Discourse Parsing Enhanced by Discourse Dependence Perception

Yuqing Xing, Longyin Zhang, Fang Kong*, Guodong Zhou
School of Computer Science and Technology, Soochow University, China
{yq_xing,zzlynx}@outlook.com
{kongfang,gdzhou}@suda.edu.cn

Abstract

In recent years, top-down neural models have achieved significant success in text-level discourse parsing. Nevertheless, they still suffer from the top-down error propagation issue, especially when the performance on the upper-level tree nodes is terrible. In this research, we aim to learn from the correlations in between EDUs directly to shorten the hierarchical distance of the RST structure to alleviate the above problem. Specifically, we contribute a joint top-down framework that learns from both discourse dependency and constituency parsing through one shared encoder and two independent decoders. Moreover, we also explore a constituency-to-dependency conversion scheme tailored for the Chinese discourse corpus to ensure the high quality of the joint learning process. Our experimental results on CDTB show that the dependency information we use well heightens the understanding of the rhetorical structure, especially for the upper-level tree layers.

1 Introduction

According to the representative Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), a text can be presented as a hierarchical discourse tree (DT) built on a set of elementary discourse units (EDUs). Given a piece of text, RST-style discourse parsing identifies such a DT with EDUs serving as terminal nodes. Moreover, it labels the rhetorical relations and nuclearity attributes associated with each non-terminal node of the DT. Due to its far-reaching effects on text understanding and downstream NLP applications, text-level discourse parsing has been drawing more and more attention in the past decade.

From the early bottom-up approaches (Feng and Hirst, 2014; Ji and Eisenstein, 2014; Heilman and Sagae, 2015; Li et al., 2016; Braud et al., 2017; Yu et al., 2018; Mabona et al., 2019) to the more recent top-down frameworks (Lin et al., 2019; Kobayashi et al., 2020; Zhang et al., 2020, 2021; Koto et al., 2021), previous studies gradually switch from feature-based machine learning methods to deep neural models and have achieved particular success. Among current neural models, top-down parsers, in most cases, perform better than bottom-up ones due to their capability of capturing global context information. Nevertheless, due to the long-distance dependencies in between textual units and the notorious lack of training data, top-down text-level discourse parsing still faces the following possible bottlenecks:

- At the initial parsing stage, top-down parsers consider each entire text to determine the upper-level DT nodes. However, the whole text segment usually consists of diverse information, too much for the machine to understand thoroughly. As a result, our experimental statistics show that the parsing performance decreases by about 30% when the DT level is greater than 5.

- In RST-style constituency trees, there are far fewer training instances for the upper-level discourse tree layers when compared with the lower-level ones. For example, just as noted by Zhang et al. (2020), among the 933 test instances in the CDTB corpus, only 13 instances have a height of 8 or greater, occupying only about 1.3%.

- According to the above two points, on the one hand, the incorrect decisions made for the upper-level nodes may seriously impact the lower-level ones due to error propagation. On the other, the lack of upper-level training instances exacerbates the impact of error propagation.

Facing the above challenges, some recent studies have done certain preliminary explorations, hoping to improve top-down parsing by expanding the original small-scale training data (Kobayashi et al.,

*Corresponding author

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 1: Long Papers), pages 354–363
November 20–23, 2022.

2021) or introducing global optimization objectives (Zhang et al., 2021). Unlike previous work, we aim to improve the accuracy of upper-level node prediction to reduce error propagation for better RST parsing performance. To achieve this goal, we set our sights on discourse-level dependencies, aiming at employing the dependencies in between EDUs to dig out clues hidden within those head EDUs that are conducive to the understanding of rhetorical structures. Specifically, we cast discourse constituency tree (DCT) parsing as the main task and discourse dependency tree (DDT) parsing as the auxiliary one and joint the two tasks through one shared encoder and two different decoders. In this way, on the one hand, we enhance the EDU representation with multi-task knowledge through the shared EDU encoder. On the other, since the converted DDTs derive from the manually annotated DCTs, perceiving the dependencies between EDUs will conversely stimulate the DCT parsing model to produce better results, especially for the upper-level DT nodes[1].

2 Related Work

In the literature, previous work on discourse parsing can be classified into two categories: bottom-up and top-down approaches.

