MCDTB: A Macro-Level Chinese Discourse TreeBank

Feng Jiang[1], Sheng Xu[1], Xiaomin Chu[1], Peifeng Li[1], Qiaoming Zhu[12], Guodong Zhou[1]
[1]School of Computer Science and Technology, Soochow University, China
[2]Institute of Artificial Intelligence, Soochow University, China
{fjiang, sxu, xmchu}@stu.suda.edu.cn
{pfli, qmzhu, gdzhou}@suda.edu.cn

Abstract

In view of the differences between the annotations of micro and macro discourse relationships, this paper describes the relevant experiments on the construction of the Macro Chinese Discourse Treebank (MCDTB), a higher-level Chinese discourse corpus. Following RST (Rhetorical Structure Theory), we annotate the macro discourse information, including discourse structure, nuclearity and relationship, and the additional discourse information, including topic sentences, lead and abstract, to make the macro discourse annotation more objective and accurate. Finally, we annotated 720 articles with a Kappa value greater than 0.6. Preliminary experiments on this corpus verify the computability of MCDTB.

1 Introduction

In the field of natural language processing, discourse analysis is becoming increasingly important as the object of research gradually shifts from the word level to sentence, event and other semantic aspects. Discourse analysis primarily examines the text coherence and cohesion, including the analysis on structure, nuclearity and relationship. In discourse analysis, discourse refers to a series of continuous clauses, sentences or paragraphs of language as a whole; it not only includes text sequences but also the text structure and the logical relationship between text sequences.

Discourse analysis aims at studying the internal structure of texts and understanding the semantic relation between different text units. The granularity of this research can be clause, sentence, sentence group, paragraph and whole article. Commonly, discourse analysis is divided into microstructure and macrostructure analysis. The former is the study of intra-sentence or inter-sentence discourse relations, while the latter is discourse relation between sentence clusters, paragraphs and chapters (Van Dijk, 1976), which highlights a higher semantic level for text comprehension.

Example 1 is the content of the article chtb0155, which has a title (T) and five paragraphs (P1-P5), i.e., five discourse units. Figure 1 shows the example's macro discourse structure tree. The leaf nodes of the macro discourse structure tree refer to discourse units, and the internal nodes refer to relational nodes, which represent the larger discourse units of the relevant children. When connecting the parent and child nodes, the directed edge indicates that the child is an important discourse unit in the relationship, and the undirected edge indicates that the child is secondary in the relationship. Therefore, in Figure 1, there is an *Evaluation* relation between the discourse units P1 and P2, and they form a higher level unit (relational node) DU1-2. The discourse units P4 and P5 exhibit the *Purpose-Behaviour* relation and form a higher level unit DU4-5. Consequently, DU1-2 and P3 have the *Elaboration* relation and form a higher level unit DU1-3. Finally, DU4-5 is supplementary to DU1-3.

Based on the discourse structure tree shown in Figure 1, the article of Example 1 is easily understood. In the task of automatic summarization, for example, according to this discourse structure tree and the directed edges from the root node to the leaf nodes, the topic sentences of the leaf nodes can be used as the text summary. The summary of Example 1 is a combination of the first paragraph and the second paragraph's topic sentence, which is more appropriate than the simple use of the first paragraph as a summary of the full text.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3493–3504
Santa Fe, New Mexico, USA, August 20-26, 2018.

(T) 中保财险公司为上海提供迄今最大一笔出口信用保险 (The Property Insurance Co. of PICC provides Shanghai with the largest amount of export credit insurance to date)
(P1) 中保财产保险有限公司今天为上海…有限公司向巴西出口…，提供了…出口信用保险。(Today, the Property Insurance Co. of PICC, Ltd., provided an export credit insurance… to the Shanghai … Company, Ltd. for them to export … to Brazil.)
(P2) 这是上海地区迄今提供的最大一笔出口信用保险。 (This is the largest export credit insurance in the Shanghai region to date.)
(P3) 这个出口项目包括…起重机。巴西方面先预付…定金，余下…款项…。中保财险公司…提供. . 服务就是为保证…安全收汇，以达到…目的。 (This export project includes …cranes. The Brazilian side will pay… in advance, with the remaining … payments... PICC …is to ensure that …exchange safely … so as to …)
(P4) 出口信用保险制度是…惯例。中国政府责成中保财产保险有限公司代政府开办…业务，为…. 保险等。(The system of export credit insurance is … internationally. The Chinese government is obliged to set up a business by the government of PICC, for... insurance, etc.)
(P5) 为此，中国中央财政于一九八八年拨…专款作为风险准备金，用于…保险业务。中保财险公司…承保了…出口收汇风险，其中…项目四十多个。 (For this, in 1998 China's Central Treasurer allocated ... as risk preparation funds, to be used … insurance services. The company has underwritten about … in the risk of foreign exchange receipts, including … more than 40 … projects.)

