Speaker-Aware Discourse Parsing on Multi-Party Dialogues

Nan Yu[1], Guohong Fu[1,2]*, Min Zhang[3]
[1]School of Computer Science and Technology, Soochow University, China
[2]Institute of Artificial Intelligence, Soochow University, China
[3]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen), China
nyu@stu.suda.edu.cn,{ghfu,minzhang}@suda.edu.cn

Abstract

Discourse parsing on multi-party dialogues is an important but difficult task in dialogue systems and conversational analysis. It is believed that speaker interactions are helpful for this task. However, most previous research ignores speaker interactions between different speakers. To this end, we present a speaker-aware model for this task. Concretely, we propose a speaker-context interaction joint encoding (SCIJE) approach, using the interaction features between different speakers. In addition, we propose a second-stage pre-training task, same speaker prediction (SSP), enhancing the conversational context representations by predicting whether two utterances are from the same speaker. Experiments on two standard benchmark datasets show that the proposed model achieves the best-reported performance in the literature. We will release the codes of this paper to facilitate future research[1].

1 Introduction

Discourse parsing on multi-party dialogues aims to identify the discourse relations between utterances in dialogues, which has received increasing attention in the natural language processing (NLP) community (Shi and Huang, 2019; He et al., 2021; Liu and Chen, 2021; Yang et al., 2021). Unlike traditional text-level discourse parsing based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), this task is performed based on the Segmented Discourse Relation Theory (SDRT) (Asher et al., 2003). It represents a multi-party dialogue by a discourse dependency tree (Afantenos et al., 2015). Figure 1 shows an example. The leaf nodes are utterances, and the arcs indicate the discourse relations between utterances. Each utterance is referred as an elementary discourse unit (EDU) in SDRT discourse parsing.

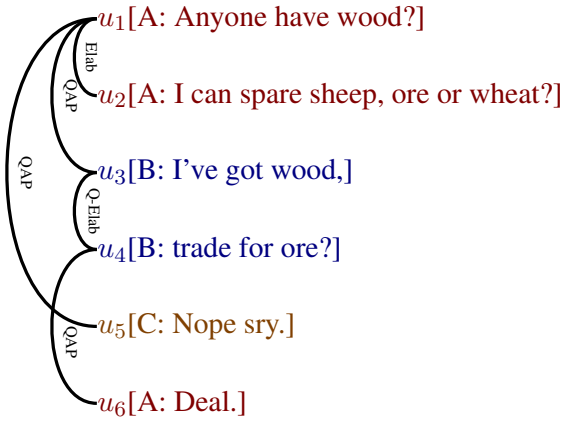


Figure 1: An example of a discourse dependency tree. u_1, u_2, u_3, u_4, u_5 refer to EDUs. "Q-Elab", "QAP", "Q-Elab", and "Elab" refer to discourse relations. "A", "B", and "C" are three speakers.

A multi-party dialogue has several aspects that make its discourse parsing more challenging than that of a written text created by one author. It involves multiple speakers who interact with each other in different roles during turn shifting and make contributions to the interactions with multiple potential threads (Afantenos et al., 2015). Therefore, in addition to conversational contexts, speaker interactions are also important cues in determining the discourse structure of a multi-party dialogue.

Most current research for discourse parsing on multi-party dialogues focuses on conversational context modeling with different methods. A pioneer study by Afantenos et al. (2015) adopts a statistical model for this task, using human-designed features extracted from conversational contexts, while an early neural research by Shi and Huang (2019) proposes a deep sequential model, using hierarchical GRUs to learn conversational contextual cues for discourse parsing. Recent research exerts more efforts on integrating rich information with context modeling and explores different techniques such as domain adaptation (Liu and Chen, 2021), edge-centric encoding (Wang et al., 2021), multi-task

*Corresponding author.
[1]https://github.com/yunan4nlp/SA-DPMD

Proceedings of the 29th International Conference on Computational Linguistics, pages 5372–5382
October 12–17, 2022.

learning (He et al., 2021), and joint model (Yang et al., 2021).

Although above approaches give competitive performances on discourse parsing on multi-party dialogues, only a few studies (Afantenos et al., 2015; Shi and Huang, 2019; Wang et al., 2021) consider speaker interactions. These studies use EDU pair features to represent speaker interactions and demonstrate that introducing speaker interactions is beneficial to this task. However, the EDU pair based speaker interaction modeling only represent whether an EDU pair is from the same speaker. The informative interaction between different speakers remains unexplored. As shown in Figure 1, the connected EDUs u_1 and u_5 are from two different speakers, but the EDU pair features will not clearly tell who said the EDUs. Since a general multi-party dialogue involves more than two speakers, the problem could be extremely serious.

To alleviate the above problem, we propose a speaker-aware model for discourse parsing on multi-party dialogues. Concretely, to handle the interactive information between the same speaker within a dialogue, we present SSP-BERT, a second-stage pre-training method based on BERT that is designed to predict whether two EDUs are from the same speaker. Based on SSP-BERT, we investigate a speaker-context interactions joint encoding (SCIJE) approach to handle the interactions between different speakers. First, we follow the node-centric based encoding approach (Shi and Huang, 2019; Liu and Chen, 2021), adopting BERT and BiGRU to represent conversational contexts. Then we embed the speaker sequence of each dialogue to vectors and feed them into BiGRU to further obtain speaker interaction representations. We finally combine them and thus obtain speaker-context interaction joint representations.

