Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 4
ICCIA 2024

A Survey of Studies on Discourse Structure and Relation

Nianyi Huang

School of Information Science, Beijing Language and Culture University, Beijing, China.

Huang_Nianyi@163.com

Abstract. Discourse Analysis aims at high-level semantic and structural analysis. Discourse structure analysis and relation recognition are two key tasks in discourse analysis research, while discourse analysis plays an important role in studying the structure and semantic content of texts. This paper firstly introduces the Rhetorical Structure Theory (RST) and the Penn Discourse TreeBank (PDTB) annotation standards and its corresponding resources establishment of each corpus. Then the mainstream models of discourse structure and relation are expounded. In the last part, the opportunities and challenges of discourse analysis are discussed by combining with the mainstream large language model ChatGPT. In addition, the development prospect of discourse analysis is explored by summarizing the related studies.

Keywords: Natural Language Processing, Discourse Analysis, Rhetorical Structure Theory (RST), The Penn Discourse TreeBank (PDTB).

1. Introduction

With the progress of AI research, the research level of discourse analysis is developing from the relatively mature and simple level of words and sentences to the more complex level of paragraphs and discourse. Discourse structure and relation have gradually become an important research area to understand the relationship between structure and semantics of a text.

Discourse Analysis was firstly proposed by American linguist Zellig Harris in his article *Discourse Analysis*, defining discourse as a body of language connected by sentences. In 1991, Liao Qiuzhong[1] defined discourse as a complete body of language used in one communication process. In this way, discourse can show the text content more clearly compared with words and sentences in article comprehension. Furthermore, discourse does not exist independently but depends on the contextual structure and the interconnection between units to accurately represent the semantics of the text. Theories of discourse analysis include Rhetorical Structure Theory (RST) and the Penn Discourse TreeBank (PDTB). After the proposal of the theories, the corresponding corpora were constructed by annotating the Chinese and English language data respectively.

Section 2 of this paper introduces RST, PDTB annotation standards, and the construction of the relevant corpora. Section 3 expounds the mainstream models of discourse structure and relation, focusing on understanding the use of classic and emerging methods. Section 4 discusses the opportunities and challenges for discourse analysis brought by ChatGPT, a product of the popular trend of the times, and tries to explore the future development of discourse analysis.

2. Introduction to the Theories and Corpora of Discourse Analysis

Theories of discourse analysis include two aspects, discourse structure analysis and discourse relation analysis. The former is represented by RST and the other is represented by PDTB.

2.1. RST

2.1.1. RST

In the discourse structure framework, RST is a form of article-level discourse analysis proposed by Mann et al.[2] in 1988 for analyzing the internal organization and rhetorical devices of texts, which lays the foundation for subsequent corpus building and computational model construction.

In RST, a complete and coherent discourse consists of different units and levels. The parts of a discourse are not stacked randomly, but are interconnected by rhetorical relations. In the discourse, structural relations are used to study the wholeness and coherence of the text, and the overall structure and semantic relations of the discourse are analyzed by describing the rhetorical relations at each level of the discourse. Rhetorical relations are the semantic relations by virtue of which fine-grained semantic units (phrases, sentences, paragraphs, etc.) in the discourse are combined into coarse-grained semantic units of discourse.

Most of the rhetorical relations have asymmetric semantic features[3], which are reflected in RST as the nucleus and satellite of discourse units. The nucleus is the unit containing important central semantic information, and the satellite is the unit conveying supporting information. In this way, satellite "works" around the nucleus. If detached from the nucleus, it is equivalent to information without central semantic, which cannot stand alone as a unit of semantic expression. Rhetorical relations are divided into nucleus-satellite relation and multinuclear relation. The former exists between two structural segments, reflecting a structural segment (nucleus) is more prominent in the discourse structure, while the other structural segment (satellite) represents the situation that supports the information. This indicates that the content of the discourse has a primary and secondary. Multinuclear relation explains the connection between the discourse units are juxtaposed. Of these, nucleus-satellite relation is predominant in the discourse.

