RST-based Discourse Coherence Quality Analysis Model for Students' English Essays

Guimin Huang, Min Tan[a], Zhenglin Sun and Ya Zhou

School of Information and Communication Engineering, Guilin University of Electronic Technology,Guilin,China

Abstract. Against the problems which can't be solved by the word-level based local coherence analysis model, we propose a new discourse coherence quality analysis model (abbreviated RST-DCQA) by analyzing the full hierarchical discourse structure of English essays. Under the framework of rhetorical structure theory (RST), firstly, we design an RST-style discourse relations parser to capture the deep hierarchical discourse structure of essays; secondly, we transform the discourse relation information into a discourse relation matrix; finally, we design an algorithm to analyze the discourse coherence quality of student's English essays. The experimental results show that the average error of our model's score and teacher's score is only 2.63, and the Pearson correlation coefficient is 0.71. Compared with the other models, our RST-DCQA model has a higher accuracy and better practicality in the field of students' essays assessment.

1 Introduction

In recent years, due to the continuous development of research in artificial intelligence and its subdivided fields, many machine output results will be provided to people. If the quality of the coherence is poor, people's reading experience will be seriously affected, and even the results may not be available. Therefore, considerable attention has been paid to the studies on how to quantify discourse coherence and its practicality. Among the researchers' studies, the most typical ones are the latent semantic analysis (LSA) method proposed by Foltz et al. [1] and the entity grid model proposed by Barzilay & Lapata[2,3]. In the LSA method, each word or sentence is represented by a vector, and the similarity between the two words or sentences is measured by the cosine similarity. However, LSA is a more mathematical approach which leads to many shortcomings, such as the poor interpretability, inability to deal with polysemy phenomena, ignoring the sequence of words and so on. Hence, many scholars have turned their research into the entity grid model which has a powerful theoretical basis and a better interpretability.

Inspired by the Centering theory [4], Barzilay & Lapata proposed an entity grid model which can analyze the original text automatically by building the entity grid. The algorithm has better portability and scalability, so many scholars have put forward their own improvement methods for the problems that apply in different fields and the model itself existing. Guinaudeau and Strube [5] converted text to a graph of sentences and entities and evaluated the quality of the text by calculating the average of the out-degrees of the entire graph. Muyu Zhang et al. [6] analyzed the semantic relevance between entity words through the knowledge base. However, these above-mentioned word-level models are unable to find the phenomena of discourse coherence through explicit or implicit discourse relations. The following sentences is an example:

[The villagers think that this road needs to be rebuilt.]S_1 [But everyone has no idea on how to do it.]S_2

These sentences are two coherence sentences that are connected by means of an explicit discourse relation—*But*. However, thanks to these two sentences do not contain relevant entities. If the existing entity grid model is used to coherence assessment, the result will be incoherent. Aiming at this problem, Lin et al. [7] use an end-to-end Penn Discourse Treebank style (PDTB-style) discourse method to encode the discourse relations of text. And they proved that the discourse relations can help better capture the discourse coherence. However, PDTB-style method only encodes very shallow discourse structures, the relations are mostly locality and adjacency, within a single sentence or between two adjacent sentences. They can't find the discourse relations existing in higher levels. In allusion to the deficiencies of the above model existing, combined with the RST we design and implement a discourse coherence quality analysis model (RST-DCQA).

In this paper, the RST-DCQA model is described in Section 2. The Section 3 introduces the model training corpus with 20,000 coherent English essays. Section 4 shows the two experiments on the model.

2 RST-DCQA Model

The theory of rhetorical structure (RST) was first proposed by American scholar Mann & Thompson [8]. It is a set of theory about the description of natural discourse structure. The research of RST began with the connection relations of the clauses, gradually transitioned to natural passages of various lengths and complete texts. During the research

[a] Corresponding author: 1098382665@qq.com

process, Mann & Thompson found that both the clauses and the larger discourse units are connected by some few, recurring discourse relations. Based on the above, we design the RST-DCQA model by capturing the full hierarchical discourse structure of text. In the following part of this section: firstly, we introduce the discourse relation tree which is generated by the parser; secondly, we describe the discourse relation matrix; finally, we analyze the discourse coherence quality of essays by the RST-DCQA model.