We conduct experiments on STAC (Asher et al., 2016) and Molweni (Li et al., 2020) to evaluate our proposed model. Experimental results show that SSP-BERT is highly competitive for discourse parsing on multi-party dialogues. When the speaker-context interaction joint representations are integrated, the proposed model is able to obtain further improvements. Our proposed model achieves the best performance among all the state-of-the-art (SOTA) models reported in the literature.

In summary, we mainly make the following three contributions in this paper:

- We propose SCIJE for discourse parsing on multi-party dialogues, which is capable of modeling the interactions between different speakers.
- We propose a second-stage pre-training approach to integrate the interaction features between the same speaker into conversational context representations.
- Our final model achieves the SOTA performance on two benchmark datasets.

2 Related Work

Text-level discourse parsing can be categorized into two types: the RST-style (Mann and Thompson, 1988) and the PDTB-style (Prasad et al., 2008) parsing. Both tasks have been intensively investigated since early (Lin et al., 2014; Li et al., 2014). Compared with text-level discourse parsing, discourse parsing on multi-party dialogues is still at its early stage. The pioneer study (Afantenos et al., 2015) mainly borrows the dependency parsing paradigm from RST-style parsing (Li et al., 2014) for this task, using human-designed features. Recently, inspired by the success of neural discourse parsing models (Braud et al., 2016, 2017; Yu et al., 2018), several neural discourse parsing models for multi-party dialogues have been proposed as well (Shi and Huang, 2019; He et al., 2021; Liu and Chen, 2021; Yang et al., 2021; Wang et al., 2021). In this paper, we follow the line of the work using neural models to this task.

It is believed that speaker interactions are helpful for modeling multi-party dialogues, giving great improvements on language modeling (Zhang and Zhao, 2021), dialogue comprehension (Ma et al., 2021, 2022). In discourse parsing, Afantenos et al. (2015) extract hand-crafted features from the EDU pair that have the same speaker, and feed them into a statistical discourse parsing model. Shi and Huang (2019) use a speaker highlight mechanism to represent speaker interactions. Wang et al. (2021) treat speaker interactions as edges of EDUs, feeding them into graph neural network (GNN) to obtain edge-centric representations. However, these speaker interaction models based on EDU pairs only indicate whether two EDUs are from the same speaker, ignoring the interactions between different speakers. In this paper, we investigate the interaction features between different speakers, using them as a strong supplementary for the context representations.

Recent research investigates pre-training on dia-

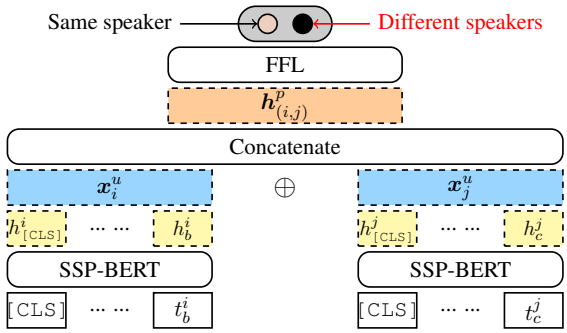


Figure 2: Framework of the SSP task.

logues intensively (Henderson et al., 2020; Zhang et al., 2020; Xu et al., 2021; Zhang and Zhao, 2021; **?**). Almost all studies focus on capturing coherance between utterances by using pre-training tasks such as dialogue generation or response selection. In this work, we enhance the conversational context representations with interaction features between the same speaker.

3 Our Proposed Model

3.1 SSP-BERT

In order to integrate same speaker interactions into contextual representations, we present SSP-BERT, a second stage pre-training method based on BERT. The approach is mainly inspired by Yu et al. (2022), which pre-train XLNet (**?**) with two EDU-level tasks in the second stage. Here we change the original approach of Yu et al. (2022) to match discourse parsing on multi-party dialogues. As shown in Figure 2, we sample the EDU pair from dialogues, and adopt SSP-BERT to predicts whether two EDUs have the same speaker. Concretely, given an EDU pair u_j and u_i, we exploit BERT to encode them respectively, obtaining corresponding token embeddings.

$$\begin{aligned} u_i &= \{\texttt{[CLS]}, t_1^i, ..., t_m^i\} \\ u_j &= \{\texttt{[CLS]}, t_1^j, ..., t_n^j\} \\ \boldsymbol{h}^i_{\texttt{[CLS]}}, \boldsymbol{h}^i_1, ..., \boldsymbol{h}^i_m &= \text{BERT}(u_i) \\ \boldsymbol{h}^j_{\texttt{[CLS]}}, \boldsymbol{h}^j_1, ..., \boldsymbol{h}^j_n &= \text{BERT}(u_j) \end{aligned} \tag{1}$$

We choose the representation of "[CLS]" as the corresponding EDU representation, and then concatenate these two EDU representations as the representation of the EDU pair:

$$\boldsymbol{h}^p_{(j,i)} = \boldsymbol{h}^i_{\texttt{[CLS]}} \oplus \boldsymbol{h}^j_{\texttt{[CLS]}} \tag{2}$$

When the EDU pair representation is ready, we feed it into a feed forward layer (FFL):

$$\boldsymbol{y}^p = \boldsymbol{W}^p \boldsymbol{h}^p_{(j,i)} \tag{3}$$

where $\boldsymbol{W}^p$ is a learnable model parameter and $\boldsymbol{y}^p$ is the output scores.

3.2 Discourse Parsing Model

Our discourse parsing model follows an encoder-decoder framework. As shown by the bottom of Figure 3, the encoder represents the speakers and the contexts to speaker-context interaction joint representations. The top of Figure 3 shows the decoder. It predicts the links and their corresponding relations between EDUs.