For a long time, many researchers manually exploited various lexical, syntactic, and semantic features (Hernault et al., 2010; Joty et al., 2013; Feng and Hirst, 2014) or automatically captured hidden information (Li et al., 2014a, 2016) to compute the probability distribution of relations between two adjacent discourse units (DUs) and then selected the two units with the highest probability to merge into an upper-level unit. Recursively in this way, a discourse constituency tree is created from bottom to up. Besides, there are also some studies that cast RST parsing as a transition action determination process, where the discourse parser makes `shift` or `reduce` action decisions in a greedy way to determine whether to merge the current two DUs or not (Ji and Eisenstein, 2014; Wang et al., 2017; Braud et al., 2017; Yu et al., 2018).

Until recent years, top-down neural architectures gained much more popularity. In the literature, Lin et al. (2019) proposed the first top-down sentence-level discourse parser based on pointer nets, which operates in a linear time. Zhang et al. (2020; 2021) cast text-level discourse parsing as a top-down split point ranking process and introduced an adversarial method to optimize the parsing steps from a global perspective. Kobayashi et al. (2020; 2021) proposed parsing a document in three levels of granularity (i.e., document-level, paragraph-level, and sentence-level) and further introduced a semi-supervised method to extend the original RST-DT corpus for performance improvement. Notably, some recent studies also proved the effectiveness of pre-trained language models on discourse parsing (Koto et al., 2021; Nguyen et al., 2021).

In general, compared with bottom-up parsing, current top-down parsers obtain more outstanding performance since they benefit from the global information of the entire article. However, the global context information is known to be multifarious and complicated. It is challenging for the top-down parsers to grasp all the textual details accurately, especially at the initial stage of parsing, which may aggravate the issue of top-down error propagation. In this work, we build our parser based on the top-down framework of Zhang et al. (2020) and explore tackling the above problem via discourse dependency information.

3 Motivation

In order to make better choices at the initial stage of discourse parsing to lay a good foundation for succedent parsing of subtrees, we consider incorporating discourse-level dependencies. To support our argument, we present an example in Figure 1 where Figure (a) shows a native DCT tree[2] and Figure (b) shows the converted DDT structure corresponding to the tree. Subsequently, our motivation comes from the following two observations:

- First, compared to the constituency structure, which joins EDUs with nuclearity and rhetorical relations, the dependency structure represents a more direct parent-child relationship between EDUs. The dependency structure is more conducive to weakening the hierarchical nature of the RST constituency tree and shortening the distance between EDUs.

[1]Although most of the existing conversion methods, including ours, have irreversible problems (Morey et al., 2018), that is, the reverse conversion of DDT to DCT structure is not unique, but in most cases, the correlation between EDUs is helpful for DCT parsing, especially for the upper nodes. This point will be further analyzed in Subsection 5.3.

[2]For brevity, we omit the discourse rhetorical relations and only present the nuclearity information (either **N**ucleus or **S**atellite) of each non-terminal node in the DCT structure.

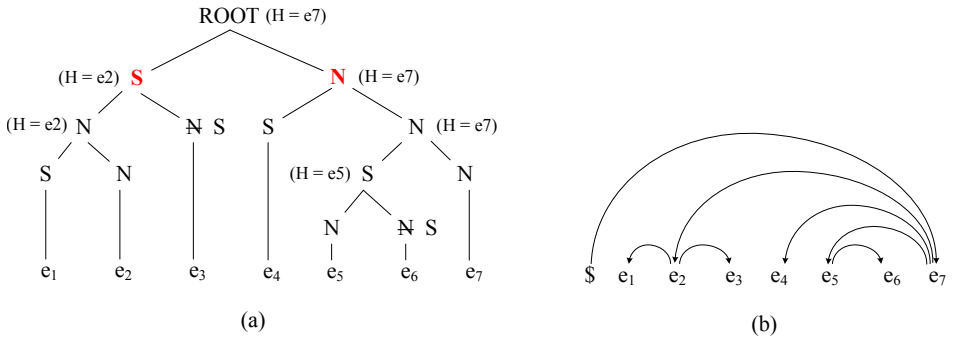


Figure 1: Figures (a) and (b) denote the example DCT structure and the converted DDT structure, respectively.

- Second, as the example shows, our constituency-to-dependency conversion method (described in Subsection 4.2.1) ensures that each sub-DCT in the tree corresponds to a unique single-rooted sub-DDT in the dependency structure. In this way, the rhetorical connection between two adjacent DUs is converted to a more straightforward correlation between two sub-DDTs, or more nuancedly, between their respective head EDUs. In this case, we believe that the direct connection between head EDUs can provide valuable structural or textual clues for better DCT parsing.

