A New Dataset and Parsing Model for Chinese Multiparty Dialogue Discourse Structure

Yuru Jiang
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
yurujiang@126.com

Yu Li
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
liyu666@bistu.edu.cn

Weikai He
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
1003512608@qq.com

Jie Chen
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
chenjie@bistu.edu.cn

Yanchao Yu
School of Computing, Engineering and Building Environment
Edinburgh Napier University
Edinburgh, United Kingdom
y.yu@napier.ac.uk

Yangsen Zhang
Intelligent Information Processing Institute
Beijing Information Science and Technology University
Beijing, China
zhangyangsen@163.com

***Abstract*—With the advancement of Natural Language Processing (NLP) technology, discourse parsing in multi-party dialogues has become increasingly significant i n a pplications such as machine reading comprehension, machine translation, and text summarization. However, there remains a lack of discourse parsing datasets specifically d esigned f or C hinese multi-party dialogues, and the fusion of speaker information has not been fully investigated. This research, based on the script of the TV series "I Love My Family", has constructed a multi-party dialogue discourse parsing dataset named DialogueDSA[1]. In this study, we also propose a speaker-aware parsing model for multi-party dialogue discourse. Our method addresses the problem of representing excessively long texts in DialogueDSA by encoding each Elementary Discourse Unit individually, and further interactions are facilitated through the use of BiGRU, attention mechanisms, and Transformer. The effectiveness and robustness of the proposed method are empirically demonstrated on both the English Molweni and STAC datasets, as well as the Chinese DialogueDSA dataset.**

***Index Terms*—Multi-party Dialogue, Discourse Parsing, Annotation Methodology**

I. Introduction

In recent years, discourse parsing for multi-party dialogue has attracted considerable attention in the field of Natural Language Processing (NLP). With the rapid advancement of NLP technologies, discourse parsing plays a critical role in various downstream tasks such as machine reading comprehension, machine translation, and text summarization. Discourse parsing aims to streamline the structure of a discourse by identifying the relationships between discourse units, thereby facilitating the understanding of the content of the discourse.

Unlike conventional text-level discourse parsing based on Rhetorical Structure Theory (RST) [1] and the Penn Discourse Treebank (PDTB) [2], discourse parsing for multi-party dialogue is conducted based on Segmented Discourse Representation Theory (SDRT) [3]. It utilizes a discourse dependency tree [4] to represent multi-party dialogues. In SDRT discourse parsing, each utterance is referred to as an Elementary Discourse Unit (EDU), and all EDUs and discourse relationship connectors ultimately form a directed acyclic graph, as illustrated in Figure 1.

Discourse parsing in multi-party dialogues traditionally focused on conversational context modeling. Early work, such as by Afantenos et al. [6], relied on maximum spanning tree techniques for EDU relationships. Shi et al. [8] later employed Hierarchical GRUs for context insights. Recent studies have harnessed technologies like domain adaptation [9], edge-centric encoding [10], and multi-task learning [11] to improve parsing. However, limitations persist, as general models like BERT struggle with long contexts due to their maximum sequence length limit, often causing information loss and performance issues in long articles or complex discourses.

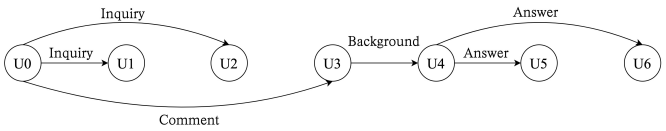
In multi-party dialogues, speaker exchanges generate various topic cues, making speaker information crucial for discourse parsing. Despite this, current research has insufficiently explored speaker information integration. In light of this, our paper presents a speaker-aware model for discourse parsing in multi-party dialogues, effectively capturing speaker interaction

[1]https://github.com/LIyu810/DialogueDSA

979-8-3503-3078-6/23/$31.00

	Speaker	Utterance
U0	圆圆 (Yuanyuan)	.....我向大家承认错误，昨天那封信..... ("...I confess my error to all, concerning the letter from yesterday...")
U1	傅老 (Fulao)	什么信.....啊？(What letter...?)
U3	傅老 (Fulao)	.....噢，那个事儿啊，小孩子的把戏，我早就看出来了。(...Ah, that incident, just child's play, I had seen through it long ago.)
U2	众人 (Zhongren)	什么信啊？（What letter?）
U4	傅老 (Fulao)	昨天我喝了一点酒，借着那件事，给大家讲了几个笑话，大家还记得么 (Yesterday, I had a bit of drink, and took that incident as an opportunity to tell a few jokes. Do you all still remember?)
U5	众人 (Zhongren)	没有没有..... (No, no...)
U6	和平 (Heping)	我也喝了点儿酒 (I also had a bit to drink.)

