Chinese Paragraph-level Discourse Parsing with Global Backward and Local Reverse Reading

Feng Jiang, Xiaomin Chu[*], Peifeng Li, Fang Kong, Qiaoming Zhu
School of Computer Science and Technology, Soochow University, China
fjiang@stu.suda.edu.cn
{xmchu, pfli, kongfang, qmzhu}@suda.edu.cn

Abstract

Discourse structure tree construction is the fundamental task of discourse parsing and most previous work focused on English. Due to the cultural and linguistic differences, existing successful methods on English discourse parsing cannot be transformed into Chinese directly, especially in paragraph level suffering from longer discourse units and fewer explicit connectives. To alleviate the above issues, we propose two reading modes, i.e., the global backward reading and the local reverse reading, to construct Chinese paragraph level discourse trees. The former processes discourse units from the end to the beginning in a document to utilize the left-branching bias of discourse structure in Chinese, while the latter reverses the position of paragraphs in a discourse unit to enhance the differentiation of coherence between adjacent discourse units. The experimental results on Chinese MCDTB demonstrate that our model outperforms all strong baselines.

1 Introduction

Discourse parsing aims to study the internal structure of texts and understand the semantic relationship between various kinds of text units, such as clauses, sentences, sentence groups, paragraphs, and chapters. Since it is fundamental to understanding the overall text semantics, discourse parsing has been widely applied to various Natural Language Processing (NLP) applications, such as sentiment analysis (Bhatia et al., 2015), question answering (Sadek and Meziane, 2016), and summarization (Cohan and Goharian, 2018; Xu et al., 2020).

As one of the most influential theories in discourse parsing, Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) represents text as a hierarchical structure known as a discourse tree. In the literature, various RST-style corpora have been built, such as RST Discourse Treebank (RST-DT) (Carlson et al., 2003) and Macro Chinese Discourse Treebank (MCDTB) (Jiang et al., 2018b). In RST-style discourse parsing, the parser first identifies whether there is a rhetorical relationship between discourse units to construct a naked tree and then recognizes the nuclearity and relation labels for each relationship, as shown in Figure 1. According to the granularity of the leaf nodes, the discourse tree is divided into three levels: clause level, sentence level and paragraph level (Kobayashi et al., 2020). This paper focuses on constructing paragraph-level Chinese discourse trees where the leaf node is a paragraph. We call a leaf node that contains only one paragraph as PDU (Paragraph-level Discourse Unit) to distinguish the elementary discourse unit (EDU) at the clause level. It is more important for downstream tasks as the upper part of a complete discourse tree.

In English, there is a series of successes (Feng and Hirst, 2014; Ji and Eisenstein, 2014; Wang et al., 2017) in RST-DT (Carlson et al., 2003), especially Lin et al. (2019) and Liu et al. (2019) got excellent performance in clause-level discourse parsing. However, owing to linguistic and cultural differences, discourse structure is different between Chinese and English: English is linear while Chinese is spiral (Kaplan, 1966). Therefore, the state-of-the-art model in English cannot be directly transformed into Chinese well. Besides, fewer studies (Sporleder and Lascarides, 2004; Zhou et al., 2019) on paragraph-level discourse parsing may suffer from data sparsity and its performance is much lower than that of

[*] Corresponding author

Proceedings of the 28th International Conference on Computational Linguistics, pages 5749–5759
Barcelona, Spain (Online), December 8-13, 2020

Corpus	MCDTB (Chinese)		RST-DT (English)	
level	clause	paragraph	clause	paragraph
#avg. tokens	22	100	8.12	51.96
#total trees	3981	720	3846	385
% connectives	23.98	2.99	22.65	13.51

Table 1: The average tokens per leaf node, the total trees and the percentage of connectives in MCDTB and RST-DT at different levels. The connective lists in RST-DT are derived from PDTB.

sentence-level or clause-level. It is more challenging for the paragraph-level discourse parsing because of its longer text length (over 4.5 times), fewer samples (less than 20%) and fewer connectives than clause-level, especially in Chinese, as shown in Table 1. This poses two critical challenges to improve the performance of paragraph-level discourse tree construction in Chinese. One is how to utilize the bias of discourse structure better in Chinese. Another challenge is how to better capture the semantic relationship of larger discourse units at the paragraph level.

In this paper, we analyzed the differences in paragraph-level discourse structure between English and Chinese and proposed a global backward reading mode that processes discourse units from the end to the beginning in a document to utilize the bias of discourse structure in Chinese. Moreover, we proposed the Triple semantic Matching model based on BERT (TM-BERT) that views three adjacent discourse units as a triangle to capture the relationship across discourse units better. To deal with longer paragraph-level discourse units, we used only 1-2 PDUs to represent a discourse unit and proposed a local reverse reading mode to reverse the order of PDUs. The experimental results on Chinese MCDTB demonstrate that our proposed model with global backward and local reverse reading outperforms all strong baselines.