In short, the converted dependency arcs can help reduce the complexity of DCT trees to some extent, and the more direct connections between EDUs could provide valuable clues for better parsing performance, especially for the upper-layer tree nodes with a deep hierarchy. On this basis, we propose a multi-task learning approach to jointly learn DCT and DDT parsing, aiming to enhance the discourse representation via discourse dependencies for a better understanding of the rhetorical structure.

4 Joint DCT and DDT Parsing

Adopting the multi-task strategy, our model simultaneously conducts discourse constituency parsing and discourse dependency parsing by sharing the EDU representations, where discourse constituency parsing is the main task, and discourse dependency parsing serves as the auxiliary one. The whole architecture can be framed as an encoder-decoder model that contains one encoder and two different decoders, as illustrated in Figure 2.

4.1 Discourse Constituency Parsing

For DCT parsing, we follow Zhang et al. (2020) to cast the discourse parsing task as a recursive top-down split point selection process. The parsing

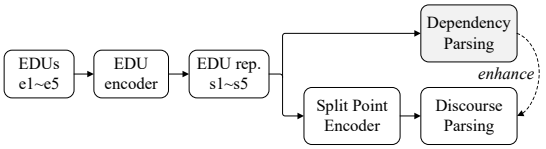


Figure 2: Joint parsing of DCT and DDT structures.

model comprises three parts, i.e., EDU encoder, split point encoder, and attention-based encoder-decoder. Firstly, a bi-GRU network and the self-attention mechanism are conducted over each EDU text to obtain EDU representation. Then, the split point encoder containing another bi-GRU network and a CNN network with a window size of 2 will work on the achieved EDU representations to model the representation for each split point between two adjacent EDUs. After that, the split point representations are further fed into a stack-augmented RNN decoder for discourse parsing. In this work, we employ the publicly-available implementation[3] of the parser of Zhang et al. (2020) for DCT parsing. For details of the parsing process, please refer to their paper.

4.2 Discourse Dependency Parsing

4.2.1 Discourse Dependency Trees Acquisition

In the literature, Hirao et al. (2013) and Li et al. (2014b) have proposed two different methods to convert from DCTs to DDTs automatically. Unlike the method of Li et al. (2014b), different EDUs in a sentence could have multiple heads outside the sentence in the DDT structure of (Hirao et al., 2013). In other words, their method often loses the single-rooted tree for each sentence. In order to reduce the complexity of DDTs, Hayashi et al. (2016) improve the method of (Hirao et al., 2013) by set-

[3] github.com/NLP-Discourse-SoochowU/t2d_discourseparser

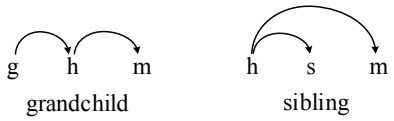


Figure 3: Diagram of grandchild and sibling structures.

ting constraints to restrict EDUs in a sentence for a single-rooted tree.

To our knowledge, all the abovementioned conversion methods are applied on the RST-DT corpus, while for the Chinese CDTB corpus, there are few related studies. Different from RST-DT, each sentence in the CDTB corpus occupies a complete sentence-level discourse tree. Under this circumstance, a discourse dependency structure that assigns each sentence with a single-rooted dependency tree is more appropriate for the CDTB corpus. Given this, we introduce a conversion method tailored for the Chinese corpus as follows:

- For each tree node $\mathcal{N}$, we take the head node of its leftmost **N**ucleus child as its head node (noted as H value); if no child is Nucleus, we take the head of the leftmost child as the head of $\mathcal{N}$.
- For each non-terminal node, if it maintains a multi-nucleus relation, we follow the principle of leftmost priority and treat the right child as a **S**atellite node.
- For each leaf node, we pick the nearest Satellite on the path from the leaf node to the root node and define the head of the Satellite node's parent as its head. If there exists no such Satellite node, the EDU is just the root of this dependency tree.

Following the above rules, the DCT structure shown in Figure 1 is finally converted into a complete dependency graph. As stated before, each sentence in the CDTB corpus corresponds to an independent sub-DCT. Similarly, using our method for conversion, each sentence, or more broadly, each sub-DCT, still yields a single-rooted sub-DDT in the converted structure, which vastly reduces the complexity of the resulting DDT structure.