(a) Dialogue example from DialogueDSA Dataset.



(b) A dependency tree example for DP task based on the dialogue in (a).

Fig. 1. Dialogue and its corresponding Dependency Tree from DialogueDSA Dataset for DP task.

to enhance discourse parsing performance.In summary, our key contributions are outlined as follows:

We introduce DialogueDSA, the first discourse parsing dataset for Chinese multi-party dialogues, providing robust data support for discourse parsing research.

We propose a discourse parsing model that effectively incorporates speaker information, addressing the long-text representation problem in the DialogueDSA dataset through individual EDU encoding.

We have demonstrated the robustness and efficacy of our proposed model across diverse linguistic contexts by conducting tests using the English Molweni and STAC datasets, as well as the Chinese DialogueDSA dataset.

II. Related work

A. Discourse Parsing

Discourse parsing, instrumental for numerous NLP tasks, has seen significant advancements recently. Li et al. introduced the Molweni dataset [5], emphasizing the intricacies of multi-party dialogues. Liu and Chen furthered this with a Transformer-based parser integrated with domain techniques for enhanced generalization [9]. Addressing error propagation, Wang et al. introduced a Structure Self-Aware model using an edge-centric Graph Neural Network for updating EDU pairs [14]. Yu et al. improved performance by considering speaker interactions in their model [10]. Chi and Rudnicky achieved leading results through structured encoding in dialogue discourse parsing [17]. He et al. highlighted discourse parsing's role in multi-party dialogues, unveiling a model adept at both question answering and discourse parsing [11].

B. Discourse Parsing Datasets

Analyzing the discourse structure of multi-party dialogues is more challenging than that of standard documents due to weaker coherence between adjacent utterances. Notably, two English datasets, namely STAC [4] and Molweni [5], are currently available for such analysis. The STAC dataset, derived from the online game "Settlers of Catan", comprises 1,091 dialogues, 10,677 EDUs, and 11,384 discourse relationships. On the other hand, the Molweni dataset, sourced from the online forum of the Ubuntu operating system, contains 10,000 dialogues, 88,303 EDUs, and 78,245 discourse relationships. Although these English datasets provide useful insights, they originate from online forums, thus lacking reflections of the Chinese context or everyday life scenarios. To address this, our study introduces the DialogueDSA dataset specifically designed for discourse parsing in Chinese multi-party dialogues. With a total of 705 dialogues, the dataset encompasses 24,451 EDUs and 25,583 discourse relationships, all rigorously annotated.

C. Speaker-Aware Modeling in Dialogues

Recent progress in multi-party dialogue modeling has emphasized speaker interactions. Afantenos et al. [6] introduced a model for same-speaker EDU pairs, which was later expanded upon by Shi et al. [8] with a speaker highlighting mechanism, and Wang et al. [14] using a GNN-based approach incorporating EDU edge to emphasize speaker interactions. However, many models focused primarily on two EDUs from the same speaker, overlooking the nuances of inter-speaker dynamics. Yu et al. [10] bridged this gap using the SCIJE model, incorporating SSP-BERT for comprehensive speaker information utilization. Our research further refines speaker-aware modeling by integrating BERT-based speaker encoding with BiGRU-driven contextual interactions, enabling a comprehensive speaker representation. This method offers fresh perspectives on discourse analysis in multi-party dialogues, marking our pivotal contribution to speaker-aware dialogue modeling.

III. Construction of the DialogueDSA Dataset

The Dialogue Discourse Structure Analysis (DialogueDSA) dataset, sourced from the scripts of the 1994 Chinese sitcom "I Love My Family", encompasses transcripts from 120 episodes involving 117 characters, including 8 main roles. The sitcom's everyday life content, diverse topics, and minimal dialect or

professional terminology make it an ideal source for analyzing everyday discourse structure.

