A Distance-Aware Multi-Task Framework for Conversational Discourse Parsing

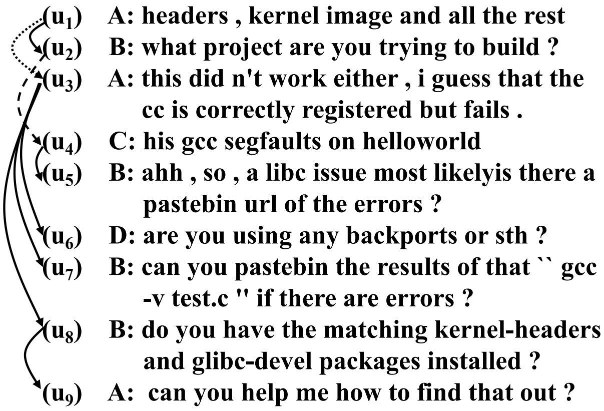
Yaxin Fan, Peifeng Li, Fang Kong, and **Qiaoming Zhu**[*]
School of Computer Science and Technology, Soochow University, Suzhou, China
yxfansuda@stu.suda.edu.cn
{pfli, kongfang, qmzhu}@suda.edu.cn

Abstract

Conversational discourse parsing aims to construct an implicit utterance dependency tree to reflect the turn-taking in a multi-party conversation. Existing works are generally divided into two lines: graph-based and transition-based paradigms, which perform well for short-distance and long-distance dependency links, respectively. However, there is no study to consider the advantages of both paradigms to facilitate conversational discourse parsing. As a result, we propose a distance-aware multi-task framework DAMT that incorporates the strengths of transition-based paradigm to facilitate the graph-based paradigm from the encoding and decoding process. To promote multi-task learning on two paradigms, we first introduce an Encoding Interactive Module (EIM) to enhance the flow of semantic information between both two paradigms during the encoding step. And then we apply a Distance-Aware Graph Convolutional Network (DAGCN) in the decoding process, which can incorporate the different-distance dependency links predicted by the transition-based paradigm to facilitate the decoding of the graph-based paradigm. The experimental results on the datasets STAC and Molweni show that our method can significantly improve the performance of the SOTA graph-based paradigm on long-distance dependency links. Our code is available at https://github.com/yxfanSuda/DAMT.

1 Introduction

The goal of conversational discourse parsing is to uncover latent conversation topics and construct an implicit utterance dependency tree to reflect the turn-taking in a multi-party conversation. Since the discourse structure is essential to understand multi-party conversations, it has been widely applied to various Natural Language Processing (NLP) applications,such as response generation (Hu et al., 2019), reading comprehension(Li et al., 2021b; Li and Zhao, 2021), meeting summarization (Feng et al., 2021), and emotion recognition(Sun et al., 2021).



(u_1) A: headers , kernel image and all the rest

(u_2) B: what project are you trying to build ?

(u_3) A: this did n't work either , i guess that the cc is correctly registered but fails .

(u_4) C: his gcc segfaults on helloworld

(u_5) B: ahh , so , a libc issue most likelyis there a pastebin url of the errors ?

(u_6) D: are you using any backports or sth ?

(u_7) B: can you pastebin the results of that `` gcc -v test.c '' if there are errors ?

(u_8) B: do you have the matching kernel-headers and glibc-devel packages installed ?

(u_9) A: can you help me how to find that out ?

Figure 1: A multi-party dialogue with its dependency structure, where the solid lines, dotted lines and dashed lines denote the relations “Clarification Question”, “Comment”, and “Result” respectively, and A, B, C, D mean different speakers.

Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) is one of the most influential theories to reveal the overall discourse structures in conversational discourse parsing. Unlike Rhetorical Structure Theory (RST) (Mann and Thompson, 1987), which limits the relationship to occur between adjacent EDUs[1] in monologue, SDRT represents multi-party conversations as dependency-based discourse structures, due to crossing dependencies. Recently, some SDRT-style corpora have been built, such as STAC (Asher et al., 2016) and Molweni (Li et al., 2020). Figure 1 shows an example of a multi-party conversation and its dependency structure from Molweni.

Existing work on conversational discourse pars-

[*] Corresponding author

[1] Elementary Discourse Units(EDUs) are the fundamental discourse units in discourse parsing. In the monologue, each EDU corresponds to a phrase or sentence. In the conversation, each EDU corresponds to an utterance.

Proceedings of the 29th International Conference on Computational Linguistics, pages 912–921
October 12–17, 2022.

