can provide some help for improving students' English writing level.

Acknowledgements
This work was supported by the National Natural Science Foundation of China (No. 61662012, No.61662015).

References
[1] M. Lapata, R. Barzilay. Automatic evaluation of text coherence: models and representations, 1085-1090(2005).
[2] Barzilay, Regina, and M. Lapata. "Modeling Local Coherence: An Entity-Based Approach." Computational Linguistics 34(2008): 1-34.
[3] Li, Jiwei, and E. Hovy. "A Model of Coherence Based on Distributed Sentence Representation." Conference on Empirical Methods in Natural Language Processing 2014.
[4] Z. Lin, H.T. Ng, M.Y. Kan. Automatically evaluating text coherence using discourse relations, 997-1006 (2011).
[5] Guinaudeau, Camille, and M. Strube. "Graph-based Local Coherence Modeling." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2013.
[6] Gotama, Putra Jan Wira, and T. Tokunaga. "Evaluating text coherence based on semantic similarity graph." 2017 Workshop on Graph Based Methods in Natural Language Processing, Annual Meeting of Association for Computational Linguistics (ACL) 2017.
[7] Soon, Wee Meng, H. T. Ng, and D. C. Y. Lim. "A Machine Learning Approach to Coreference Resolution of Noun Phrases." Computational Linguistics 27.4(2001):521-544.
[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pages 1532–1543.
[9] Xu, Fan, et al. "An entity-driven recursive neural network model for chinese discourse, coherence modeling." International Journal of Artificial Intelligence & Applications(2017).
[10] G. Shichun and Y. Huizhong, Chinese Learner English Corpus. Shanghai, CA: Shanghai Foreign Language Education Press, 2003.