Example 1: Content of chtb0155 (simplified version).

This paper uses the RST style to annotate macro discourse structure, nuclearity and relationship and constructs a Macro Chinese Discourse Treebank (MCDTB) including 720 articles. Compared with micro discourse annotation, macro discourse annotation has different characteristics. Macro discourse units are longer and fuzzier, and the relationship between discourse units is looser and less logical. To this end, we have formulated detailed annotation guidelines and quality assurance strategies. When marking macro discourse structure, nuclearity and relationship, we also annotate the annotation of the macro discourse information, such as topic sentences, lead and abstract, to make the annotation of macro discourse relations more objective and accurate.

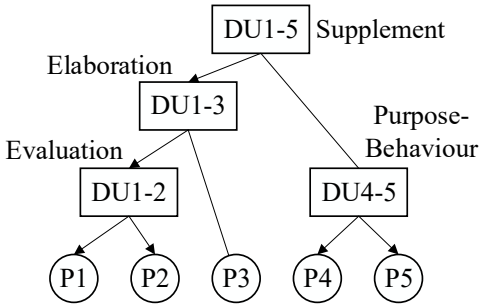

Figure 1: Macro discourse structure tree of chtb0155.

The rest of this paper is organized as follows. Section 2 overviews the related work. Section 3 describes the annotation system of MCDTB. Section 4 introduces our annotation process. Section 5 shows the corpus statistics on MCDTB. Section 6 provides a preliminary experiment. Finally, we present the study's conclusions and outline future work in Section 7.

2 Related Work

Discourse analysis is mainly divided into micro and macro discourse analysis. The theory of micro discourse primarily includes Hobbs model (Hobbs, 1985), rhetorical structure theory (RST) (Mann and Thompson, 1987; Mann et al., 1992), sentence group theory (Wu and Tian, 2000) and complex sentence theory (Fu, 2001), and the macro discourse theory has macro hyper-theme theory (Martin and Rose,

2003) and macro structure theory (Van Dijk, 1980). Under the guidance of these theories, various corpora have been developed accordingly.

In micro discourse analysis, the most popular corpora are RST-DT and PDTB. Based on RST, RST-DT (Marcu et al., 1999; Carlson et al., 2003) contained 385 articles from the Wall Street Journal annotated by 16 categories and 78 classes' rhetorical relations, to represent the relationship between two or more discourse units. Most of the elements of discourse units in RST-DT are phrases including partial clauses. To ensure the universality of the corpus, PDTB (Parsad et al., 2008) uses the LTAG (Lexicalized Tree-Adjoining Grammars) theory, and the annotation is entirely based on lexicalization. This approach primarily annotated the argument structure, including the conjunction and the semantic distinction information. The basic unit of argument is the event or state, and the specific form is the clause or sentence. The corpus is relatively large with 2,159 articles and approximately 1 million words.

Regarding Chinese micro discourse corpus, Zhou et al. (2015) used the PDTB annotation style to annotate 164 Chinese documents from CTB (Xue, 2005; Xue et al., 2005). The experiment demonstrated the transferability of the English discourse analysis theory in Chinese language. Lv et al. (2015) constructed a Chinese corpus of 496 news articles from People's Daily under the framework of the Chinese Framing Network (CFN) (Hao et al., 2007). Each news article annotated the discourse frame, structure and relationship. However, the discourse units are also smaller with a minimum of one sentence and a maximum of five sentences, and their research object is primarily at the word level. Li et al. (2014) integrated RST-DT and PDTB and used the representation method based on the connection dependence tree to annotate the Chinese Dependency Treebank (CDTB) containing 500 Chinese Xinhua news articles. However, this method only considered the discourse relationship annotation inside the paragraph and did not annotate the discourse relationship between paragraphs and the macro discourse information of the whole text.

Compared with the well-studied micro discourse analysis, macro discourse analysis is in the trial stage both in English and Chinese. In English, Sporleder and Lascarides (2004) attempted to perform an experiment on macro discourse analysis on RST-DT. However, RST-DT focuses on the micro discourse level, which is only tagging the relations on the sentence level and ignoring the relations on the paragraph level. Thus, it leads to some tailoring and correction when Sporleder carries out macro discourse analysis on the paragraph level. In Chinese, Chu et al. (2017) carried out relevant research on the nuclearity of macro discourse and made corresponding attempts to construct a macro discourse corpus.

In the field of Chinese discourse annotation, the number of relevant corpora is still relatively small, and its scale is also small. In addition, the existing micro discourse corpora still lack macro discourse information. For the task of analysis of macro discourse relationships, it is necessary to construct a macro discourse corpus, especially in Chinese.