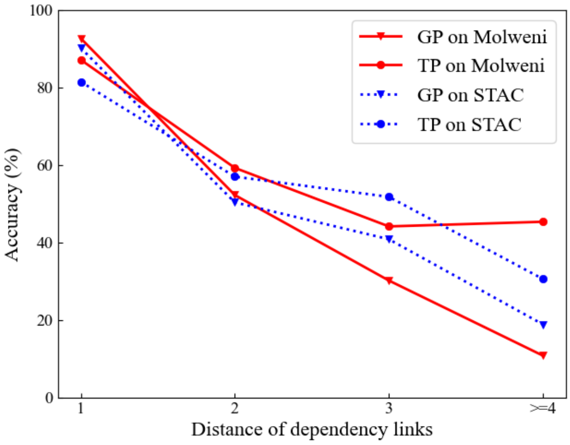


Figure 2: The performance of dependency links at different distances predicted by the graph-based and transition-based paradigms on the testing set of Molweni and STAC, respectively. The x-axis is the relative distance between EDUs (For example, the distance of (u_1, u_2) and (u_3, u_8) are 1 and 5 in Figure 1.). And the y-axis is the accuracy of dependency links. "TP" indicates Transition-based Paradigm and "GP" means Graph-based Paradigm.TP comes from (Shi and Huang, 2019), which utilizes hierarchical GRU to encode the conversations, while GP comes from (Wang et al., 2021a) and we also utilize hierarchical GRU to encode the dialogues for a fair comparison.

ing can be divided into two lines: graph-based and transition-based paradigms. The graph-based paradigm (Muller et al., 2012; Afantenos et al., 2015; Perret et al., 2016; Yang et al., 2021; Wang et al., 2021a) first obtains the probability of the discourse relation for each EDU pair, then a global decoding method is applied to construct the discourse structure. The transition-based paradigm (Shi and Huang, 2019; Wang et al., 2021b) first obtains the probability of the discourse relation between the current EDU and all previous EDUs, then discourse structure is constructed incrementally. Due to the discrepancy in the parsing process, both of them have different strengths in predicting the dependency links at various distances.

As shown in Figure 2, we analysis the performance of dependency links at different distances predicted by the above two paradigms. The results show that the graph-based paradigm performs better for dependency links with the distance 1, while the transition-based paradigms performs better when the distance greater than 1. As a result, it is a great challenge to combine the advantages of both two paradigms to facilitate conversational discourse parsing.

Previous work (Falenska et al., 2020) has demonstrated the effectiveness of multi-task learning to integrate both two paradigms in a similar task dependency syntactic parsing. Through the shared encoding layer, both two paradigms can facilitate each other implicitly. However, it is not sufficient to apply this approach to conversational discourse parsing, because the advantages of one paradigm cannot be explicitly exploited to facilitate the other one.

To alleviate the above issues, we propose a Distance-Aware Multi-Task framework (DAMT) for conversational discourse parsing that allows one paradigm to explicitly facilitate the decoding process of the other from the encoding and decoding process, respectively. Specially, we introduce an Encoding Interactive Module (EIM) to enhance the flow of semantic information between both two paradigms during the encoding step. And then we apply a Distance-Aware Graph Convolutional Network (DAGCN) in the decoding process, which can incorporate the different-distance dependency links predicted by the transition-based paradigm to facilitate the decoding of the graph-based paradigm. The experimental results on two datasets STAC and Molweni show that our DAMT outperforms the SOTA baselines, especially the significant improvement on those dependency links with long distances.

2 Related Work

Most previous studies for overall discourse structure parsing are based on Rhetorical Structure Theory Discourse TreeBank (RST-DT) (Carlson et al., 2003), including greedy bottom-up approach (Feng and Hirst, 2014), CYK-based approaches (Joty et al., 2015; Liu and Lapata, 2017) and transition-based methods (Wang et al., 2017; Lin et al., 2019; Kobayashi et al., 2020; Zhang et al., 2021).

In this paper, we focus on parsing conversational dependency structures that allow crossing dependencies. Recently, there are two available corpora, i.e., STAC (Asher et al., 2016) and Molweni (Li et al., 2020) defined 16 relation types. STAC collected from an online game *The Settlers of Catan*, which contains 1,062 and 111 dialogues for training and testing, respectively. Molweni is based on Ubuntu Chat (Lowe et al., 2015), which contains 9,000, 500 and 500 instances for training, validating and testing, respectively.

Up to now, only a few studies focused on con-

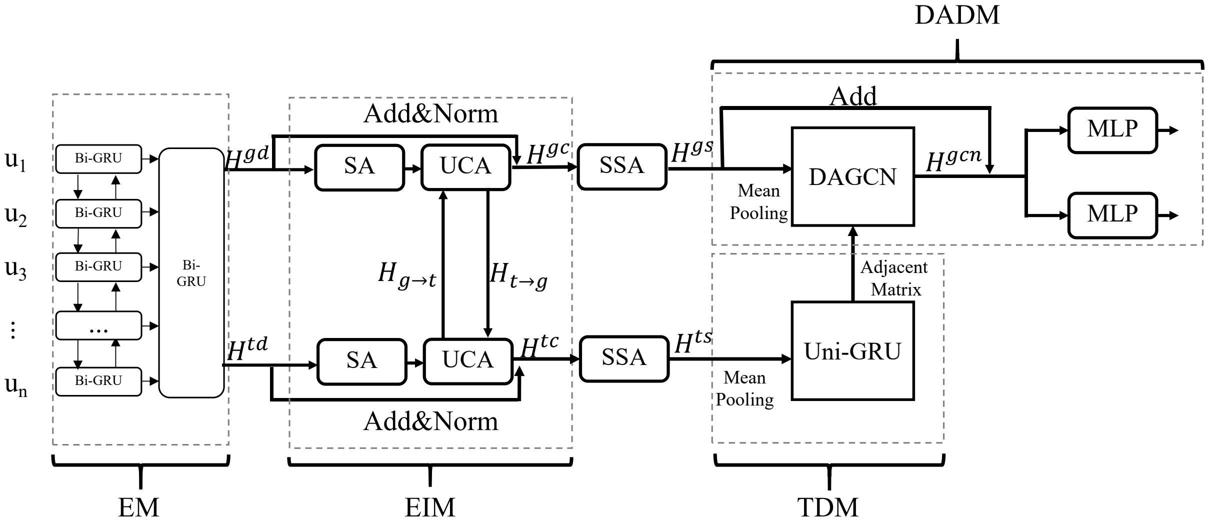


Figure 3: The architecture of our DAMT framework.