3.2.1 Encoder

Speaker Interaction Representation Here we introduce the approach of obtaining the speaker interaction representations. Given a dialogue with n turn, we first gather the speaker sequence with n length. For instance, we can obtain the corresponding speaker sequence $\{\text{A}, \text{A}, \text{B}, \text{B}, \text{C}, \text{A}\}$ from the dialogue in Figure 1. Then we embed the speaker sequence to the speaker vectors, and use BiGRU to encode these speaker vectors, obtaining speaker representations:

$$\begin{aligned} \boldsymbol{x}^s_{\text{A}}, ..., \boldsymbol{x}^s_{\text{C}}, \boldsymbol{x}^s_{\text{A}} &= \text{A}, ..., \text{C}, \text{A} \\ \boldsymbol{h}^s_1, ..., \boldsymbol{h}^s_{n-1}, \boldsymbol{h}^s_n &= \text{BiGRU}(\boldsymbol{x}^s_{\text{A}}, ..., \boldsymbol{x}^s_{\text{C}}, \boldsymbol{x}^s_{\text{A}}) \end{aligned} \tag{4}$$

We concatenate two speaker representations to further obtain the speaker interaction representation:

$$\boldsymbol{h}^s_{(j,i)} = \boldsymbol{h}^s_j \oplus \boldsymbol{h}^s_i \tag{5}$$

where $\oplus$ is a concatenate operation, $\boldsymbol{h}^s_{(i,j)}$ denotes the speaker interaction representation.

Context Interaction Representation We borrow the node-centric encoding approaches (Shi and Huang, 2019; Liu and Chen, 2021) to represent the conversational contexts. It consists of BERT and BiGRU. The BERT layer is used to represent sequential tokens in EDUs, and the BiGRU layer is used to represent sequential EDUs. Concretely, for each input EDU u_i, first we tokenize it by byte pair encoding (BPE) and then place a [CLS] before it. By this way, the input tokens of the first layer BERT are $\{\texttt{[CLS]}, t_1^i, ..., t_m^i\}$. Thus we adopt BERT to represent these input tokens:

$$\begin{aligned} u_i &= \texttt{[CLS]}, t_1^i, ..., t_m^i \\ \boldsymbol{h}^i_{\texttt{[CLS]}}, \boldsymbol{h}^i_1, ..., \boldsymbol{h}^i_m &= \text{BERT}(u_i) \end{aligned} \tag{6}$$

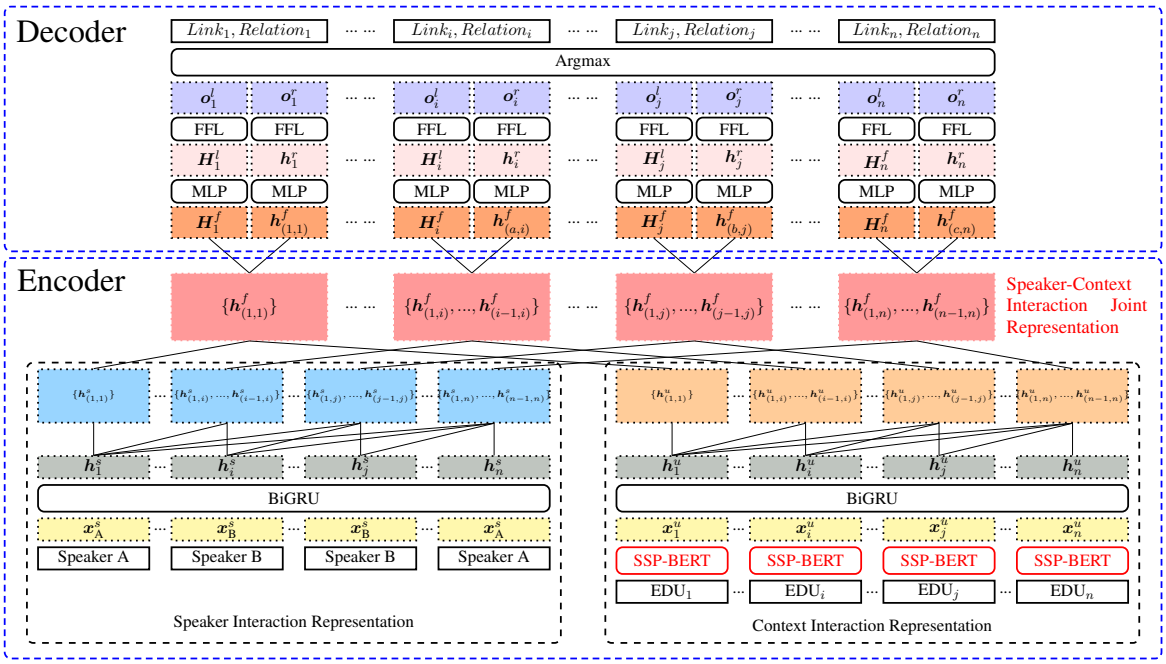


Figure 3: Framework of our proposed speaker-aware discourse parsing model.

