RST-style Discourse Parsing Guided by Document-level Content Structures

Ming Li Texas A&M University liming@tamu.edu

Texas A&M University huangrh@cse.tamu.edu

Ruihong Huang

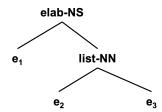
Abstract

Rhetorical Structure Theory based Discourse Parsing (RST-DP) explores how clauses, sentences, and large text spans compose a whole discourse and presents the rhetorical structure as a hierarchical tree. Existing RST parsing pipelines construct rhetorical structures without the knowledge of document-level content structures, which causes relatively low performance when predicting the discourse relations for large text spans. Recognizing the value of high-level content-related information in facilitating discourse relation recognition, we propose a novel pipeline for RST-DP that incorporates structure-aware news content sentence representations derived from the task of News Discourse Profiling. By incorporating only a few additional layers, this enhanced pipeline exhibits promising performance across various RST parsing metrics.

1 Introduction

Rhetorical Structure Theory based Discourse Parsing (RST-DP) (Mann and Thompson, 1988) aims to elucidate a hierarchical representation of the rhetorical structure within discourse by constructing a tree. Each leaf node in an RST tree represents an elementary discourse unit (EDU) while each internal node represents the relation between two text spans (an example shown in Figure 1). It explores how clauses, sentences, and large text spans compose a whole discourse, which is useful for many NLP applications (Kraus and Feuerriegel, 2019; Isonuma et al., 2019; Spangher et al., 2021).

Early RST models (Joty et al., 2013; Li et al., 2014, 2016; Wang et al., 2017) mostly utilize bottom-up approaches which limit the tree construction to consider only local context while recent models (Kobayashi et al., 2020; Zhang et al., 2020; Koto et al., 2021) mostly follow a top-down paradigm to fully utilize the global context. Despite the achievements of previous models, they might



- e₁: it offered to buy \$ 500 million of its convertible preferred stock from the Manville Personal Injury Settlement Trust in a move
- e2: that would improve the trust's liquidity
- e₃: and reduce the potential number of Manville shares outstanding.

Figure 1: An example discourse tree from English RST Discourse Treebank (Carlson et al., 2001) development set with small text spans. e_1 , e_2 and e_3 are EDUs, *elab* and *list* are rhetorical relations. *NS*, *NN* are nuclearity relations where *N* represents the nucleus and *S* represents the satellite.

overlook the performance discrepancy in predicting rhetorical relations between small and large text spans. The failure to address this disparity resulted in an inadequate exploration of the potential benefits offered by document-level content structures in guiding the recognition of rhetorical relations.

Since RST discourse parsing aims to present the whole document into one tree, there is a significant disparity in length and amounts between the smallest text spans and relatively larger text spans. The smallest text spans (EDUs) are at clause level with large amounts in one discourse while larger text spans can be paragraph-level with only a few in each discourse. This substantial difference in length and amounts poses challenges for existing methods in maintaining their performance when confronted with exceptionally lengthy text spans. To accurately identify the discourse relation between two relatively large text spans, the model must possess a comprehensive understanding of

¹Quantitive analysis on the performance disparity between small and large text spans can be found in Appendix A.

Left Text Span:

Television news, of course, has always been part show-biz. Broadcasters have a healthy appreciation of the role entertainment values play in captivating an audience. But, as CBS Broadcast Group president Howard Stringer puts it, the network now needs to broaden the horizons of nonfiction television, and that includes some experimentation. (Expectation)

Right Text Span:

Since its premiere Sept. 16, the show on which Ms. Chung appears has used an actor to portray the Rev. Vernon Johns, a civil-rights leader, and one to play a teenage drug dealer. It has depicted the bombing of Pan Am flight 103 over the Scottish town of Lockerbie. On Oct. 21, it did a rendition of the kidnapping and imprisonment of Associated Press correspondent Terry Anderson, who was abducted in March 1985 and is believed to be held in Lebanon. The production had actors playing Mr. Anderson and former hostages David Jacobsen, the Rev. Benjamin Weir and Father Lawrence Jenco. (Background Information / Previous Event)

Figure 2: An example with relatively large text spans. The model needs to predict the discourse relation and nuclear relation between these two large text spans. The golden label of this example is *Elaboration* and *Nuclear-Satellite*. In parentheses (green) are the possible news content type labels predicted by the news discourse profiling system. The incorporation of news content type labels serves as a valuable aid in comprehending the core information encapsulated within lengthy textual contexts, thereby facilitating informed decision-making.

the content contained within both segments which is a nontrivial task.

