A Structure Self-Aware Model for Discourse Parsing on Multi-Party Dialogues

Ante Wang[1,2*], **Linfeng Song**[3*], **Hui Jiang**[1,2], **Shaopeng Lai**[1,2], **Junfeng Yao**[1,2], **Min Zhang**[4], **Jinsong Su**[1,2,5†]

[1]Center for Digital Media Computing and Software Engineering, School of Informatics, Xiamen University

[2]Institute of Artificial Intelligence, Xiamen University

[3]Tencent AI Lab, Bellevue, WA

[4]Institute of Artificial Intelligence, School of Computer Science and Technology, Soochow University

[5]Pengcheng Lab, Shenzhen

{wangante, hjiang, splai}@stu.xmu.edu.cn, freesunshine0316@gmail.com, {yao0010, jssu}@xmu.edu.cn, minzhang@suda.edu.cn

Abstract

Conversational discourse structures aim to describe how a dialogue is organised, thus they are helpful for dialogue understanding and response generation. This paper focuses on predicting discourse dependency structures for multi-party dialogues. Previous work adopts incremental methods that take the features from the already predicted discourse relations to help generate the next one. Although the inter-correlations among predictions are considered, we find that the error propagation is also very serious and hurts the overall performance. To alleviate error propagation, we propose a Structure Self-Aware (SSA) model, which adopts a novel edge-centric Graph Neural Network (GNN) to update the information between each Elementary Discourse Unit (EDU) pair layer by layer, so that expressive representations can be learned without historical predictions. In addition, we take auxiliary training signals (e.g. structure distillation) for better representation learning. Our model achieves the new state-of-the-art performances on two conversational discourse parsing benchmarks, largely outperforming the previous methods.

1 Introduction

As a common dialogue scenario, multi-party dialogues have lots of potential applications, attracting increasing research attentions recently. To understand multi-party dialogues, conversational discourse parsing was proposed, which aims at discovering the inter-dependencies between EDUs[1]. In this aspect, most of dominant approaches study dependency-based structures. Figure 1 shows a multi-party dialogue involving five speakers (A, B, C, D, E) and the corresponding discourse relations. We can observe that it effectively includes relations between non-adjacent utterances, such as the "*Comment*" relation between the first turn and the last turn.

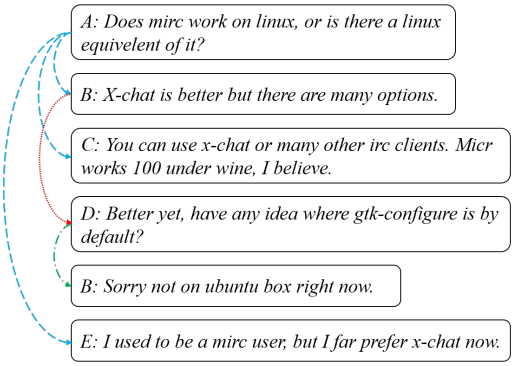


Figure 1: A multi-party dialogue from the *Molweni* [Li *et al.*, 2020] dataset with its discourse structure, where the links in slahsed blue, dotted red and slash-dotted green denote "Comment", "Clarification Question", and "Question-Answer Pair" respectively.

Initial efforts [Muller *et al.*, 2012; Li *et al.*, 2014; Afantenos *et al.*, 2015] for discourse parsing are mainly based on handcrafted features, where the decoding process is modeled in a pipeline manner. In this process, the probability of the discourse relation for each EDU pair is firstly estimated, and then a discourse structure is inferred by a search algorithm such as maximum spanning tree. Inspired by the success of deep learning on other NLP tasks, Shi and Huang [2019] proposed a neural model, i.e. *DeepSequential*, for discourse parsing on multi-party dialogues. Typically, *DeepSequential* simultaneously constructs and utilizes the discourse structure for each dialogue: it first extracts features from the already predicted discourse structure, then makes the next prediction before merging it into the partial discourse structure.

