Topic-driven Distant Supervision Framework for Macro-level Discourse Parsing via Transferring Models

Anonymous ACL submission

Abstract

Discourse parsing, the task of analyzing the internal rhetorical structure of texts, is a challenging problem in natural language processing due to the complex linguistic structure and lack of large-scale and high-quality corpora, especially at the macro level. Recent studies have attempted to overcome this limitation by utilizing results from other NLP tasks (source task) to distantly supervise the discourse parsing (target task). However, most of them only consider shallow connections across tasks using result-converting methods. It brings more cascading errors and makes it difficult to continue training due to the heterogeneity of the source and target task. To address these issues, we propose a topic-driven distant supervision framework via transferring models. The key recipe of this framework is to transfer the topic segmentation model into a discourse parser by additionally considering the global structural correlation instead of a simple converting result algorithm for transferring knowledge. The experimental results on two RST-style datasets, in both Chinese (MCDTB) and English (RST-DT), demonstrate that our method outperforms strong baselines not only in distant-supervised scenarios but also in fully supervised settings.

1 Introduction

In coherent documents, every discourse unit, ranging from clauses and sentences to paragraphs, is semantically interconnected. Discourse parsing, the process of uncovering the internal rhetorical structure formed by these units, plays a pivotal role in enhancing numerous Natural Language Processing (NLP) applications. These include automatic summarization (Cohan and Goharian, 2018), reading comprehension (Mihaylov and Frank, 2019), and machine translation (Tan et al., 2022), where understanding the document's discourse structure can contribute to performance improvements.

As one of the most popular theories of discourse parsing, Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) represents a document as a hierarchical Discourse Tree (DT) that can be split into micro and macro levels (Van Dijk and Kintsch, 1983). This paper mainly focuses on the macro-level, analyzing inter-paragraph relationships as shown in Figure 1, because it offers insights into the document's overall rhetorical organization at a higher level and provides a more comprehensive understanding critical for NLP applications' effectiveness (Kobayashi et al., 2021).

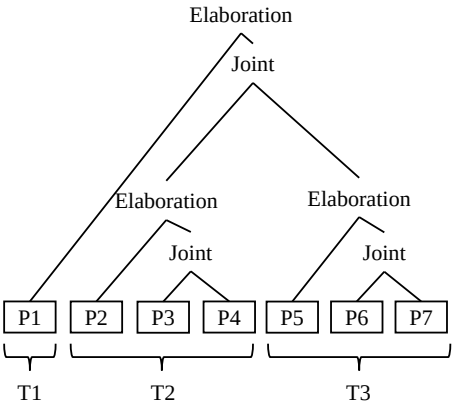


Figure 1: The example of part macro-level discourse tree of a document with seven paragraphs (P1-P7) (Jiang et al., 2021). Seven paragraphs belong to three topics (T1-T3): the Congress adopted the Arbitration Law and the Audit Law; the purpose and content of Arbitration Law; the purpose and content of the Audit Law.

Although supervised deep learning methods (Zhang et al., 2021; Jiang et al., 2021; Yu et al., 2022; Kobayashi et al., 2022) have made significant progress in discourse parsing, they are restricted by the limited size of high-quality manually annotated corpora (Carlson et al., 2003; Subba and Di Eugenio, 2009; Zeldes, 2017; Jiang et al., 2018; Peng et al., 2022). The intricate granularity required for annotations and the complexity of the annotation process severely limit the expansion of supervised discourse parsing research, particularly at the macro level.

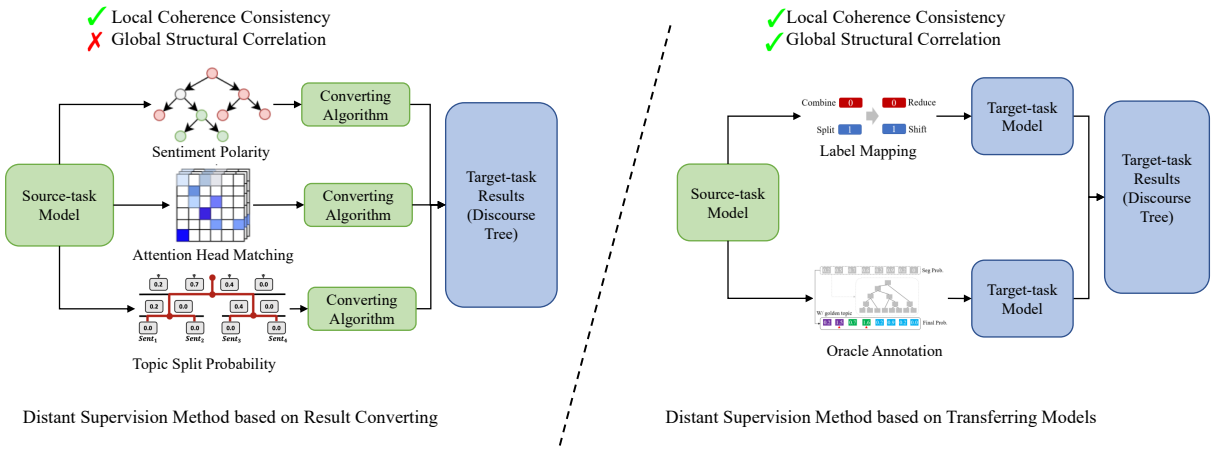


Figure 2: The overview of distant supervision framework for discourse parsing. The left part is the existing works that use result converting for distant supervision. The right part is our proposed method that uses transferring models for distant supervision.

Therefore, mainstream research shift to utilizing other NLP tasks (source task) (Huber and Carenini, 2019, 2020; Xiao et al., 2021; Huber et al., 2022) to distantly supervise the discourse parsing (target task), thereby mitigating the need for target task annotation data, as shown in Figure 2. Most of them convert the output results from sentiment polarity (Huber and Carenini, 2020), attention head matching (Xiao et al., 2021), and topic split probability (Huber et al., 2022) into a discourse structure tree (named **result converting method**), utilizing the local discourse coherence consistency of two discourse units [1].