versational discourse parsing and most of them can be divided into two paradigms, i.e., graph-based paradigm and transition-based paradigm.

Graph-based paradigm Muller et al. (2012), Afantenos et al. (2015) and Perret et al. (2016) adopted traditional manual features to calculate the probabilities of all EDU pairs and then global decoding algorithm (e.g., Maximum Spanning Trees, A* and Integer Linear Programming) was used to construct dependency structures. With the development of deep learning, some advanced methods are used to obtain the semantic representation of each EDU pair. Wang et al. (2021a) proposed a novel edge-centric graph neural network to enhance the semantic representation of EDUs. Yang et al. (2021) first used the dependency syntactic graph to obtain a better EDU representation and then a biaffine relation prediction layer was applied to obtain the probability of each EDU pair.

Transition-based paradigm Shi and Huang (2019) proposed a Deep Sequence Model (DSM) to predict dependency links and corresponding relation types jointly and alternately. Their model not only consider the local information of the concerned EDUs but also utilizes the historical structure. Based on DSM, Wang et al. (2021b) adopted the graph attention network by incorporating cohesion information including lexical chain and coreference chain to enhance the semantic representation of EDUs.

3 DAMT

Our framwwork DAMT is shown in Figure 3, which includes four components: Encoding Module (EM), Encoding Interaction Module (EIM), Transition-based Decoding Module (TDM), and Distance-Aware Decoding Module (DADM). In EM, the hierarchical GRU is applied to obtain the semantic representation of dialogues. In EIM, the semantic representations of different paradigms can interact explicitly to promote the multi-task learning. In TDM, we adopt a pointer network for transition-based decoding to obtain the dependency structures. In DADM, the dependency structures predicted by TDM are incorporated by DAGCN for the final graph-based decoding.

3.1 Encoding Module

In the encoding module EM, we adopted hierarchical GRU to obtain the semantic representation of dialogues for both paradigms. For each EDU u_i in dialogue D=$\{u_1, u_2, \cdots, u_n\}$, a bidirectional GRU (bi-GRU) encoder is applied on the word sequence, and the last hidden states in two directions are concatenated as the EDU-level semantic representation, denoted as $\boldsymbol{h}_e^i \in \mathbb{R}^d$. Then another bi-GRU is applied on the EDU-level representation to obtain the dialogue-level representation. We use $\boldsymbol{H}_{td}$ and $\boldsymbol{H}_{gd}$ to denote the semantic representation of transition-based paradigm and graph-based paradigm respectively, where $\boldsymbol{H}_{td}, \boldsymbol{H}_{gd} \in \mathbb{R}^{n\times d}$.

3.2 Encoding Interaction Module

Several studies (E et al., 2019; Qin et al., 2021; Li et al., 2021a) have demonstrated that the explicit interaction between encoding representations of different tasks can better improve each other in multi-task learning. Inspired by this, we propose an encoding interaction module EIM, which can build a bidirectional connection between two paradigms.

Our EIM consists of two Self Attention (SA) layers and two Unidirectional Cross Attention (UCA) layers. We first feed the semantic representation $\boldsymbol{H}_{td}$ and $\boldsymbol{H}_{gd}$ from EM into the SA sub-layers to obtain the internal semantic information for each paradigm as follows.

$$\begin{aligned} \boldsymbol{H}_t &= \mathrm{SA}(\boldsymbol{W}_q^s \boldsymbol{H}_{td}, \boldsymbol{W}_k^s \boldsymbol{H}_{td}, \boldsymbol{W}_v^s \boldsymbol{H}_{td}) \\ \boldsymbol{H}_g &= \mathrm{SA}(\boldsymbol{W}_q^s \boldsymbol{H}_{gd}, \boldsymbol{W}_k^s \boldsymbol{H}_{gd}, \boldsymbol{W}_v^s \boldsymbol{H}_{gd}) \end{aligned} \tag{1}$$

where $\mathrm{SA}(.)$ denotes multi-head attention as (Vaswani et al., 2017) and $\boldsymbol{W}_q^s, \boldsymbol{W}_k^s, \boldsymbol{W}_v^s$ are weight matrix, which map vectors to the same feature space.