Although taking the previously predicted structure can provide richer information, *DeepSequential* is confronted with severe error propagation. Figure 2 gives the prediction accuracy of *DeepSequential* and its baseline without historical predictions at different dialogue turns. We can see that utilizing predicted structure has a negative effect on EDUs after

*Equal contribution

†Corresponding author

[1]EDUs are the fundamental discourse units in discourse parsing. Each EDU corresponds to an utterance in a dialogue.

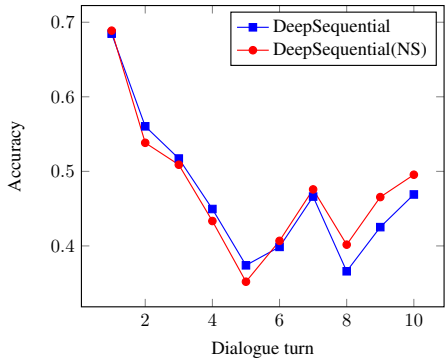


Figure 2: The prediction accuracy of *DeepSequential* [Shi and Huang, 2019] and *DeepSequential(NS)* at different dialogue turns on the *Molweni* test set. Unlike *DeepSequential*, *DeepSequential(NS)* does not use any features from already predicted discourse structure.

the 6th dialogue turn. One likely reason for the severe error propagation is that the current state-of-the-art performance is not accurate enough (less than 60% in accuracy), thus such utilization will introduce more noises than benefits.

In this work, we propose a novel edge-centric Structure Self-Aware Graph Neural Network (SSA-GNN) for discourse parsing of multi-party dialogues. With this model, we explore another direction that learns effective representations without the features from historical actions (thus no error propagation is introduced). Unlike previous work that focuses on learning EDU-specific representations, our model directly uses an edge-specific vector to capture the implicit structural information between each EDU pair. Benefiting from the message passing of Graph Neural Networks [Kipf and Welling, 2017; Marcheggiani and Titov, 2017; Velickovic *et al.*, 2018], edge-specific vectors in SSA-GNN can gradually capture implicit correlation and global information via the semantic interactions with their connected EDU nodes. As a result, our model can learn better representations using implicit structural information instead of explicit historical predictions.

To further enhance representation learning, we introduce two auxiliary loss terms (i.e. relation recognition loss and structure distillation loss) that provide orthogonal training signals into the overall objective function. The first one is calculated by conducting relation recognition at each intermediate layer of SSA-GNN. The second one is an MSE loss function for knowledge distillation [Hinton *et al.*, 2014; Zhang *et al.*, 2019]. It transfers the knowledge of a teacher model that accesses ground-truth discourse relations except for the relation which needs predicting to our model.

To summarize, our contributions in this work mainly include the following three aspects:

- We propose a novel SSA-GNN model for discourse parsing on multi-party dialogues. It directly learns the representation for each EDU pair, yielding stronger performances than previous node-centric GNN models.
- We explore relation recognition and structure distillation to further enhance the robustness of our model for learning better representations.
- Extensive experiments and analysis on two benchmarks demonstrate the effectiveness of our model.

2 Related Work

2.1 Discourse Parsing

Most previous studies for discourse parsing are based on Penn Discourse TreeBank (PDTB) [Prasad *et al.*, 2008] or Rhetorical Structure Theory Discourse TreeBank (RST-DT) [Mann and Thompson, 1988]. PDTB mainly focuses on shallow discourse relations while ignoring the overall discourse structure [Yang and Li, 2018]. As for RST, there have been many approaches including transition-based methods [Braud *et al.*, 2017; Wang *et al.*, 2017; Yu *et al.*, 2018], CYK-based approaches [Joty *et al.*, 2015; Li *et al.*, 2016; Liu and Lapata, 2017] and greedy bottom-up approach [Feng and Hirst, 2014]. However, constituency-based RST does not allow structures with crossing dependencies [Afantenos *et al.*, 2015].

To deal with this issue, other approaches [Prasad *et al.*, 2008; Li *et al.*, 2014] take dependency-based structures to represent discourse relations. The dependency-based formalism is especially prevalent on dialogues [Holmer, 2008; Perret *et al.*, 2016], where the non-adjacent can frequently occur. We follow this line of research and propose a novel edge-centric GNN with several auxiliary losses to learn better representations, which alleviates error propagation to make further improvement.