However, such distant supervision methods encounter two challenges when applied to cross-task scenarios: (1) The additional cascading errors. These result-converting methods cannot leverage deep, explicit connections between source and target tasks, leading to the accumulation of errors that stem from the cross-task alignment when converting the results from the source task to the target task. (2) The difficulty in continuing training. Using result converting only gets the target-task result and does not transfer the model. It will suffer from the mismatch between the learning goal of the source-task model and the annotation form of the target-task training data, leading to not continuing learning from these data.

To address the challenges mentioned above, we introduce a topic-driven distant supervision framework via transferring models, which operate in a basic **transfer learning model** and a **teacher-student model**. The basic transfer learning model transfers the topic segmentation model into a discourse parser via mapping labels. Besides, in the teacher-student model, we first use the teacher model to generate a silver rhetorical structure corpus by oracle annotation. We then let the student model learn from such corpus to become a discourse parser. This framework not only inherits leveraging the local coherence consistency found in previous works (Huber et al., 2022) but also leveraging global discourse structure correlation (Jiang et al., 2021) between topic and rhetorical structures to distant supervision, thereby facilitating more accurate and effective discourse parsing.

The biggest difference from previous distant supervision methods is that our method seeks to construct a native target-task model (discourse parser) leveraging the source-task corpus instead of a converting algorithm for the source-task model. It has the following advantages achieved by bridging the two tasks through transfer models rather than converting prediction results. First, it harnesses the deeper relationships between topic structure and rhetorical structure, thereby reducing cascading errors when crossing tasks. Second, the transferred model can effectively utilize the source-task training data in the distant supervision scenario, while also benefiting from the target-task training data for continuing training.

We conduct the experiments on two RST-style corpora, the Chinese MCDTB and English RST-DT. The experimental results demonstrate that our method outperforms other strong baselines for macro-level discourse parsing in both distant su-

[1]The example of result converting method (Huber et al., 2022) can be seen in Appendix A.

Distant Supervision	Source Task	LCC.	GSC.	Can Continue Training on Target Task?
Result Converting (Huber and Carenini, 2020)	Sentiment Analysis	Sentiment Polarity	-	✗
Result Converting (Xiao et al., 2021)	Automatic Summarization	Attention Head Matching	-	✗
Result Converting (Huber et al., 2022)	Topic Segmentation	Topic Split Probability	-	✗
Transfer Model (our)	Topic Segmentation	Topic Split Probability	Label Mapping	✓
Transfer Model (our)	Topic Segmentation	Topic Split Probability	Oracle Annotation	✓

Table 1: The comparison of our methods and existing distant supervision methods in discourse parsing. LCC is short for Local Coherence Consistency and GSC is short for Global Structural Correlation.

pervision and supervised scenarios, affirming our framework's effectiveness.

2 Related Work

2.1 Topic Segmentation

Topic segmentation (Hearst, 1997) aims at identifying topic transitions within a text, distinguishing between topic maintenance and shifts. Typically, it involves determining whether each part of a text sequence represents a topic boundary. With the availability of large-scale topic corpora (Koshorek et al., 2018), supervised methods based on the pre-trained models have gained popularity.

Li et al. (2018) first proposed the SegBot model, which encodes text using a gated recurrent unit (GRU) module and employs a pointer network to determine topic segmentation points. Lukasik et al. (2020); Liu et al. (2023) separately framed topic segmentation as a sequential labeling task and modeling topic with hierarchical two-layer models. Xing et al. (2020); Yu et al. (2023); Gao et al. (2023) combined sequential labeling for topic segmentation enhanced by local coherence modeling. Jiang et al. (2021) introduced the TM-BERT model, which incorporates a local model with a sliding window to predict topic boundaries. Lee et al. (2023) further used the local BERT model segmenting topic via multi-task learning.

2.2 Distant Supervision Discourse Parser

Compared to flat topic structures, hierarchical rhetorical structures are more complex. Due to large-scale manually annotated corpora scarcity, recent researchers have turned to distant supervision methods for constructing discourse trees.

Huber and Carenini (2019, 2020) employed distant supervision to generate discourse trees based on sentiment analysis, utilizing the relationship between the sentiment polarity of child and parent nodes. Xiao et al. (2021) constructed distant supervision discourse trees using summarization. They established associations between Elementary Discourse Units (EDUs) through attention matrices in a transformer-based summarization model and then parsed the discourse tree using the CYK and CLE algorithms. Huber et al. (2022) utilized distant supervision based on topic segmentation to construct discourse trees. It greedily constructs a discourse tree from top to bottom, following the order of topic segmentation probabilities. The above three methods all convert the output cues for source-task models to the target task results.

3 Topic-driven Distant Supervision Framework for Macro-level Discourse Parsing via Transferring Models

As mentioned in the Introduction Section, previous work mainly utilized simpler and easier annotated NLP tasks (such as sentiment analysis (Huber and Carenini, 2020), automatic summarization (Xiao et al., 2021), topic segmentation (Huber et al., 2022), etc.) for distant supervision of discourse structure analysis, as shown in Table 1. Their approach is to use shallow cross-task connections (i.e., local coherence consistency) to convert the results of other tasks into discourse structure trees by converting algorithms.

However, these result-converting methods cannot leverage deep, explicit connections between source and target tasks, leading to the accumulation of errors that stem from two main sources. Except for the internal errors of source-task models that are inherent inaccuracies present in the source-task predictions, alignment issues occur when converting the results from the source task to the target task with **local coherence consistency** they used, where *two discourse units have a rhetorical relation if they are semantically closely related in other tasks*. Only considering local coherence consistency makes it difficult to fill the gap caused by the heterogeneity between the source task (classification) and the target task (hierarchical tree construction), leading to additional cascading errors when transferring knowledge.