2.1 Discourse relation tree

In the framework of RST, the text can be represented as a discourse relation tree. Its leaves are clauses and RST consider these clauses are the basic text units which called Elementary Discourse Units (EDUs). The adjacent leaf nodes can be related by discourse relations, forming a discourse relation subtree, and then associated with other neighbor nodes to form a higher-level discourse relation tree until a complete discourse tree generated. In RST-style relation structure, RST adopts a “nucleus-satellite” structure where “nucleus” is the core sentence which plays an important role in the text, and “satellite” is subordinate to the core sentences.

Based on the previous work of Surdeanu et al. [9], we design a parser by following the architecture introduced by Hernault et al. [10]and Feng & Hirst [11] to parse the text deeply. The parser first uses an independent identically distributed classifier to split the text into EDUs. Then the parser uses a bottom-up pattern to construct the discourse relation tree through two classifiers: the one detects which adjacent EDUs can be associated together; the other one is used to mark the discourse relations between associated EDUs. As for the features of the parser, we build on the work of Joty et al. [12] and extend it. We utilize a richer feature space which uses the Stanford Toolkit to syntactic parse and coreference resolution, and implements each syntactic feature using both constituent and dependency syntax. Compared with other parsers, this parser can provide more detailed and accurate parsing results. Figure 2 shows the discourse relation tree for the text in Figure 1. Each leaf node e_i in Figure 2 represents an EDU, each EDU being linked together by the rhetorical discourse relations (*Summary, Temporal and contrast, etc.*); the vertical lines represent and the arrows point to the “nucleus”. For example, in Figure 2, e_1 and e_2 are related by the discourse relation *Temporal*, with e_1 is the core unit— “nucleus”, e_2 is the “satellite”. And the units *(e_1-e_2)* and *(e_3-e_7)* related by *Summary* on higher levels.

S1: [The dollar finished lower yesterday,]$\boldsymbol{e_1}$ [after tracking another rollercoaster session on Wall Street.]$\boldsymbol{e_2}$
S2: [Concern about the volatile U.S stock market had faded in recent sessions,]$\boldsymbol{e_3}$ [and traders appeared content to let the dollar languish in a narrow range until tomorrow,]$\boldsymbol{e_4}$ [when the preliminary report on third-quarter U.S gross national product is eased.]$\boldsymbol{e_5}$
S3: [But seesaw gyrations in the Dow Jones Industrial Average yesterday put Wall Street back in the spotlight]$\boldsymbol{e_6}$ [and inspired participants to bid the U.S unit lower.]$\boldsymbol{e_7}$

Figure 1. An example text.