In addition, RST assumes that the discourse structure is hierarchical in nature, and two neighboring discourse units can be connected to form a more macroscopic structure through rhetorical relations, which leads to a situation that all the discourse units can be connected to each other to form a discourse structure tree. Therefore, RST assumes that each document has a hierarchical discourse structure tree containing a root node and markers, as shown in Figure 1. This theory also lays the foundation for the establishment of hierarchical corpora such as RST-DT.

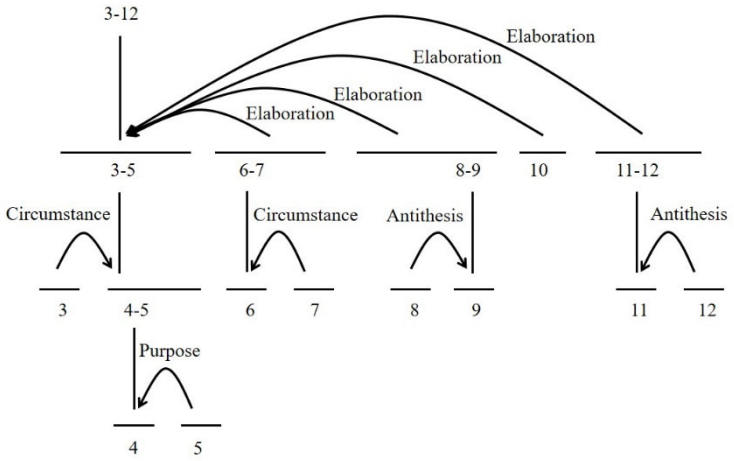

Figure 1. Example of A RST Discourse Structure Tree

2.1.2. Relevant Corpora of RST

The English RST Discourse Treebank (RST-DT) is the most important corpus of RST, annotating the corpus selected from 385 articles in the *Wall Street Journal* from Penn Treebank[4] (PTB) which totals more than 176,000 words. It mainly annotates the primary, secondary and rhetorical relations

of the elementary discourse units (EDUs) with a result of 16 groups of 78 kinds of relations. The corpus construction process of RST-DT consists of two main parts.

Recognition of EDUs: A piece of original text is identified, annotated and cut to obtain the smallest segment that can show a certain semantic meaning, which is called an EDU. EDUs are usually consecutive and non-overlapping segments of the text and the basic and important building blocks of discourse analysis.

Construction of the Discourse Structure Tree: The discourse structure of a text is represented as a tree defined by four dimensions. First, leaf nodes represent EDUs. Second, internal nodes mean consecutive structural segments with specific rhetorical relations. Third, each node follows the asymmetric nature of RST, with the nucleus representing the more important units of information, and the satellite representing the contextual or supportive units of information. Fourth, each node is characterized by a rhetorical relation that exists between two or more adjacent structural segments that can be intentional, semantic, or textual in nature. The rhetorical relations between these structural segments are cascaded to the root node, which is the overall discourse, to eventually form a discourse structure tree.

The English corpora GUM and SciDTB have also reached considerable scale for RST parsing and downstream task research. GUM, constructed and extended by Georgetown University, is a multilevel corpus containing 12 genres of news, interviews, etc. SciDTB is a domain-specific corpus annotating abstracts from scientific papers, which is not composed entirely according to RST. Its discourse structure is based on a dependency structure tree. In addition, the COVID19-DTB corpus has a similar role to SciDTB in the area of COVID-19 related academic papers.

Due to the language differences between Chinese and English, the application of RST in Chinese is challenging. In 2008, Le[5] integrated RST into Chinese for the first time with the CJPL corpus constructed under the guidance of RST. Since then, the corpora GCDT and MCDTB have both gained new developments in the RST task. GCDT is a multi-type RST corpus for Chinese medium-length documents, which contains more than 60,000 tokens and 9,000 EDUs. MCDTB is a discourse treebank for Chinese macro discourse relations, which is based on RST to annotate macro discourse information including discourse structure, core and relations. The basic information of the above corpora is shown in Table 1.