The second layer BiGRU is built over sequential EDUs. We should first obtain a suitable representation for each EDU, which is composed of a span of tokens inside a certain EDU. Assuming an EDU u_i with its tokens by $\{[\texttt{CLS}], t_1^i, ..., t_m^i\}$, after applying the first layer BERT, we obtain their representations by $\{\boldsymbol{h}^i_{[\texttt{CLS}]}, \boldsymbol{h}^i_1 ..., \boldsymbol{h}^i_m\}$, then we select the representation of `[CLS]` as the EDU representations $\boldsymbol{x}^u$. When the EDU representations are ready, we apply the BiGRU layer, resulting:

$$\boldsymbol{h}^u_1, ..., \boldsymbol{h}^u_n = \text{BiGRU}(\boldsymbol{x}^u_1, ..., \boldsymbol{x}^u_n) \quad (7)$$

We concatenate $\boldsymbol{h}^u_i$, and $\boldsymbol{h}^u_j$ to obtain the corresponding context interaction representation.

$$\boldsymbol{h}^u_{(j,i)} = \boldsymbol{h}^u_j \oplus \boldsymbol{h}^u_i \quad (8)$$

Speaker-Context Interaction Joint Encoding When the speaker interaction and the context interaction representations are ready, we combine them jointly to obtain the speaker-context interaction joint representations.

$$\boldsymbol{h}^f_{(j,i)} = \alpha \boldsymbol{h}^s_{(j,i)} + (1-\alpha)\boldsymbol{h}^u_{(j,i)} \quad (9)$$

where the α is a learnable parameter, $\boldsymbol{h}^f_{(j,i)}$ denotes the speaker-context interaction joint representation.

3.2.2 Decoder

The decoder performs the link prediction and the relation classification. Concretely, given two EDUs u_i and u_j $(j < i)$, the link prediction task predicts whether u_j is the parent node of u_i. If u_j is the parent node of u_i, the relation classification task would further predicts the discourse relation type between u_i and u_j.

Link Prediction As mentioned before, the encoder represents u_i and u_j to corresponding speaker-context interaction joint representations. We gather the sequence of input representations of $\{(u_1, u_i), ..., (u_{(i-1)}, u_i)\}$, and thus apply a multi-layer perceptron (MLP) layer to obtain link hidden representations as inputs of the link prediction task:

$$\begin{aligned} \boldsymbol{H}_i &= \boldsymbol{h}^f_{(1,i)}, ..., \boldsymbol{h}^f_{(i-1,i)} \\ \boldsymbol{H}^l &= \tanh\left(\boldsymbol{W}^l_2 \tanh\left(\boldsymbol{W}^l_1 \boldsymbol{H}_i + \boldsymbol{b}^l_1\right) + \boldsymbol{b}^l_2\right) \end{aligned} \quad (10)$$

where $\boldsymbol{W}^l_1$, $\boldsymbol{W}^l_2$, $\boldsymbol{b}^l_1$, and $\boldsymbol{b}^l_2$ are model parameters, "tanh" is an activation function, $\boldsymbol{H}^l$ denotes the link hidden representations. Then we apply a feed-forward layer (FFL) to obtain the parent EDU scores:

$$\boldsymbol{o}^l = \boldsymbol{U}^l \boldsymbol{H}^l \quad (11)$$

where $\boldsymbol{o}^l$ is the parent EDU scores and $\boldsymbol{U}^l$ is a model parameter.

Relation Classification We also apply a MLP layer to obtain relation hidden representations:

$$\boldsymbol{h}^r = \tanh\left(\boldsymbol{W}_2^r \tanh\left(\boldsymbol{W}_1^r \boldsymbol{h}_{(a,i)}^f + \boldsymbol{b}_1^r\right) + \boldsymbol{b}_2^r\right) \tag{12}$$

where $\boldsymbol{W}_1^r$, $\boldsymbol{W}_2^r$, $\boldsymbol{b}_1^r$, and $\boldsymbol{b}_2^r$ are model parameters, $\boldsymbol{h}^r$ denotes the relation hidden representation. We also apply a FFL to obtain discourse relation scores:

$$\boldsymbol{o}^r = \boldsymbol{U}^r \boldsymbol{h}^r \tag{13}$$

where $\boldsymbol{o}^r$ is the discourse relation scores and $\boldsymbol{U}^r$ is a model parameter.

3.3 Training

Following previous studies (Shi and Huang, 2019; Wang et al., 2021), we use cross-entropy as the optimization objectives of the link prediction and the relation classification tasks. We add these two objective terms together as the final optimization objective of our discourse parser:

$$\mathcal{L}(\Theta) = -[\log\left(p_{u_g}\right) + \log\left(p_{r_g}\right)] \tag{14}$$

where p_{u_g} and p_{r_g} are probabilities of the gold parent EDU and the gold discourse relation, respectively. Θ is the set of model parameters of our discourse parser.

Given an EDU u_i, its gold parent EDU u_g, and gold discourse relation r_g, we first calculate the link and the relation outputs using Equation 11 and 13, respectively, and then apply softmax to obtain the gold parent probability $p_{u_g} = \frac{\exp\left(\boldsymbol{o}_{u_g}^l\right)}{\sum_1^j \exp(\boldsymbol{o}_{u_k}^l)}$, and the gold relation probability $p_{r_g} = \frac{\exp\left(\boldsymbol{o}_{r_g}^r\right)}{\sum_1^q \exp\left(\boldsymbol{o}_{r_k}^r\right)}$.

4 Experiment Settings

Data We evaluate our proposed model on STAC[2] (Asher et al., 2016) and Molweni[3] (Li et al., 2020). STAC has annotated 1,173 dialogues, where 1,062 for training and the remaining 111 dialogues for testing. All dialogues are collected from an online game trading corpus. To facilitate parameter tuning, we randomly select 10% of the training dialogues as a development corpus. Molweni has annotated 10,000 dialogues, where 9,000 for training, 500 for development, and the remaining 500 dialogues for testing, respectively. All dialogues are collected from the Ubuntu dialogue corpus (Lowe et al., 2015). For fair comparison, we preprocess two datasets following Shi and Huang (2019), and all experiments are conducted based on manually segmented EDUs.