2 Related Work

In English, RST-DT (Carlson et al., 2003) is one of the popular discourse corpora (Subba and Di Eugenio, 2009; Zeldes, 2017; Kolhatkar and Taboada, 2017), which annotates the discourse structure, nuclearity, and relationship of a document. Most previous studies have focused on complete discourse parsing and can be mainly categorized into the shift-reduce algorithm (Ji and Eisenstein, 2014; Wang et al., 2017; Yu et al., 2018; Jia et al., 2018), the probabilistic CKY-like algorithm (Joty et al., 2013; Li et al., 2014a; Li et al., 2016), and the bottom-up algorithm (Hernault et al., 2010; Feng and Hirst, 2014; Kobayashi et al., 2019; Kobayashi et al., 2020). Recently, the generative algorithm (Mabona et al., 2019) and the top-down algorithm (Liu et al., 2019; Lin et al., 2019) tried out discourse parsing. At the high level (paragraph-level), Sporleder and Lascarides (2004) built paragraph-level discourse trees by bottom-up algorithm after pruning and revising the original discourse trees in the RST-DT.

In the literature, there are three kinds of strategies for representing larger discourse units. Most traditional machine learning methods (Feng and Hirst, 2014; Joty et al., 2015; Wang et al., 2017) extracted

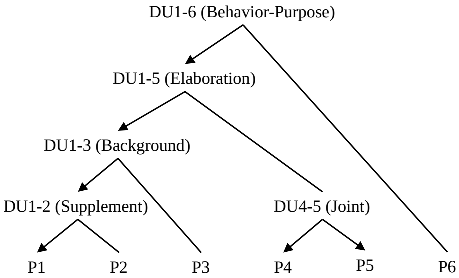
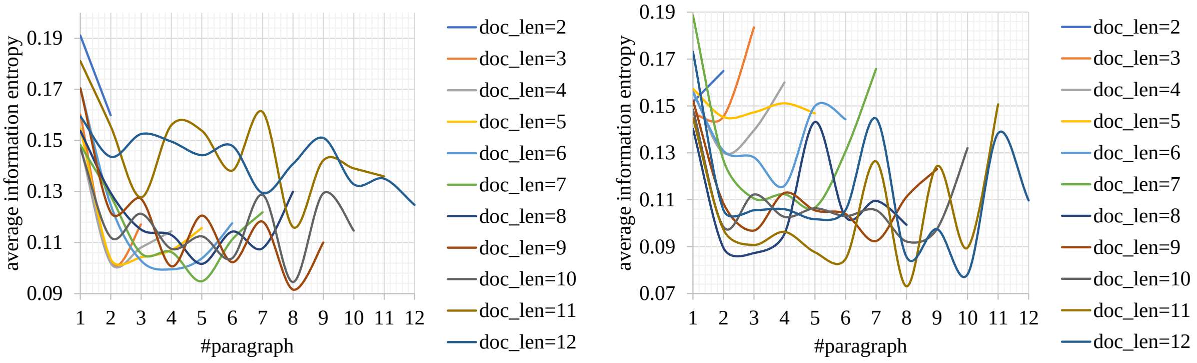


Figure 1: The paragraph-level discourse tree of chtb 0022 in MCDTB. P1-P6 refer to six PDUs and directed edges indicate nucleus discourse units. The relation (e.g., $Elaboration$ and $Joint$) of two related discourse units is shown in the bracket.



(a) Chinese MCDTB

(b) English RST-DT

Figure 2: Distribution of average information entropy in documents of different lengths in MCDTB and RST-DT at the paragraph-level.

semantic features from the first and last EDU of a large discourse unit. Neural network methods (Braud et al., 2016; Li et al., 2016) tended to choose all EDUs of the larger discourse unit as its semantic representation. Other works (Sagae, 2009; Braud et al., 2017) selected only one EDU (e.g. the first one) as the representation according to nuclearity.

In Chinese, a few discourse corpora (Li et al., 2014b; Zhou and Xue, 2015) were annotated at the clause-level. To the best of our knowledge, the MCDTB (Jiang et al., 2018b) is the only available paragraph-level Chinese corpus. The bottom-up algorithm (Chu et al., 2018; Jiang et al., 2018a) was the earliest applied to construct paragraph-level discourse structure trees in MCDTB. Recently, Zhou et al. (2019) built discourse structure trees by the shift-reduce algorithm in MCDTB. All of them used all PDUs for the semantic representation of a larger discourse unit.