Table 1. Representative Corpora of RST

Corpus	Language	Annotation Structure	Theory	Size (Articles)	Genre
RST-DT	English	Hierarchical Structure Tree	RST	385	News Reports
GUM	English	Hierarchical Structure Tree		193	Multi-Genre
SciDTB	English	Dependency Structure Tree		798	Academic Papers
COVID19-DTB	English	Hierarchical Structure Tree		300	Academic Papers
CJPL	Chinese	Hierarchical Structure Tree		97	News Comments
GCDT	Chinese	Hierarchical Structure Tree		50	Multi-Genre
MCDTB	Chinese	Hierarchical Structure Tree		720	News Reports

2.2. PDTB

2.2.1. PDTB

In contrast to RST, PDTB is the annotation standards based on the D-LTAG theory, which does not focus on the level of discourse structure. Therefore, the presentation of PDTB is a flat discourse relation corpus. The current popular PDTB-2.0[6] and PDTB-3.0[7] corpora mainly follow the lexical-based annotation of discourse relations proposed by Webber et al. in 2003. PDTB-2.0 introduces the lexical annotation standards for the *Wall Street Journal* from the PTB corpus, which contains 40,600 relations. Based on PDTB-2.0, the newly publicized PDTB-3.0 has been improved and enhanced to contain a total of 53,631 relations.

PDTB mainly focuses on the semantic relations between arguments, dividing the discourse into a "connective-argument" structure. It highlights the role of connectives and takes them as the core to annotate relevant discourse units. The two discourse units connected by a connective are called arguments. The discourse unit guided by a connective is Arg2, and the other discourse element is Arg1. The structure composed of the two arguments is called an "argument pair". The difference between PDTB annotation and RST is that PDTB no longer focuses on phrase-level linguistic units as discourse units. In addition, PDTB categorizes discourse relations into explicit and implicit ones based on the presence or absence of explicit connectives within a discourse. The explicit relation is when there is an explicit connective between two parts, and the implicit one is the opposite.

2.2.2. Other Relevant Corpora of PDTB

The Chinese corpus CDTB-0.5 is based on the PDTB annotation standards, which investigates the syntactic and statistical distributions of discourse connective annotations in the Chinese discourse treebank, and improves the annotation strategy for the Chinese corpus. Another Chinese corpus, HIT-CDTB, also follows the PDTB annotation standards, which contains 525 annotated texts from OntoNotes 4.0, and analyzes and annotates the Chinese corpus by distinguishing between explicit and implicit discourse relations.

Table 2. Representative Corpora of PDTB

Corpus	Language	Annotation Structure	Theory	Size	Genre
PDTB-2.0	English	Flat Relation Annotation	PDTB	40600	News Reports
PDTB-3.0	English			53631	News Reports
CDTB-0.5	Chinese			5500	News Reports
HIT-CDTB	Chinese			525	Netnews

3. Introduction to Mainstream Models of Discourse Analysis

3.1. RST Analysis

RST discourse analysis consists of two subtasks, EDU recognition and discourse structure tree construction. The EDU recognition is the first step of RST analysis. The task is to cut the text into EDUs, and then the obtained EDUs will be used to construct the discourse structure tree according to the hierarchy through the rhetorical relationship among them.

There are two main approaches to EDU recognition, including statistically-based quadratic classifier and sequence-annotated discourse segmentation models. The first approach is to classify whether a morpheme is the boundary of the EDU by a quadratic classifier. The approach introduces two probabilistic models that assign probabilities to each word by combining syntactic and lexical features. The main studies include that Subba et al.[8] in 2007 first applied neural networks to discourse segmentation and trained a multilayer perceptron quadratic classifier using lexical and contextual features. In the same year, Fisher et al.[9] trained a classifier with the help of finite state and context-independent derived features. The second approach treats discourse segmentation as a sequence annotation task. Hernault et al.[10] proposed a method using a Conditional Random Field (CRF) probabilistic model for discourse segmentation based on sequence annotation in 2010. The results of this study showed that the model achieved a good performance with a F_1 score of 94%.

The research on automatic recognition of EDU has reached a relatively mature level, and the F_1 scores of some studies have exceeded 95%, which is very close to the accuracy of manual annotation. In this way, the research trend of RST analysis has been shifted to the task of analyzing discourse coherence relations and constructing discourse structure trees, which have more room for improvement.