3 Annotation System

3.1 Discourse Structure, Nuclearity and Relationship

Macro discourse analysis includes the analysis on discourse structure, nuclearity and relationship, similar to micro discourse analysis. However, macro discourse relationships are different from those of micro discourse. First, since the element of a macro discourse unit is a paragraph, macro discourse mainly considers the structure of paragraphs or its upper levels (e.g., DU1-2 in Figure 1). As shown in Figure 1, the annotation goal is to construct a discourse structure tree, where the leaf nodes are the element of the discourse units, i.e., paragraphs. Second, different from micro discourse relations, there are no connectives between the macro discourse units with lengths that are longer than those of micro discourse units. Consequently, it is difficult to grasp the theme of macro discourse units and the relationship between them. To ensure the correctness of macro discourse structure annotation, we first determine the theme of discourse units.

Category	Nuclearity
Mononuclear	Nucleus Ahead, Nucleus Behind
Multi-Nuclear	Multi-Nucleus

Table 1: Classification of discourse nuclearity.

In the annotation of discourse nuclearity, we are consistent with RST, which is divided into hypotactic ("Mononuclear") and paratactic ("Multi-Nuclear"). Specifically, it is divided into three categories: Nucleus Ahead, Nucleus Behind and Multi-Nucleus. Nucleus Ahead and Nucleus Behind belong to the Mononuclear, and Multi-Nucleus belongs to the Multi-Nuclear as Table 1.

Different from the discriminant method of micro discourse nuclearity, we not only compare importance among candidate macro discourse units but also consider which unit is more important with the full-text writing intention. In example 1, discourse units P1 and P2 have the *Evaluation* relation, where P1 tells the main story, and P2 is its evaluation. If the global information is not considered, P1 is more important than P2, and it should be annotated as *Nucleus Ahead*. However, given the full text, the title of which is "The Property Insurance Co. of PICC provides Shanghai with the largest amount of export credit insurance to date", P1 and P2 jointly form the subject of the full text such that P1 and P2 are equally salient; therefore, we annotated it by the relation *Multi-Nucleus*.

Following previous corpora on discourse analysis, we also construct a macro discourse corpus on news articles. In the annotation of macro discourse relationships, the theme of the paragraph as the element of the discourse unit is vaguer; the logic between discourse units is more insignificant. Considering the overall intentions of a news article, it is rare to have a *Transition* relation on the macro-level, and most of them have a *Contrast* relation. Therefore, based on the micro discourse relation representation that has 4 categories and 17 relations defined by CDTB, we delete the relations, such as *Transition*, *Concession* and other relations in the *Transition* category, and finally form the macro discourse relationship representation with 3 categories and 15 relations, as shown in Table 2.

Category	Relations
Coordination	Joint, Sequence, Progression, Contrast, Supplement
Causality	Cause-Result, Result-Cause, Background, Behaviour-Purpose, Purpose-Behaviour
Elaboration	Elaboration, Summary, Evaluation, Statement-Illustration, Illustration-Statement

Table 2: Classification of macro discourse relation.

3.2 Additional Macro Discourse Information

Van Djik (1980) notes that theme, gist, keystone and main points all belong to the overall macro discourse structure. The topic sentence of a paragraph, the lead and the abstract of a full text are related to the theme and key points of the text and are the overall structure of semantics, which also belong to the macro discourse structure. Therefore, in addition to the traditional annotation of micro discourse corpus, we also annotated the additional macro discourse information, including the topic sentence of a paragraph, the lead and the abstract of a full text. In addition, we automatically annotate the pragmatic functions of discourse elements.

The discourse structure tree can play a crucial role in building a natural language generation system (Carlson et al., 2003). A good macro discourse structure tree can preserve the semantic integrity of the text better (Liu and Zou, 2017). In discourse-level automated summarization tasks, a more natural and complete chapter summary can be generated in conjunction with annotated macro discourse information.

4 Annotation Process

Our annotation team consists of a doctoral candidate, two senior master degree candidates and three junior master degree candidates. All annotators are engaged in research on Natural Language Processing or Computational Linguistics and have a certain theoretical foundation of linguistics. Such an annotation team has increased professionalism and improved the efficiency of labelling. To ensure annotation quality, the entire annotation process has four phases:

1) Initialization phase. We annotated 10 articles per cycle. In each cycle, the doctoral candidate and two senior master degree candidates were annotating the same article at the same time, that is, triple tagging. At this stage, we had annotated 97 documents.
2) Revision phase. Different from the previous stage, the doctoral candidate double-tagged with two senior master degree candidates in each cycle, respectively. Finally, we annotated 147 documents.