We pre-train BERT on a large-scale unlabeled dialogue corpus in the second stage. It is collected from the Ubuntu dialogue corpus (Lowe et al., 2015), containing 930,000 unlabeled dialogues.

Evaluation We adopt two standard metrics to evaluate our proposed model, including `Link` and `Link&Rel` metrics. The `Link` metric evaluates the capability of link prediction only, and the `Link&Rel` metric evaluates link prediction together with discourse relations. We follow Shi and Huang (2019), reporting the micro F_1 scores.

Hyper-Parameters There are several hyper-parameters in our proposed speaker-aware discourse parsing model.

In the SSP-BERT model, we use *PyTorch* (Paszke et al., 2019) to implement our neural modules, and BERT is implemented by *Transformers* (Wolf et al., 2020). We use *bert-base-uncased*[4] to initialize the model parameters of BERT, and other model parameters are initialized randomly. We optimize model parameters by the Adam algorithm (Kingma and Ba, 2015). The learning rate of BERT is set to 1e-7 and the learning rate of the linear layer is set to 1e-3. We train our SSP-BERT by online learning with mini-batch, and the batch size is set to 8. Several key hyper-parameters are set according to the development experiments in Section 5. We randomly sample 100,000 dialogues for the SSP task with 4 epochs on Molweni, and 100,000 dialogues with 5 epochs on STAC.

In the discourse parsing model, most of hyper-parameters are same on STAC and Molweni. The hidden size of the BiGRU layer is set by 250, and the hidden size of the MLP layer is set by 1,000. The batch size is set to 8, and the maximum training interaction is set to 5. The learning rate of BERT is set differently on STAC and Molweni, 1e-5 and 2e-5 respectively. The learning rate of BiGRU, MLP, and FFL is set to 1e-3.

[2] https://www.irit.fr/STAC/corpus.html
[3] https://github.com/HIT-SCIR/Molweni
[4] https://huggingface.co/bert-base-uncased

Inputs	Dev		Test	
	Link	Link&Rel	Link	Link&Rel
Molweni				
EDUs	79.5	57.8	77.8	56.5
Texts	77.2	56.5	77.0	55.5
STAC				
EDUs	71.4	52.2	72.4	55.4
Texts	71.1	52.4	72.1	54.4

Table 1: Influence of different input methods of BERT.

5 Development Experiments

In this section, we conduct development experiments to examine the effectiveness of some important factors on our proposed model.

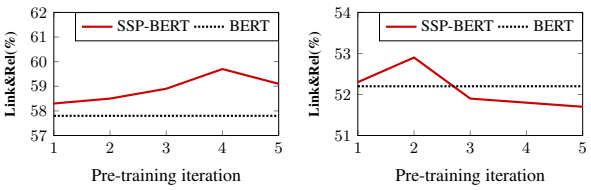
Input Methods First, we investigate the influence of different input methods of BERT. There are two different methods to encode the dialogues with BERT. The first method inputs an EDU sequence into BERT, encoding each EDU independently, It is widely used in previous studies (Shi and Huang, 2019; Liu and Chen, 2021; Yang et al., 2021). The second method treats a dialogue as a whole text, and feeds it into BERT to obtain corresponding EDU representations (He et al., 2021). Table 1 shows the comparisons. We can see that using EDUs as inputs is better than using whole texts.

Pre-Trained Language Models Then we examine how different PLMs influence the performance of our proposed model. It is believed that pre-trained language models (PLMs) are promising for discourse parsing on multi-party dialogues (Wang et al., 2021; Liu and Chen, 2021; Yang et al., 2021). As mentioned before, we use BERT to represent conversational contexts. The BERT layer can be replaced by other PLMs, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), and XLM-R (Conneau et al., 2020). Table 2 shows the development results. When we use SSP to enhance these PLMs, these discourse parsing models are able to obtain further improvements. We find that the SSP-BERT discourse parsing model achieves the best performance among these models on two development sets. Thus we use SSP-BERT in our subsequent experiments.

BiGRU vs Transformer As mentioned before, we use BiGRU to obatin EDU representations in Equation 7. Exploiting transformer (**?**) is an alternative method for obtaining EDU representation, and it may capture the longer dependence in an EDU sequence than BiGRU. Here we further investigate the influence of different EDU representations based on the BiGRU and transformer models. As shown in Table 3, we find that the BiGRU models outperform the transformer models. It may due to that the turn of dialogues in two corpora is short, and BiGRU is enough for capturing the long dependence in these dialogues.

Models	Dev		Test	
	Link	Link&Rel	Link	Link&Rel
Molweni				
BERT	79.5	57.8	77.8	56.5
ELECTRA	79.9	57.7	77.3	55.5
RoBERTa	79.9	57.3	77.4	55.2
XLM-R	79.8	57.8	76.7	54.5
SSP-B	81.6	59.1	79.1	57.7
SSP-E	80.5	58.4	78.1	55.8
SSP-R	80.3	59.0	78.9	57.0
SSP-X	80.2	58.9	78.3	56.7
STAC				
BERT	71.4	52.2	72.4	55.4
ELECTRA	70.7	50.3	72.5	55.4
RoBERTa	71.2	50.7	71.8	54.6
XLM-R	70.1	51.0	71.3	54.1
SSP-B	70.0	52.9	72.6	57.0
SSP-E	71.6	51.9	71.8	55.5
SSP-R	71.3	51.1	71.5	55.5
SSP-X	70.8	51.6	72.2	54.1

Table 2: Effect of different PLMs. "SSP-B", "SSP-E", "SSP-R", and "SSP-X" refer to "BERT", "ELECTRA", "RoBERTa", and "XLM-R" with SSP, respectively.