Based on the above discussion, the primary objective of this paper is to integrate the news-specific content information into the framework of RST discourse parsing as extra guidance on recognizing the relations between text spans. Specifically, the content information is derived from the news discourse profiling (NDP) task, which assigns one of eight distinct content types to each sentence within a news article. The content types include the main news event description, its immediate consequences, its direct context informing contents, and other supporting contents. ² These content types effectively capture the common discourse roles of sentences when describing a news story, which requires a comprehensive understanding of the underlying document-level content structures. This document-level content can serve as valuable guidance for the classification of nuclearity and rhetorical relations, especially for large text spans.

Figure 1 presents an illustrative discourse tree from the English RST Discourse dataset (Carlson et al., 2001). This particular discourse tree comprises three EDUs. *elab* and *list* are rhetorical relations and *NS*, *NN* are nuclearity relations. Given their simplicity, RST models can easily comprehend the primary content of each EDU and proceed to analyze their interrelations. However, complexities arise when the textual spans escalate in size. Figure 2 showcases an instance involving relatively large text spans. Unlike the recognition of small spans, accurately predicting large spans necessitates a profound comprehension of the textual con-

tent. With the provided discourse-level content information in parentheses³, it reduces the overload for the model itself to extract its main content but can concentrate more on relation analysis. The incorporation of this content-type information serves as a valuable aid in comprehending the core content encapsulated within lengthy textual contexts, thereby facilitating informed decision-making.

Therefore, we incorporate content structures and introduce a new pipeline for RST discourse parsing called C2RNet, (C2R is short for "from Content structure to Rhetorical structure"), where two learning branches conduct news discourse profiling (NDP branch) and RST discourse parsing (RST branch) respectively and sentence embeddings derived from the NDP branch will be sent to the RST branch for RST discourse parsing. The two branches operate in parallel, ensuring that additional processing time is not required. To our knowledge, we are the first to incorporate extra news content structure to facilitate RST parsing. Comparative evaluations against a baseline system that does not leverage content structures demonstrate the promising performance of C2RNet on the English RST Discourse dataset.

2 The Architecture of C2RNet

The pipeline illustrated in Figure 3 comprises three key components: a shared language model, an NDP branch, and an RST branch. Given that our primary objective is to incorporate content-type information into the RST discourse parsing task, instead of

²NDP task is introduced in detail in Appendix C.

³Labels in parentheses are used to illustrate how content information help grasp the key points of text spans. The final predicted categories are not needed in our system.

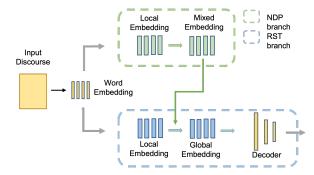


Figure 3: Overall C2RNet pipeline. The NDP branch is represented by the upper dashed box, while the RST branch is depicted by the lower dashed box. Initially, the input discourse is fed into a pretrained language model to generate shared word embeddings. Subsequently, both the NDP branch and RST branch operate in parallel. The combined sentence embedding derived from the NDP branch is then transmitted to the RST branch, enabling the incorporation of comprehensive event-related knowledge.

designing intricate model architectures, we adopt existing off-the-shelf model structures to validate the effectiveness of our proposed method.

2.1 NDP branch

To maintain the overall efficiency of the pipeline, we adopt the NDP branch proposed by LimNet (Li et al., 2022). This branch incorporates two self-attention modules (Bahdanau et al., 2014; Chorowski et al., 2015) to generate embeddings at the EDU level. The initial self-attention module computes attention weights for each word within the local EDU, resulting in local EDU embeddings. Subsequently, the second self-attention module calculates attention weights across the entire discourse, yielding the global EDU embeddings. Finally, the mixed EDU embeddings are obtained by combining the local and global EDU embeddings through addition.

2.2 RST branch

The RST branch in our approach is based on the hierarchical structure proposed by (Koto et al., 2021). We re-implement this structure, which consists of two BiLSTM layers, to capture the discourse structure. The first BiLSTM layer is applied to the local Elementary Discourse Units (EDUs), followed by average pooling to obtain the local EDU embeddings. These local EDU embeddings are then combined with the mixed EDU embeddings obtained from the NDP branch. The concatenated embeddings serve as the input for the second BiLSTM

layer, which produces the global EDU embeddings for the RST segmenter.