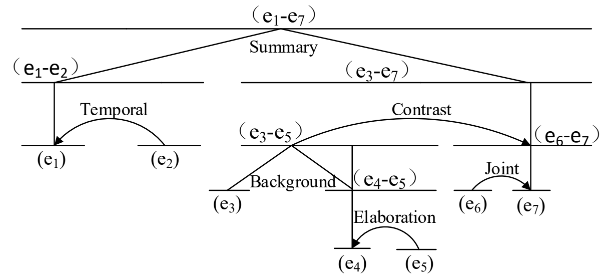

Figure 2. The RST discourse relation tree

2.2 Discourse relation matrix

In order to capture discourse coherence information from the discourse relation tree, we use a series of entities with related discourse relations to encode the essay. And in the course of the study, we found that the relations of coherent essay will follow a certain pattern. But using only these patterns to evaluate the coherence of the essay will lead to feature sparseness. To resolve this, we transform these relation sequences into a discourse relation matrix. We encode RST-style discourse relation in a similar PDTB-style encoding. The parameter information in the encoding style of PDTB (Arg1 and Arg2) is replaced by the RST “Nucleus – Satellite” information. Table 1 is a discourse relation matrix generated by the short text in Figure 1. The columns of the matrix represent the entities, the rows represent the location information, and each unit <*Ei,Sj*> represents the discourse relation set of entity *Ei* in sentence *Sj*. For example, the entity *wall* appears in the sentence S_1 and forms a time relation in the context, so the cell <*wall*, S_1> represents the discourse relation *Temporal.S*. A cell maybe empty (denoted by *nil*, such as in <*product, S1*>) or have multiple discourse roles (as in <*trader, S2*>). *Temporal.S ->nil* represents a length-2 discourse relation transition sub-sequence.

Table 1. The discourse relation matrix

	wall	**product**	**gyrations**	**yesterday**	**stock**	**participation**	**trader**
S1	Temporal.S	nil	nil	Temporal.N Summary.S	nil	nil	nil
S2	nil	Elaboration.S	nil	nil	Backgruund.S	nil	Elaboration.N Background.N Contrast.S
S3	Summary.N Contrast.N	nil	Summary.N Contrast.N	Summary.N Contrast.N	nil	Joint.S	nil

2.3 Discourse coherence quality analysis

Most of the existing models are based on the method of coherent quality analysis of support vector machine [13]. Only essays can be classified into discourse coherent and incoherent, and the coherence quality of essays cannot be quantified. We find that the discourse relation transitions are captured locally in each EDU, but all the EDUs in the whole context converge the possibility of the discourse relation transitions. And the overall distribution of discourse relation transitions is different in coherent and incoherent texts. Therefore, we treat the discourse relation matrix as a distribution of transformation types and capture the discourse relation transition information in it to analyze an unknown essay's discourse coherence quality. The main steps are as follows:

Firstly, a large number of texts with good coherence are used as the training corpus, and the coherent texts' distribution features of discourse relations and discourse relation transitions are captured in training. Secondly, extract the discourse relation and discourse relation transition information of the essay which need to be analyzed. Finally, combined the rhetorical structure theory with Barzilay and Lapata's grid computing method, based on the joint probability distribution and the first-order Markov model, an algorithm is designed to calculate the discourse coherence quality. The *DCQA-SCORE(E)* indicates the discourse coherence quality of the essays *E*.

$$DCQA-SCORE(E)=\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\log P(r_{i,j}\,|\,r_{(i-h),j}...r_{(i-1),j})}{m\times n}$$

where *h* is the length of the history mode, $r_{i,j}$ is the discourse relation of the entity *j* in the sentence *i*. $P(r_{i,j}|r_{1,j}...r_{(i-1),j})$ can be obtained from the training corpus. Since the probability of discourse role transition occurrence may decrease with the longer of the discourse, we take the number of sentences *m* and the number of entities *n* to normalize making the value of *DCQA-SCORE(E)* between 0 to 1.

3 RST-DCQA Model training corpus

When we were doing model training, we found that the current corpus which can be used for discourse coherence quality training are small. And the number of training texts is positively related to the performance of the model, so we need to build a larger training corpus. For the training texts, we have collected about 15,000 articles from Middle school English textbooks, college English textbooks, and professional English textbooks, and 20,000 articles from news of Reuters, a well-known western news media. On the one hand, we believe that the articles of teaching materials are coherent. On the other hand, Reuters is one of the top four news agencies in the West. The news content provided by Reuters are rich and colorful. We use Reuters news as training set because as an internationally renowned news media their news content's quality of the coherence is unquestionable. From the experiment results (Figure 3), we see that as the number of training texts increases, the accuracy of the model also increases. However, when the number of texts increases to 10000, the accuracy rate begins to increase very slowly and reaches its peak (72.31%) on 20,000. Hence, we separately select 10000 articles from the textbook corpus and Reuters corpus to form a training corpus which contain 20,000 articles. Each article in the training corpus will be preprocessed. After removing unnecessary tags and generating plain text, the discourse relation and discourse relation transition distribution characteristics in each article will be captured.