2.2 Edge-centric GNN

As one type of effective approaches for processing structural inputs, GNNs have attracted increasing attentions in recent years. Most previous GNN models [Marcheggiani and Titov, 2017; Beck *et al.*, 2018; Song *et al.*, 2019] mainly resort to learning representations for nodes. Until recently, some studies [Zhu *et al.*, 2019; Yin *et al.*, 2019; Cai and Lam, 2020] introduce edge representations. However, their edge representations are generated only from edge labels and kept constant to serve as additional inputs for enriching node representations.

Compared with previous GNNs, our edge-centric GNN has the following advantages: (*i*) it directly learns the representation for each edge, thus it can work better for problems that involve a pair of nodes as an input (e.g. discourse parsing); (*ii*) our GNN iteratively updates the edge hidden states and it allows information exchange both from node states to edge states and vice versa within each iteration. Thus, it can generate more accurate representations for both edges and nodes. To our knowledge, this is the first attempt to apply a GNN model on conversational discourse parsing.

3 Our Model

In this section, we first give a brief description about the task definition before introducing our proposed model in detail.

3.1 Problem Definition

Unlike previous work that considers this task as a "resolution" problem, we formulate it as a classification problem for each utterance pair. Given a sequence of EDUs (utterances) $x_1, x_2, ..., x_N$ from a dialogue, we aim to predict all relations $\{(x_j, x_i, l_{ji}) \mid i>j\}$ between EDU pairs, where (x_j, x_i, l_{ji}) stands for a discourse link of the relation type l_{ji} from x_j

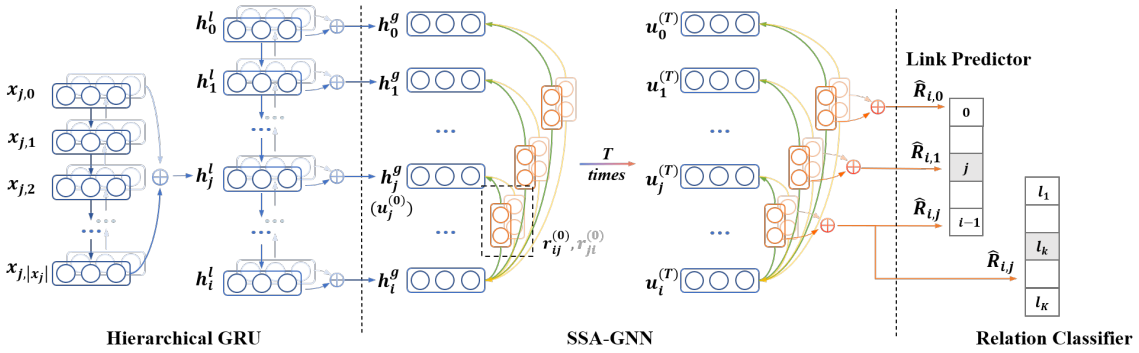


Figure 3: The architecture of our model, which includes a hierarchical GRU layer, our proposed SSA-GNN layer, a link predictor and a relation classifier. We take $(x_j \rightarrow x_i, l_k)$ as an example and only show the edges between x_i and other EDUs in SSA-GNN for clarity.

to x_i.[2] Generally, the prediction of each triple (x_j, x_i, l_{ji}) is divided into *link prediction* $P\left(x_j \rightarrow x_i \mid x_0, x_1, ..., x_i\right)$ and *relation classification* $P\left(l_{ji} \mid x_j \rightarrow x_i\right)$.

3.2 Structure Self-Aware Graph Neural Network

Figure 3 illustrates the architecture of our model. We first employ a hierarchical GRU consisting of two bidirectional GRU (BiGRU) layers to learn vector representations of EDUs. The bottom layer consumes each EDU x_i, where the last hidden states in two directions are concatenated to form the local EDU representation $\boldsymbol{h}_i^l$, and the top layer acts on $\boldsymbol{h}_0^l, \boldsymbol{h}_1^l, \ldots, \boldsymbol{h}_n^l$ to learn the global context-aware EDU representations $\mathbf{H}^g = [\boldsymbol{h}_0^g, \boldsymbol{h}_1^g, \ldots, \boldsymbol{h}_n^g]$ in a dialogue.