For the RST decoder part, we follow the same implementation from Koto et al. (2021), where implicit paragraph boundary features are concatenated and the RST trees are constructed iteratively. During each iteration, the input sentence is split at the position with the highest probability, which is determined by the model. To obtain representative vectors for the left and right text spans at each splitting iteration, average pooling operations are applied to the corresponding vectors. These left and right vectors are then concatenated together and passed through a final fully-connected layer. This final layer generates the joint probability distribution over the nuclearity and rhetorical relations.

3 Experiments

3.1 Dataset

We use the English RST Discourse Treebank (Carlson et al., 2001) for our C2RNet. It is based on the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993), which contains 347 documents for training, and 38 documents for testing. For a fair comparison, we use the same development set as Koto et al. (2021) which contains 35 documents from the training set.

3.2 Implementation Details

The NDP model LiMNet (Li et al., 2022) was trained in advance and the weights are transferred to the NDP branch of C2RNet as the initial weights. The weights of the NDP branch are fixed for the initial 40 epochs and then optimized under only RST data. Our pipeline is trained using Adam optimizer (Kingma and Ba, 2014) with the initial learning rate of 5e-4, epsilon of 1e-6 for 150 epochs. The dropout rate (Srivastava et al., 2014) is 0.5 for all experiments. By default, we employ the t5-large model (Raffel et al., 2020) from the huggingface library (Wolf et al., 2019) as our pretrained language model. The pretrained language model is not fine-tuned but rather kept fixed to ensure the efficiency and scalability of our model. 4

3.3 Evaluation

Table 1 presents the results of various models, examining their performance in RST discourse pars-

⁴The performance of our model with other language models is provided in Appendix B, where we present the results and analyses for different language model configurations.

)riginal	Parsev	al	RST Parseval					
	S	N	R	F	S	N	R	F		
(Hayashi et al., 2016)	65.1	54.6	44.7	44.1	82.6	66.6	54.6	54.3		
(Li et al., 2016)	64.5	54.0	38.1	36.6	82.2	66.5	51.4	50.6		
(Braud et al., 2017)	62.7	54.5	45.5	45.1	81.3	68.1	56.3	56.0		
(Yu et al., 2018)	71.4	60.3	49.2	48.1	85.6	72.9	59.8	59.3		
(Mabona et al., 2019)	67.1	57.4	45.5	45.0	-	-	-	_		
(Kobayashi et al., 2020)	-	-	-	-	87.0	74.6	60.0	-		
(Zhang et al., 2020)	67.2	55.5	45.3	44.3	-	-	-	_		
(Nguyen et al., 2021)	74.3	64.3	51.6	50.2	87.6	76.0	61.8			
(Koto et al., 2021)	73.1	62.3	51.5	50.3	86.6	73.7	61.5	60.9		
(Zhang et al., 2021)	76.3	65.5	55.6	53.8	-	_	-	_		
(Yu et al., 2022)	76.4	66.1	54.5	53.5	-	-	-	-		
Baseline	75.4	64.1	53.6	52.1	87.7	75.0	63.1	62.3		
C2RNet (ours)	76.8	66.2	55.4	53.8	88.4	76.5	64.5	63.8		
Human	78.7	66.8	57.1	55.0	88.3	77.3	65.4	64.7		

Table 1: RST discourse parsing results on the test set of the RST dataset, using original Parseval and RST Parseval metrics (Marcu, 2000). S, N, R, and F represent Span, Nuclearity, Relation, and Full. The results of ours are averaged over 3 random runs and the highest performances are in bold. *Baseline* represents the model without an NDP branch and other configurations are kept the same.

ing. The *Baseline* model refers to the configuration where the NDP branch is excluded, while maintaining all other settings identical to C2RNet, including the pretrained language model. Our *C2RNet* model consistently surpasses the *Baseline* across all metrics. These findings affirm that incorporating content structure-aware sentence representations provides notable advantages in all aspects of RST discourse parsing, encompassing span identification, nuclearity prediction, and rhetorical relation recognition. ⁵

3.4 Result Analysis

We compared the performance of our model and the baseline model on spans of different lengths. The length of a span is defined as the number of EDUs subsumed by a subtree. We divided the spans into three groups based on their lengths. The first group consists of minimal length spans containing exactly 2 EDUs, (for example, e_2 and e_3 in Figure 1), which accounts for 34.4% of all data. The remaining spans were split into medium-length spans (3, 4, and 5 EDUs) and long spans (more than 5 EDUs), (for example, two text spans in Figure 2), representing 35.5% and 30% of all spans, respectively. Table 2 displays the accuracy gaps between our model and the baseline model, depicting the variances when predicting different text spans within the test set. The results reveal that the incorpora-