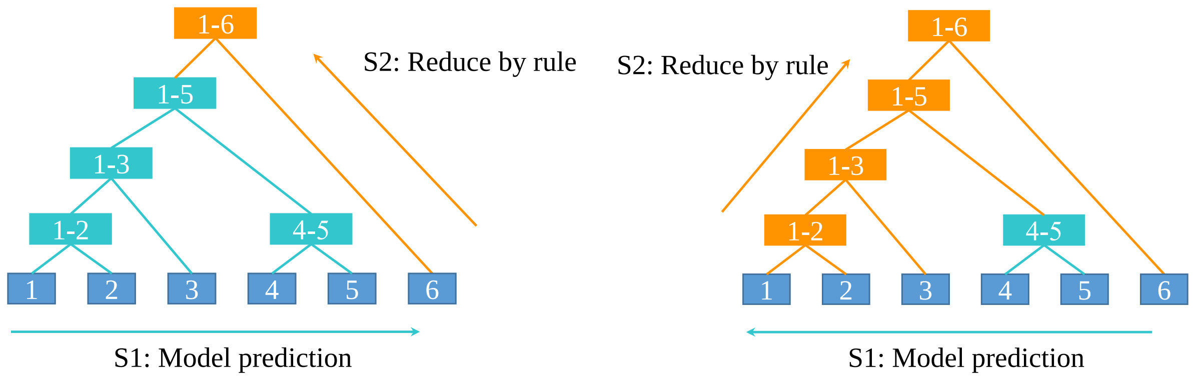
3 Motivation and Methods

In this section, we propose a global backward reading mode to construct a paragraph-level discourse tree by the shift-reduce algorithm to utilize the bias of Chinese discourse structure. In particular, we propose a triple semantic matching model based on BERT as a local model in the shift-reduce algorithm to capture the across discourse relationship better. In addition, we propose a local reverse reading mode to enhance the differentiation of discourse units coherence and obtain eight kinds of representation strategies to deal with the longer discourse units at the paragraph-level.

3.1 Global Backward Reading

Kaplan (1966) pointed out that different languages have different discourse structures, such as English is linear while Chinese is spiral. It means that the important topic of a document tends to be introduced at the beginning discourse unit in Chinese, and the following discourse units are all around this topic, while the important topics are relatively scattered in English. This difference is more significant at the paragraph level. Figure 2 shows the distribution of information entropy (Shannon, 1948) of documents in MCDTB and RST-DT at the paragraph-level. The gap of the average information entropy between the beginning paragraph and the end paragraph is greater in MCDTB than that in RST-DT. It indicates a discourse unit at the front part (closer to the beginning) tends to be more important and informative in MCDTB than RST-DT at the paragraph-level.

This difference will lead to the discourse structure trees in MCDTB being more biased to the left-branching trees, while those in RST-DT are more inclined to right-branching trees. We use the $P_l(t)$ value (Sampson, 1997), a production-based measure of left-branching for parsing trees, to measure the bias of branching. If a parsing tree is a complete left-branching tree, the $P_l(t)$ is close to 1. Statistically, the $P_l(t)$ value is 0.5044 and 0.3901 in MCDTB and RST-DT at the paragraph-level, respectively. More specifically, there are 67.22% left-branching trees ($P_l(t) \geq 0.5$) and 32.78% right-branching trees in MCDTB, while these figures are 35.23% and 64.77% in RST-DT at the paragraph level.



(a) The two stages of the forward reading

(b) The two stages of the backward reading

Figure 3: The two stages of the shift-reduce algorithm in two global reading modes.

Existing methods on RST-DT often process discourse units from the beginning to the end of a document to build discourse trees, such as the shift-reduce algorithm (Wang et al., 2017). However, these successful methods in English may not be transformed into Chinese well because of the bias of discourse structure mentioned above. Therefore, inspired by Bi-LSTM (Hochreiter and Schmidhuber, 1997) in processing text streams and bi-direction in syntax parsing (Sagae and Lavie, 2006), we propose a global backward reading mode that processes discourse units from the end to the beginning by the shift-reduce algorithm to utilize this branching bias of Chinese better.

In a typical shift-reduce algorithm, the parsing process is modeled as a sequence of *shift* and *reduce* actions on a stack and a queue; the stack is initially empty and the queue contains all PDUs in a document. At each step, the parser performs either *shift* or *reduce*: while *shift* pushes the first PDU in the queue to the top of the stack, *reduce* pops and merges the top 2 elements in the stack to yield a new sub-tree that is then pushed back to the top of the stack. When there are no candidate discourse units in the queue, the *reduce* action continues until a complete discourse tree is constructed.

Therefore, there are two stages in building the structure tree by the traditional shift-reduce algorithm in forward reading mode. As shown in Figure 3a, in the first stage, the parser needs a local model process discourse units from the beginning to the end of a document to predict the actions (*shift* or *reduce*). In the second stage, when there are no candidate discourse units in the queue but still candidate discourse units in the stack, the parser will operate a series of *reduce* on two discourse units at the top of the stack by rule instead of model, processing discourse units from the opposite direction.