3) Trial tagging. We annotated 20 articles per cycle. The tagging method is in line with the second stage, and we finally cumulatively tagged 305 documents. At the same time, three junior master degree candidates participated in the annotation.
4) Formal annotation phase. We annotated 30 articles per cycle. During each cycle, three senior annotators (one doctoral candidate and two senior master degree candidates) were matched with three junior master degree candidates, respectively, and formed three groups. After each cycle, we exchanged the partner between different groups. Finally, 720 articles have been tagged.

4.1 Annotation criteria

To ensure consistency and reliability of annotations in the macro discourse relationship labelling, the following annotation criteria are used in our annotations:

1) The annotation of nuclearity and relationship is independent of each other, and we do not consider the correlation between them. In Example 1, when judging the nuclearity of the two discourse units P1 and P2, we pay more attention to their relevance to the topic, not considering their relation Evaluation. Therefore, P1 and P2 are equally salient.
2) We annotate only one type of discourse relations to two discourse units. For example, both the *Sequence* and *Cause-Result* relations have the chronological order. When two discourse units have the chronological order, and there is a clear causal relation between them, they are annotated as *Cause-Result*; otherwise, it will be annotated as *Sequence*.
3) We adopt the way of majority voting. An article may have different discourse structure trees from the different perspectives. When these structures are all reasonable, we use a majority vote to choose a more widely accepted understanding to ensure the objectivity of annotation.
4) We adopt incremental annotation. At the time of macro discourse relationship tagging, our annotation strategy is to determine the topic sentence of each paragraph first. After grasping the intentions of each paragraph, we first divide the whole article into several independent regions from top to bottom and build sub-trees from bottom to top in each area. Next, we consider the structure of each region and eventually form a complete discourse structure tree. According to the complete discourse structure tree, we also annotate the lead and the abstract of an article. In example 1, when we read this article, if we can first find that P4 and P5 have the relationship of *Purpose-Behaviour*, we could directly connect them without considering the specific structure of P1 to P3.

4.2 Annotation sample

To facilitate labelling and accelerating the annotation speed, we have developed a platform for macro discourse annotation. In this platform, an article to be annotated should go through the following steps: data preprocessing, annotating paragraph topic sentence, annotating discourse structure, nuclearity and relationships, annotating lead and abstract, and automatically annotating pragmatic functions. The final generated annotation results of example 1 are shown in Table 3. An annotated document includes three parts: DISCOURSE, RELATION and TEXT. DISCOURSE refers to the abstract information of a document, including date (Dateline), topic (DiscourseTopic), lead (LEAD) and abstract (ABSTRACT). RELATION provides all discourse relationships in this document and its attributes include nuclearity and relation. TEXT shows all elementary discourse units (EDUs). The labels used in MCTDB are as follows.

- ID: the ID of the relation between two discourse units.
- Centre: the nucleairity relation (1-*Nucleus Ahead* / 2-*Nucleus Behind* / 3-*Multi-Nucleus*).
- ChildList: the ID list of its children relations (nodes). In this example, this relation has two children relations whose IDs are 2 and 3.
- Function: the pragmatic functions are derived from Van Dijk (1990)'s Hypothetical structure of the news schema and we modified them to suit the characteristics of Chinese macro discourse. Details of the labels are shown in Table 10.
- Layer: the layer in a macro discourse structure tree.
- ParagraphPosition: the positions of two discourse units which are split by "|". In this example, "1...3" indicates that the beginning paragraph and ending paragraph of the first discourse unit

are 1 and 3, respectively, while “4...5" indicates those of the second paragraph are 4 and 5, respectively.

```
<DOC>
<DISCOURSE Dateline="新华社上海三月六日电" DiscourseTopic="中保财险公司为上海提供迄今最大一笔出口信用保险">
<LEAD>中保财产保险有限公司……的最大一笔出口信用保险。</LEAD>
<ABSTRACT>中保财产保险有限公司……的目的。</ABSTRACT>
</DISCOURSE>
<RELATION>
<R ID="1" Center="1" ChildList="3|2" Function="NewsReport" Layer="1" ParagraphPosition="1...3|4...5" ParentId="-1" RelationType="Supplement" StructureType=" Hierarchical segmentation" />……
 <TEXT>
<P Function="Lead" ID="1" ParagraphTopic="中保财产保险有限公司……的出口信用保险。">中保财产保险有限公司…. . . 的出口信用保险。</P>……
<P Function="Behaviour" ID="5" ParagraphTopic="中国中央财政于一九八八年拨一亿美元专款作为风险准备金，用于中保集团公司开办出口信用保险业务。">为此，中国中央财政……四十多个。（完）</P>
</TEXT>
</DOC>
```

Table 3: Annotation format of chtb0155.