With the learned EDU representations, we then apply a Structure Self-Aware Graph Neural Network (SSA-GNN) to capture the implicit structural information between EDUs. The input of SSA-GNN is a fully connected graph, where each EDU or each edge connecting two EDUs is represented as a vector. The basic intuition behind our learnable edge representations is to explicitly capture and exploit the implicit structural information within the input dialogue. To initialize hidden states of SSA-GNN, we directly use $\mathbf{H}^g$ as the initial node representations $\mathbf{u}^{(0)}$. Besides, we form the initial vector representation $\boldsymbol{r}_{ij}^{(0)}$ for each EDU pair (x_j, x_i) by concatenating three learnable embeddings: $\boldsymbol{s}_{ij}$ indicating whether x_i and x_j are from the same speaker, $\boldsymbol{t}_{ij}$ meaning whether x_i and x_j are continuous utterances of the same speaker, and $\boldsymbol{d}_{ij}$ denoting the relative distance between x_i and x_j.

Afterwards, inspired by the recent work [Zhu *et al.*, 2019; Cai and Lam, 2020; Wang *et al.*, 2020], we perform Structure-Aware Scaled Dot-Product Attention operation to update node hidden states. With the t-th layer node representations (e.g. $\mathbf{u}_i^{(t)}$) and edge representations (e.g. $\mathbf{r}_{ij}^{(t)}$), we obtain the node representations $\mathbf{u}^{(t+1)}$ at the next layer as follows:

$$
\begin{aligned}
\boldsymbol{u}_i^{(t+1)} &= \sum_{j=1}^{N} \boldsymbol{\alpha}_{ij} \left(\boldsymbol{u}_j^{(t)} \boldsymbol{W}^V + \boldsymbol{r}_{ij}^{(t)} \boldsymbol{W}^F \right), \\
\boldsymbol{\alpha}_{ij} &= \frac{\exp\left(\boldsymbol{e}_{ij}\right)}{\sum_{j'=1}^{N} \exp\left(\boldsymbol{e}_{ij'}\right)}, \\
\boldsymbol{e}_{ij} &= \frac{\left(\boldsymbol{u}_i^{(t)} \boldsymbol{W}^Q \right) \left(\boldsymbol{u}_j^{(t)} \boldsymbol{W}^K + \boldsymbol{r}_{ij}^{(t)} \boldsymbol{W}^R \right)^T}{\sqrt{d_u}},
\end{aligned}
\tag{1}
$$

where $\boldsymbol{W}^*$ ($* \in \{Q, K, V, R, F\}$) are learnable model parameters[3], and d_u is the dimension of the node representations. Meanwhile, we also update edge representations, which enables our model to capture implicit structural information gradually. Specifically, we adopt a GRU-style gating mechanism to update the edge representation $\boldsymbol{r}_{ij}^{(t)}$:

$$
\begin{aligned}
\boldsymbol{\gamma}_{ij} &= \sigma([\boldsymbol{u}_i^{(t)}; \boldsymbol{u}_j^{(t)}] \boldsymbol{W}^r), \\
\boldsymbol{z}_{ij} &= \sigma([\boldsymbol{u}_i^{(t)}; \boldsymbol{u}_j^{(t)}] \boldsymbol{W}^z), \\
\tilde{\boldsymbol{r}}_{ij} &= \tanh([\boldsymbol{\gamma}_{ij} \odot \boldsymbol{r}_{ij}^{(t)}; \boldsymbol{u}_i^{(t)}; \boldsymbol{u}_j^{(t)}] \boldsymbol{W}^h), \\
\boldsymbol{r}_{ij}^{(t+1)} &= (1 - \boldsymbol{z}_{ij}) \odot \boldsymbol{r}_{ij}^{(t)} + \boldsymbol{z}_{ij} \odot \tilde{\boldsymbol{r}}_{ij},
\end{aligned}
\tag{2}
$$

where $\odot$ represents the dot-product operation, and $\boldsymbol{\gamma}_{ij}$ and $\boldsymbol{z}_{ij}$ are reset gate and update gate, respectively.