(a) Molweni-Dev (b) STAC-Dev

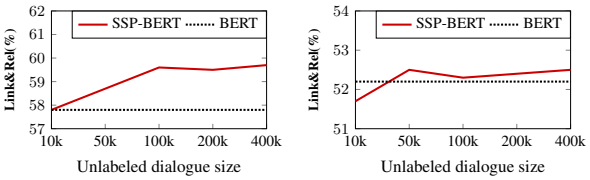
Figure 4: Influence of pre-training iteration.

Pre-Training Iteration Here we investigate the influence of training iteration in second-stage pre-training. Figure 4 shows the development performances with respect to the training iteration. On Molweni, the performance has been improving when the iteration increases from 1 to 4. However the performance does not improve when the iteration exceeds 4. The experiment over STAC shows a similar trend but the critical iteration is 2. Thus we use iteration 4 and 2 for the subsequent experiments on Molweni and STAC, respectively.

Unlabeled Dialogue Size We also study the influence of the size of unlabeled dialogues in second-stage pre-training. As shown in Figure 5, the

Models	Dev		Test	
	Link	Link&Rel	Link	Link&Rel
Molweni				
BiGRU	81.6	59.1	79.1	57.7
Transformer	80.0	57.8	78.0	56.2
STAC				
BiGRU	70.0	52.9	72.6	57.0
Transformer	71.2	52.4	70.7	53.9

Table 3: Influence of different EDU representations.



(a) Molweni-Dev (b) STAC-Dev

Figure 5: Influence of unlabeled dialogue size.

Models	Dev		Test	
	Link	Link&Rel	Link	Link&Rel
Molweni				
BERT	79.5	57.8	77.8	56.5
+STs	81.8	59.2	79.6	57.6
+GNN	84.3	59.8	83.0	58.9
+SCIJEC	83.3	59.2	82.6	58.3
+SCIJE	82.9	59.9	83.3	59.4
STAC				
BERT	71.4	52.2	72.4	55.4
+STs	71.2	52.7	72.4	56.4
+GNN	71.6	51.9	72.7	55.8
+SCIJEC	71.2	52.0	71.4	54.9
+SCIJE	72.8	53.0	73.1	56.1

Table 4: Influence of different speaker interaction representation integration methods.

Models	Link	Link&Rel
Molweni		
Li et al. (2020)	78.1	54.8
Wang et al. (2021)	81.6	58.5
Liu and Chen (2021)	80.2	56.9
He et al. (2021)*	80.0	57.0
BERT	77.8	56.5
SSP-BERT + SCIJE	**83.7**	**59.4**
STAC		
Shi and Huang (2019)	73.2	55.7
Wang et al. (2021)	73.5	57.3
Yang et al. (2021)	74.1	57.0
Liu and Chen (2021)	**75.5**	57.2
BERT	72.4	55.4
SSP-BERT + SCIJE	73.0	**57.4**

Table 5: Main results on two test sets. "*" means that we report the performance by rerunning their model.

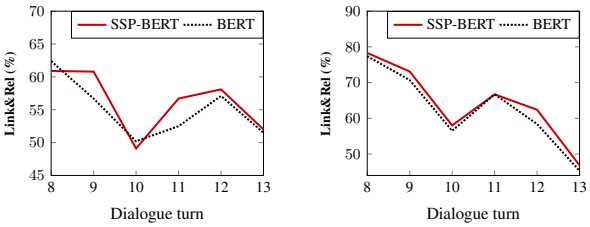
`Link&Rel` F-measure of our discourse parsing model increases apparently, when the size increases from 100k to 400k, and more unlabeled dialogues does not bring significant improvements. Thus we use 400k dialogues in second-stage pre-training.

Speaker-Context Joint Representation There are several choices for integrating speaker and context interaction representations. Here we compare three approaches with our SCIJE approach. The first approach is simple, which adds speaker tags (STs) with conversational texts as the concatenated texts and uses PLMs to model speaker interaction. In the second approach, we use a graph neural network (GNN) (Wang et al., 2021) to model speaker interaction. In the third approach, we use concatenation to replace the Equation 9. Table 4 shows the results. First, we find that the speaker interaction information is effective for discourse parsing on multi-party dialogues, which is consistent with previous observations (Afantenos et al., 2015; Shi and Huang, 2019; Wang et al., 2021). Second, the SCIJE approach is slightly better than applying GNN. Furthermore, SCIJE can achieve the best performance using a learnable model parameter, better than using concatenation (SCIJEC).

6 Main Results and Analysis

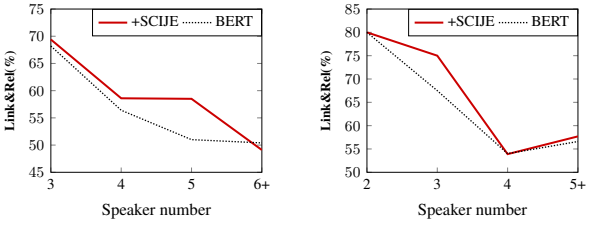
Main Results Here we report the final results of the proposed model over the Molweni and the STAC test sets. As shown in Table 5, our discourse parsing model achieves a `Link` F-measure of 77.8 and a `Link&Rel` F-measure of 56.5 on the Molweni test set, and a `Link` F-measure of 72.4 and a `Link&Rel` F-measure of 55.4 on the STAC test set. We find that the performance of our discourse parsing model on the Molweni test set outperforms most performances of previous SOTA systems. When both SSP-BERT and SCIJE are adopted, our final model achieves a `Link` F-measure of 83.7 and a `Link&Rel` F-measure of 59.4 in the Molweni test set, resulting improvements 83.7 - 77.8 = 5.9 on `Link` and 59.4 - 56.5 = 2.9 on `Link&Rel`. On STAC, our final model achieves a `Link` F-measure of 73.0 and a `Link&Rel` F-measure of 57.4, resulting improvements 73.0 - 72.4 = 0.6 on `Link` and 57.4 - 55.4 = 2.0 on `Link&Rel`.