- ParentId: the ID of the relation’s parent.
- RelationType: the type of this relation, defined as RST, and details of the labels are shown in Table 2.
- StructureType: the annotation is divided into 2 categories: *Hierarchical segmentation* and *Parallel segmentation*. When the relationship is Mononuclear and has only two children, it is marked as *Hierarchical segmentation*. Otherwise, it will be marked as *Parallel segmentation* (usually the relation is *Joint*).
- ParagraphTopic: the topic sentence of the paragraph.

4.3 Quality Assurance

We guarantee the annotation quality of the corpus from two aspects. The first one is the tree verification, which is to ensure the correctness of corpus annotation. The second one is consistent assessment, which is designed to ensure the objectivity of corpus annotation.

In tree validation, we take two steps to verify the annotated corpus. First, the annotated corpus is imported to the macro discourse annotation platform to verify its correctness. Second, an automated pragmatic annotation tool is used to annotate the pragmatic functions of each discourse unit. Through pragmatic annotation rules, it can be used to calibrate the annotation of discourse relations and structures. When a problem occurs in the validation process, we manually verify the error type and then correct it. Compared with the traditional manual calibration, we use a double automatic verification method to improve the speed and ensure quality.

The evaluation of consistency can be used to measure the objectivity of corpus annotation. For evaluating the consistency of the discourse structure tree, tree consistency can more objectively reflect the quality of annotation. CDTB and PDTB only evaluated the consistency of certain indicators, such as discourse relation, and did not evaluate the tree consistency. RST-DT proposed a method of evaluating the complete consistency of a discourse tree, but it also had several shortcomings. The evaluation objects of RST-DT are the discourse structure, nuclearity and relationship annotated on the children nodes, while the evaluation object of the standard syntax tree (Black et al., 1991) is the annotations on the parent node. In this way, RST-DT will be at least double the standard syntax tree evaluation of the sample, and

because of the mutual constraint relationship between children nodes, the consistency of corpus annotation of RST-DT is objectively higher.

In the annotation of a macro discourse relationship, the leaf nodes are natural paragraphs and do not need to be manually divided. Therefore, unlike the assessment method of the RST-DT agreement, we do not use the leaf nodes as an example of a consistent assessment. In addition, we adopted the standard syntax tree method using the annotation on the parent node as the evaluation object. Although this may lead to a reduction in evaluation examples and a decline in the indicators of consistency, we believe that this evaluation method is more objective.

Indicators	Structure	Nuclearity	Relation
Agreement	86.24%	83.88%	80.46%
Kappa	0.68	0.66	0.61

Table 4: Consistency indicators of MCDTB.

To eliminate coincidental annotations, we also used the Kappa value (Siegel and Castellan, 1988) as an assessment of the consistency of annotations while using agreement rates. The average value of the agreement assessments of the 200 articles sampled served as the ultimate evaluation indicators of the corpus. As shown in Table 4, the corpus has an agreement rate of greater than 80% and a Kappa value greater than 0.6 in discourse structure, nuclearity and relationships. Krippendorff (1980) noted that the Kappa value of the annotation data is above 0.6, indicating that it has good annotation quality.

Confusing Relationships	Proportion
Elaboration and Supplement	21.90%
Elaboration and Joint	9.52%
Sequence and Joint	8.57%
Supplement and Joint	7.62%
Elaboration and Evaluation	5.71%
Elaboration and Sequence	5.71%

Table 5: Confusing Relationships in double-tagged.

To find the most confusion annotations among the annotators, we sampled 100 articles and obtained the double-tagged information (720 articles are double-tagged). The top six confusion relationships are shown in Table 5. Apart from the difference in the structure annotation, the primary confusing relationships are *Elaboration* and *Supplement*, accounting for 21.90%. Both of these relationships belong to the *Elaboration* category; therefore, it is easier to confuse them. The second confusing relationship is *Elaboration* and *Joint*, accounting for 9.52%. The confusion results from significant differences in the understanding of the article.

Confusing Nucleus	Proportion
Nucleus Ahead and Multi-Nucleus	89.19%
Nucleus Ahead and Nucleus Behind	8.11%
Nucleus Behind and Multi-Nucleus	2.70%

Table 6: Confusing Nucleus in double-tagged.

As shown in Table 6, the main confusion nucleus are primarily Nucleus Ahead and Multi-Nucleus, accounting 89.19%; this finding may be due to the confusion between *Elaboration* and *Joint* in the relationship. This result is also consistent with previous experimental results (Jiang et al., 2018), which show that humans and models are confused in the same place.

5 Statistics on MCDTB

Our macro discourse corpus MCDTB[1] annotated 720 news articles from CTB 8.0. As shown in Table 7, the entire annotated corpus has 3,981 paragraphs, and the average number of paragraph is 5.53 per document. Each article contains at most 22 paragraphs and at least 2 paragraphs.