It is natural because English has a right-branching bias of discourse structure. However, considering the left-branching bias in Chinese, we propose the backward reading mode, which processes discourse

S2	S1	Q1	Action
1(1)	2(2)	3(3)	Reduce
Null(Null)	1-2(2-1)	3(3)	Shift
1-2(2-1)	3(3)	4(4)	Reduce
Null(Null)	1-3(3-1)	4(4)	Shift
1-3(3-1)	4(4)	5(5)	Shift
4(4)	5(5)	6(6)	Reduce
1-3(3-1)	4-5(5-4)	6(6)	Reduce
Null(Null)	1-5(5-1)	6(6)	Shift
1-5(5-1)	6(6)	Null(Null)	**Reduce**

(a) An example of the forward reading

S2	S1	Q1	Action
6(6)	5(5)	4(4)	Shift
5(5)	4(4)	3(3)	Reduce
6(6)	5-4(4-5)	3(3)	Shift
5-4(4-5)	3(3)	2(2)	Shift
3(3)	2(2)	1(1)	Shift
2(2)	1(1)	Null(Null)	**Reduce**
3(3)	2-1(1-2)	Null(Null)	**Reduce**
5-4(4-5)	3-1(1-3)	Null(Null)	**Reduce**
6(6)	5-1(1-5)	Null(Null)	**Reduce**

(b) An example of the backward reading

Table 2: Examples of the two global reading modes for chtb 0022 in MCDTB, and the examples of local reverse reading are in the bracket.

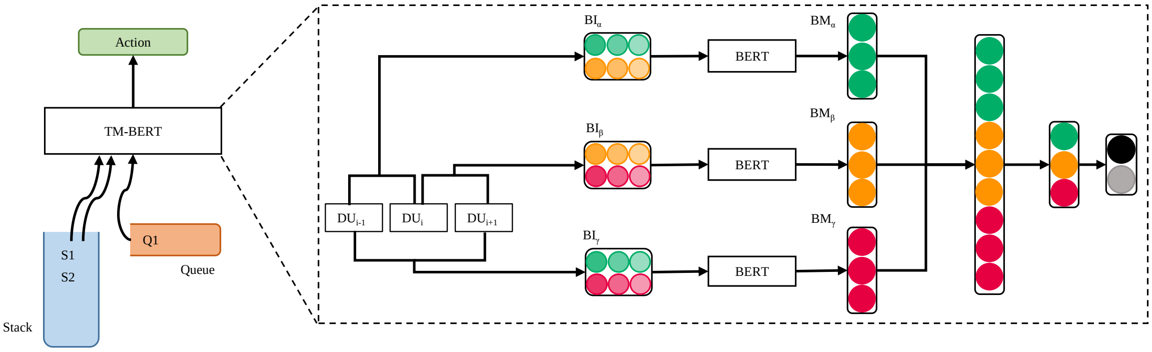


Figure 4: The TM-BERT model for predicting action by three discourse units in the shift-reduce algorithm. The input of BERT is BI that consists of three parts: Token, Segment and Position of two matching discourse units. BM is the embedding of $[CLS]$ position in the output of BERT.

units from the end to the beginning of a document by the local model and uses rules to merge discourse units from the beginning to the end in the second stage. Figure 3b shows that it will reduce the prediction deviation of the local model in the shift-reduce algorithm. As exemplified in Table 2a and Table 2b, it shows the building process of the forward reading mode and backward reading mode for chtb 0022, respectively. S2, S1, Q1, and Action refer to the second element of the stack, the first element of the stack, the first element in the queue and the decision under the current step, respectively. It can be seen that there are more decisions (in bold) made by rules in backward reading than in forward reading for a left-branching tree (chtb 0022), which can reduce the cascading errors of the local model because it is more consistent with the left-branching bias.

3.2 Local Model: TM-BERT

In a typical discourse parser using the shift-reduce algorithm, a local model is needed to predict the next step actions (*shift* or *reduce*) by three discourse units (two discourse units at the top of the stack and one at the head of the queue). The popular pre-training models such as BERT (Devlin et al., 2018), achieve excellent performance on semantic matching tasks. However, they only match the relationship between two discourse units. Therefore, we propose the Triple semantic Matching model based on BERT (TM-BERT), which can capture the relationship between multiple discourse units, as shown in Figure 4. We also use a similar TM-BERT model in the nuclearity recognition task and relation classification task.

In the encoding layer, previous studies (Wang et al., 2017; Kong and Zhou, 2017) relied on various kinds of language-specific sentence-level features (e.g., syntactic information, nuclear information, lexical chains, and tree information). However, their features cannot be directly applied to other languages or the paragraph level. Thanks to the excellent performance of BERT, we first encode the two adjacent discourse units into the input of BERT as BI that is fed to match the semantics of two discourse units. Then we select the embedding of $[CLS]$ position as the output (BM). Moreover, we propose a triple semantic matching mechanism based on this module, that views three adjacent discourse units as a triangle. That is, the model not only employs the pairs of $\alpha(DU_{i-1}, DU_i)$ and $\beta(DU_i, DU_{i+1})$ for semantic matching but also matches the across discourse unit pairs $\gamma(DU_{i-1}, DU_{i+1})$, as shown in Equation 1. It is worthwhile that we do not add any handcrafted features into the model to ensure its universality.