	Span = 2	$2 < \text{Span} \le 5$	Span > 5
Nuclearity	-0.4%	+2.0%	+1.9%
Rhetorical relations	0%	-0.7%	+1.6%

Table 2: The accuracy gaps between our model and the baseline model, when parsing text spans of different lengths in the test set.

tion of content structures proves to be particularly advantageous in the parsing of longer text spans, as observed in both nuclearity and rhetorical relation prediction. Specifically, our model achieves superior performance in nuclearity prediction for both medium-length and very long spans, while for rhetorical relation prediction, our model demonstrates a significant advantage primarily in long spans. ⁶

4 Conclusion

In this paper, we propose a novel pipeline C2RNet for RST discourse parsing that leverages document-level content knowledge to enhance the recognition of rhetorical relations. Our approach incorporates content structure-aware sentence representations, which prove to be beneficial for enhancing discourse parsing, encompassing span identification, nuclearity prediction, and rhetorical relation prediction, especially for relatively large text spans. With only a few layers added, our model achieves promising performance.

⁵Detailed analysis on the effects of NDP labels is presented in Appendix D.

⁶Detailed table and analysis is provided in Appendix A.

Limitations

A key limitation of our proposed approach lies in its dependence on advanced language models. In order to maintain efficiency throughout the pipeline, we employ a single shared pretrained language model for both branches, necessitating the selection of a sufficiently powerful language model capable of supporting both tasks. Our additional results, presented in the appendix, demonstrate that employing a weaker language model leads to a decrease in performance. Nonetheless, our C2RNet model consistently outperforms the baseline models by a considerable margin, highlighting its robustness and effectiveness.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.
- Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc.
- Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.
- 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.
- Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles. Association for Computational Linguistics.
- Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2142–2152, Florence, Italy. Association for Computational Linguistics.
- Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multisentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria. Association for Computational Linguistics.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Topdown rst parsing utilizing granularity levels in documents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106.
- Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.
- Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory:learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79.
- Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.
- Ming Li, Sijing Yu, and Ruihong Huang. 2022. Less is more: Simplifying feature extractors prevents overfitting for neural discourse parsing models. *arXiv* preprint arXiv:2210.09537.

- Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371, Austin, Texas. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure 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 2284–2295, Hong Kong, China. Association for Computational Linguistics.
- William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. RST parsing from scratch. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1613–1625, Online. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

- 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv* preprint *arXiv*:1910.03771.
- Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- 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.
- 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.
- Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.

A Performance on different text spans

In this section, we present a detailed analysis of the performance of the baseline model and our proposed C2RNet model in predicting different text spans. We evaluate their performance on nuclear recognition and rhetorical relation recognition, which are reported in Table 3 and Table 4, respectively. Both the baseline model and our C2RNet model employ the fixed T5 language models. The metric \leq *threshold* represents the overall accuracy when predicting text spans that are less than or equal to the specified length threshold. Conversely, the metric > *threshold* represents the overall accuracy when predicting text spans that are greater than the specified length threshold.

Text Span Length	3	4	5	6	7	8	9	10	11	13	15
Baseline (> threshold)	43.3	37.9	33.3	30.8	30.5	30.2	28.1	27.2	26.8	27.3	26.9
Baseline (\leq threshold)	75.1	72.8	71.7	70.1	68.9	67.8	67.4	66.9	66.5	65.3	64.6
C2RNet (> threshold)	44.9	39.7	35.8	34.2	33.5	32.1	31.0	30.1	30.0	29.2	28.2
$C2RNet (\leq threshold)$	75.8	73.5	72.0	70.3	69.4	68.8	68.3	67.8	67.3	66.4	65.9
Difference (> threshold)	1.6	1.8	2.5	3.4	3.0	1.9	2.9	2.9	3.2	1.9	1.3
$Difference \ (\leq threshold \)$	0.7	0.7	0.3	0.2	0.5	1.0	0.9	0.9	0.8	1.1	1.3

Table 3: Nuclear relation recognition accuracies in predicting text spans with different lengths. > threshold represents the overall accuracy when predicting the text spans greater than the given length. \le threshold represents the overall accuracy when predicting the text spans less or equal to the given length. Difference represents the gap between the baseline model and our C2RNet.