4.2.2 Discourse Dependency Parsing

Concerning the dependency parsing module, we refer to (Ma et al., 2018) on parsing syntactical dependency based on a top-down neural architecture and view the EDUs in a text as words in a sentence. Unlike the parsing procedure in (Zhang et al., 2020) which employs pointer nets to select split points from top to down to build the DCT structure, DDT parsing utilizes the pointer nets to select EDUs directly. Therefore, the split point encoding phase is omitted during DDT parsing.

Having obtained the EDU representation vectors, $s_1, \ldots, s_n$, through the shared EDU encoder described before, we use the stack-pointer network with two kinds of subtree information (grandchild and sibling) integrated for discourse dependency parsing. The definitions of the grandchild and sibling structures are described as follows, and their diagrams are shown in Figure 3.

- **grandchild structure:** a pair of dependencies connected head-to-tail. For the modifier $\mathbf{m}$, the parent of its head $\mathbf{h}$ is noted as its grand node $\mathbf{g}$.
- **sibling structure:** a head word with two successive modifiers. For the modifier $\mathbf{m}$, the most recent child $\mathbf{s}$ of its head node $\mathbf{h}$ is noted as its sibling.

Figure 4 illustrates partial of the decoding procedure. At the very beginning of the parsing process, the stack only contains the root node. For the convenience of calculation, we set a virtual root node \$ pointing to the first node of the dependency tree, and its representation is zero-initialized. At each step of decoding, we pop out the top element of the stack, noted as e_h, and lookup for its sibling node e_s and grandparent node e_g from the converted DDT structure, then the input of decoder is created by summing up the representation vectors of them, as shown in Equation 1. If there exists no sibling or grandparent of e_h, the value of s_s or s_g will be assigned with zero vectors.

$$S_t = s_h + s_s + s_g \tag{1}$$

We use a uni-directional RNN as the decoder. At each time step t, it receives the structure information S_t as input and outputs the hidden vector noted by h_t. Then, the biaffine attention mechanism is utilized to calculate the probability score e_i^t of each EDU as the dependence of the current unit. Equations 2-4 show the details, where $\mathbf{w}$, $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{b}$ are parameters, denoting the attention weight of the bi-linear term, the two linear terms, and the bias term, respectively. It is worth noting that before attention calculation, we let h_t and s_i go through a one-layer perceptron with elu activation function for dimension reduction to reduce the risk of overfitting. We choose the most probable EDU e_c as the

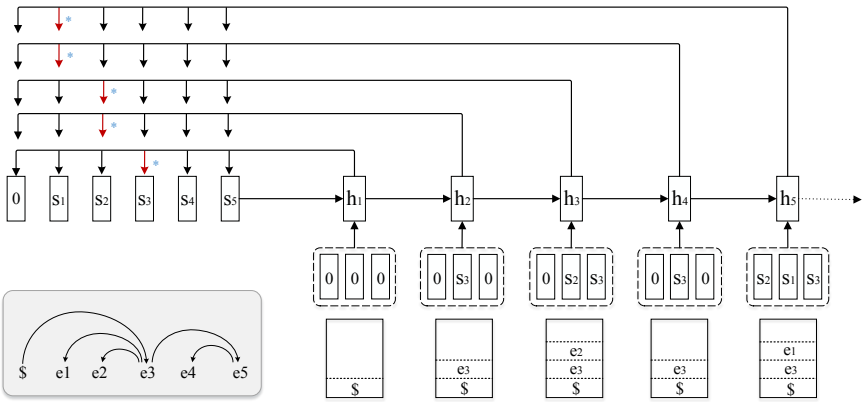


Figure 4: The decoding architecture used for discourse dependency parsing.

dependence of e_h, thus one dependency arc is obtained, (e_h, e_c). Then we push the newly selected element e_c onto the stack for the following steps. Moreover, a self-directed dependency arc will appear when c equals h. In this case, all the children of the head node e_h have been successfully found. Then we pop e_h out of the stack and go into the next parsing period. The parsing process will be terminated when the stack becomes empty.

$$s'_i = \text{elu}\left(\text{w}_1 s_i + \text{b}_1\right) \tag{2}$$

$$h'_t = \text{elu}\left(\text{w}_2 h_t + \text{b}_2\right) \tag{3}$$

$$e_i^t = {h'_t}^{\text{T}} \mathbf{w} s'_i + \mathbf{u}^{\text{T}} h'_t + \mathbf{v}^{\text{T}} s'_i + \mathbf{b} \tag{4}$$

Considering that one head node may have multiple child nodes, we follow the inside-out strategy to order the child nodes according to the distances between these nodes and the head node, the left side first and then the right side, which ensures that the parsing path is unique. Taking the instance in Figure 4 for example, the ordered parsing path is $\{(\$, e_3), (e_3, e_2), (e_3, e_1), (e_3, e_5), (e_5, e_4)\}$.