A. Data Preprocessing

The DialogueDSA dataset, an extension of the CRECIL corpus [15]. Typographical and punctuation errors in the "I Love My Family" sitcom scripts were corrected, following CRECIL's method of excluding non-essential narrations. The dataset construction introduces a novel dialogue segmentation approach, building on CRECIL's foundation. We formed "dialogue groups" based on shared scene and topic parameters from the original scripts. These groups, dialogues in the same scene around a shared topic, offer a structured way to segment the dataset. This method retains the original scripts' context and facilitates detailed interpretation in later analyses. Consequently, DialogueDSA exemplifies an advancement in processing and organizing multi-party dialogue datasets, preserving original dialogue essence while innovating in segmentation and analysis.

B. Annotation Guidelines

The study commenced with pre-annotation of 17% of the DialogueDSA dataset using 16 discourse relations from the Molweni dataset. Afterward, the discourse relations were refined to better suit the DialogueDSA dataset. Specifically, "Q-A Pair" was split into "Inquiry" and "Answer" for enhanced detail; "Q-Clarify" and "Narration" were replaced with "Question-Elaboration" and "Continuation", respectively, due to their similarity and lower occurrence; "Alternation" was omitted due to lack of corresponding text; "Causation" and "Response" were introduced to capture frequent daily-life situations. Importantly, discourse relations were designed as one-way, pointing from a later to an earlier utterance, reflecting their inherent dependency. Statistics show 95% of relations have a speech interval of 4 or fewer EDUs. Thus, for efficiency, relations exceeding this interval are excluded.

C. Annotation Process

Each dialogue group in the DialogueDSA dataset was manually segmented into Elementary Discourse Units (EDUs) by an annotator, who also marked discourse relations between EDUs. To validate data quality, 30 dialogue groups underwent double-blind annotation, yielding an 84.1% consistency for the presence of discourse relations and 50.4% for their specific type. In contrast, the Molweni dataset, derived from a technical forum with clearer logic and frequent use of the '@' symbol to denote recipients, recorded consistencies of 91% and 56%, respectively. Given the 16 discourse relation types, a consistency above 50% is deemed acceptable. The rest of the annotations were done by one annotator, with two reviewers checking post the initial round.

D. Statistical Results

The annotated dataset, DialogueDSA, comprises 705 dialogues and 24,451 EDUs. As seen in Table I, it contrasts with the Molweni and STAC datasets. DialogueDSA has a notably higher average EDUs and relation pairs per dialogue compared to both. From Table II, 70.6% of discourse relations in DialogueDSA occur between adjacent EDUs, with 90.9% having an interval of 1 or less. In Molweni, 86.4% exhibit a similar interval trend. This suggests in multi-party dialogues, participants typically address recent utterances rather than engage in in-depth discussions, leading to fewer relations with widely separated EDUs.

TABLE I
COMPARATIVE STATISTICS OF DIALOGUEDSA, MOLWENI, AND STAC DATASETS.

Statistics Item	**DialogueDSA**	**Molweni**	**STAC**
Dialogues total	705	10000	1081
EDUs total	24451	88303	10678
Relations total	25583	78245	10513
Avg EDUs/dialogue	34.6	8.8	9.9
Avg relations/dialogue	36.3	7.8	9.7

TABLE II
DISTRIBUTION OF SEPARATION DISTANCE BETWEEN EDU PAIRS IN DIALOGUEDSA, MOLWENI, AND STAC DATASETS.

EDU Pairs Distance	**DialogueDSA**(%)	**Molweni** (%)	**STAC** (%)
Distance = 0	70.6	64.9	54.5
Distance = 1	20.3	21.5	21.2
Distance = 2	6.2	7.4	10.6
Distance $\geq$ 3	2.9	6.2	13.7

Table III outlines the discourse relations across the DialogueDSA, Molweni, and STAC datasets. In DialogueDSA, "comment", "inquiry", "elaboration", and "answer" together account for 52.8%. In contrast, Molweni's leading relations are "comment", "Q-Clarification", "Q&A pair", and "continuation", summing up to 82.5%. These differences underscore the distinct nature of technical forum exchanges versus everyday conversations.