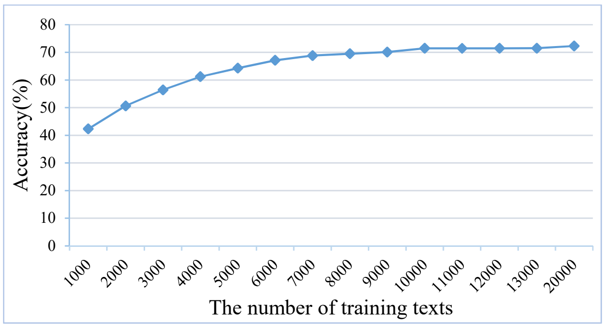

Figure 3. The change of accuracy with increased training texts

4 Experiments

To evaluate the RST-DCQA model, we conduct two parts of experiments on two different test sets. Firstly, we test the performance of the RST-DCQA model and compare with other three discourse coherence analysis models on the Chinese English Learner Corpus (CELC). Secondly, we use the Ten-thousand English Compositions of Chinese Learners (TECCL) corpus to evaluate effectiveness in actual use of the RST-DCQA model by correcting students' English essays.

In the first part of the experiment, we selected 150 articles from the CELC as our test set. The essays in this corpus are all from Chinese students' exam essays. They are highly representative and in line with the areas of our model to be applied. And the essays in the corpus contain the teachers' correcting scores (0-15 points). In the experiments, the RST-DCQA model will compare with the entity grid model, latent semantic analysis algorithm and Lin's PDTB-style method on the test set. Finally, the Pearson correlation coefficients of the four different models are compared and analyzed. Pearson correlation coefficient is a quantity that is widely used to evaluate the similarity between machine generating results and artificial results. The score of this coefficient is closer to 1, the higher the degree of correlation between the two scores, whereas the score is closer to 0, the higher the degree of uncorrelation between them. The experimental results are shown in Table 2.

According to the experimental results, we can find that the performance of word-level based coherence analysis model—the entity grid model and the LSA algorithm on this test is relatively poorer than the discourse relation-based model—Lin's PDTB method and the RST-DCQA model. Their Pearson correlation coefficients are only 0.2159 and 0.3394. The result of Lin's PDTB-style method which uses the shallow discourse relation encoding is 0.5685 that is better than the entity grid model and LSA. In addition, from the results we can find the RST-DCQA model's correlation reaches 0.7147 which is higher than the other three models. It's closer to the manual scores, proving that our model in the student's English essays scoring have a higher accuracy.

Table 2. The Pearson correlation coefficient of the four models

	Entity Grid	**LSA**	**PDTB**	**RST-DCQA**
Manual	0.2159	0.3394	0.5685	0.7147

In the second part of the experiment, the test set is TECCL corpus which are collected from the online writing platform. At the same time, we invited four professional English teachers to score the discourse coherence quality of 1000 essays in the TECCL corpus according to the CET4(College English Test 4) and CET6 test scoring standard. Then take the average score of the four teachers scoring as the essays' teacher score. As a part of the essays' score, the discourse coherence quality accounts for 25% of the essays' full score and the full score of essay is 100 points. Therefore, we convert the score of discourse coherence ranging from 0 to 25. We use scatterplots to show the distribution of teacher score and our RST-DCQA model score. The experimental results are shown in Figure 4. In the figure, the blue circles are teacher score and the yellow forks represent the RST-DCQA model score.

From the distribution of scores, there are some differences between teacher and RST-DCQA model. Since the scoring standard is generalized and discourse coherence quality assessment itself is subjective, even different teachers score differently on the same essay, the difference is unavoidable and acceptable. And due to the strictness of teacher's correction, the model score is generally higher than teacher score. After that we calculated the error scores of teacher and model for each essay and the average of the error scores is 2.6315 which means the difference between our model score and teacher score is within 3 points. And the Pearson correlation coefficient between them reaches 0.71, which is a strong correlation. It indicates that the accuracy rate of the discourse coherence quality analysis of the RST-DCQA model is very high, and the result is credible.