$$BM_k = BERT(BI_k), where\ k \in \{\alpha, \beta, \gamma\} \tag{1}$$

In the decoding layer, the concatenating output $(BM_\alpha, BM_\beta, BM_\gamma)$ is fed to a Softmax layer, as shown in Equation 2. During training, we use the Adam optimizer to optimize the network parameters by maximizing the log-likelihood loss function between the predicted label and the true label. The key hyper-parameters are following: batch-size=2, epoch=5, max-length=512, and learning-rate=1e-5.

$$y = Softmax(con(BM_\alpha, BM_\beta, BM_\gamma)) \tag{2}$$

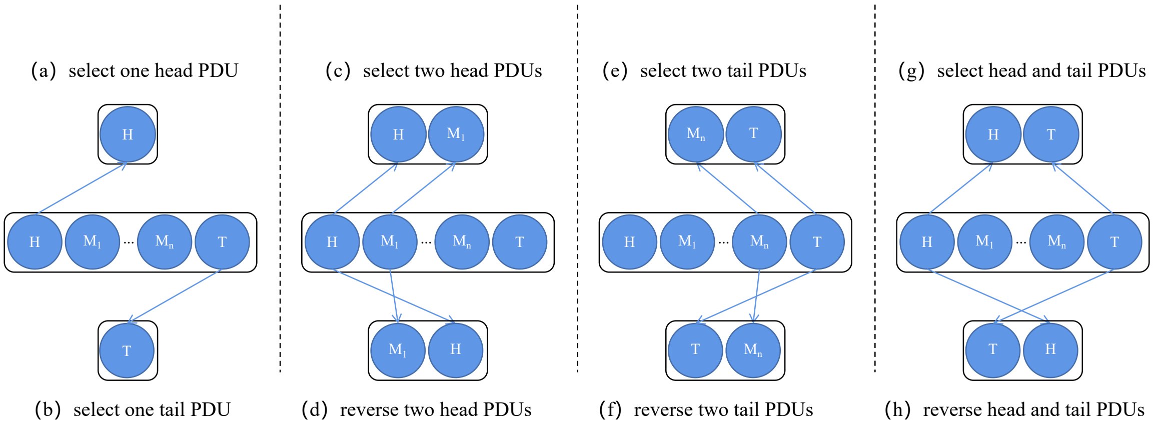

(a) select one head PDU
(b) select one tail PDU
(c) select two head PDUs
(d) reverse two head PDUs
(e) select two tail PDUs
(f) reverse two tail PDUs
(g) select head and tail PDUs
(h) reverse head and tail PDUs
H
M1
Mn
T

Figure 5: The eight strategies to represent a discourse unit by its PDUs. H (head), M (middle) and T (tail) mean the position of PDUs in a discourse unit.

3.3 Representation Strategies of the Discourse Unit

The paragraph-level discourse parsing is more difficult than that in clause-level in terms of the semantic representation of discourse units. At the paragraph-level, the elementary discourse units (PDUs) are paragraphs, which have longer text length (over 4.5 times) and fewer samples (less than 20%) than clause-level. On the other hand, the thematic unity of paragraphs (Longacre, 1979) reduces the explicit connection (such as fewer connectives as shown in Table 1) between two paragraphs that will make the differentiation of coherence between the discourse units smaller.

Therefore, we try to find a better representation of paragraph-level discourse units. There are three strategies to represent a discourse unit mentioned in Section 2 (Braud et al., 2017; Wang et al., 2017; Li et al., 2016): using a single PDU, using two PDUs (head and tail PDUs), and using all PDUs. Considering the longer text length of PDU at the paragraph-level, we select at most two PDUs to represent a discourse unit. In particular, we propose a local reverse reading mode to reverse PDUs in a discourse unit to represent itself. It can enhance the differentiation of the coherence between the discourse unit and adjacent discourse units. For example, if P1-P3 is coherent with P4-P5 (such as the discourse tree in Figure 1), it is still coherent after reversing (P3-P1 and P5-P4) because P1 is coherent with P5. However, if P1-P3 is not coherent with P4-P5, it is more non-coherent after reversing.

Finally, we get eight kinds of strategies based on local reverse reading: select one head PDU (H); select one tail PDU (T); select two head PDUs (HH); reverse two head PDUs (HHR); select two tail PDUs (TT); reverse two tail PDUs (TTR); select head and tail PDUs (HT); reverse head and tail PDUs (HTR), as shown in Figure 5.

4 Experimentation

4.1 Experimental Settings and Baselines

In Chinese, we evaluate our model on MCDTB with 720 documents annotated paragraph-level discourse tree. Following the previous work (Zhou et al., 2019; Wang et al., 2017), we transform the non-binary trees of the original data into right-binary trees in MCDTB. In particular, to balance the training set and the test set, we divide the documents containing different numbers of paragraphs into the training set (576 documents) and the test set (144 documents) according to the proportion. We randomly select 10% of the training set as the validation set.