Our statistical analysis of the DialogueDSA dataset reveals a significant challenge: 64% (454 samples) of dialogues exceed 500 tokens, with the remaining 251 samples falling short of this length. This complexity is emphasized when compared to the Molweni and STAC datasets, where all dialogues fit within 512 tokens. Current natural language processing models, like BERT, cap sequence lengths at 512 tokens. Texts that exceed this limit can result in ineffective data incorporation and potentially compromise model performance. Given that a majority of the DialogueDSA samples surpass this limit, their processing presents a substantial challenge.

IV. CONTEXT-SPEAKER ENHANCED DISCOURSE PARSING MODEL

Our discourse parsing model, the Context-Speaker Enhanced Discourse Parsing Model (CSE-DPM), comprises four main modules: the Context Encoding Layer, Speaker Encoding Layer, Interaction Layer, and Classifier Layer.

A. Context Encoding Layer

The Context Encoding Layer's main role is to capture the context information of input discourse statements. This layer

TABLE III
DEFINITIONS AND PROPORTIONS OF DISCOURSE RELATIONS IN DIALOGUEDSA, MOLWENI, AND STAC DATASETS.

No.	Discourse Relation	Definition	DialogueDSA (%)	Molweni (%)	STAC (%)
1	Comment	Arg2 expresses opinion on Arg1.	18.6	31.7	17.6
2	Inquiry	Arg2 inquires about Arg1.	13.8	–	–
3	Elaboration	Arg2 supplements Arg1.	10.5	2.2	8.3
4	Answer	Arg2 responds to Arg1.	9.9	–	–
5	Causation	Arg2 makes Arg1 act.	8.5	–	–
6	Continuation	Arg2 and Arg1 share a topic.	7.7	6.7	9.4
7	Explanation	Arg2 explains Arg1.	5.6	1.6	4.2
8	Confirmation	Arg2 acknowledges Arg1.	5.5	3.2	–
9	Response	Arg2 responds to Arg1.	4.0	–	–
10	Background	Arg2 narrates Arg1's background.	3.2	0.4	0.6
11	Correction	Arg2 corrects Arg1.	3.0	1.2	2.0
12	Parallel	Arg2 and Arg1 are identical.	2.8	0.2	2.0
13	Q-Elab	Arg2 supplements Arg1.	2.0	3.0	5.7
14	Contrast	Arg2 contrasts Arg1.	2.0	1.2	4.7
15	Result	Arg2 is Arg1's result.	1.6	2.6	5.5
16	Conditional	Arg2 is Arg1's condition.	1.3	1.0	1.2
17	Q-A Pair	Arg2 inquires or answers Arg1.	–	20.1	24.2
18	Q-Clarify	Arg2 clarifies Arg1.	–	24.0	–
19	Narration	Arg2 narrates Arg1.	–	0.3	1.2
20	Alternation	Arg2 alternates Arg1.	–	0.2	1.4
21	Clarification_question	Arg2 asks for clarification about Arg1	–	–	2.5
22	Acknowledgment	Arg2 acknowledges Arg1	–	–	9.6

first employs BERT to encode each EDU, denoted as u_i, where i represents the sequence position, and m symbolizes the maximum sequence length. Equation 1 shows the representation of u_i, and Equation 2 elaborates on how BERT is used for encoding u_i.

$$U_i = [CLS], t_1^i, ..., t_m^i \tag{1}$$

$$x_{[CLS]}^u, x_1^u, ..., x_m^u = BERT(U_i) \tag{2}$$

After acquiring the BERT-encoded EDU vectors, we utilize a BiGRU for sequential representation of EDUs. The sequence h_1^u, h_2^u,...,h_n^u, derived via Equation 3, includes h_i^u as the BiGRU representation of each i-th EDU.

$$h_1^u, h_2^u, ..., h_n^u = BiGRU(x_{[CLS]}^u, x_1^u, ..., x_m^u) \tag{3}$$

Following the BiGRU layer, a self-attention mechanism is employed to refine the context representation further. The final output representation, denoted as Λ, is determined via Equations 4 and 5.

$$A_1^u, A_2^u, ..., A_n^u = Attention(h_1^u, h_2^u, ..., h_n^u) \tag{4}$$

$$\Lambda = A_1^u, A_2^u, ..., A_n^u \tag{5}$$

Following the attention mechanism, the output Λ is fed into the Transformer encoder [13]. Through multi-layer self-attention and positional encoding, the Transformer enhances the model's understanding of the input context. The output T_1^u, T_2^u,...,T_n^u is then concatenated with the BERT-encoded output to yield the global context representation, ζ^u, as described in Equations 6 and 7.