Text Span Length	3	4	5	6	7	8	9	10	11	13	15
Baseline (> threshold)	32.1	26.6	23.8	22.3	22.1	21.1	20.7	19.9	19.6	20.4	20.1
Baseline (\leq threshold)	62.4	60.3	58.6	56.9	55.8	55.0	54.3	53.9	53.6	52.1	51.9
C2RNet (> threshold)	33.2	28.2	25.1	23.5	23.0	22.3	22.1	21.8	21.9	20.7	20.7
$C2RNet (\leq threshold)$	62.0	60.0	58.4	56.9	56.0	55.2	54.6	54.1	53.6	52.6	52.4
Difference (> threshold)	1.1	1.6	1.3	1.2	0.9	1.2	1.4	1.9	2.3	0.8	0.6
$Difference \ (\leq threshold \)$	-0.4	-0.3	-0.2	0.0	0.2	0.2	0.3	0.2	0.0	0.4	0.5

Table 4: Rhetorical relation recognition accuracies in predicting text spans with different lengths. > threshold represents the overall accuracy when predicting the text spans greater than the given length. \le threshold represents the overall accuracy when predicting the text spans less or equal to the given length. Difference represents the gap between the baseline model and our C2RNet.

The *Difference* column quantifies the performance gap between the baseline model and our C2RNet model.

From both tables, a consistent trend is observed: as the length of the text span increases, the accuracies for both nuclear recognition and rhetorical relation recognition tend to decrease. This finding is consistent for both the baseline model and our proposed model, indicating the challenge of accurately recognizing subtrees with relatively larger text spans. Our C2RNet demonstrates improved performance compared to the baseline model, particularly for subtrees with moderately long text spans. As presented in Table 3, when predicting nuclear relations, our model consistently achieves better performance across all span lengths. Notably, for subtrees with lengths greater than 3, our model surpasses the baseline model by 1.6%, and for subtrees with lengths greater than 6, the margin increases to 3.4%. A similar trend is observed in the rhetorical relation recognition results shown in Table 4. The observed enhancement in performance can be attributed to the integration of documentlevel content knowledge. This integration aids the model in acquiring and processing pivotal information more efficiently.

To provide a simplified overview, we present Table 2 in Section 3.5, where the text spans are

divided into three groups with balanced representation, allowing for a clearer comparison of performance differences between our model and the baseline.

B Effect on different language model

There are several reasons why we choose to freeze the parameters of pretrained language models. Firstly, our objective is to introduce a high-level content representation to enhance the RST task. By fixing the language model parameters, we can focus on leveraging the content information provided by the NDP branch without introducing additional complexity from joint training on language models. This allows us to assess the performance gain specifically attributable to the content structureaware representations, rather than the joint learning paradigm. Furthermore, freezing the language model parameters enhances the flexibility of our pipeline. If there are other tasks that could potentially benefit RST parsing or other tasks, we can seamlessly integrate their decoders into our system without the need for extensive retraining of language models using data from all these tasks. This saves significant time and effort in adapting the pipeline to new tasks or expanding its capabilities. Lastly, by freezing the language model parameters, we maintain the generalization ability

	Original Parseval				RST Parseval				
	S	N	R	F	S	N	R	F	
Baseline (RoBERTa)	72.1	61.2	50.8	49.5	86.0	73.2	61.1	60.4	
NCRNet (RoBERTa-all)	72.7	62.4	51.8	50.4	86.4	74.1	61.7	61.0	
C2RNet (RoBERTa-RST)	75.6	64.3	53.2	51.8	87.8	75.8	63.1	62.6	
Baseline (BERT)	70.2	58.6	48.1	46.8	85.1	72.0	59.3	58.6	
C2RNet (BERT-all)	71.4	59.5	49.3	48.1	85.7	72.3	60.0	59.3	
C2RNet (BERT-RST)	76.4	64.6	52.8	51.6	88.2	75.7	62.6	62.1	

Table 5: RST discourse parsing results with different language models, on the test set of the RST dataset, using original Parseval and RST Parseval metrics. S, N, R, and F represent Span, Nuclearity, Relation, and Full. *Baseline (RoBERTa)* and *Baseline (BERT)* represent the baseline models without the NDP branch and using the corresponding language models. *C2RNet (RoBERTa-all)* and *C2RNet (BERT-all)* represent C2RNet using RoBERTa or BERT as the shared language model. *C2RNet (RoBERTa-RST)* and *C2RNet (BERT-RST)* represents C2RNet using RoBERTa or BERT in RST branch.

of the pretrained models and ensure that the trainable parameters remain lightweight. This helps to reduce computational requirements and facilitates efficient deployment of the system.