In addition, the above methods still use the source-task model and only design converting al-

gorithms for converting the result across tasks. It causes them to be unable to use high-quality data from the target task for continuing training due to a mismatch between the learning goal of the source-task model and the annotation form of the target-task training data.

Therefore, we propose a topic-driven distant supervision framework via transferring models, which contain two variants: the transfer learning model based on label mapping and the teacher-student model based on oracle annotation, as shown in Figure 2. It reduces cascading errors by additionally considering the **global structural correlation** of topic and rhetorical structures (Jiang et al., 2021), which refers to *the topic structure reveals the skeleton of the rhetorical structure tree globally, and each topic contains a discourse sub-tree* (the build discourse tree with the golden topic structure can achieve about 83% F1-score), as shown in Figure 1. It transfers models into the target task, which can leverage the target-task data for continuing training.

3.1 Transfer Learning Model Based on Label Mapping

We first propose a basic transfer learning model based on label mapping. Instead of converting results directly, it maps the labels of the topic segmentation model into that of the rhetorical tree construction model using the global structure correlation, as shown in Figure 3.

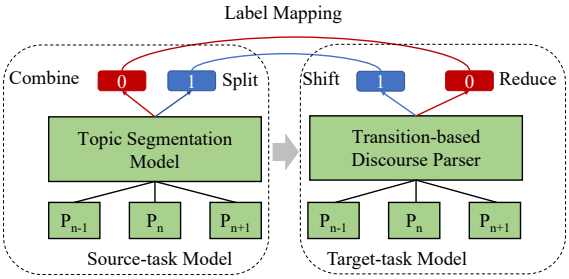


Figure 3: The architecture of transfer learning model based on label mapping.

Specifically, we adopt the sequential labeling model (Jiang et al., 2021) in the source task (topic segmentation), which uses a local TM-BERT model to segment topics through sliding windows. For each discourse unit (P_n), the model needs to predict whether it is the boundary of the topic according to the context, and the predicted results are labeled as *combine* or *split*. Then we map the label of this model based on the global structural correlation to make it a transition-based discourse parser (Wang et al., 2017), which views the discourse tree construction into a sequence of actions containing the *shift* and *reduce*. The labels *combine* and *reduce* are mapped to 0, and the *split* and *shift* are mapped to 1. Different from the Result Converting method (Huber et al., 2022), this label mapping not only uses the local coherence consistency but also uses the global structural correlation because the action label in transition-based discourse parser can further reveal the whole discourse tree from a global view. Additionally, it transfers the model into a native discourse parser which can be trained on the rhetorical structure corpus.

3.2 Teacher-Student Model Based on Oracle Annotation

Furthermore, we propose a teacher-student model based on oracle annotation, considering the deeper connections between rhetorical structures and topic structures. Leveraging the golden topic structure information from the source-task corpus, the teacher model first constructs a silver rhetorical structure corpus. Then, a student model is trained as a target-task model on this corpus for distant supervision, as shown in Figure 4.

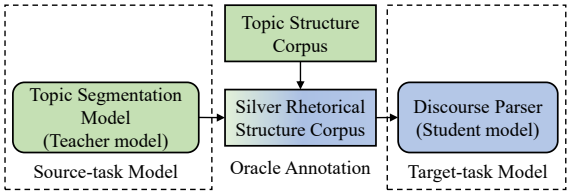


Figure 4: The architecture of the teacher-student model based on oracle annotation.

3.2.1 Teacher Model

In the teacher model, we follow the previous success model (Huber et al., 2022) that offers the possibility of using the topic segmentation model to construct rhetorical structure trees, but we add global structure correlation into it to build the rhetorical tree more accurately, as shown in Figure 5. Inspired by Jiang et al. (2021) using golden topic structures to assist discourse parsing in the rhetorical corpus and achieving much higher accuracy (about 83%), we propose the oracle annotation to build a silver rhetorical structure corpus by fusing these two methods.

Specifically, we first use a topic segmentation

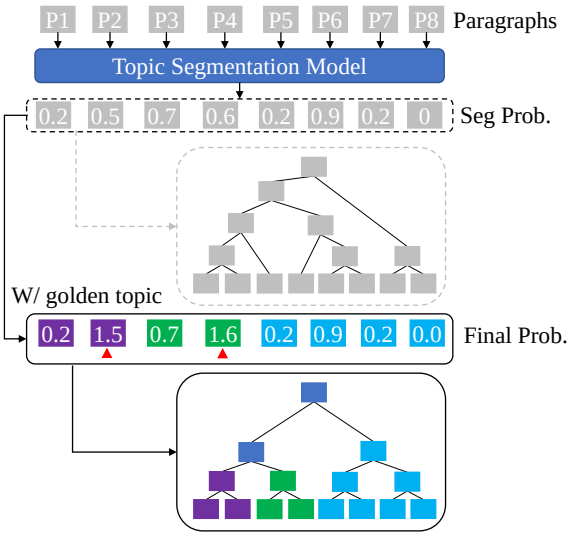


Figure 5: The example of creating the silver rhetorical structure tree by oracle annotation. Discourse units within the same color indicate belonging to the same topic, and the red triangle indicates that the discourse unit is the last one on that topic.

model[2] to predict the probability of each topic segment point (Seg Prob.), following previous work (Huber et al., 2022). Instead of directly using Seg Prob., we use the golden topic boundary as the constraint condition to build the discourse tree (Jiang et al., 2021). It means if the discourse unit is the last one in a topic section, it's Seg Prob. will be added 1 to have priority in building the discourse tree. Then, we greedily build the silver discourse rhetorical structure tree from top to down by the final probability (Final Prob.). It can ensure that the constructed discourse rhetorical tree is better with the golden topic structure.

Leveraging this oracle annotation, we conduct a ten-fold cross-validation of the source-task topic structure corpus to create the silver rhetorical structure corpus. In each fold, we use nine parts as training sets and the rest one part as the test set[3].