$$T_1^u, T_2^u, ..., T_n^u = Transformer(\Lambda) \tag{6}$$

$$\zeta^u = concat[x_i^u : T_i^u] \tag{7}$$

B. Speaker Encoding Layer

In a context of n EDUs, we form a speaker sequence of equivalent length, represented as $\{A, B, C, B, B, C, D\}$ in Figure 1.Each speaker in this sequence is processed through BERT, resulting in speaker-specific encodings, as delineated in Equation 8. To better capture speaker interactions, these encodings are further processed by a Bi-directional Gated Recurrent Unit (BiGRU) as seen in Equation 9. The output forms our final speaker representation ζ^s (Equation 10), encapsulating both speaker attributes and their interplay within dialogues. This representation enriches our understanding of multi-party discourse structure.

$$x_A^s = BERT(A) \tag{8}$$

$$h_1^s, h_2^s, ..., h_n^s = BiGRU(x_A^s, x_B^s, ..., x_D^s) \tag{9}$$

$$\zeta^s = h_1^s, h_2^s, ..., h_n^s \tag{10}$$

C. Interaction Layer

In the Interaction Layer, speaker and context information are merged using a weighted strategy, as detailed in Equation 11. The learnable weight parameter, γ, dynamically adjusts the contributions of both types of information to accommodate varying dialogue environments and task requirements.

$$H_i = \gamma\zeta^s + (1 - \gamma)\zeta^u \tag{11}$$

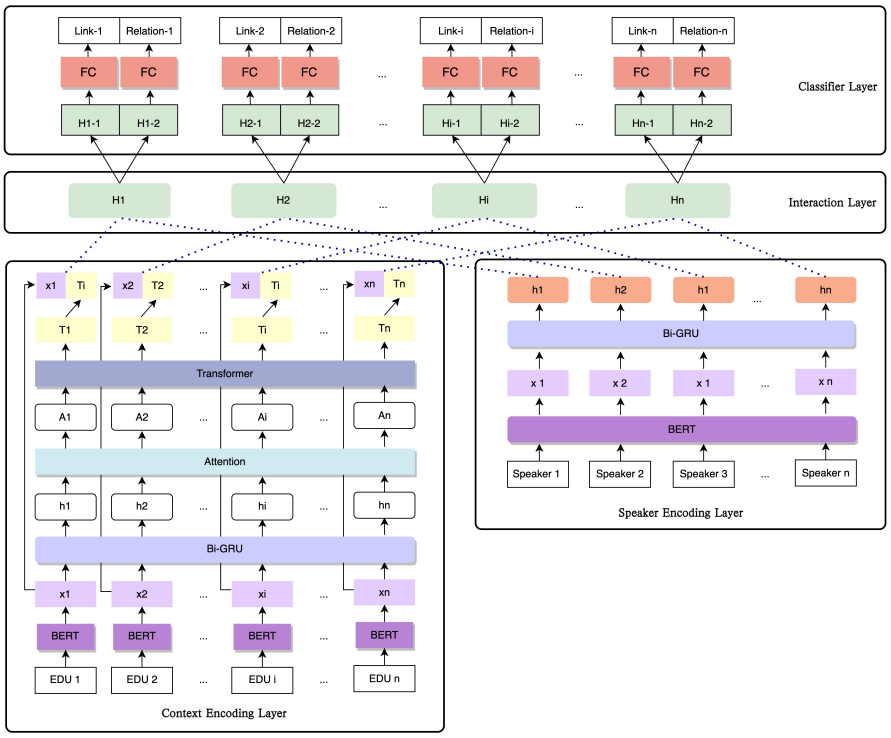


Fig. 2. The Framework of the Context-Speaker Enhanced Discourse Parsing Model.