We iterate the above hidden state updating process for T times, where the top-layer hidden states are then used for conversational discourse parsing. Concretely, for each EDU x_j preceding x_i in the dialogue, we adopt $\widehat{\boldsymbol{R}}_{i,j} = [\boldsymbol{r}_{ij}^{(T)}; \boldsymbol{r}_{ji}^{(T)}]$, the concatenated vector of $\boldsymbol{r}_{ij}^{(T)}$ and $\boldsymbol{r}_{ji}^{(T)}$ to conduct link prediction and relation classification.

4 Structure Self-Aware Training

Given the training data $\mathcal{D}$, we train our model according to the following training objective:

$$
\mathcal{L}(\mathcal{D}; \theta) = \sum_{d \in \mathcal{D}} \mathcal{L}_{ce}(d; \theta) + \alpha \mathcal{L}_{cls}(d; \theta) + \beta \mathcal{L}_{skd}(d; \theta), \tag{3}
$$

[2] Following previous work, we add a dummy root x_0 to represent the beginning of a dialogue.

[3] We use $\boldsymbol{W}^*$ to denote learnable model parameters in this work.

where d is a multi-party dialogue from training corpus $\mathcal{D}$, $\mathcal{L}_{ce}(d;\theta)$ is the standard loss term of conversational discourse parsing based on cross entropy, $\mathcal{L}_{cls}(d;\theta)$ and $\mathcal{L}_{skd}(d;\theta)$ are two auxiliary loss terms on multiple granularities, both of which are used to further enhance representation learning, and α, β are hyperparameters used to balance the preference among loss terms.

The intuition behind our auxiliary losses is as follows: SSA-GNN adopts a multi-layer architecture and leverages edge hidden states to capture implicit structural information. However, without sufficient supervision information, SSA-GNN may heavily rely on its top layer to make the final predictions, resulting in extra challenge for our model to be fully trained. To deal with this issue, we augment the conventional training objective with two loss terms of *relation recognition* and *structure distillation*, aiming to guide each SSA-GNN layer effectively learn the implicit structural knowledge. In essence, $\mathcal{L}_{cls}(d;\theta)$ provides the coarse-grained structural supervision information at label level, while $\mathcal{L}_{skd}(d;\theta)$ exploits the fine-grained structural supervision information at neuron level. Thus, they have the potential to be used together to improve model training.

4.1 Discourse Parsing Loss $\mathcal{L}_{ce}(d;\theta)$

Formally, $\mathcal{L}_{ce}(d;\theta)$ is composed of the loss term $\mathcal{L}_{link}(d;\theta)$ for link prediction and the loss term $\mathcal{L}_{rel}(d;\theta)$ for relation classification:

$$\begin{aligned}\mathcal{L}_{ce}(d;\theta) &= \mathcal{L}_{link}(d;\theta) + \mathcal{L}_{rel}(d;\theta),\\ \mathcal{L}_{link}(d;\theta) &= -\sum_{i=1}^{|d|} \log P\left(x_i^* \mid \widehat{\mathbf{R}}_{i,<i}\right),\\ \mathcal{L}_{rel}(d;\theta) &= -\sum_{i=1}^{|d|} \log P\left(l_{ji}^* \mid \widehat{\boldsymbol{R}}_{i,j}, x_j = x_i^*\right),\end{aligned} \tag{4}$$

where $|d|$ indicates the EDU number of d, x_i^* and l_{ji}^* denote the gold parent and corresponding relation for x_i respectively, and $\widehat{\mathbf{R}}_{i,<i} = [\widehat{\boldsymbol{R}}_{i,0}, \widehat{\boldsymbol{R}}_{i,1}, ..., \widehat{\boldsymbol{R}}_{i,i-1}]$ denotes the relations of x_i with its previous EDUs. Note that if EDU x_i does not depend on any preceding EDU, x_i^* is the added dummy root x_0.