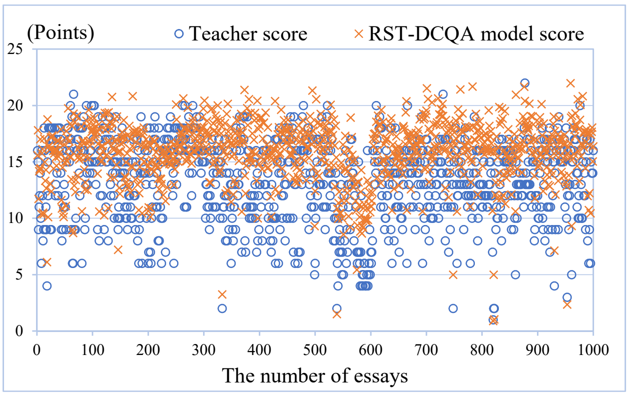

Figure 4. The scatter plots of the Teacher score and RST-DCQA model score

5 Conclusion

Starting from the rhetorical structure theory, this paper proposes a discourse coherence quality analysis（RST-DCQA） model by capturing the discourse structure information of essays in-depth. For improving the performance of the discourse coherence model, on the one hand, we capture the full hierarchical discourse relation information of essays by adopting the bottom-up parsing mode and enriching the feature space, which solves the problem that the parser of the existing model can only capture discourse relations in local context. On the other hand, we combine the discourse relations with the entity grid to construct the discourse relation matrix, achieving a coherence analysis method based on sentences and paragraphs instead of the traditional word-level. What's more, we capture the distribution characteristics of discourse relation information in coherent texts through extensive training. When applied to students' essays assessment to analyze the coherence patterns, our RST-DCQA model significantly outperforms the previous local coherence model and will provide technical support for the overall quality analysis of the student's English essays.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61662012) as well as the Foundation of Key Laboratory of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology, No. CRKL150105).

References

1. T.K. Landauer, S.T. Dumais. *The latent semantic analysis theory of acquisition, induction, and representation of knowledge*, 211-240(1997)
2. M. Lapata, R. Barzilay. *Automatic evaluation of text coherence: models and representations*,1085-1090 (2005)
3. R. Barzilay, M. Lapata. *Modeling local coherence: an entity-based approach*, 141-148(2008)
4. B.J. Grosz, S. Weinstein, A.K. Joshi. Centering: a framework for modeling the local coherence of discourse, 203-225(2002)
5. C. Guinaudeau, M. Strube. *Graph-based Local Coherence Modeling*, 93-103(2013)
6. M. Zhang, V.W. Feng, B. Qin, G. Hirst, T. Liu. *Encoding World Knowledge in the Evaluation of Local Coherence*, 524-533(2015)
7. Z. Lin, H.T. Ng, M.Y. Kan. *Automatically evaluating text coherence using discourse relations*, 997-1006 (2011)
8. W.C. Mann, S.A. Thompson. Rhetorical Structure Theory: Toward a functional theory of text organization, **8(3)**, 243-281(2009)
9. M. Surdeanu, T. Hicks, M.A. Valenzuela-Escarcega. Two Practical Rhetorical Structure Theory Parsers, (2015)
10. H. Hernault, H. Prendinger, D.A. Duverle. HILDA: A Discourse Parser Using Support Vector Machine Classification, **1(3)**, (2010)
11. V.W. Feng, G. Hirst. *Text-level discourse parsing with rich linguistic features*.60-68 (2012)
12. S. Joty, G. Carenini, R. Ng. *Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis*, 486-496(2013)
13. T. Joachims. *Optimizing search engines using clickthrough data*, 133-142(2002)