Second, two UCA layers are applied to build the connection between the two paradigms, where one from $\boldsymbol{H}_t$ to $\boldsymbol{H}_g$ and one from $\boldsymbol{H}_g$ to $\boldsymbol{H}_t$ as follows.

$$\begin{aligned} \boldsymbol{H}_{g\to t} &= \mathrm{UCA}(\boldsymbol{W}_q^c \boldsymbol{H}_g, \boldsymbol{W}_k^c \boldsymbol{H}_t, \boldsymbol{W}_v^c \boldsymbol{H}_t) \\ \boldsymbol{H}_{t\to g} &= \mathrm{UCA}(\boldsymbol{W}_q^c \boldsymbol{H}_t, \boldsymbol{W}_k^c \boldsymbol{H}_g, \boldsymbol{W}_v^c \boldsymbol{H}_g) \end{aligned} \tag{2}$$

$\mathrm{UCA}(.)$ is a variant of $\mathrm{SA}(.)$, which uses $\boldsymbol{H}_t(\boldsymbol{H}_g)$ as query vectors and $\boldsymbol{H}_g(\boldsymbol{H}_t)$ as the context vector, thus enabling an explicit interaction between the two vectors. The UCA layer is used to make the encoding semantics of one paradigm updated with the guidance of the other one, achieving a bidirectional connection between both two paradigms.

Then, we add a residual connection and layer normalization function $\mathrm{LayerNorm}(.)$ to obtain the semantic representations of the two paradigms as follows.

$$\begin{aligned} \boldsymbol{H}_{tc} &= \mathrm{LayerNorm}(\boldsymbol{H}_t + \boldsymbol{H}_{t\to g}) \\ \boldsymbol{H}_{gc} &= \mathrm{LayerNorm}(\boldsymbol{H}_g + \boldsymbol{H}_{g\to t}) \end{aligned} \tag{3}$$

Finally, following the previous work (Wang et al., 2021a), Structure Self Attention (SSA) is applied to enhance the semantic representation of dialogues by incorporating the structural information of conversations. By feeding $\boldsymbol{H}_{tc}$ and $\boldsymbol{H}_{gc}$ to the SSA, we can obtain the semantic representation of all EDU pairs which incorporate the structural information of dialogues, denotes as $\boldsymbol{H}_{ts}$ and $\boldsymbol{H}_{gs}$, where $\boldsymbol{H}_{ts}, \boldsymbol{H}_{gs} \in \mathbb{R}^{n \times n \times d}$.

3.3 Transition-Based Decoding Module

We adopt a pointer network for transition-based decoding. After obtaining the semantic representations $\boldsymbol{H}_{ts}$, we first applied mean pooling on $\boldsymbol{H}_{ts}$ to obtain the semantic representation of EDUs containing structural information, denotes as $\boldsymbol{H}_{tm} \in \mathbb{R}^{n \times d}$. As shown in Figure 4, $\boldsymbol{H}_{tm}$ are fed into a uni-directional GRU for transition-based decoding, and the initial state of the decoder is taken from the combination of the last states of hierarchical GRU in both directions[2].

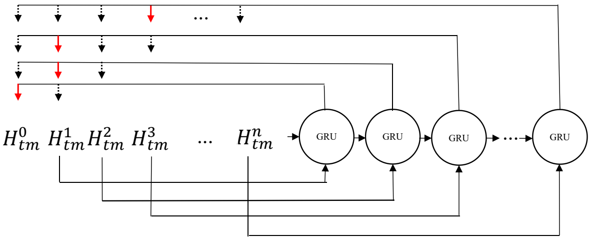


Figure 4: The decoding process of pointer network, where the solid red arrows indicate dependency links of current EDUs.

At each decoding step, it supposes the current EDU index is i at the k-th step. Then, the semantic representation $\boldsymbol{H}_{tm}$ is fed to the decoder and its output at the k-th step is $\boldsymbol{h}_{dk}$. After that, we adopted the Biaffine Attention mechanism to the representation $\boldsymbol{H}_{tm}$ and the output $\boldsymbol{h}_{dk}$ to obtain the probability between the current EDU and all previous EDUs that existing the dependency links and the corresponding relation type as follows.

$$\boldsymbol{s}_i^j = \boldsymbol{H}_{tm}{}^T \boldsymbol{W} \boldsymbol{h}_{dk} + \boldsymbol{U} \boldsymbol{H}_{tm} + \boldsymbol{V} \boldsymbol{h}_{dk} + \boldsymbol{b} \tag{4}$$

where $\boldsymbol{s}_i^j \in \mathbb{R}^m$ refers to the probability between the current EDU u_i and the previous EDU u_j and m is set to 1 when there is a dependency link between them, or set to the number of relation types when there is a dependency type between them. Besides, $\boldsymbol{W}$ denotes the weight matrix of the bi-linear term, $\boldsymbol{U}, \boldsymbol{V}$ are the two weight vectors of the linear terms, and $\boldsymbol{b}$ is the bias vector.

For a conversation with n utterances, we use the adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ to represent the transition-based dependency structure, where $\boldsymbol{A}_{ij} = 1$ if there is a dependency link between the EDUs u_i and u_j. Then, the adjacency matrix $\boldsymbol{A}$ is fed to DAGCN to incorporate the transition-based dependency structure.

3.4 Distance-Aware Decoding Module

In the distance-aware decoding module DADM, DAGCN is first applied to capture the dependency structure predicted by the transition-based

[2] Following previous work, we add a dummy root u_0 to represent the beginning of a dialogue.

paradigm. Then the output of DAGCN and the semantic representation of the graph-based paradigm are fused together for final decoding.