4.3 Model Training

Our training objective is composed of two parts, i.e., jointly minimizing the discourse constituency parsing loss and the discourse dependency parsing loss. Since both tasks can be recognized as multi-step classification problems, we employ the negative log-likelihood (NLL) loss to calculate and optimize the two loss terms.

On the one hand, for discourse constituency parsing, we need to identify three parts, including the bare tree structure, the rhetorical relation, and the nuclearity category. Therefore, the loss function consists of three parts, i.e., split point prediction loss L_s, relation prediction loss L_r, and nuclearity prediction loss L_n. Supposing that the correct index of the gold standard split point at the t-th step is i, the value of L_s is calculated as follows:

$$L_s = \sum\nolimits_{steps} -\log\left(\hat{p}_t^s \mid \theta\right) \tag{5}$$

$$\hat{p}_t^s = \frac{a_{t,i}^s}{\sum a_t^s} \tag{6}$$

where a_t^s denotes the probability distribution of split points at the current time step and $\hat{p}_t^s$ denotes the probability of selecting the i-th one as the predicted split point. The calculation of L_r and L_n is similar to that of L_s. In consideration of the different convergence rates of the three loss terms, we obtain the overall discourse rhetorical structure parsing loss through weighted summation:

$$L_c = \alpha_s L_s + \alpha_n L_n + \alpha_r L_r \tag{7}$$

On the other hand, the discourse dependency tree is essentially converted from the original discourse constituency tree according to the nuclearity property while ignoring the internal relations. So we only need to consider the correctness of dependency arcs. The calculation of discourse dependency parsing loss L_d is similar to that of split point prediction in DCT parsing. Finally, we merge the weighted dependency loss to the original constituency loss, and the final optimization objective is formalized as follows:

$$L = L_c + \alpha_d L_d \tag{8}$$

5 Experimentation

This section systematically evaluates our top-down discourse parser and primarily focuses on the impact of the dependency information on DCT parsing. We merely focus on the performance of the

main task of DCT parsing, while the auxiliary DDT parsing task only works for representation enhancement. Therefore, we do not discuss the performance of DDT parsing in the following parts.

5.1 Experimental Settings

Datasets. In this paper, we employ the Chinese connective-driven discourse treebank (CDTB[4]) (Li et al., 2014c) as the benchmark data set. The corpus consists of 500 newswire articles, divided into 2336 paragraphs, and each paragraph yields an independent CDT tree. Following (Zhang et al., 2020), we divide the corpus into three parts, i.e., 425 training documents containing 2002 discourse trees and 6967 rhetorical relations, 25 development documents containing 105 discourse trees and 396 relations, and 50 test documents containing 229 discourse trees and 993 relations.

Evaluation metrics. The metrics of discourse parsing evaluation include bare tree structure referred to as span (**S**), tree structure with nuclearity (**N**) indication, and tree structure with relation (**R**) indication. We use Full (**F**) to evaluate the overall tree structure with both nuclearity and relation considered. For a fair comparison, same as Zhang et al. (2020), we adopt the original Parseval procedure to evaluate the performance of our parser and report the micro-averaged F1 scores as our parsing performance. Following previous work, we evaluate our system with gold EDU segmentation and binarize those non-binary subtrees with right-branching (Sagae and Lavie, 2005).

Hyper-parameters. For hyper-parameters, we keep consistency with (Zhang et al., 2020) in the shared EDU encoder, the split point encoder, and the DCT parsing module. While for the DDT parsing module, we set the size of hidden states after dimension reduction to 64 and the weight α_d in the joint loss objective to 2. For other hyper-parameter details, please refer to (Zhang et al., 2020).