#documents	720	#paragraphs	3,981
# paragraphs of the longest document	22	Average length (paragraphs /document)	5.53
# paragraphs of the shortest document	2	Average length (sentences/paragraph)	2.09

Table 7: MCDTB corpus details.

The distribution of document numbers on paragraph numbers in MCDTB is shown in Figure 2. 54.72% of the documents have 4 to 6 paragraphs, while this number is 92.08% when a document has 2 to 9 paragraphs, which indicates that news articles primarily focus on the middle passage.

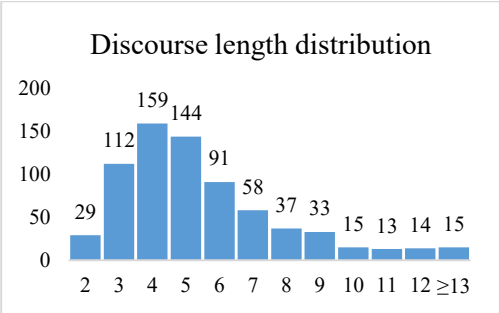

Figure 2: Discourse length distribution in MCDTB.

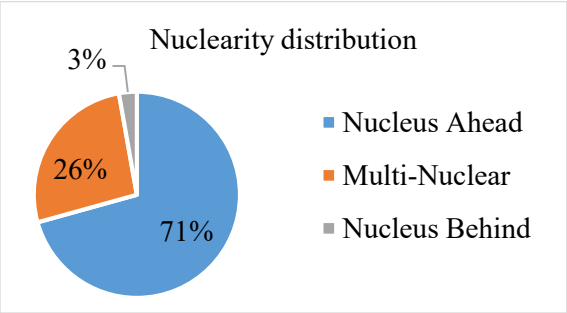

Figure 3: Nuclearity distribution in MCDTB.

The nuclearity distribution in MCDTB is shown in Figure 3. Among these numbers, the number of the *Nucleus Ahead* is 2,026, accounting for 70.69%. There are 760 *Multi-Nuclear* relations, accounting for 26.52%, while there are only 80 *Nucleus Behind*, accounting for only 2.79%. This finding shows that news articles tend to explain important events first, which is in line with the conclusion of linguistic statistics by Li and Liao (2001).

Relation	**Number**	**Proportion**	**Relation**	**Number**	**Proportion**
Elaboration	990	34.54%	Joint	634	22.12%
Supplement	493	17.20%	Background	237	8.27%
Evaluation	127	4.43%	Result-Cause	103	3.59%
Sequence	99	3.45%	Cause-Result	49	1.71%
Statement-Illustration	39	1.36%	Summary	23	0.80%
Illustration-Statement	20	0.70%	Contrast	17	0.59%
Behaviour-Purpose	14	0.49%	Purpose-Behaviour	11	0.38%
Progression	10	0.35%			

Table 8: Statistics on discourse relations.

As shown in Table 8, MCDTB contains 2,866 discourse relations, of which the top six relations are *Elaboration*, *Joint*, *Supplement*, *Background*, *Evaluation* and *Result-Cause*. These six types of relations accounted for 90.16% of the total number of relations.

[1] The Macro Chinese Discourse TreeBank is available at https://figshare.com/s/250474dba44e4161b040.

Furthermore, we analysed the correlation between the nuclearity and the relationship. The nuclearity distributions on different discourse relationships are shown in Table 9. This finding shows that most of the *Elaboration* and *Causality* relations are *Nucleus Ahead*, while those of the *Coordination* relations are more inclined to be equally salient in each discourse unit. The three relationships are about the same in *Nucleus Behind*.

Category	Nucleus Ahead	Nucleus Behind	Multi-Nucleus
Elaboration	1208	36	9
Causality	365	29	20
Coordination	453	15	731

Table 9: Nuclearity distribution of discourse relationships.

In particular, we analysed the distribution of pragmatic functions. MCDTB has a total of 6,847 pragmatic functions, including 3,981 annotations on the paragraph level, and 2,866 on the larger discourse unit level, as shown in Table 10. Apart from annotating the necessary pragmatic functions, such as *Situation* and *Story*, the annotations which are helpful to the automatic summarization and information extraction tasks, such as *Lead*, *Summary*, *Background*, *Comment*, *Cause* and *Result*, make up the majority.

Pragmatic function (in paragraph)	Number	Proportion	Pragmatic function (in discourse units)	Number	Proportion
Situation	1902	27.78%	Story	1437	20.99%
Supplement	439	6.41%	News Report	720	10.52%
Summary-Lead	354	5.17%	Summary	300	4.38%
Lead	284	4.15%	Sub-Summary	95	1.39%
Background	196	2.86%	Result	73	1.07%
Sub-Summary	193	2.82%	Cause	69	1.01%
Story-Situation	183	2.67%	Supplement	54	0.79%
Comment	117	1.71%	Background	41	0.60%
Cause	83	1.21%	Statement	22	0.32%
Result	79	1.15%	Behaviour	17	0.25%
Others	151	2.21%	Others	38	0.55%

Table 10: MCDTB pragmatic function statistics.