To incorporate the information of dependency links with different distances, the adjacency matrix $\boldsymbol{A}$ converted from the dependency structure, which is predicted by the transition-based paradigm, is fed to DAGCN to facilitate the decoding process of the graph-based paradigm. Inspired by Meng et al. (2020), we hypothesize that the dependency links with different distances have a different impact on the decoding process of the graph-based paradigm. Hence, the trainable weights is applies to the adjacency matrix and each weight is determined by the distance of the dependency link and the corresponding relation type from the transition-based paradigm.

Let $\boldsymbol{W}_d$ be a matrix of $\mathbb{R}^{N_d \times d'}$ where N_d is the number of different distances [3] and d' is the dimension of the embedding space. Let $\boldsymbol{W}_r$ be a matrix of $\mathbb{R}^{N_r \times d'}$ where N_r is the number of dependency relation types. The feature combination weight over the element $\boldsymbol{A}_{ij}$ in the adjacency matrix $\boldsymbol{A}$ can be represented as follows.

$$\boldsymbol{\alpha}_{ij} = \boldsymbol{W}[\text{dis}(\boldsymbol{d}_{ij}) : \text{rel}(\boldsymbol{r}_{ij})] \tag{5}$$

Where $\boldsymbol{d_{ij}}$ is the distance from the EDU u_i to u_j, and $\boldsymbol{r}_{ij}$ is the corresponding relation type. The functions dis(.) and rel(.)are vector mapping functions, which map the one-hot vector $\boldsymbol{d}_{ij}$ and $\boldsymbol{r}_{ij}$ into the corresponding column of $\boldsymbol{W}_d$ and $\boldsymbol{W}_r$, respectively. [:] indicates the concatenation operation, and $\boldsymbol{W} \in \mathbb{R}^{2d'}$ is weight matrix.

Let $\boldsymbol{A}^{'}$ be the final adjacent matrix for DAGCN, then each element of $\boldsymbol{A}^{'}$ can be computed as:

$$\boldsymbol{A}^{'}_{ij} = \boldsymbol{\alpha}_{ij} \boldsymbol{A}_{ij} \tag{6}$$

Then, we add an identity matrix to $\boldsymbol{A}^{'}$, which makes each node can connect to itself.

After obtaining the adjacent matrix $\boldsymbol{A}^{'}$, we first apply mean pooling on the semantic representation of graph-based paradigm $\boldsymbol{H}_{gs}$ to obtain $\boldsymbol{H}_{gm} \in \mathbb{R}^{n \times \frac{d}{2}}$, then the calculation of DAGCN is as follows.

$$\boldsymbol{H}_{gcn} = \text{ReLU}(\boldsymbol{A}^{'} \boldsymbol{H}_{gm} \boldsymbol{W}) + \boldsymbol{H}_{gm} \tag{7}$$

where $\boldsymbol{H}_{gcn} \in \mathbb{R}^{n \times \frac{d}{2}}$ is the output of DAGCN, which incorporate the dependency structure predicted by TDM, and $\boldsymbol{W}$ is the parameter matrix. Then, we broadcast the output of DAGCN and add it to the semantic representations $\boldsymbol{H}_{gs}$ to obtain the final representation $\boldsymbol{H}_f \in \mathbb{R}^{n \times n \times d}$ for EDU pairs.

Lastly,$\boldsymbol{H}_f$ is fed into two multi-layer perceptrons to obtain the probability distribution of each EDU pair's existing dependency link and the corresponding relation type as follows.

$$\begin{aligned} \boldsymbol{S}_l &= \text{Softmax}(\text{MLP}(\boldsymbol{H}_f)) \\ \boldsymbol{S}_r &= \text{Softmax}(\text{MLP}(\boldsymbol{H}_f)) \end{aligned} \tag{8}$$

where $\boldsymbol{S}_l \in \mathbb{R}^{n \times n \times 1}$ and $\boldsymbol{S}_r \in \mathbb{R}^{n \times n \times m}$ and m is the number of relation type. To find the highest-scoring tree, we apply greedy decoding method on $\boldsymbol{S}_l$. After determining the dependency link from u_j to u_i, the relation type can be obtained by the probability $\boldsymbol{S}_r^{ij}$.

3.5 Multi-Task Learning

For multi-task learning, we have two goals: (i) optimizing the Transition-based Paradigm (TP) and (ii) optimizing the Graph-based Paradigm (GP).

To optimize TP, we minimize the sum of the loss for constructing the right dependency structure and the loss for predicting the correct relation type, which is calculated as follows.

$$\begin{aligned} L_t(\theta_t) = &- \sum_{t=1}^{n} logP_{\theta_t}(y_t | y_{<t}, X) \\ &- \sum_{t=1}^{n} \sum_{j=1}^{m} r_{t,j} logP_{\theta_t}(r_t, X) \end{aligned} \tag{9}$$

where θ_t denotes the parameters of TP to be optimized, $y_{<t}$ represents the historical structure that has been generated at previous steps,n is the total number of dependency links, m is the total number of relation types and $r_{t,j}$ is the golden relation type.