[4]It should be noted that our proposed approach is language-independent. Although previous studies on the English RST-DT corpus (Carlson and Marcu, 2001) are much more affluent, the corpus is not well suited to validate our approach. The RST-DT corpus consists of 385 documents, and each document is represented as a single DT. According to our statistics, the heights of trees in the corpus range from 1 to 26. No matter for training or testing, there are too few instances. In addition, the quality of the high-level annotation is not good, which may lead to poor performance of the converted dependency tree. Considering the abovementioned quality and quantity issues, we only conduct experiments on the CDTB corpus.

Systems	S	N	R	F
Sun and Kong (2018)*	84.8	55.8	52.1	47.7
Zhang et al. (2020)*	85.2	57.3	53.3	45.7
Ours (Joint)	**86.4**	**60.5**	**54.3**	**49.5**

Table 1: Performance comparison. Sign "*" denotes the results are borrowed from (Zhang et al., 2020).

TLs (#)	S (B/O)	N (B/O)	R (B/O)	F (B/O)
1 (385)	339/**340**	251/**255**	**233**/232	213/**216**
2 (220)	183/**191**	117/**126**	116/**121**	94/**103**
3 (139)	119/**120**	71/**80**	71/**74**	59/**69**
4 (88)	**75**/73	**52**/47	**44**/39	**39**/35
5 (44)	34/**38**	17/**26**	16/**23**	10 /**21**
6 (26)	**18**/17	**13**/12	6/**8**	6/**8**
7 (18)	16/**17**	7/**10**	**6**/5	2/**5**
8+ (13)	**11**/10	0/**8**	0/**5**	0/**5**

Table 2: Performance over different tree levels (TLs) of the DTs. Signs "B" and "O" denote the results of the baseline system (Zhang et al., 2020) and our proposed joint method, respectively.

5.2 Experimental Results

In this part, we compare our system with two previous state-of-the-art (SoTA) systems on CDTB using the same evaluation metrics.

- Sun and Kong (2018): a transition-based system that parses the discourse rhetorical structure in a bottom-up way.
- Zhang et al. (2020): a top-down text-level discourse parser based on the pointer networks. In this paper, our system directly inherits from their system on DCT parsing. Therefore, we take their implemented system as our baseline.

Table 1 presents the performances of our method and the two previous SoTA systems. The results show that our joint model significantly outperforms the two SoTA systems on all four indicators. In comparison with the bottom-up parser of Sun and Kong (2018), the top-down approaches (the parser of Zhang et al. (2020) and ours) show better performance, on the whole, benefiting from global information. In addition, with the help of dependency information, our joint model achieves the gains of 1.2, 3.2, 1.0, and 3.8 on the four evaluation indicators, respectively, when compared with (Zhang et al., 2020). Moreover, to our knowledge, the top-down parser of Zhang et al. (2020) shows terrible performance on the Full metric because of using three independent classifiers for span, nuclearity, and relation classification. With the global

dependency graph harnessed for representation enhancement, our parser can significantly make up for this problem.

As mentioned before, we aim at improving the parsing performance of the upper-level discourse tree nodes in this work. Here, we further count the correctly identified nodes over different DT levels, and the results are shown in Table 2. Comparing the statistical results of the baseline system (Zhang et al., 2020) and ours, we find that

- Our joint model performs better than the baseline system at most levels. Among the three aspects, the improvement on nuclearity is significant, and that on bare tree structure is the weakest;
- When the height is larger than 5, our joint model performs much better in nuclearity and relation identification. This also contributes to the improvements on the Full metric;
- When the height is equal to or greater than 8, our joint model fulfills the zero breakthroughs in nuclearity, relation, and Full identification.

Same as Zhang et al. (2020), we also divide the discourse trees into six groups by EDU number and evaluate our joint model over different groups. From the results in Table 3 we find that

- On the structure indicator, except for the case with EDU number larger than 25, the contribution of dependency information is not apparent;
- On the nuclearity indicator, in most cases, our joint model performs better. For the case when the EDU number is larger than 25, the improvement is very significant;
- On the relation indicator, our joint model is equal to or better than the baseline system in all groups of discourse trees.

In addition to how many EDUs a tree contains, the tree height is another perspective to measure the complexity of tree structures. Thus we further divide the DTs into different groups according to their heights and evaluate our model over different tree groups using a macro-averaged evaluation, i.e., calculating the F1 score for each DT solely and reporting the averaged F1 score in the test set. The results in Table 4 show that the contribution to structure building varies over different heights. For nuclearity and relation detection, our joint model performs better than the baseline system in most cases.