To investigate the impact of different language models, we evaluate the performance of C2RNet using RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019) (large version) in the RST branch. The results are presented in Table 5. Specifically, C2RNet (RoBERTa-all) and C2RNet (BERTall) denote the models where RoBERTa or BERT serves as the shared fixed language model for both branches. In this configuration, the sentence embeddings obtained from the NDP branch may not be as representative, thus potentially affecting the overall performance of RST. We discuss this issue in the Limitation section. On the other hand, C2RNet (RoBERTa-RST) and C2RNet (BERT-RST) represent the models where RoBERTa or BERT is employed in the RST branch, while fixed word embeddings from T5 are used in the NDP branch.

Although these language models lead to a performance decline, our C2RNet still surpasses the baseline models. This finding suggests that incorporating high-level content-related knowledge benefits RST discourse parsing, particularly in terms of nuclearity and rhetorical relation predictions.

C Introduction of News Discourse Profiling

The NewsDiscourse dataset (Choubey et al., 2020) was created for news discourse profiling, which consists of 802 news articles. Following Choubey and Huang (2021), 502 documents are used for training, 100 documents for validation and 200 documents for testing. These data are only used for

training the NDP branch.

The news discourse profiling task aims to analyze the news event structure of news articles. The motivation behind this task lies in the significance of detecting and integrating discourse structures to enhance language comprehension. It categorizes the contents of news articles based on the main event and studies genre-specific discourse structures. Unlike existing tasks that primarily focus on understanding rhetorical aspects or detecting shallow topic transition structures, this task emphasizes the comprehension of global content organization structures with the main event as the central element.

The content types in the news discourse profiling task revolve around the main news event as the central focus. These content types encompass various aspects, such as the description of the main news event (Main Event) and its immediate consequences (Consequence). Additionally, there are content types related to providing contextual information, including previous events (Previous Events) and the current context (Background)(Current Context), which may serve as causes or preconditions for the main event. The further supportive information consists of historical events (Historical Event), anecdotal events (Anecdotal Event), evaluations from different parties (Evaluation), and speculations about the potential impacts of the main event (Expectation).

D Effect of NDP labels

The model represented by the second row, denoted as *C2RNet* (*one-hot*) in Table 6, incorporates the final classification layer of the NDP branch and utilizes the NDP one-hot predictions as input for the

	C	riginal	Parsev	'al		RST P	arseval	
	S	N	R	F	S	N	R	F
Baseline	75.4	64.1	53.6	52.1	87.7	75.0	63.1	62.3
C2RNet (one-hot)	75.9	65.2	54.8	53.3	88.0	76.3	64.3	63.5
C2RNet (ours)	76.8	66.2	55.4	53.8	88.4	76.5	64.5	63.8
Human	78.7	66.8	57.1	55.0	88.3	77.3	65.4	64.7

Table 6: RST discourse parsing results on the test set of the RST dataset, using original Parseval and RST Parseval metrics (Marcu, 2000). S, N, R, and F represent Span, Nuclearity, Relation, and Full. The results of ours are averaged over 3 random runs and the highest performances are in bold. *Baseline* represents the model without an NDP branch and other configurations are kept the same. *C2RNet (one-hot)* represents the model where the NDP final one-hot predictions are sent to the RST branch.

RST branch. This experiment aims to examine the effectiveness of different forms of information for enhancing RST performance. Comparing *C2RNet* (one-hot) with the *Baseline* model, it is observed that the former outperforms the latter, indicating that the direct utilization of event content labels in the RST parsing task through simple one-hot vector representations is advantageous. However, despite its improved performance, *C2RNet* (one-hot) still exhibits lower performance compared to our final C2RNet, suggesting that the sentence embeddings generated by the NDP branch contain additional essential information that is beneficial for the RST branch.

To further evaluate what information is the main contribution to the RST performance increase, we conducted a probing test towards sentence embeddings generated by the NDP branch in the context of the news discourse profiling task. Specifically, we feed the test data of the NDP task to the trained C2RNet and put back the final classification layer of the original NDP model to see how the performance changes. And we find its performance merely drops a little. These performance drops can be attributed to the fact that the weights of the final prediction layer were originally trained for NDP and remain fixed, while the preceding layers have been updated with RST data. Despite the observed drops, the relatively modest decline in performance suggests that the NDP branch still provides valuable event content information and it is this information that facilitates the RST task.