To optimize GP, we minimize the cross-entropy of gold dependency links between EDUs pairs as follows.

$$\begin{aligned} L_g(\theta_g) = &- \sum_{i=1}^{n} y^* logP_{\theta_g}(y_t, X) \\ &- \sum_{i=1}^{n} \sum_{j=1}^{m} r_{i,j} logP_{\theta_g}(r_i, X) \end{aligned} \tag{10}$$

where θ_g denotes the parameters of GP to be optimized, y^* represents the golden dependency links,

[3] In our implementation, we set the distance to 2 for all distances greater than or equal to 2.

Model		Molweni		STAC	
		Link	Link&Rel	Link	Link&Rel
w/o pre-trained model					
TP	DSM*	77.32	54.15	72.10	53.56
	LCCC	-	-	72.50	55.20
	PN	81.02	56.47	72.56	54.40
GP	SSAM*	81.15	56.93	72.92	54.83
Ours	DAMT	**82.25**	**57.35**	**73.54**	**55.32**
w/ pre-trained model					
GP	DiscProReco	-	-	74.10	57.00
	SSAM*	81.52	57.90	73.09	56.57
Ours	DAMT	**82.50**	**58.91**	**73.64**	**57.42**

Table 1: F1 scores (%) for different models where Link refers to link prediction and Link&Rel refers to that a correct prediction must predict dependency link and relation type correctly at the same time. We used the t-test with a 95% confidence interval for the significance test and all improvements of DAMT over SSAM are significant ($p < 0.05$). TP is short for Transition-based Paradigm, and GP is short for Graph-based Paradigm. $*$ indicates that we reproduce the scores of models using their released code.

Distance	Percentage(%)	
	Molweni	STAC
1	64.94	55.63
2	21.49	21.26
3	7.38	10.63
>=4	6.19	12.48

Table 2: Distribution of dependency links at different distances in the training set of Molweni and STAC.

n is the total number of dependency links, m is the total number of relation types and $r_{t,j}$ is the golden relation type.

For multi-task learning of the TP and GP, we add the above loss terms as follows.

$$L = L_t + L_g \tag{11}$$

4 Experimentation

In this section, we first introduce the datasets, hyper-parameters and baselines used in our evaluation, and then report the experimental results.

4.1 Datasets

We conduct experiments on two publicly available multi-party dialogue datasets: Molweni (Li et al., 2020) and STAC (Asher et al., 2016) and we preprocess these two datasets following Shi and Huang (2019). The distribution of dependency links at different distances on the training set of Molweni and STAC are shown in Table 2.

We can find that the dependency links with the distance 1 dominate both two corpora (about 65% and 56% in Molweni and STAC, respectively). The small percentage of long-distance dependency links poses challenges for models to predict. Our method DAMT can explicitly leverage the strengths of the transition-based paradigm to facilitate the graph-based decoding and further improve the performance on long-distance dependency links.

4.2 Hyper-Parameters

Following the previous work, we initialize words with GloVe embeddings (Pennington et al., 2014) that are fine-tuned during training. For Molweni and STAC corpus, the embedding dimension is set at 200 and 100, respectively. The dimension of the hidden representation d is set to 256 and the layer of EIM is set to 1 with 4 heads. The layer of SSA is set to 2 with 4 heads. The dimension of the embedding space d' in DAGCN is set to 8. And we set the dropout rate to 0.5 and employ Stochastic Gradient Descent (SGD) to train the model. The batch size is set to 150 and 70 for Molweni and STAC, respectively, and the initial learning rate is set to 3e-2. For the experiments using the pre-trained model, we apply XLNet-based (Yang et al., 2019) to obtain the semantic representations of EDUs and adopt AdamW to optimizer the model. The learning rate is set to 3e-4 and the dimension of the hidden representation d is set to 768. Besides, in this paper, the micro-averaged F1 score is adopted for evaluation.

4.3 Baselines

To verify the effectiveness of our DAMT, we conduct the following strong baselines for comparison.

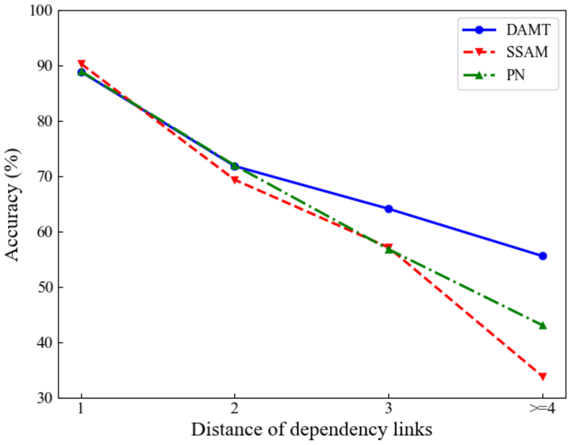


Figure 5: Comparison of dependency link accuracy at the different distance between our approach and two baselines.