We compare our final model with previous SOTA systems as well. Shi and Huang (2019) propose a deep sequential discourse parsing model, using local information of EDUs and global information of predicted discourse structures. Yang et al. (2021) propose a joint model for discourse parsing and dropped pronoun recovery. Liu and Chen (2021) propose a domain information enhanced dis-



(a) Molweni (b) STAC

Figure 6: `Link&Rel` against dialogue length.



(a) Molweni (b) STAC

Figure 7: `Link&Rel` against speaker number.

course parsing model. He et al. (2021)[5] propose a multi-task framework for performing discourse parsing and dialogue comprehension jointly. As shown in Table 5, we find that our proposed model achieves the SOTA performances on two benchmark datasets.

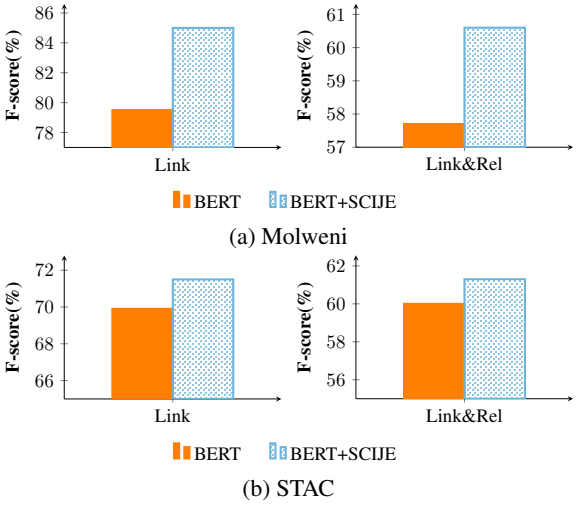
Ablation Studies Here we investigate our proposed model by ablation studies. Table 6 shows the results of ablation studies on two test sets. On Molweni, both SCIJE and SSP-BERT are effective for this task. Without SCIJE, the `Link&Rel` F-measure decreases by close to 1.7%. Without SSP-BERT, the `Link` F-measure decreases by close to 0.4%. On STAC, the results have the same tendencies. Without SCIJE, the `Link&Rel` F-measure decreases by close to 0.4%. Without SSP-BERT, the `Link&Rel` F-measure decreases by close to 1.3%. Based on above results, we find that our proposed speaker-aware model are more effective on Molweni. It may be due to that the dialogues in Molweni involve more different speakers.

Influence of Dialogue Turn As mentioned before, the SSP task predicts whether two EDUs are from the same speaker. It is able to integrate the speaker interaction features between the same

[5]It should be noted that Molweni contains two datasets, one for dialogue comprehension (100 dialogues) and other for discourse parsing (500 dialogues). He et al. (2021) only report their results on dialogue comprehension test set. For fair comparison, here we rerun their model on the discourse parsing test data .

Models	Link	Link&Rel
Molweni		
SSP-BERT + SCIJE	83.7	59.4
SSP-BERT	79.1	57.7
BERT + SCIJE	83.4	59.4
BERT	77.8	56.5
STAC		
SSP-BERT + SCIJE	73.0	57.4
SSP-BERT	72.6	57.0
BERT + SCIJE	73.1	56.1
BERT	72.4	55.4

Table 6: Ablation study on two test sets.



(a) Molweni

(b) STAC

Figure 8: Influence of SCIJE on connected EDU pairs from different speakers.

speaker into BERT. Therefore, it is expected that the introduce of SSP-BERT may bring better performance for longer dialogues. As such, here we investigate the discourse parsing model with SSP-BERT by the capability of modeling dialogue turns. Figure 6a shows the results on Molweni. The discourse parser with SSP-BERT performs better when dialogue lengths are 9, 11, and 13. It performs slightly worse when the dialogue lengths are 8 and 10. The tendency is different on STAC. As shown in Figure 6b, the discourse parser with SSP-BERT consistently outperforms the original parser for dialogues of different lengths.

Influence of Speaker Number As mentioned before, our speaker-aware model exploits a SCIJE approach to encode the speaker and the context interactions of dialogues. We believe that it is able to integrate the different speakers interaction information into the discourse parsing model. Therefore, it is expected that exploiting SCIJE may bring better performance for multi-party dialogues with more speakers. As such, here we plot `Link&Rel`