EDU	S		N		R	
Num.	Base	Joint	Base	Joint	Base	Joint
1-5	97.7	96.7	67.1	64.8	56.6	57.0
6-10	86.0	88.5	57.3	63.2	59.9	60.5
11-15	75.2	74.9	50.3	55.9	41.4	43.3
16-20	56.2	56.2	25.0	37.5	25.0	25.0
21-25	76.6	73.5	57.7	51.6	40.8	45.5
26-30	69.2	76.9	42.3	50.0	19.2	19.2

Table 3: Performance over different EDU numbers. Here, "Base" and "Joint" denote the baseline system and our proposed joint model, respectively.

	S		N		R	
Height	Base	Joint	Base	Joint	Base	Joint
1	100	100	66.7	64.9	56.1	56.1
2	94.8	94.8	77.3	70.8	61.8	62.8
3	90.8	91.5	55.7	59.2	54.0	54.4
4	84.6	88.3	56.9	62.7	58.3	59.3
5	84.2	84.5	50.9	54.8	56.2	59.0
6	81.8	76.8	50.1	44.6	46.1	38.7
7	82.9	87.3	62.8	67.8	55.9	61.2
$\geqslant 8$	72.0	70.5	55.0	60.5	42.3	40.0

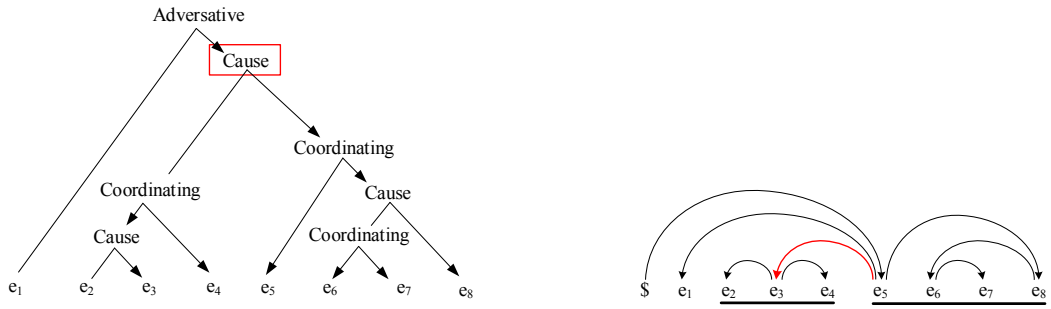
Table 4: Performance over different DT heights.

As described in Subsection 4.2.1, during the acquisition of DDT structures, we only consider the bare structure and nuclearity of each constituency tree. So the incorporation of dependency information can reasonably improve the performance of tree structure and nuclearity detection. Curiously, how can the discourse dependencies improve the performance of relation prediction? To figure it out, we give a further analysis in the following part.

5.3 Further Analysis

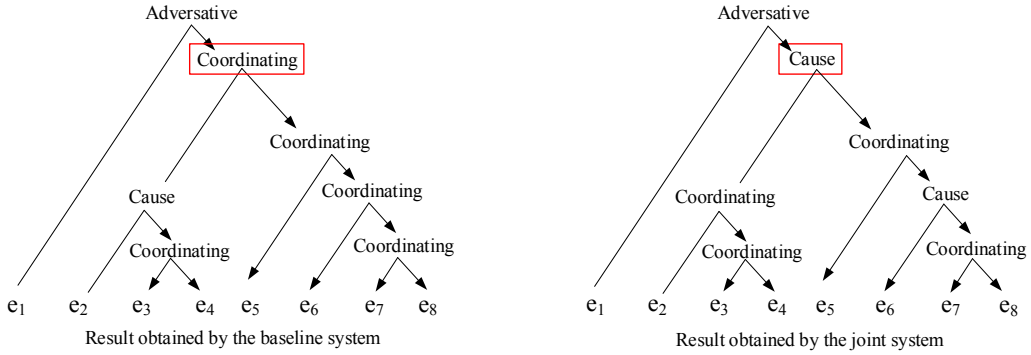
A certain number of cases have shown that the dependency arcs between long-distance EDUs may provide practical and explicit clues for predicting the rhetorical relation between the upper tree nodes. Here, we use an example in Figure 5 to analyze the effects of RST dependencies on rhetorical relation prediction.