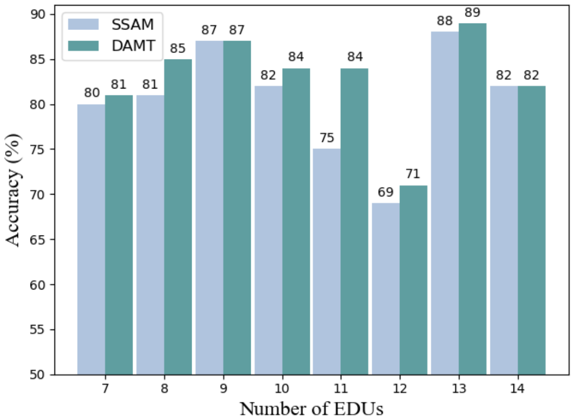


Figure 6: Accuracy of dependency links at various numbers of EDUs between our DAMT and SSAM.

Transition-based models 1) **DSM** (Shi and Huang, 2019): it predicted links and corresponding relation types jointly and alternately by considering the historical structure predicted; 2) **LCCC** (Wang et al., 2021b): it is based on DSM and enhanced the semantic representation of EDUs by incorporating cohesion information; 3) **PN**: it is the pointer network we proposed for transition-based decoding. As shown in Figure 3, our framework becomes PN after discarding the modules EIM and DADM.

Graph-based models 1) **DiscProReco** (Yang et al., 2021): it used the syntactic dependency graph to enhance the semantic representation of EDUs pairs; 2) **SSAM** (Wang et al., 2021a): it used the structure self attention network and two auxiliary training signals to enhance the semantic representation of EDU pairs.

4.4 Results

Table 1 shows the performance comparison between our DAMT and all the Transition-based and Graph-based baselines. Besides, the results without the pre-trained model and with the pre-trained model are also shown in Table 1. We can find out that our DAMT outperforms all baselines without the pre-trained model.

Especially, compared with the SOTA transition-based PN, our DAMT improves the F1-score by 1.23 and 0.88 in link and relation on Molweni, respectively, and improves them by 0.98 and 0.92 on STAC, respectively. Compared with the SOTA graph-based SSAM, our DAMT improves the F1-score by 1.10 and 0.42 in link and relation on Molweni, respectively, and improves them by 0.62 and 0.49 on STAC, respectively. Our approach DAMT leverages the strengths of different paradigms from two perspectives with multi-task learning and therefore achieves better performance.

Besides, with the application of the pre-trained model, our approach DAMT still improves the F1-score on Molweni and STAC compared to the SSAM, which illustrates the effectiveness of our approach.

5 Analysis

In this section we first conduct a detailed analysis on the dependency links with different distances and the different lengths of dialogues, and then provide the ablation study and case study.

5.1 Performance on Dependency Links with Different Distances

To further analyze the improvements of our model DAMT, we investigate the accuracy of dependency links at different distance between DAMT and two SOTA baselines on Molweni, as shown in Figure 5, where PN is a transition-based model and SSAM is a graph-based model.

Comparing with SSAM, we can find that PN performs better when the distance of dependency links is greater than one, while SSAM performs better when the distance of dependency links is one. This result shows that even though the semantic representations of EDUs are obtained using more advanced methods, both paradigms still have different performances on dependency links with different distances due to the discrepancy in dependency structure construction.

Overall, all models have a similar downward

Model	Link Accuracy (%)	
	Molweni	STAC
DAMT	66.93	50.20
-EIM	64.60(↓ 2.33)	47.60(↓ 2.60)
-DAGCN	62.91(↓ 4.02)	44.40(↓ 5.80)
-EIM,DAGCN	62.27(↓ 4.66)	43.00(↓ 7.20)

Table 3: The Accuracy of dependency links with distance greater than one in the test set of Molweni and STAC.

trend. Compared with PN and SSAM, our DAMT improves the performance of dependency links with a distance greater than one significantly. It indicates the effectiveness of our model that explicitly integrates both paradigms from two perspectives.

5.2 Performance on Different Lengths of Dialogues

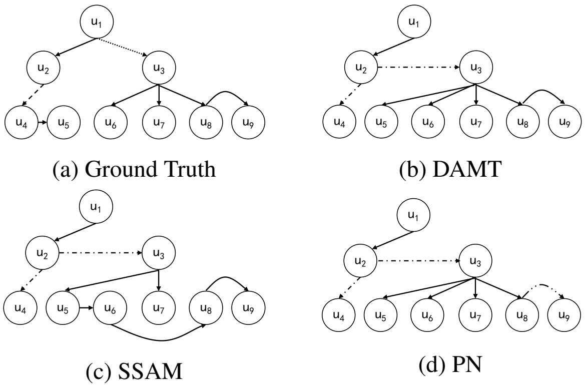
We further analyze the performance of DAMT and SSAM in terms of the number of EDUs on Molweni. Figure 6 shows the accuracy of dependency links at various numbers of EDUs in a document. Compared with SSAM, our model DAMT improved performance on almost all documents with different numbers of EDUs, especially those documents containing 11 and 8 EDUs with an improvement of 9 % and 4 %, respectively.

We analyze the distribution of dependency links with different distances at various numbers of EDUs. We find that the percentages of long dependency links (>1) in those conversations with 8 and 11 EDUs are the highest with 43% and 41%, respectively. This can reflect that our DAMT mainly improves the performance of the long dependency links.

5.3 Ablation Study

To investigate the impacts of the proposed modules on the performance of the dependency links with the longer distances, we conduct an ablation study on two modules EIM and DAGCN in DAMT, as illustrated in Table 3, where “-” indicates the removal of the single or several modules.