D. Classifier Layer

The Classifier Layer, utilizing the information representation from the Interaction Layer, performs link and relation predictions. Link prediction is shown in Equation 12 and relation prediction in Equation 13, with w_{link}, b_{link} and w_{rel}, b_{rel} representing the corresponding weight and bias parameters.

$$P_{link} = softmax(w_{link}H_i + b_{link}) \tag{12}$$

$$P_{rel} = softmax(w_{rel}H_i + b_{rel}) \tag{13}$$

V. Experimental Result and Analysis

A. Datasets

We evaluate our model on three datasets, namely Molweni, DialogueDSA, and STAC, each of which provides unique and representative multi-party dialogue scenarios. These datasets, chosen for their linguistic and contextual variance, enable a robust evaluation of our model's performance and generalization in both English and Chinese multi-party dialogues.

B. Evaluation Metrics

Our model's performance is evaluated using the F1 score, a measure balancing precision and recall, for both link and relation prediction. For link prediction, metrics are computed based on the existence of links between predicted EDUs. Relation prediction accuracy requires both correct link identification and relation type, providing a precise evaluation of the model's prediction capabilities.

C. Model Training and Configuration

1) Loss Function: We use the cross-entropy function as the optimization target for link and relation prediction tasks. In Equation 14, P_{link} and P_{rel} denote the correct probability for link and relation prediction, with θ representing the model parameters. By co-training these subtasks, we not only enhance training efficiency and conserve computational resources but also foster information complementarity between tasks, which is crucial for improving model performance.

$$L_{(\theta)} = -(logP_{link} + logP_{rel}) \tag{14}$$

2) Configuration: Experiments were conducted on a server with an NVIDIA RTX A6000 GPU using PyTorch 1.13.1. The model utilized a single-layer GRU (hidden size: 250, dropout: 0.5) and a Transformer encoder for context encoding. Configurations included an embedding size of 768, four attention heads, and additional layers with 12 attention heads each. The feed-forward network had a dimension of 768. Optimization was done using the Adam optimizer with L2 regularization. We employed BERT pretrained checkpoints, namely *bert-base-uncased* and *bert-base-chinese*, setting BERT's learning rate to 1e-5, decay rate of 0.5, decay steps of 500, and gradient clipping at 2.0. Training was for 20 epochs with a batch size of 4. Maximum lengths for discourse analysis were 256 for sequence and 300 for basic discourse units.

D. Main Results and Analysis

1) Main Results: In Table IV, we present the comparative analysis of the performance of our proposed model, namely CSE-DPM, against several state-of-the-art models.

On the Molweni dataset, our CSE-DPM model achieved a Link F1 score of 83.0 and a Link&Rel F1 score of 59.5, outperforming other evaluated models. In contrast, the state-of-the-art model, which leverages joint training of MRC and DP on $BERT_{wwm}$[2], registers a Link F1 of 88.1 and a Link&Rel F1 of 66.9 [11]. This combined training approach potentially furnishes a richer representation. Additionally, the latter's superior performance might be attributed to the utilization of the more extensive $BERT_{wwm}$ architecture. Despite these disparities, the robustness and efficacy of our CSE-DPM model in parsing multi-party dialogues stand prominently highlighted.

On the DialogueDSA dataset, our CSE-DPM model achieved a Link F1 score of 76.2 and a Link&Rel F1 of 46.7. Although slightly trailing the Hierarchical+longformer in Link F1 at 80.1, it surpassed in the Link&Rel metric. Remarkably, CSE-DPM utilized the resource-efficient PLM bert-base-chinese, in contrast to the Hierarchical+longformer's larger longformer model[3]. This underscores CSE-DPM's operational efficiency with comparable performance but reduced computational demand. Its strategy of encoding each EDU with bert, as opposed to longformer's document-level approach, further economizes memory. Consequently, CSE-DPM offers a compelling balance of performance and efficiency, crucial for resource-constrained settings. This study also introduces the DialogueDSA dataset as a benchmark for such evaluations.

On the STAC dataset, our CSE-DPM model achieved a Link F1 of 72.5 and a Link&Rel F1 of 55.6. Though the Hierarchical+longformer model, utilizing a larger-parameter longformer, reported superior scores of 81.6 for Link F1 and 58.5 for Link&Rel F1, our model displayed competitive performance, surpassing many other methods. This consistency across datasets highlights CSE-DPM's generalization capabilities in diverse NLP contexts.