The discourse structure tree on RST is further built on the basis of EDU recognition according to discourse coherence relations. Soricut and Marcu[11] proposed a sentence-level SPADE discourse

structure analysis algorithm based on syntactic information in 2003. Since the algorithm is aimed at analyzing fine-grained sentences, it has a great limitation of its recognition effect on coarse-grained discourse text. Hernault et al.[12] proposed a discourse structure analysis method using support vector machines on RST, and the structure tree was constructed in a bottom-up manner after extracting features through SVM. Wang et al.[13] also used the SVM model to the construct discourse structure tree. The difference is that the local model construction of Wang's team based on the transfer approach develops on the previous research by inputting multiple features into the model to get better results in recognizing the relations of discourses between the levels of discourse. In addition, experiments show the effectiveness of dependent syntactic analysis for RST discourse syntactic analysis. The CODRA model proposed by Joty et al.[14] employs a quadratic classifier to detect EDU boundaries and two CRFs for intra- and inter-sentence discourse relation recognition and construction of the discourse structure tree.

In summary, the current research on RST discourse structure is developing rapidly. Various researches try to use neural networks, pre-training models and other methods for the establishment of discourse trees, and have made obvious progress. However, the research effect in discourse relation recognition has not been able to achieve the effect of manual construction for the time being, with much room for improvement.

3.2. PDTB Analysis

Currently PDTB discourse analysis contains the following tasks, connective detection, argument tagging, discourse relation recognition and attribute tagging. The studies on PDTB corpus can be basically divided into two categories, explicit and implicit discourse analysis. While the current research has a good performance on explicit discourse relation recognition, so that the focus is mainly on implicit discourse relation recognition, which is a task given two arguments with the aim of classifying and recognizing implicit discourse relations.

There are three main approaches for implicit discourse relation recognition. The first one is to model the two arguments separately. With the development of neural networks, Ji et al.[15] and Rutherford et al.[16] proposed recursive and recurrent models to learn the representation of arguments respectively. Braud et al.[17] compared several different representations of implicit discourse classification. The effect of adding feature input was improved. The second approach adds modeling of the interrelationship between the two arguments to the first one. To address data sparsity and utilize lexical features, Chen et al.[18] proposed the Gate-Relevance Network GRN for discourse relation recognition, and later a new generative-discriminative framework that utilizes a new method to represent semantics and achieved good results. The third approach uses joint learning and multitasking architecture. In 2013, Lan et al.[19] first proposed a multitasking-based approach for implicit discourse relation classification. Inspired by this, Liu et al.[20] trained a multitasking neural network using PDTB as the experimental data and utilizing other data such as RST-DT as an auxiliary task. In 2020, Guo et al.[21] introduced knowledge information from WordNet to help classify discourse relations and demonstrated the role of knowledge. Due to the introduction of the pre-trained model, the efficiency of discourse relation recognition was significantly improved.

4. Impact of ChatGPT on Discourse Analysis

Conversation generative pre-training model ChatGPT is a large language model developed by OpenAI, an American AI research company. As a product of the times under the rapid development of AI technology, ChatGPT has sparked widespread attention in various research fields and even social practice since its official release in 2022. ChatGPT, through a large number of parameter trainings, is able to learn human language behaviors and make answers, achieving a strong natural language generation capability[22]. This also means that ChatGPT will certainly bring some influence to the research of discourse analysis.

Large language models such as ChatGPT can understand and generate texts more accurately, providing more powerful language comprehension and generation capabilities for discourse analysis, and offering new ways and opportunities to solve complex problems in discourse analysis. Instead of being limited to local contexts, large language models can take into account a wider range of contextual information including full text information, which helps to understand semantic relations in discourse in a more comprehensive way. In addition, large language models can be combined with domain knowledge to achieve discourse analysis of texts from different domains through methods such as adaptive learning or transfer learning, thus extending the scope of application.

By fully utilizing the capabilities of large language models, the field of discourse analysis can be advanced and the performance and effectiveness of related tasks can be improved. However, large language models may suffer from overfitting, data deviation, insufficient interpretability, etc., all of which need to be considered comprehensively and optimized continuously. Therefore, studies on discourse analysis need to continuously explore better solutions, which can reasonably utilize ChatGPT to improve the efficiency. It can also be used as an inspiration to conduct in-depth investigations in the application of downstream tasks, etc., in order to improve the effectiveness of discourse analysis.