6 Preliminary Experiment

To demonstrate the computability of MCDTB, we conducted a preliminary experiment on the identification of macro discourse structures. Discourse structure identification is the first and most critical step in the related tasks of discourse analysis. We use a Conditional Random Field (CRF) model to experiment with the parameter C of 4, the feature window of 3, and the rest of the parameters as a default value. There were 8,863 samples in total, including 3,261 positive samples and 5,602 negative samples.

Because of the language differences between English and Chinese, as well as the microscopic and macroscopic differences in features, such as the absence of syntax trees and the absence of tense features, we select only several of the structural features of the previous study (Feng and Hirst, 2012) as our Feature Set 1 as follows:

- The position of the beginning and end of a discourse unit;
- The number of sentences and the number of paragraphs contained in a discourse unit;
- The comparison of the number of sentences in the discourse unit to the previous unit;
- The comparison of the number of paragraphs in the discourse unit to the previous unit.

Considering the process of constructing discourse structure trees, we may need some procedural characteristics. We regard whether the discourse unit is a leaf node or whether the discourse unit is merged in the previous round as organizational characteristics. In terms of semantic features, we use semantic similarity, the connectives, and their part-of-speech in the first sentence of the discourse unit. On the calculation of semantic similarity, we used the word2vec model to train on CTB8.0 and calculate the semantic similarity between the two discourse units according to the semantic similarity algorithm proposed by Xu (2009). In particular, we discretized the final semantic similarity into 10 levels. Finally, we use the semantic similarity, the conjunctions of the first sentence of the discourse unit with their part-of-speech, whether they are the leaf nodes and whether they were merged in the previous round as the characteristics of Feature Set 2 as follows:

- The semantic similarity between the discourse unit and the previous unit;
- The conjunction of the first sentence and its part of speech in the discourse unit;
- Whether the discourse unit is a leaf node;
- Whether the discourse unit was merged in the previous round.

Feature selection	Feature Set 1	Feature Set 2	Feature combination
Acc.	76.09%	77.01%	**77.56%**

Table 11: Comparison of the accuracy of the various models.

In addition, we use the most probabilistic way to fuse two feature sets, that is, we use two feature sets to model and finally select the predictive labels with the larger probability value as the final annotation result. In the five-fold cross-validation experiment of the MCDTB corpus, the performance of each model is shown in Table 11.

The experimental results show that the performance of the model using the structural features and semantic features is similar, and the best performance is obtained by using the joint model based on the maximum probability of the feature combination, which improves the accuracy by 1.47% and 0.55% over Feature Set 1 and Feature Set 2, respectively.

7 Conclusions

In this paper, we describe relevant experiments concerning the construction of a Macro Chinese Discourse Treebank (MCDTB). In view of the differences between the annotations of micro discourse relationships and macro discourse relationships, we annotate the corpus following RST and provide the higher level macro discourse information, such as topic sentences of paragraphs, lead and abstract of discourse, to make the macro discourse annotation more objective and accurate. To speed the annotation process and ensure the annotation quality, we formulated detailed annotation criteria and quality assurance strategies and developed a platform for macro discourse annotation. Finally, we annotated 720 Chinese news articles, and we achieved a better consistency of labelling with an improved Kappa value of greater than 0.6 based on the improved consistency assessment standard. Preliminary experiments on this corpus verify the computability of MCDTB. In future work, we will conduct research on identifying macro discourse structure, recognizing nuclearity and classifying relationship and eventually form a complete end-to-end macro discourse analyser.

Acknowledgements

The authors would like to thank three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant Nos. 61751206, 61772354 and 61773276, and was also supported by the Strategic Pioneer Research Projects of Defense Science and Technology under Grant No. 17-ZLXDXX-02-06-02-04.

Reference

Cohan Arman, and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Pages 390-400, Lisbon, Portugal, September. Association for Computational Linguistics.

Ezra Black, Steven Abney, Dan Flickinger, C. Gdaniec, R. Grishman, P. Harrison, D. Hindle, R. Ingria, F. Jelinek, J. Klavans, M. Liberman, M. Marcus, S. Roukos, B. Santorini and T. Strzalkowski. 1991. Procedure for quantitatively comparing the syntactic coverage of English grammars. In Proceedings of the workshop on Speech and Natural Language. Pages 306-311. Pacific Grove, California, February .Association for Computational Linguistics.