2) Ablation Study: We conducted an ablation study on the DialogueDSA dataset to assess the significance of specific components in our CSE-DPM model. Two model variants were evaluated: CSE-DPM - Transformer (without the Attention and core Transformer layers crucial for handling EDU interactions) and CSE-DPM - Speaker (lacking the speaker information module). As shown in Table V, the complete CSE-DPM model outperforms its variants in both Link and Link&Rel predictions on the DialogueDSA dataset. The diminished performance of the variants underscores the importance of the Transformer layer and speaker information when processing complex multi-party dialogues in DialogueDSA.

3) Case studies: In our case study involving speakers Heping, Lidama, and Zhiguo, our model CSE-DPM accurately aligned with the Gold Tree as shown in Figure 3, showcasing its efficacy in predicting discourse relations in intricate dialogues. Examining model variants, CSE-DPM-Transformer misclassified the relation between U0 and U2 as background, indicating challenges in handling cross-utterance complexity without the Transformer. Similarly, CSE-DPM-Speaker, devoid of speaker data, misjudged the same relation as elaboration, emphasizing speaker information's significance in discourse prediction. This study underscores the CSE-DPM model's prowess in discerning discourse relations in multi-party dialogues.

4) Analysis of Relation Type Distribution and Accuracy: Figure 4 illustrates the distribution and accuracy of discourse relation types in the DialogueDSA dataset. Notably, the "Comment" category is the most prevalent at 18.6% but achieves an accuracy of only 0.489. Similarly, "Inquiry" occupies 13.8% with an accuracy of 0.585. On the other hand, the "Confirmation" category, constituting 5.5% of the dataset, boasts the highest accuracy at 0.805. This suggests that prevalence in the dataset does not always correlate with classification performance. Some less represented categories, such as "Confirmation" and "Answer", exhibit superior performance, indicating possible distinctiveness in their features. However, discourse relations like "Conditional", "Contrast", "Parallel", and "Result" report zero accuracy, likely due to their sparse representation (all below 2.0%). This underscores the need for balanced data collection and highlights the potential benefits of over-sampling under-represented categories.

VI. Conclusion and Future Work

In this paper, we introduce DialogueDSA, an innovative dataset resource designed for Chinese multi-party dialogue discourse parsing. Additionally, we propose a model that excels in integrating speaker information and managing long texts, achieved through the unique encoding of EDUs. The robust performance of our model, substantiated by evaluations on the Molweni, DialogueDSA, and STAC datasets, underscores its superior ability in link and relation prediction. Considering the intricate discourse structures and strategies inherent in real-world multi-party dialogues, our future research endeavours will be directed towards exploring these dimensions to enhance the performance of our model further.

[2]https://huggingface.co/bert-large-uncased

[3]https://huggingface.co/schen/longformer-chinese-base-4096

References

[1] Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: Current directions in discourse and dialogue, vol. 22, pp. 85–112 (2003)

[2] Miltsakaki, E., Prasad, R., Joshi, A.K., Webber, B.L.: The Penn Discourse Treebank. In: LREC (2004)

[3] Asher, N., Lascarides, A.: Logics of conversation. Cambridge University Press (2003)

[4] Asher, N., Hunter, J., Morey, M., Benamara, F., Afantenos, S.: Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In: 10th International Conference on Language Resources and Evaluation (LREC 2016), pp. 2721–2727 (2016)

[5] Li, J., Liu, M., Kan, M.Y., Zheng, Z., Wang, Z., Lei, W., Liu, T., Qin, B.: Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In: arXiv preprint arXiv:2004.05080 (2020)

[6] Afantenos, S., Kow, E., Asher, N., Perret, J.: Discourse parsing for multi-party chat dialogues. In: Conference on Empirical Methods on Natural Language Processing (EMNLP 2015), pp. pp–928 (2015)

TABLE IV
F1 SCORES OF LINK AND RELATION PREDICTION WITH MODELS TRAINED ON DIALOGUEDS, MOLWENI AND STAC."*" MEANS THAT WE REPORT THE PERFORMANCE BY RERUNNING THEIR MODEL.