Removing any of the two modules makes the performance worse, while discarding DAGCN has the greatest impact on the performance. This shows that DAGCN can explicitly aggregate the information of dependency links with different distances in the discourse structure predicted by the PN, thus improving the performance of dependency links with the longer distance.



(a) Ground Truth

(b) DAMT

(c) SSAM

(d) PN

Figure 7: Discourse structures of the example in Figure 1. (a) refers to the ground truth structure and (b)-(d) refer to the structures predicted by DAMT, SSAM and PN. Different lines indicate different relation types, where the solid lines, dotted lines, dashed lines, dashed dotted lines and dashed double-dotted lines denote the relations “Clarification Question”, “Comment”, “Result”, "Question-answer_pair" and "Q-Elab", respectively.

5.4 Case Study

Figure 7 show the dependency structures of the example in Figure 1, which are the golden truth and three predicted results by our DAMT and two baselines SSAM and PN.

Compared with the transition-based PN, we find out that SSAM can better predicts the dependency links and their corresponding relation types when the distance is one, such as $u_1 \rightarrow u_2$ and $u_8 \rightarrow u_9$. Although PN can correctly predicts the dependency link $u_8 \rightarrow u_9$, it cannot correctly predict its corresponding relation type. This shows the advantage of graph-based approach over transition-based approach for dependency links when the distance is one. On the contrary, for those dependency links with distance greater than one, PN performs better than SSAM, such as $u_3 \rightarrow u_6$ and $u_3 \rightarrow u_8$.

Compared with PN, DAMT can further correctly predict the relation type of dependency link $u_8 \rightarrow u_9$. And compared with SSAM, DAMT can correctly predict long-distance dependency links $u_3 \rightarrow u_6$, $u_3 \rightarrow u_8$. These phenomena show that our distance-aware multi-task framework is able to combine the advantages of both two paradigms to predict those long-distance dependency links.

Besides, for some dependency links, such as $u_1 \rightarrow u_3$ and $u_4 \rightarrow u_5$, all models fail to identify their links. This indicates that conversational discourse paring is still a challenging task.

6 Conclusion

In this paper, we propose a distance-aware multi-task framework DAMT to facilitate conversational discourse parsing. First, we propose an encoding interaction module to enhance the information flow between the graph-based paradigm and transition-based paradigm to promote multi-task learning. Second, we propose a distance-aware graph convolutional network DAGCN incorporating the dependency structure predicted by one paradigm to explicitly facilitate the decoding process of another paradigm. The experimental results on two public datasets show the effectiveness of our proposed DAMT. In the future, we will further explore how to recognize relation types more effectively.

7 Acknowledgments

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 61836007, 62276177 and 62006167), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Stergos Afantenos, Éric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *EMNLP*, pages 928–937.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *LREC*, pages 2721–2727.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Haihong E, Peiqing Niu, Zhongfu Chen, and Meina Song. 2019. A novel bi-directional interrelated model for joint intent detection and slot filling. In *ACL*, pages 5467–5471.

Agnieszka Falenska, Anders Björkelund, and Jonas Kuhn. 2020. Integrating graph-based and transition-based dependency parsers in the deep contextualized era. In *IWPT*, pages 25–39.

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

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *IJCAI*, pages 3808–3814.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *IJCAI*, pages 5010–5016.

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

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

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021a. Modularized interaction network for named entity recognition. In *ACL-IJCNLP*, pages 200–209.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *COLING*, pages 2642–2652.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021b. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In *IJCNN*, pages 1–8.

Yiyang Li and Hai Zhao. 2021. Self- and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension. In *Findings of EMNLP*, pages 2053–2063.

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

Yang Liu and Mirella Lapata. 2017. Learning contextually informed representations for linear-time discourse parsing. In *EMNLP*, pages 1289–1298.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, pages 285–294.

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

Fanyu Meng, Junlan Feng, Danping Yin, Si Chen, and Min Hu. 2020. A structure-enhanced graph convolutional network for sentiment analysis. In *Findings of EMNLP*, pages 586–595, Online.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *COLING*, pages 1883–1900.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *NAACL*, pages 99–109.

Libo Qin, Tailu Liu, Wanxiang Che, Bingbing Kang, Sendong Zhao, and Ting Liu. 2021. A co-interactive transformer for joint slot filling and intent detection. In *ICASSP*, pages 8193–8197.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *AAAI*, pages 7007–7014.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of ACL*, pages 2949–2958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, page 5998–6008.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021a. A structure self-aware model for discourse parsing on multi-party dialogues. In *IJCAI*, pages 3943–3949.

Jinfeng Wang, Longyin Zhang, and Fang Kong. 2021b. Multi-level cohesion information modeling for better written and dialogue discourse parsing. In *NLPCC*, pages 40–52.

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

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Nianwen Xue, and Ji-Rong Wen. 2021. A joint model for dropped pronoun recovery and conversational discourse parsing in Chinese conversational speech. In *ACL*, pages 1752–1763.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NIPS*, volume 32.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021. Adversarial learning for discourse rhetorical structure parsing. In *ACL-IJCNLP*, pages 3946–3957.