Following Morey et al. (2017) and Jiang et al. (2018b), we use micro-F_1 as the evaluation metric, which evaluates how likely discourse units are correctly merged. This evaluation method is fairer though it will get lower scores than what was reported by the existing systems.

To verify the effectiveness of our model, we use five strong baselines. **Rule (left/right)** refers to producing a complete left-branching/right-branching tree by always merging the leftmost/rightmost two

Model	Span	Nuclearity	Relation
Rule (right)	39.88	-	-
Rule (left)	52.55	-	-
JPC18	54.71	48.38	26.28
ZCP19	56.11	47.76	27.67
LS19	56.25	46.21	28.75
BERT	57.19	48.38	28.44
TM-BERT (F)*	61.82	51.62	32.30
TM-BERT (F) + R*	63.37	54.10	35.70
TM-BERT (B) + R*	**67.08**	**56.41**	**37.09**

Table 3: The performance comparison (micro-F_1) on discourse tree construction. * denotes the model is significantly superior to BERT with p-value $<$ 0.05 (t-test).

discourse units. **JPC18** (Jiang et al., 2018a) is the first model to identify the paragraph-level discourse structure but has not built a discourse tree for a document yet. We modified it to build a paragraph-level discourse tree from bottom to top. **ZCP19** (Zhou et al., 2019) is the state-of-the-art model for constructing paragraph-level discourse trees by the shift-reduce algorithm. Since the above two systems did not recognize the nuclearity and classify the relation of a span, we did that through STGSN (Jiang et al., 2019) after generating new spans. Due to the overwhelming impact of BERT in many NLP applications, **BERT** is selected as the local model for building discourse trees with the shift-reduce algorithm, to be another baseline. **LS19** (Lin et al., 2019) is the state-of-the-art clause-level discourse parser by the shift-reduce algorithm in RST-DT, which is transformed to process Chinese.

4.2 Experimental Results

Table 3 presents the performance of TM-BERT and baselines on MCDTB. The **Rule (left)** that forms a complete left-branching discourse tree got 12.67 higher than the **Rule (right)**. It demonstrates that the discourse tree in Chinese MCDTB is biased toward left-branching. It also can be seen that **LS19**, a state-of-the-art model in English clause-level discourse parsing, did not perform well in Chinese. Moreover, three baselines (**JPC18**, **ZCP19** and **LS19**) get similar performance and do not improve much than textbfRule. It shows the paragraph-level discourse tree construction is still a challenge due to longer discourse units and fewer explicit connections. Besides, the pre-training model **BERT** can reduce dependence on samples and achieves the best performance of all baselines.

Because our TM-BERT can better match the relationship of multiple discourse units, our base model **TM-BERT (F)** (global forward reading) that achieved 61.82% micro-F_1 score, outperforms all baselines without any handcrafted features and our better model **TM-BERT (F) + R** (global forward reading and local reverse reading) yielded further improvements (+1.55 micro-F_1).

Notably, **TM-BERT (B) + R** that is the TM-BERT model with global backward reading and local reverse reading achieved the best performance. The key component contributing to this improvement is that the global backward reading mode can utilize the left-branching bias of discourse structure in Chinese to reduce the cascading errors of the local model and local reverse reading can enhance the differentiation of the coherence between discourse units. Besides, the improvements in Nuclearity and Relation mainly derive from the success of our model in Span.

5 Analysis

5.1 Analysis on Representation Strategies

Table 4a shows the performance of various representation strategies in global forward reading and backward reading. It can be seen that the best representation of a discourse unit is to select the head and tail PDUs. It is a similar conclusion to that of the research on the representation of long text by BERT (Sun et al., 2019). Additionally, taking only two consecutive head PDUs or tail PDUs to represent a discourse unit (HH or TT) will get worse performance due to the semantic representation being more biased in the

Strategies	Forward	Backward
H (T)	60.28 (60.12)	63.99 (62.13)
HH (HHR)	56.57 (60.90)	57.19 (62.60)
TT (TTR)	56.88 (58.89)	58.58 (60.28)
HT (HTR)	61.82 (**63.37**)	63.37 (**67.08**)

(a) Different representation strategies

Model	LB	RB
TM-BERT (F) + R	62.58	64.06
TM-BERT (B) + R	68.54	65.80

(b) Left-branching (LB) and Right-branching (RB) trees

Table 4: The micro-F_1 scores of the global forward and backward reading mode.

case of incomplete information. When using head PDU and tail PDU to represent a discourse unit, the local reverse reading (HTR) can improve the model performance through enhancing the differentiation of the coherence between discourse units, and it brings 1.55 and 3.71 improvements in global forward reading (achieving 63.37 micro-F_1) and global backward reading (achieving 67.08 micro-F_1), respectively. Moreover, the models with backward reading are better than the models with forward reading as we expected, which demonstrates the effectiveness of global backward reading.