Methods	DialogueDSA		Molweni		STAC	
	Link	Link&Rel	Link	Link&Rel	Link	Link&Rel
Deep sequential [5]	73.1	42.2	78.1	54.8	73.2	55.7
Hierarchical [9]	-	-	80.1	56.1	75.5	57.2
DP [11]	-	-	75.9	56.2	-	-
Struct-Aware [14]	-	-	81.6	58.4	73.4	**5**7.3
HG-MDP [18]	-	-	81.5	58.5	-	-
Hierarchical+longformer* [9]	**80.1**	42.4	75.1	58.6	**81.6**	**58.5**
SSP-BERT + SCIJE* [10]	75.5	45.9	82.8	58.9	72.3	55.1
CSE-DPM	76.2	**46.7**	**83.0**	**59.5**	72.5	55.6

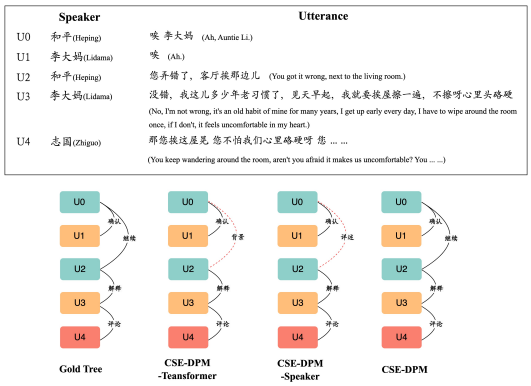


Fig. 3. Case studies of CSE_DPM model.

TABLE V
ABLATION EXPERIMENT RESULTS FOR DISCOURSE PARSING.

Methods	Link	Link&Rel
DialogueDSA		
CSE-DPM - Transformer	74.8	44.1
CSE-DPM - Speaker	74.3	43.5
CSE-DPM	**76.2**	**46.7**

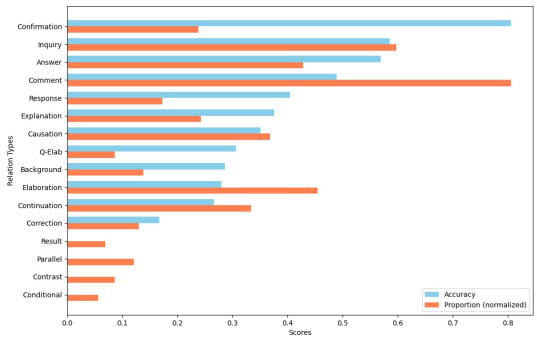


Fig. 4. Accuracy vs Proportion of Each Relation Type (Sorted by Accuracy).

[7] Perret, J., Afantenos, S., Asher, N., Morey, M.: Integer linear programming for discourse parsing. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 99–109 (2016)

[8] Shi, Z., Huang, M.: A deep sequential model for discourse parsing on multi-party dialogues. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 7007–7014 (2019)

[9] Liu, Z., Chen, N.F.: Improving multi-party dialogue discourse parsing via domain integration. In: arXiv preprint arXiv:2110.04526 (2021)

[10] Yu, N., Fu, G., Zhang, M.: Speaker-Aware Discourse Parsing on Multi-Party Dialogues. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 5372–5382 (2022)

[11] He, Y., Zhang, Z., Zhao, H.: Multi-tasking dialogue comprehension with discourse parsing. In: arXiv preprint arXiv:2110.03269 (2021)

[12] Yang, J., Xu, K., Xu, J., Li, S., Gao, S., Guo, J., Xue, N., Wen, J.R.: A joint model for dropped pronoun recovery and conversational discourse parsing in chinese conversational speech. In: arXiv preprint arXiv:2106.03345 (2021)

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems, vol. 30 (2017)

[14] Song, L.: Structure self-aware model for discourse parsing on multi-party dialogues. Google Patents. US Patent App. 17/181,431 (2022)

[15] Jiang, Y., Xu, Y., Zhan, Y., He, W., Wang, Y., Xi, Z., Wang, M., Li, X., Li, Y., Yu, Y.: The CRECIL Corpus: a New Dataset for Extraction of Relations between Characters in Chinese Multi-party Dialogues. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 2337–2344 (2022)

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805 (2018)

[17] Chi, T., Rudnicky, A.I.: Structured Dialogue Discourse Parsing. arXiv preprint arXiv:2306.15103 (2023)

[18] Li, J., Liu, M., Wang, Y., Zhang, D., Qin, B.: A speaker-aware multiparty dialogue discourse parser with heterogeneous graph neural network. In: Cognitive Systems Research, vol. 79, pp. 15–23 (2023)