3.2.2 Student Model

After obtaining the silver rhetorical structure corpus, we train a target-task student model for distant supervision. Since we already oracle annotated the source-task topic corpus with the silver rhetorical structure, we can easily take any supervised discourse parser as the student model without any changes. Inspired by previous work (Kobayashi et al., 2019; He et al., 2022), we use a simple bi-directional pointer network (BLINK) as the student model [4]. The BLINK model consists of two popular pointer networks: a top-down split network (PT (Down)) and a bottom-up merge network (PT (Up)). When building a discourse rhetorical tree, the final operation of each step is determined by the maximum probability of the prediction of two networks, as shown in Figure 6. More details of BLINKs can be seen in Appendix C.

Therefore, the teacher-student model transfers the source-task model (teacher model) to the target-task model (student model) for distant supervision. Through the oracle annotation, the teacher model creates a more accurate discourse tree with the golden topic tree for student model learning. It transfers knowledge better than existing works by leveraging both local coherence consistency (Huber et al., 2022) and global structural correlation (Jiang et al., 2021). Since the student model is a native discourse parser, it can also use the target-task data for continuing training.

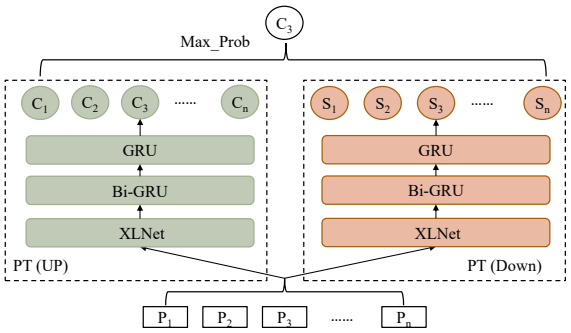


Figure 6: The architecture of student model (BLINK) for rhetorical structure tree construction.

4 Experimentation

4.1 Datasets and Evaluation Metrics

Source-Task Corpus. As the source corpus of building source-task silver macro rhetorical structure corpus, we select the CPTS (Jiang et al., 2023) and WIKI727 (Koshorek et al., 2018) as the macro topic structure Corpus in Chinese and English. CPTS is a macro-level topic structure corpus annotated 14393 Xinhua news documents from the

[2]Here, we use TM-BERT (Jiang et al., 2021). Although we have tried other models (e.g., BERT+Bi LSTM and Hier. BERT), TM-BERT achieves the highest performance.

[3]More details about the silver rhetorical corpus are shown in Appendix B.

[4]In English, we use one of the latest SOTA models (Kobayashi et al., 2022) as the student model, which is a top-down discourse parser based on DeBERTA (He et al., 2020). All hyperparameters are defaulted in the published paper.

Gigaword corpus[5]. In English, we randomly extract 5000 documents from WIKI727 as the English macro-level topic structure corpus. Similar to previous work (Huber et al., 2022), we use the first- and second-level section names as topic boundaries and lower-level section names as paragraph boundaries for the macro topic structure.

Target-Task Corpus. We verify the effectiveness of our distant supervision framework performance on Chinese MCDTB and English RST-DT. The former contains 720 documents annotated with macro discourse rhetorical structure where 80% of it is the train set and 20% is the testing set, following the previous work (Jiang et al., 2021). The latter contains 385 documents annotated with discourse rhetorical structure. Following previous works (Sporleder and Lascarides, 2004; Jiang et al., 2021; Huber et al., 2022), we prune and modify the original discourse tree in RST-DT to the macro level to evaluate discourse parsing performance.

It is worth noting that in the distant supervision scenario, we only use the source-task corpus as the training set to train our model and transfer it to the target task. In the supervised scenario, we further train the transferred model on the training sets of MCDTB and RST-DT, aligning with the practices of other supervised baselines.

Evaluation Metrics. The evaluation method is consistent with previous work (Morey et al., 2017; Jiang et al., 2021), which evaluates the span Micro-F1 score, which is equal to span accuracy when the discourse tree has already been converted to a complete binary tree. The details of the experimental setup are shown in Appendix E.

4.2 Baselines

We select the following models as the baselines, and more details can be seen in Appendix D.

Distant Supervision Method.

Chinese Baselines. Since there is no distant supervision method in Chinese, we select the **Result Converting** method (Huber et al., 2022), the SOTA English one, as a strong baseline and reproduce it with Chinese TM-BERT for a fair comparison.

English Baselines. Excepted for **Result Converting** method (Huber et al., 2022), we also add two other task distant supervision methods (**Parser**$_{senti.}$ (Huber and Carenini, 2020) and **Parser**$_{summ.}$ (Xiao et al., 2021)) as the baselines.

Supervised Method.

Chinese Baselines. We select the popular model **BERT** (Devlin et al., 2019), and the one of SOTA model **PDParser (w/o TS)** and **PDParser (w/ auto TS)** (Jiang et al., 2021) as strong baselines.

English Baselines. We select the classic models (**SL04** (Sporleder and Lascarides, 2004) and **WL17** (Wang et al., 2017)) and the SOTA models (**PDParser (w/ auto TS)** (Jiang et al., 2021), **SpanBERT** (Guz and Carenini, 2020), and **DeBERTa** (Kobayashi et al., 2022)) as the baselines.

4.3 Results on MCDTB

The experimental results are shown in Table 2. In distant supervision methods, our transfer learning model and the teacher-student model achieve 56.41% and 61.51%, which are 1.08% and 6.18% higher than the baseline Result Converting. Moreover, by utilizing the consistency of global discourse structure between topic structure and rhetorical structure, the teacher-student model is significantly improved than Result Converting and even close to the supervised method PDParser (w/ auto TS) (61.51 vs. 63.06), demonstrating that our proposed method can reduce cascading errors when crossing tasks.

Scenario	Method	Span
Distant supervision	Result Converting	55.33
Distant supervision	Transfer Learning (ours)	56.41
Distant supervision	Teacher-Student (ours)	61.51
Supervised	BERT	57.19
Supervised	PDParser (w/o TS)	63.06
Supervised	PDParser (w/ auto TS)	66.31
Supervised	Transfer Learning (ours)	66.15
Supervised	Teacher-Student (ours)	**68.01**

Table 2: The performance on MCDTB.