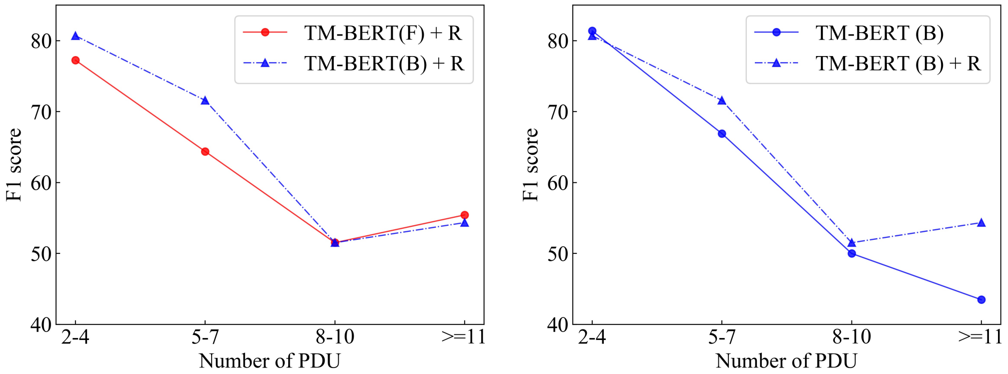
5.2 Analysis on Global and Local Reading

To reveal the reasons for performance improvement, we analyzed the performance of global backward reading and global forward reading on the left-branching trees and right-branching trees, as shown in Table 4b. It shows that due to the backward reading that can better utilize the bias of discourse structure in Chinese, the **TM-BERT (B) + R** model got more improvements (+5.96 micro-F_1) than **TM-BERT (F) + R** on left-branching trees than right-branching trees.

Moreover, we compared our three best models using the head and tail PDUs, as shown in Figure 6. Figure 6a shows a comparison of the global forward reading and the global backward reading. **TM-BERT (B) + R** is higher than **TM-BERT (F) + R** in short documents (2-7 PDUs in a document) because the discourse trees of short documents are more biased to left-branching trees and the global backward reading can utilize this bias better. Figure 6b shows the comparison of the global backward reading with or without local reverse reading. There is a huge gap between **TM-BERT (B)** and **TM-BERT (B) + R** in long documents (more than 11 PDUs). It proved the local reverse reading mode that reverses the position of PDUs in a discourse unit can enhance the differentiation of the coherence between the discourse unit and its adjacent discourse units, especially in long documents.

5.3 Evaluation on RST-DT

We also evaluate our model on English RST-DT. Following previous works, we use the same division as used by Li et al. (2014a), where 342 documents are in the training set and 38 documents are in the test set. We also transform the non-binary trees of the original data into the right-binary trees.



(a) Different global reading modes

(b) Different local reading modes

Figure 6: Micro-F_1 scores on different PDU numbers (Span).

Type	Model	Span
Statistic	FH14	**68.6**
	WLW17	68.0
	JE14	64.1
Neural	LLC16	64.5
	BCS17	62.7
	TM-BERT (F)	66.1
	TM-BERT (B)	64.9

(a) Full-level discourse tree construction

Model	Span
Rule (left)	28.57
Rule (right)	31.17
WLW17	37.40
TM-BERT (F)	**43.70**
TM-BERT (B)	42.86

(b) Paragraph-level discourse tree construction

Table 5: The micro-F_1 scores of our models on RST-DT.

We first evaluate our models on RST-DT to build a full-level structure tree and use **FH14** (Feng and Hirst, 2014), **JE14** (Ji and Eisenstein, 2014), **LLC16** (Li et al., 2016), **BCS17** (Braud et al., 2017) and **WLW17** (Wang et al., 2017) as the baselines. Table 5a illustrates the performance comparison on RST-DT in the full-level discourse tree construction. Our TM-BERT models with the global forward or backward reading modes got 66.1 and 64.9 (micro-F_1), respectively. Both of them outperformed all existing neural models and were comparable with the state-of-the-art models (**FH14** and **WLW17**).

Furthermore, we also use our model to build discourse trees on RST-DT at the paragraph-level to find out the difference between ours and the state-of-the-art model. Since the RST-DT does not explicitly distinguish sentence-level and paragraph-level discourse structures, about 21% of the paragraphs did not correspond to a discourse segment. In this case, we use the majority voting process to extract their paragraph-level structures, as done in Sporleder and Lascarides (2004). In addition to using **Rule (left)** and **Rule (right)** as baselines, we reproduced and evaluated the state-of-the-art system **WLW17** (Wang et al., 2017) as another baseline.