4.2 Relation Recognition Loss $\mathcal{L}_{cls}(d;\theta)$

This loss term is calculated by taking the edge hidden states of each intermediate SSA-GNN layer to predict the corresponding discourse relations. By using this loss term, we expect the edge hidden states of every layer can effectively capture all discourse relations. Formally, $\mathcal{L}_{cls}(d;\theta)$ is defined as

$$\mathcal{L}_{cls}(d;\theta) = -\sum_{t=1}^{T-1}\sum_{i=0}^{|d|}\sum_{j=0}^{|d|} \log P\left(l_{ji} \mid \boldsymbol{r}_{ij}^{(t)}\right), \tag{5}$$

where T is the layer number of SSA-GNN, l_{ji} is the relation type of EDU pair (x_j, x_i). In particular, if $x_j \rightarrow x_i$ does not exist in the discourse structure, we set its target label l_{ji} as "*None*". Comparing with this type of loss, the standard loss $\mathcal{L}_{ce}(d;\theta)$ has to be propagated from the last layer, and thus the supervision can be weakened throughout this process.

4.3 Structure Distillation Loss $\mathcal{L}_{skd}(d;\theta)$

This loss term is used to exploit the knowledge of a structure-aware model (i.e. teacher) for enhancing our model training. Different from our model, the teacher model takes the whole dialogue and all gold relations except for the relation being predicted as additional inputs. Inspired by previous work [Romero *et al.*, 2014], formally, we take the following Mean-Square Error (MSE) loss to reduce the distance of the edge hidden states of each intermediate layer between the teacher and our model:

$$\mathcal{L}_{skd}(d;\theta) = \sum_{t=1}^{T}\sum_{i=0}^{|d|}\sum_{j=0}^{|d|} \mathrm{MSE}\left(\boldsymbol{r}_{ij}^{(t)}\boldsymbol{W}^{(t)} \parallel \boldsymbol{r}_{ij}^{*,(t)}\right), \tag{6}$$

where $r_{ij}^{*,(t)}$ denotes the edge hidden state of (x_j, x_i) from the teacher model.

5 Experiments

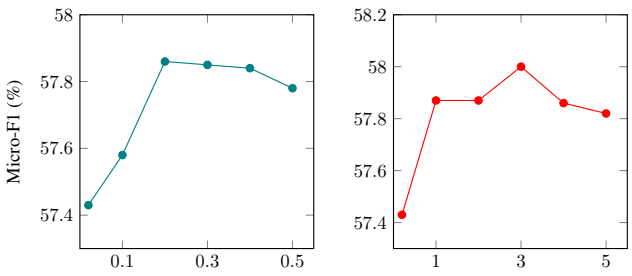
5.1 Setup

Datasets. We conduct experiments on two benchmark datasets: (*i*) **Molweni**. It is a multi-party dialogue corpus manually annotated based on Ubuntu Chat Corpus [Lowe *et al.*, 2015], which contains 9,000, 500 and 500 instances for training, development and testing, respectively. (*ii*) **STAC**. This dataset is collected from an online game. It is much smaller than *Molweni* and only contains 1,062 and 111 dialogues for training and testing, respectively. We preprocess datasets following Shi and Huang [2019].

Settings. For experiments on *STAC*, we follow previous work to represent words with 100-dimensional GloVe embeddings [Pennington *et al.*, 2014] that are fine-tuned during training. We adopt a 3-layer SSA-GNN module with 4 heads, where all layers share the same parameters. The dimensions for the edge and the node states in SSA-GNN are 128 and 256, respectively. We set the dropout rate to 0.5 and employ Stochastic Gradient Descent to train all models with batch size and initial learning rate set to 40 and 0.1, respectively. For the *Molweni* corpus, we use 200-dimensional pretrained GloVe embeddings to initialize word vectors, since it is much larger than *STAC*. Besides, the batch size is set to 100 for more stable training. For fair comparison, we use the same settings as ours for *DeepSequential* and its variants on both datasets. Performance of using either $\mathcal{L}_{cls}(d;\theta)$ or $\mathcal{L}_{skd}(d;\theta)$ with different coefficients are showed in Figure 4. We set our coefficients α and β to 0.2 and 3, respectively. For the experiments using pretrained model, we apply *ELECTRA-small* [Clark *et al.*, 2020] that has been proved effective on several tasks with a small model size. In this work, micro-averaged F_1 score is adopted for evaluation. Our code is available at https://github.com/DeepLearnXMU/Structure-Self-Aware