Figure (a) shows the gold standard DCT and DDT structures of the paragraph consisting of eight EDUs. In the DCT structure, the relation "Cause" shown in the red rectangle is associated with two sub-trees, i.e., the left sub-tree with EDUs from e2 to e4 and the right sub-tree with EDUs from e5 to e8. From the corresponding DDT structure, we can find that the two sub-DCTs also correspond to two independent single-rooted sub-DDTs, respectively, where the head EDU of the left sub-DDT is e3, and



e1 一九九五年广东制定“九五”规划时曾提出汽车作为支柱产业之一。/ W*hen Guangdong formulated the "Ninth Five-Year Plan" (1996-2000) in 1995, automobiles were mentioned as one of the pillar industries.*
e2 但从目前来看，广东不具备汽车制造的优势和条件，/ *However, from the current point of view, Guangdong does not have the advantages and conditions for automobile manufacturing,*
e3 难以形成支柱产业，/ *it is difficult to form a pillar industry,*
e4 全国也有重复建设问题。/ *and it also has the problem of repeated construction across the country.*
e5 因此，省里已明确汽车制造不再作为支柱产业，/ *Therefore, the province has made it clear that automobile manufacturing is no longer a pillar industry,*
e6 而电子信息产业是广东省的优势，/ *the electronic information industry is Guangdong Province's advantage*
e7 也是新的增长优势，/ *and it is also a new growth advantage.*
e8 应作为支柱产业加以重点扶持。/ *It should be given priority support as a pillar industry.*

(a) Gold DCT and DDT structures of the given example.



(b) DCTs predicted by the baseline system and our joint model.

Figure 5: Case study of the impact of DDTs on discourse rhetorical relation prediction.

the head EDU of the right sub-DDT is e5. Between the two sub-DDTs, an explicit arc pointing from e5 to e3 connects the two parts, which strongly suggests that there should be some relation between the two parts. Looking into the two head EDUs, e3 expresses that “it is difficult to form a pillar industry”, and e5 says that “Therefore, the province has made it clear that automobile manufacturing is no longer a pillar industry”. Obviously, the connective “因此 / therefore” in e5 is crucial in determining the “Cause” relation. This example indicates that the DDT structure will build a unique arc between two adjacent sub-DDTs (sub-DCTs), and their respective head EDUs may provide valuable clues for the upper-level sub-DCTs to determine the rhetorical relation between them. This result explains our performance improvement in relation prediction.

6 Conclusion

This paper contributes a multi-task learning architecture that jointly learns discourse-level constituency and dependency parsing through one shared encoder and two independent decoding modules. Moreover, we introduce a constituency-to-dependency conversion method tailored for the Chinese corpus to ensure the quality of the joint learning process. The experimental results on the CDTB corpus show that the discourse dependency information is efficient in improving the performance of discourse constituency parsing on all metrics, especially for the upper-level tree layers.

The results of this paper show that the use of textual knowledge such as rhetorical dependencies can effectively improve the machine’s understanding

of discourse parsing. Inspired by this, in our future work, we will explore the use of meta-learning techniques to learn the knowledge of dependencies such as reference chains and topic chains to achieve the ability to parse various discourse dependency structures including the rhetorical dependencies.

Acknowledgements

This research was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600, Projects 61876118 and 61976146 under the National Natural Science Foundation of China and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

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

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

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

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.

Michael Heilman and Kenji Sagae. 2015. Fast rhetorical structure theory discourse parsing. *arXiv preprint arXiv:1505.02425*.

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

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

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

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

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down rst parsing utilizing granularity levels in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence 2020*, pages 8099–8106.

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.

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.

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

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

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014b. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Yancui Li, wenhe Feng, jing Sun, Fang Kong, and Guodong Zhou. 2014c. Building chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of EMNLP 2014*, pages 2105–2114.

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

Xuezhe Ma, Zecong Hu, Jingzhou Liu, Nanyun Peng, Graham Neubig, and Eduard Hovy. 2018. Stack-pointer networks for dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1403–1414, Melbourne, Australia. Association for Computational Linguistics.

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.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, pages 198–235.

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.

Kenji Sagae and Alon Lavie. 2005. A classifier-based parser with linear run-time complexity. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 125–132, Vancouver, British Columbia. Association for Computational Linguistics.

Cheng Sun and Fang Kong. 2018. A transition-based framework for Chinese discourse structure parsing. *Journal of Chinese Information Processing*, 32(12):26–34.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of ACL 2017: short paper*, pages 184–188.

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.

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 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.