Turning to supervised learning methods, our transfer learning model and teacher-student model utilize target-task annotated data for continuing training based on the distant supervision model, achieving 66.15 and 68.01 and improving by 9.74 and 6.5, respectively. Our best model (Teacher-Student model) also exceeds the strongest baseline (PDParser (w/ auto TS)) by 1.7%. This proves that our method can simultaneously utilize a large amount of distant supervision silver data in the source-task corpus and high-quality manually annotated data in the target-task corpus, achieving better performance.

[5]https://catalog.ldc.upenn.edu/LDC2009T2

4.4 Results on RST-DT

Table 3 shows the performance of our proposed methods and baselines on English RST-DT. Similar to that in Chinese MCDTB, our proposed teacher-student model achieves the best performance (44.42%) among a variety of distant supervision models. Moreover, the teacher-student model with oracle annotation also achieves the best performance among supervised models in English, especially 3.37% higher than the latest SOTA model (DeBERTa). It demonstrates the generalization of our proposed method in English.

Scenario	Method	Span
Distant supervision	Parser$_{senti.}$	31.62
Distant supervision	Parser$_{summ.}$	32.09
Distant supervision	Result Converting	41.90
Distant supervision	Teacher-Student (ours)	44.42
Supervised	SL04	34.29
Supervised	WL17	37.40
Supervised	PDParser (w/ auto TS)	40.52
Supervised	SpanBERT	52.75
Supervised	DeBERTa	54.81
Supervised	Teacher-Student (ours)	**58.18**

Table 3: The performance on RST-DT.

5 Analysis

In this paper, we take the Chinese experiments as an example for analysis since there are few works focused on non-English languages.

5.1 Ablation Study of Transferring Models

We first perform an ablation study of our transfer-model-based framework to demonstrate its effectiveness, as shown in Table 4. All three distant supervision models utilize identical source-task data (ST Data) and models (ST Models). PT(down) and PT(up) represent two components of BLINK, each operating as a unidirectional model to construct a discourse tree. It is evident that the target-task model and data significantly boost the overall performance of our framework. Transfer Learning outperforms the Result-converting model due to the introduction of the TT Model. Moreover, PT(up) and PT(down) achieve superior outcomes compared to Transfer Learning, attributed to its training on the external TT Data via oracle annotation. BLINK, a bidirectional discourse parser, further contributes to performance enhancements, achieving improvements of 2.62 and 3.55 over the unidirectional PT(down) and PT(up) models.

Similar to the distant supervision scenario, our models maintain competitiveness when further trained on manually annotated target-task data (MCDTB). It is noteworthy that our best model, the Teacher-Student model integrated with BLINK, not only achieves a 6.5 improvement over the distant supervision counterpart but also surpasses the supervised BLINK model that was trained on MCDTB (68.01 vs. 63.37). It shows the efficacy of our method in leveraging both the additional silver data we newly generated and the high-quality gold data that already exists, by continuing training through a transferred model across different tasks.

Model	ST Data	ST Model	TT Data	TT Model	Span
Result Converting	CPTS	TM-BERT	-	-	55.33
Transfer Learning	CPTS	TM-BERT	-	TM-BERT (Map.)	56.41
Teacher-Student	CPTS	TM-BERT	CPTS_Dist	PT(up)	57.96
Teacher-Student	CPTS	TM-BERT	CPTS_Dist	PT(down)	58.89
Teacher-Student	CPTS	TM-BERT	CPTS_Dist	BLINK	**61.51**
PDParser(w/o TS)	-	-	MCDTB	TM-BERT	63.06
Base Model	-	-	MCDTB	BLINK	63.37
Teacher-Student	CPTS	TM-BERT	CPTS_Dist +MCDTB	PT(down)	64.14
Teacher-Student	CPTS	TM-BERT	CPTS_Dist +MCDTB	PT(up)	66.00
Teacher-Student	CPTS	TM-BERT	CPTS_Dist +MCDTB	BLINK	**68.01**

Table 4: The ablation study of our framework.

5.2 The Effect of Transferring Models in Different Layers of the Discourse Tree

Since our method transfers the model from different tasks, it cannot only work on distant supervision scenarios when lacking manually annotated data but also further leverage them via continuing training when we provide them. Therefore, we analyze the effect of transferring models from the performance of the model in different layers of the discourse tree, as shown in Figure 7.

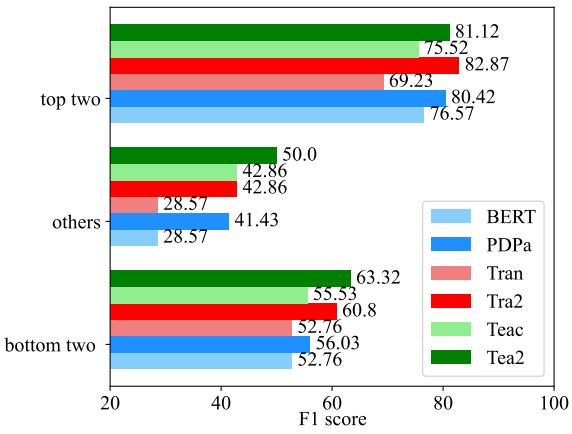


Figure 7: The performance of various models in different layers.

First of all, the distant supervision models based

on Transfer Learning (Tran) and Teacher-Student (Teac) are comparable to the two supervised learning models (BERT and PDParser (w/o TS) (PDPa)) on the bottom two layers and the middle layers, while it is slightly weaker than the supervised learning model on the top two layers. The main reason is that these transfer models are only trained on the topic structure corpus, which only guarantees the correctness of middle-level boundaries of discourse rhetorical structure according to the global structural correlation.