Baselines. We compare ours with the following baselines:

- **DeepSequential** [Shi and Huang, 2019]: It adopts an incremental predicting method. Note that it is confronted with error propagation as discussed above.
- **DeepSequential(NS)**: It is a variant of *DeepSequential* that does not use any features from already predicted discourse structures.



(a) relation recognition (b) structure distillation

Figure 4: Experimental results on the *Molweni* development dataset.

- **DeepSequential(Share)**: It is a variant of *DeepSequential* that shares the parameters of link prediction and relation classification modules (e.g. embedding layer and hierarchical GRU) except the prediction layer.
- **HGRU**: It only adopts the hierarchical GRU for representations learning and possesses the same structure as *DeepSequential(NS)*.

Besides, we also conduct experiments with pretrained language model (i.e. *ELECTRA-small*) to further verify the effectiveness of our model.

5.2 Main Results

Table 1 lists the test results on *Molweni* and *STAC*. Here, we can draw the following conclusions. **First**, *DeepSequential* which adopts incremental predicting method does not outperform *DeepSequential(NS)* on *Molweni*. This indicates that the severe error propagation causally hurts the model performance. **Second**, different from the discovery by Shi and Huang [2019], we find that *DeepSequential(share)* using less parameters is comparable with *DeepSequential*. **Third**, our model outperforms all baselines on *Molweni* and *STAC* (bootstrap test, $p < 0.01$), demonstrating its effectiveness and robustness. **Fourth**, with the better local representations generated by the pretrained model, *Our+ELECTRA* can be further enhanced. The improvement mainly comes from the gains of relation classification accuaries, which shows that token-level information is very important for capturing the knowledge about relation types.

5.3 Ablation Study

To evaluate the effectiveness of different components in our model, we compare ours with the following variants:

- **HGRU+SSA(FixEdgeRep)**: It is a simplified node-centric SSA-GNN, where edge representations of each layer are always set as the initial vectors $\mathbf{r}^{(0)}$.
- **HGRU+SSA(NodeRep)**: It uses the concatenated node representations to predict discourse structures.
- **HGRU+SSA(ShareEdgeRep)**: In this varient, edge representations r_{ij} and r_{ji} of each SSA-GNN layer share the same vector. For fair comparison, we extend its edge hidden size to 256.

[4]https://github.com/shizhouxing/DialogueDiscourseParsing

Model	Molweni		STAC	
	Link	Link&Rel	Link	Link&Rel
DeepSequential	0.7694	0.5349	0.7199	0.5362
DeepSequential(NS)	0.7657	0.5360	0.7074	0.5280
DeepSequential(Share)	0.7680	0.5403	0.7158	0.5377
HGRU	0.7623	0.5336	0.7145	0.5258
Our	0.8142	0.5689	**0.7379**	0.5513
HGRU+ELECTRA	0.7672	0.5531	0.7068	0.5386
Our+ELECTRA	**0.8163**	**0.5854**	0.7348	**0.5731**

Table 1: Main test results (F_1 scores). *Link* shows the performance regarding link prediction only, and *Link&Rel*, the ***main*** metric, indicates the performance when both link and relation are correctly predicted at the same time. We reproduce the scores of *DeepSequential* and its variants using their released code[4].