5. Conclusion

In summary, discourse analysis plays an important role in today's more mature study of fine-grained texts and has more room for development. This paper addresses both the structural and relational aspects of discourse analysis. Based on this paper, it is believed that discourse analysis has become popular and the large language model can bring new inspiration for discourse analysis. However, there are still many challenges in the theory of discourse analysis and corpus construction. In the future research, it is necessary to further improve the corpus construction and explore the computational models more applicable to discourse analysis, so as to provide more favorable help for the field of natural language processing.

References

[1] Liao Q.Z., Studies in Discourse, Pragmatics and Syntax [J].Language Teaching and Linguistic Studies, 1991(04):16-44.

[2] Mann W C, Thompson S A. Rhetorical structure theory: Toward a functional theory of text organization[J]. Text-interdisciplinary Journal for the Study of Discourse, 1988, 8(3): 243-281.

[3] Liu S.Z., Zhang Z., Rhetorical Structure Theory and the RST Tools [J]. Technology Enhanced Foreign Languages, 2003(04):20-23.

[4] Marcus M, Santorini B, Marcinkiewicz M A. Building a large annotated corpus of English: The Penn Treebank[J]. Computational linguistics, 1993, 19(2): 313-330.

[5] Le M., An Annotation Study of the Rhetorical Structure of Chinese Discourses [J]. Journal of Chinese Information Processing, 2008, (04):19-23+42.

[6] Prasad R, Miltsakaki E, Dinesh N, et al. The penn discourse treebank 2.0 annotation manual[J]. December, 2007, 17: 2007.

[7] Webber B, Prasad R, Lee A, et al. The penn discourse treebank 3.0 annotation manual[J]. Philadelphia, University of Pennsylvania, 2019, 35: 108.

[8] Subba R, Di Eugenio B. Automatic discourse segmentation using neural networks. In: Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue. 2007, 189–190

[9] Fisher S, Roark B. The utility of parse-derived features for automatic discourse segmentation. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007, 488–495

[10] Hernault H, Bollegala D, Ishizuka M.A sequential model for discourse segmentation. In: Proc.of the CICLing 2010.2010.315-326.

[11] Soricut R, Marcu D.Sentence level discourse parsing using syntactic and lexical information. In: Proc.of the NAACL-HLT 2003.2003.149-156.

[12] Hernault H, Prendinger H, du Verle DA, Ishizuka M. HILDA: A discourse parser using support vector machine classification. Dialogue & Discourse, 2010, 1(3): 1–33.

[13] Wang YZ, Li SJ, Wang HF. A two-stage parsing method for text-level discourse analysis. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver: ACL, 2017. 184-188.

[14] Joty S, Carenini G, Ng R T. Codra: A novel discriminative framework for rhetorical analysis. Computational Linguistics, 2015, 41(3): 385-435

[15] Ji Y, Eisenstein J. One vector is not enough: Entity-augmented distributed semantics for discourse relations. Transactions of the Association for Computational Linguistics, 2015, 3: 329-344

[16] Rutherford A T, Demberg V, Xue N. Neural network models for implicit discourse relation classification in english and chinese without surface features. 2016, arXiv preprint arXiv: 1606.01990

[17] Braud C, Denis P. Comparing word representations for implicit discourse relation classification. In: Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2015). 2015

[18] Chen J, Zhang Q, Liu P, Qiu X, Huang X. Implicit discourse relation detection via a deep architecture with gated relevance network. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1726−1735

[19] Lan M, Xu Y, Niu Z Y. Leveraging synthetic discourse data via multi task learning for implicit discourse relation recognition. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013, 476–485

[20] Liu Y, Li S, Zhang X, Sui Z. Implicit discourse relation classification via multi-task neural networks. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. 2016, 2750−2756

[21] Guo F, He R, Dang J, Wang J. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 7822−7829

[22] Feng Z.W., Zhang D.K., Rao G.Q., From Turing Test to ChatGPT: Milestones and Implications for Man-Machine Conversation [J]. Chinese Journal of Language Policy and Planning, 2023,8(02): 20-24. DOI:10.19689/j.cnki.cn10-1361/h.20230202.