Secondly, after continuing training, these two transferred models (Tra2 and Tea2) can fully use high-quality manual annotation information to make up for this defect, achieving better performance. Specifically, the transfer learning model (Tra2) based on label mapping makes further improvement in the middle layer and the top two layers, with an increase of 13.64% and 14.29%, respectively, while the teacher-student model (Tea2) makes further improvement in the middle layer and the bottom two layers, with an increase of 7.14% and 7.79% respectively.

5.3 The Effect of Source-task Corpus in Different Lengths of the Document

Since our proposed methods also gain a significant improvement after continuing training under the supervised scenario, we further analyze the effect of source-task corpus on different length documents in the supervised scenario, as shown in Figure 8. The transfer learning model only directly takes the source-task topic structure corpus (CPTS) as the additional training data. Meanwhile, the teacher-student model enhances the source-task topic structure corpus (CPTS) with oracle annotation to create the silver rhetorical structure corpus (CPTS_dist) as the additional training data.

It can be seen that the transfer learning model achieves 81.38%, 67.99%, and 56.08% in documents with 2-10 paragraphs. Compared with the supervised baseline model (PDParser (w/o TS)), it has improved significantly in shorter documents with 2-4 paragraphs, reaching 5.52%. We believe that the transfer learning model learns the topic structure better through label mapping due to the short documents usually have a clearer topic structure, outperforming the baseline model on these documents.

In addition, the teacher-student model reaches 69.67%, 68.09%, and 52.99%, respectively, in documents with more than five paragraphs and increases significantly in 5-10 paragraph documents by 5.28% and 14.30% than the baseline model. One reason for this significant improvement is the large-scale silver rhetorical structure corpus (CPTS_dist) oracle annotated by golden topic structure can better cover complex discourse rhetorical structures. In the MCDTB corpus, there are 27-37 types of discourse rhetorical trees annotated in paragraphs 6-10, which do not increase with the number of paragraphs. However, in the CPTS_dist corpus, the types of discourse rhetorical structure trees in 6-10 paragraphs have increased from 35 to 437, covering complex discourse rhetorical structure trees better. More details are shown in Appendix F.

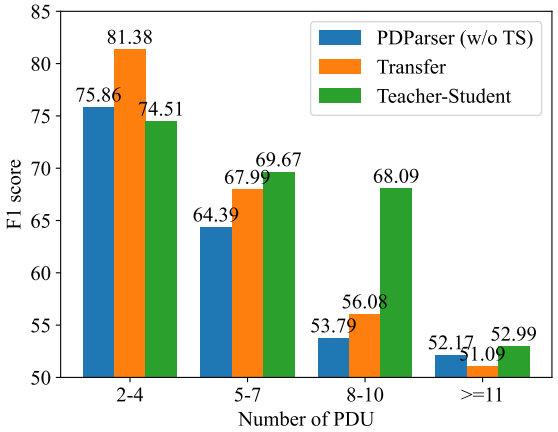


Figure 8: The performance of our models on different length documents in the supervised scenario.

6 Conclusion

In this paper, we propose a topic-driven distant supervision framework for macro-level discourse parsing via transferring models instead of result converting. The experiments in Chinese MCDTB and English RST-DT corpora have shown that our framework, through transferring models, can better utilize the deep connection between rhetorical structures and topic structures (global structural correlation) compared to the result-converting method, reducing cascading errors across tasks in distant supervision. Moreover, since our method involves transferring the model to the target task, we can further utilize the target-task corpus for further continuing training, which previous distant supervision methods are unable to do. We have also demonstrated the effectiveness of each part of the framework through analysis and ablation studies. In the future, we will jointly learn the rhetorical and topic structure and analyze the discourse structure more comprehensively.

Limitations

In this paper, we are mainly concerned about the completeness of the silver rhetorical structure corpus we constructed. Despite being annotated with both topic and rhetorical structure, the CPTS_dist and WIKI_dist corpus is not entirely correct, as its rhetorical structure was constructed through oracle annotation. We aim to improve its quality and incorporate human input in future work. Furthermore, we plan to expand its unannotated attributes, such as nuclearity and the rhetorical relationship between discourse units, to better represent the discourse structure of the text.

Another concern we think is that this framework can also be adapted to the micro-level, even full-level discourse parsing. However, this paper focuses on the macro-level which is more important and the performance of the model is still much lower. Also, since there are many successes at the micro-level, we are working on a better combination between the micro-level and macro-level.

Ethics Statement

We acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. Discourse parsing is a fundamental aspect of natural language processing that has many downstream benefits. It enables an understanding of the internal rhetorical structure of the text and does not generate any harmful or biased content. Additionally, the data we collect comes from open sources and is freely accessible to anyone. We will provide all details of the dataset and models to ensure reproducibility.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. *Current and New Directions in Discourse and Dialogue*, pages 85–112.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haoyu Gao, Rui Wang, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Unsupervised dialogue topic segmentation with topic-aware utterance representation. *arXiv preprint arXiv:2305.02747*.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.

Longwang He, Feng Jiang, Xiaoyi Bao, Yaxin Fan, Weihao Liu, Peifeng Li, and Xiaomin Chu. 2022. Bidirectional macro-level discourse parser based on oracle selection. In *Pacific Rim International Conference on Artificial Intelligence*, pages 224–239. Springer.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2020. MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.

Patrick Huber, Linzi Xing, and Giuseppe Carenini. 2022. Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Fang Kong. 2021. Hierarchical macro discourse parsing based on topic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 13152–13160.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2023. Advancing topic segmentation and outline generation in chinese texts: The paragraph-level topic representation, corpus, and benchmark. *arXiv preprint arXiv:2305.14790*.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A macro-level Chinese discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2022. A simple and strong baseline for end-to-end neural rst-style discourse parsing. *arXiv preprint arXiv:2210.08355*.

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 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5797–5802, Hong Kong, China. Association for Computational Linguistics.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Jeonghwan Lee, Jiyeong Han, Sunghoon Baek, and Min Song. 2023. Topic segmentation model focusing on local context. *arXiv preprint arXiv:2301.01935*.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4166–4172.