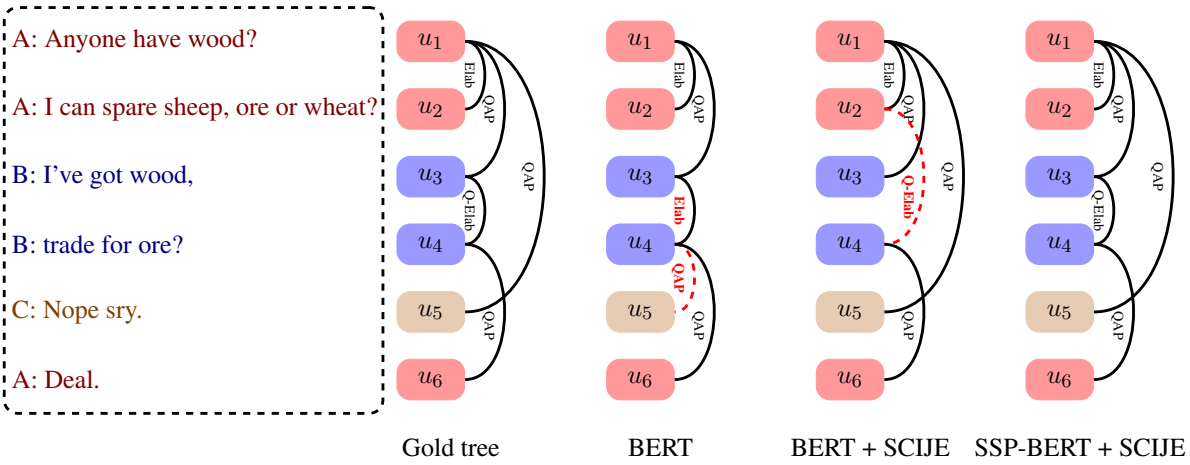


Figure 9: Case studies of the proposed speaker-aware discourse parsing model.

F-measures with respect to speaker number of dialogues. As shown in Figure 7a, we find that the discourse parsing model with speaker-context interaction joint representations performs better on dialogues with 3 to 5 speakers. The tendency is different on STAC. As shown in Figure 7b, the discourse parsing model with speaker-context interaction joint representations performs better apparently when the speaker number is 3. It may be due to that the dialogues in STAC have less connected EDU pairs from different speakers.

EDU Pairs from Different Speakers Furthermore, we investigate the performances in the connected EDU pairs from different speakers. We filter the connected EDU pairs from the same speaker, and only investigate the performances with respect to the connected EDU pairs from different speakers. As shown in Figure 8, we find that the BERT discourse parsing with SCIJE performs better for the EDU pairs from different speakers on both Molweni and STAC. The findings indicate that SCIJE could integrate the interaction information from different speakers to discourse parsing model.

Case Studies Here we present several case studies to demonstrate the advantages of the proposed speaker-aware discourse parsing model. As shown in Figure 9, the first tree is the gold tree of the dialogue, and other predicted trees are provided by the our proposed models. We find that the BERT-based parser is incapable of handling the arc from different speakers (i.e. u_1 and u_5) and the relation from the same speakers (i.e. u_3 and u_4). In the third tree, we show how the BERT-based parser benefits from SCIJE. We find that the BERT-based parser with SCIJE correctly recognizes the arc between u_1 and u_5, as SCIJE integrate the different speakers interaction information for discourse parsing. In the forth tree, we show how SSP further enhance the proposed model. We find that the final model corresponding recognizes the relation between u_3 and u_4, as SSP offers the same speaker interaction information for this task.

7 Conclusion

In this paper, we proposed a speaker-aware model for discourse parsing on multi-party dialogues, It is able to better model the speaker interactions for this task. First, we proposed SCIJE to incorporate the interaction features between the different speakers. Second, we integrated the interaction features between the same speaker to the conversational context representations by exploiting SSP-BERT. We conducted experiments and analysis on two standard benchmark datasets, namely STAC (Afantenos et al., 2015) and Molweni (Li et al., 2020). Results show that our proposed speaker-aware discourse parsing model significantly outperforms previous SOTA systems in the literature.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No. 62076173, U1836222), the High-level Entrepreneurship and Innovation Plan of Jiangsu Province (No. JSSCRC2021524), and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

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

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST Discourse Parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Conference on Computational Linguistics (CoLing)*, pages 1903 – 1913, Osaka, Japan.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*. ArXiv: 2003.10555.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-tasking Dialogue Comprehension with Discourse Parsing.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and Accurate Conversational Representations from Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *ICLR*.

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 *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive Deep Models for Discourse Parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Technical report. Publication Title: arXiv e-prints ADS Bibcode: 2019arXiv190711692L Type: article.

Zhengyuan Liu and Nancy Chen. 2021. Improving Multi-Party Dialogue Discourse Parsing via Domain Integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2021. Enhanced Speaker-aware Multi-party Multi-turn Dialogue Comprehension.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural Characterization for Dialogue Disentanglement. *arXiv:2110.08018 [cs]*. ArXiv: 2110.08018.

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. Publisher: De Gruyter Mouton Section: Text & Talk.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *In Proceedings of the 6th Language Resources and Evaluation Conference (LREC*.

Zhouxing Shi and Minlie Huang. 2019. A Deep Sequential Model for Discourse Parsing on Multi-Party Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7007–7014.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A Structure Self-Aware Model for Discourse Parsing on Multi-Party Dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3943–3949, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14158–14166. Number: 16.

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 *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 1752–1763, Online. Association for Computational Linguistics.

Ze Yang, Liran Wang, Zhoujin Tian, Wei Wu, and Zhoujun Li. 2022. TANet: Thread-Aware Pretraining for Abstractive Conversational Summarization. *arXiv:2204.04504 [cs]*. ArXiv: 2204.04504.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based Neural RST Parsing with Implicit Syntax Features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. RST Discourse Parsing with Second-Stage EDU-Level Pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. *arXiv:1911.00536 [cs]*. ArXiv: 1911.00536.

Zhuosheng Zhang and Hai Zhao. 2021. Structural Pre-training for Dialogue Comprehension. *arXiv:2105.10956 [cs]*. ArXiv: 2105.10956.