Xiaomin Chu, Qiaoming Zhu and Guodong Zhou. 2017. Discourse primary-secondary relationships in natural language processing. Chinese Journal of Computers, 40(04): 842-860.

Teun A. Van Dijk. 1976. Narrative macrostructures. PTL: A journal for descriptive poetics and theory of literature, 1: 547-568.

Teun A. Van Dijk. 1980. Macrostructure: An interdisciplinary study of global structures in discourse, interaction, and cognition. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Teun A. Van Dijk. 1990. News as discourse. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1(ACL 2012). Pages 60-68. Jeju, Korea, July. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2014). Pages 511-521, Baltimore, USA, June. Association for Computational Linguistics.

Fuyi Fu. Study of complex sentences in Chinese. 2001. Beijing: Commercial Press.

Xiaoyan Hao, Wei Liu, Ru Li and Kaiying Liu. 2007. Description systems of the Chinese framenet database and software tools. Journal of Chinese Information Processing, 21(5): 96-100.

Jerry R. Hobbs 1985. On the coherence and structure of discourse. Center for the Study of Language and Information. Pages 1-36.

Klaus Krippendorff. 1980. Content analysis: An introduction to its methodology. Seikeigeka Orthopedic Surgery, 79(385): 204.

Feng Jiang, Xiaomin Chu, Sheng Xu, Peifeng Li and Qiaoming Zhu. 2018. A Macro Discourse Primary and Secondary Relation Recognition Method. Journal of Chinese Information Processing, 32 (01): 43-50.

Jin Li and Kaihong Liao. Comparison of theme patterns and paragraph structures in Chinese and English. 2001. Jinan Journal (Philosophy & Social Science Edition), 23(05): 89-93.

Yancui Li, Wenhe Feng, Jing Sun, Kong Fang and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Pages 2105-2114. Doha, Qatar, October. Association for Computational Linguistics.

Yancui Li. 2015. Research on the structure of Chinese text structure and the construction of resources. Soochow University.

Yancui Li, Jing Sun and Guodong Zhou. 2015. Automatic recognition and classification on Chinese discourse connective. Acta Scientiarum Naturalium Universitatis Pekinensis, (02): 307-314.

Jiayi Liu and Yimin Zou. 2017. A review of automatic text summarization research in recent 70 years. Information Science. 35 (07), pages 154-161.

Guoying Lv, Na Su, Ru Li, Zhiqiang Wang and Qinghua Chai. 2015. Frame-based discourse structure modeling and relation recognition for Chinese sentence. Journal of Chinese Information Processing, 29 (06): 98-109.

Carlson Lynn, Daniel Marcu and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Current and new directions in discourse and dialogue. Pages 85-112. Springer, Dordrecht.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task.

In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Pages 747-757, Seattle, USA, October. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: a theory of text organization. University of Southern California, Information Sciences Institute.

William C. Mann, Christian M. I. M. Matthiessen and Sandra A. Thompson. 1992. Rhetorical structure theory and text analysis. Discourse description: Diverse linguistic analysis of a fund-raising text. Pages 39-78.

Daniel Marcu, Estibaliz Amorrortu and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In Proceedings of the ACL'99 Workshop on Standards and Tools for Discourse Tagging. Pages 48-57.

James Robert Martin and David Rose. 2003. Working with discourse: Meaning beyond the clause. London: Continuum.

Rashmi Prasad, Nikhil Dinesh, Alan Lee and Eleni Miltsakaki. The Penn Discourse Treebank 2.0. 2008. In International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco, May. Pages 2961-2968.

Sidney Siegel and N. John Castellan. 1988. Nonparametric statistics for the behavioral sciences. 2. New York: McGraw-Hill Book Company.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In Proceedings of the 20th international conference on Computational Linguistics (Coling 2004). Pages 43. Association for Computing Machinery.

Yizhong Wang, Sujian Li and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2017), pages 184-188, Vancouver, Canada, July. Association for Computational Linguistics.

Weizhang Wu and Xiaolin Tian. 2000. Chinese sentence group. Beijing: Commercial Press.

Shuai Xu. 2009. Research and implementation of paraphrase recognition for question answering. Harbin Institute of Technology.

Nianwen Xue. 2005. Annotating discourse connectives in the Chinese Treebank. In Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky. Association for Computational Linguistics. Pages 84-91.

Nianwen Xue, Fei Xia, Fudong Chiou and Marta Palmer. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural language engineering, 11(2): 207-238.

Yuping Zhou and Nianwen Xue. 2015. The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. Language Resources & Evaluation, 49(2): 397-431.

Bowei Zou, Guodong Zhou and Qiaoming Zhu. 2014. Negation focus identification with contextual discourse information. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014). Pages 522-530, Baltimore, USA, June. Association for Computational Linguistics.