Table 5b shows the results of our models and three baselines. Our models with the global forward or backward reading modes got 6.3 and 5.46 higher than **WLW17** on the micro-F_1 score, which demonstrates the effectiveness of our models. Notably, **Rule (right)** and **TM-BERT (F)** are superior to **Rule (left)** and **TM-BERT (B)** on RST-DT at paragraph-level, while **Rule (left)** and **TM-BERT (B)** are better on MCDTB. This opposite result is because the cultural and linguistic differences bring the bias of discourse structure as we mentioned before. Besides, the performance of **WLW17** has a huge decline from full-level discourse parsing (68.0 micro-F_1) to the paragraph-level (37.40 micro-F_1) on RST-DT, which verifies that the task of paragraph-level discourse tree construction is more difficult than other levels.

6 Conclusions

In this paper, we propose a global backward reading mode in the shift-reduce algorithm, which processes discourse units from the end to the beginning of a document to utilize the left-branching bias of discourse structure better in Chinese. Besides, we propose a TM-BERT as the local model and eight representation strategies with the local reverse reading mode to enhance the differentiation of the coherence between discourse units. The experiment results on MCDTB show that our method achieved the state-of-the-art in discourse tree construction. In the future, we will explore how to construct an end-to-end paragraph-level discourse parser that well integrates discourse structure learning with nuclearity recognition and relation classification in Chinese.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (No. 61836007, 61772354 and 61773276.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *EMNLP*, pages 2212–2218.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of rst discourse parsers. In *COLING 2016*, pages 1903–1913.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. *arXiv preprint arXiv:1701.02946*.

Lynn Carlson, 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.

Xiaomin Chu, Feng Jiang, Yi Zhou, Guodong Zhou, and Qiaoming Zhu. 2018. Joint modeling of structure identification and nuclearity recognition in macro Chinese discourse treebank. In *COLING*, pages 536–546.

Arman Cohan and Nazli Goharian. 2018. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, 19(2-3):287–303.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *ACL*, volume 1, pages 511–521.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80, 12.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *ACL*, volume 1, pages 13–24.

Yanyan Jia, Yuan Ye, Yansong Feng, Yuxuan Lai, Rui Yan, and Dongyan Zhao. 2018. Modeling discourse cohesion for discourse parsing via memory network. In *ACL*, volume 2, pages 438–443.

Feng Jiang, Peifeng Li, Xiaomin Chu, Qiaoming Zhu, and Guodong Zhou. 2018a. Recognizing macro Chinese discourse structure on label degeneracy combination model. In *NLPCC*, pages 92–104.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018b. MCDTB: A Macro-level Chinese Discourse TreeBank. In *COLING*, pages 3493–3504.

Feng Jiang, Peifeng Li, and Qiaoming Zhu. 2019. Joint modeling of recognizing macro Chinese discourse nuclearity and relation based on structure and topic gated semantic network. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 276–286. Springer.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *ACL*, volume 1, pages 486–496.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Robert B Kaplan. 1966. Cultural thought patterns in intercultural education. *Language Learning*, 16:1–20.

Naoki Kobayashi, Tsutomu Hirao, Kengo Nakamura, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2019. Split or merge: Which is better for unsupervised rst parsing? In *EMNLP-IJCNLP*, pages 5801–5806.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down RST parsing utilizing granularity levels in documents. In *AAAI*. Association for the Advancement of Artificial Intelligence.

Varada Kolhatkar and Maite Taboada. 2017. Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online*, pages 11–17.

Fang Kong and Guodong Zhou. 2017. A CDT-styled end-to-end Chinese discourse parser. *TALLIP*, 16(4):26.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014a. Recursive deep models for discourse parsing. In *EMNLP*, pages 2061–2069.

Yancui Li, Fang Kong, Guodong Zhou, et al. 2014b. Building Chinese discourse corpus with connective-driven dependency tree structure. In *EMNLP*, pages 2105–2114.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *EMNLP*, pages 362–371.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *ACL*, pages 4190–4200.

Linlin Liu, Xiang Lin, Shafiq Joty, Simeng Han, and Lidong Bing. 2019. Hierarchical pointer net parsing. In *EMNLP-IJCNLP*, pages 1006–1016.

Robert E Longacre. 1979. The paragraph as a grammatical unit. In *Discourse and syntax*, pages 113–134. Brill.

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. In *EMNLP-IJCNLP*, pages 2284–2295.

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

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. In *EMNLP*, pages 1325–1330.

Jawad Sadek and Farid Meziane. 2016. A discourse-based approach for arabic question answering. *TALLIP*, 16(2):11.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *NAACL-HLT*, pages 129–132.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 81–84. Association for Computational Linguistics.

Geoffrey Sampson. 1997. Depth in english grammar. *Journal of Linguistics*, 33(1):131–151.

C. E Shannon. 1948. A mathematical theory of communication. *Bell Labs Technical Journal*, 27(4):379–423.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *COLING*, page 43.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *NAACL-HLT*, pages 566–574.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *ACL*, volume 2, pages 184–188.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *ACL*, pages 5021–5031, Online. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *COLING*, pages 559–570.

Amir Zeldes. 2017. The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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

Yi Zhou, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2019. Constructing chinese macro discourse tree via multiple views and word pair similarity. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 773–786. Springer.