Zhengyuan Liu, Siti Umairah Md Salleh, Hong Choon Oh, Pavitra Krishnaswamy, and Nancy Chen. 2023. Joint dialogue topic segmentation and categorization: A case study on clinical spoken conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 185–193.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. 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.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

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 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.

Siyao Peng, Yang Janet Liu, and Amir Zeldes. 2022. GCDT: A Chinese RST treebank for multigenre and multilingual discourse parsing. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Caroline Sporleder and Alex Lascarides. 2004. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 43–49, Geneva, Switzerland. COLING.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574, Boulder, Colorado. Association for Computational Linguistics.

Xin Tan, Longyin Zhang, Fang Kong, and Guodong Zhou. 2022. Towards discourse-aware document-level neural machine translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4383–4389. Main Track.

Teun A Van Dijk and Walter Kintsch. 1983. *Strategies of discourse comprehension*. Acadamic Press.

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)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2023. Improving long document topic segmentation models with enhanced coherence modeling. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5592–5605.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST discourse parsing with second-stage EDU-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

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

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3946–3957, Online. Association for Computational Linguistics.

A The Process of the Topic-driven Distant Supervision by Result Converting

By calculating the probability of topic segmentation points of two discourse units, Huber et al. (2022) design a converting algorithm (if a topic segment point between two discourse units has a higher likelihood, they are likely not to have a rhetorical relationship.) to convert the outputs from the source-task model into whether there is a relationship between the two discourse units and then use a top-down greedy algorithm to construct a discourse structure tree.

Figure 9 shows an example of topic-driven distant supervision by result converting. The topic segmentation model could predict the sequence of EDU to get the segmentation probability (Seg Prob.). Then, the result-converting method will split the sequence according to the order of the probability of segment points. For example, sentence 2 ($Sent_2$) is the highest probability (0.7) that is split first. Then is sentence 4 ($Sent_4$) and sentence 3 ($Sent_3$). Therefore, it uses the top-down parsing method to convert the topic segmentation result into a rhetorical structure tree.

B The Details of Silver Rhetorical Structure Corpus

We select the CPTS (Jiang et al., 2023) and WIKI727K (Koshorek et al., 2018) corpora as our source-task data. In Chinese, we use all 14,393 news documents with annotated macro topic structures from CPTS. In English, we use 5,000 wiki documents with section names, following previous work (Huber et al., 2022), and use the first- and second-level section names as topic boundaries and lower-level section names as paragraph boundaries.

In our transfer learning method, we train a source-task topic segmentation model using the topic structure corpus and then map the labels to convert it into a rhetorical tree construction model.

In the teacher-student model, we use a ten-fold cross-validation method to oracle annotate the topic structure corpus into a silver rhetorical structure corpus (CPTS_dist and WIKI_dist). It means that we split the dataset into 10 folds, and the silver rhetoric structure on each fold is obtained by the topic segmentation model trained by the remaining nine datasets through the oracle annotation method.

C The Details of BLINK Model

PT (Down) and PT (Up) have the same architecture. In the encoder, we first use the pre-trained model XLNet to encode all paragraphs of the document. Then, we use XLNet to obtain the vector representation of each word in the input $W = \{w_1, w_2, ..., w_m\}$, where m represents the number of words input in the document. After that, we use the Bi-GRU module to encode W to obtain the overall semantic representation of the document $E = \{e_1, e_2, ..., e_m\}$, as shown in Eq. 1. Then, at each step t, we obtain the vector of each <*SEP*> token as the representation of paragraphs $P = \{p_{<t,1>}, p_{<t,2>}, ..., p_{<t,n>}\}$, where n is the number of paragraphs included in a document.

$$E, h_f = f_{Bi-GRU}(W, h_0) \quad (1)$$

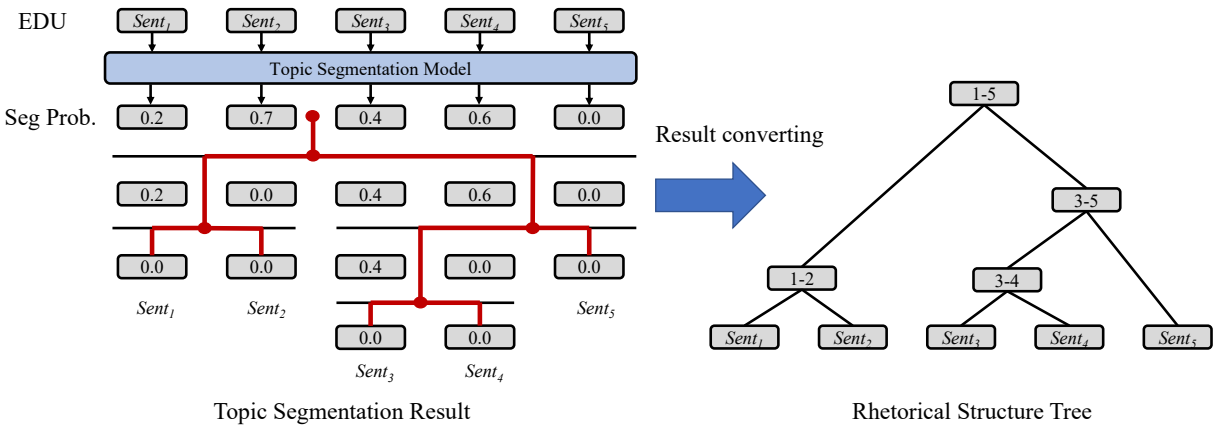


Figure 9: An example of the topic-driven distant supervision by result converting.

At the decoding step t, we feed the vector of the last paragraph ($p_{<t,l>}$) and the hidden layer vector h_{t-1} of the decoder in the previous time step into the decoder (GRU) to obtain the decoding representation (d_t) of the current discourse units sequence, as shown in Eq. 2.

$$d_t, h_t = f_{GRU}(p_{<t,l>}, h_{t-1}) \quad (2)$$

Finally, we calculate the attention score ($score$) between d_t and each paragraph through dot product (δ) to obtain the probability distribution of the split point or merge point at the current time step (t), as shown in Eqs. 3, 4 and 5.

$$score^m_{<t,i>} = \delta(d^m_t, p^m_{<t,i>}) \quad m \in c, s \quad (3)$$

$$C_p = argmax_p(Softmax(score^c_t)) \quad (4)$$

$$S_p = argmax_p(Softmax(score^s_t)) \quad (5)$$

where $C_p = \{c_1, c_2, ..., c_n\}$ represents the probability distribution of each paragraph as the combination point, $S_p = \{s_1, s_2, ..., s_n\}$ indicates the probability distribution of each paragraph as the split point. We select the highest probability value from C_p and S_p as the final action. For example, as shown in Figure 6, at this step, the BLINK model finally selects the maximum probability value (C_3), which means that paragraph 3 (P_3) and paragraph 4 (P_4) should be combined.

D The Details of Baselines

D.1 Supervision model in Chinese

BERT (Devlin et al., 2019) is a popular model in various NLP tasks, and we take it as the simple classification local model in the parser. **PDParser (w/o TS)** model and **PDParser (w/ auto TS)** model (Jiang et al., 2021). They are two SOTA models in Chinese discourse parsing, which are based on a triple semantic matching BERT model (TM-BERT). Their difference is that **PDParser (w/ auto TS)** model has the predicted topic boundaries to help build discourse trees while **PDParser (w/o TS)** did not have that.

D.2 Supervision model in English

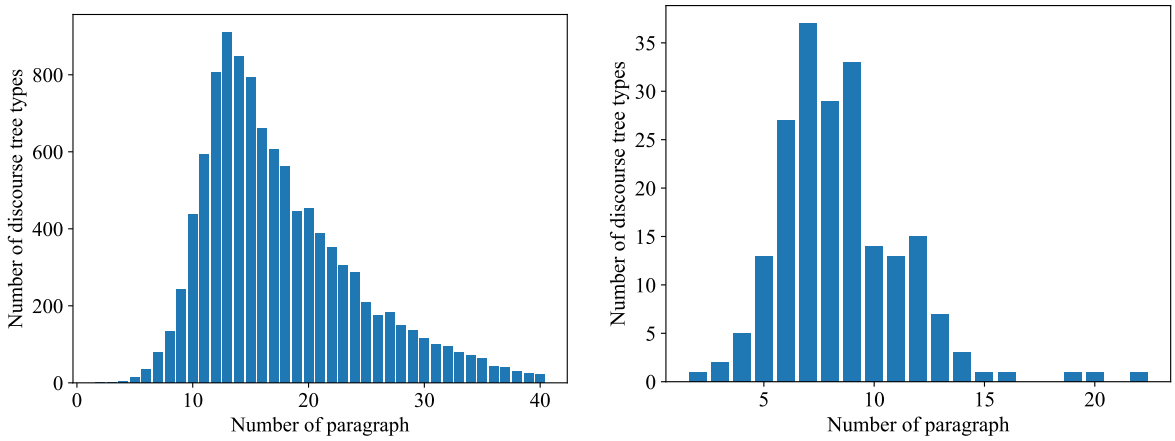
SL04 (Sporleder and Lascarides, 2004) is the first greedy bottom-up method to build macro-level discourse trees on the RST-DT. **WL17** (Wang et al., 2017) is a discourse parser based on the traditional SVM model and builds the discourse tree with the shift-reduce algorithm. **PDParser (w/ auto TS)** (Jiang et al., 2021) is a discourse parser using the synthetic topic structure to build the discourse tree. **SpanBERT** (Guz and Carenini, 2020) is one SOTA method based on the pre-trained language model (SpanBERT). It also uses the shift-reduce algorithm to build the discourse tree. **DeBERTa** (Kobayashi et al., 2022) is the latest SOTA model, which uses the DeBERTa as the local model to build the discourse tree from top to down.

E The Details of Experimental Settings

E.1 MCDTB

The hyper-parameters of the topic segmentation model used are following the previous work (Jiang et al., 2021): batch-size=2, epoch=10, max-length=512, and learning rate=1e-5. The pre-trained language model is the bert-base model (https://huggingface.co/bert-base-chinese).

In the teacher-student model we proposed, the main hyper-parameters of the student model (BLINK) are the following: the batch-size=2, epoch=50, the hidden size of GRU



(a) CPTS_dist (b) MCDTB

Figure 10: The main distribution of discourse tree types in CPTS_dist and MCDTB.

is 64, the layer number of GRU is 4, and the learning rate=1e-6. The pre-trained language model is the chinese-xlnet-mid model (https://huggingface.co/hfl/chinese-xlnet-mid).

We use an NVIDIA Tesla V100 GPU with 32GB to conduct the experiment.

E.2 RST-DT

The hyper-parameter of the topic segmentation model used is the same as the model in MCDTB, except that the pre-trained language model is an English bert-base-uncased model (https://huggingface.co/bert-base-uncased).

In the teacher-student model we proposed, the main hyper-parameters of the student model (Deberta) are the same as previous work (Kobayashi et al., 2022).

We use an NVIDIA RTX 3090 GPU with 24GB to conduct the experiment.

F The Main Distribution of Discourse Tree Types in CPTS_dist and MCDTB

Figure 10 shows the main distribution of discourse tree types in CPTS_dist and MCDTB. In CPTS_dist corpus, the discourse tree types increase with the number of paragraphs when the document has less than 13 paragraphs. Utilizing various types of discourse rhetorical structure trees can lead to a more robust structure tree construction model and improved performance. Additionally, even though there may be a decline in diversity in longer documents (#paragraphs > 13), it is still significantly more than the types in manually annotated MCDTB. For instance, documents with 25 paragraphs still contain over 200 different types of discourse structure trees in CPTS_dist, while MCDTB is basically not cover that.