Model	Link	Link&Rel
HGRU	0.7623	0.5336
HGRU+SSA(FixEdgeRep)	0.8038	0.5581
HGRU+SSA(NodeRep)	0.8069	0.5591
HGRU+SSA	0.8102	0.5631
HGRU+SSA+$\mathcal{L}_{cls}$	0.8138	0.5665
HGRU+SSA+$\mathcal{L}_{skd}$	0.8136	0.5673
HGRU+SSA+$\mathcal{L}_{cls}$+$\mathcal{L}_{skd}$	0.8142	0.5689
HGRU+SSA(ShareEdgeRep) +$\mathcal{L}_{cls}$+$\mathcal{L}_{skd}$	0.8106	0.5665
HGRU+SSA(Teacher)	0.8331	0.5950

Table 2: Ablation study on the test dataset of *Molweni*.

- **HGRU+SSA+$\mathcal{L}_*$**: It denotes a kind of models trained with our proposed auxiliary loss terms. The following variants are considered: *HGRU+SSA+*$\mathcal{L}_{cls}$, *HGRU+SSA+*$\mathcal{L}_{skd}$ and *HGRU+SSA+*$\mathcal{L}_{cls}$*+*$\mathcal{L}_{skd}$.
- **HGRU+SSA(Teacher)**: It is the teacher model used to guide the training of our model via structure distillation loss term.

Table 2 reports the ablation experimental results on *Molweni*, where we have the following observations. **First**, the extra information learned by our SSA-GNN module is very important for context-aware representations learning. Compared with *HGRU*, *HGRU+SSA(FixEdgeRep)* achieves better performance, because our SSA-GNN module helps extract better dialogue features. **Second**, predicting with edge representations directly can reach better overall performance than that with node representations. Comparing *HGRU+SSA* with *HGRU+SSA(NodeRep)*, both *Link* and *Link&Rel* scores are improved, which indicates the gain of using edge-centric representations. **Third**, both $\mathcal{L}_{cls}$ and $\mathcal{L}_{skd}$ can improve the performance of our model. The results of +$\mathcal{L}_{skd}$ are better than those of +$\mathcal{L}_{cls}$ since the distillation loss ($\mathcal{L}_{skd}$) provides more fine-grained (neuron-level) supervision than the label-level classification loss ($\mathcal{L}_{cls}$). Nevertheless, training with $\mathcal{L}_{cls}$ and $\mathcal{L}_{skd}$ can give further improvements, indicating that they can still provide complementary information. **Fourth**, by comparing *HGRU+SSA* with *HGRU+SSA(ShareEdgeRep)*, we

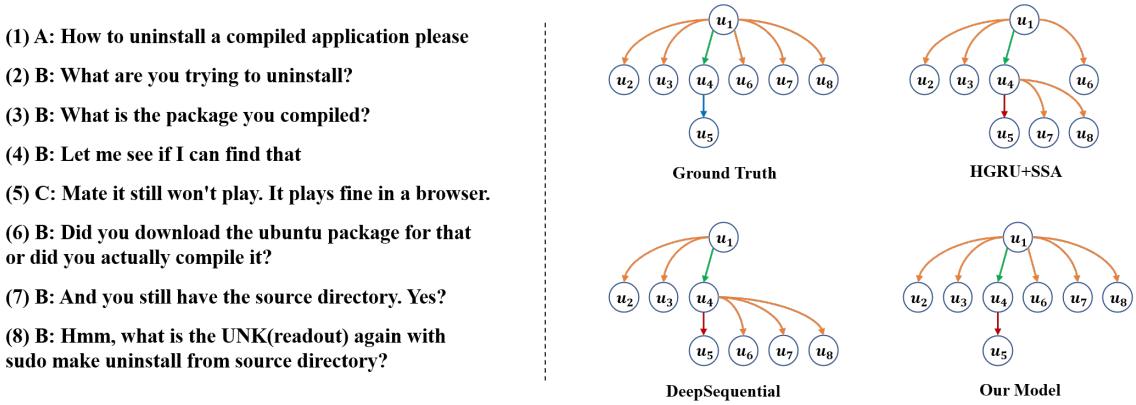


Figure 5: An example involving three speakers (A,B,C) from the *Molweni* corpus. Different relation types are showed in different colors, where orange denotes "Clarification Question", green is "Question-Answer Pair", blue is "Result" and red is "Comment".

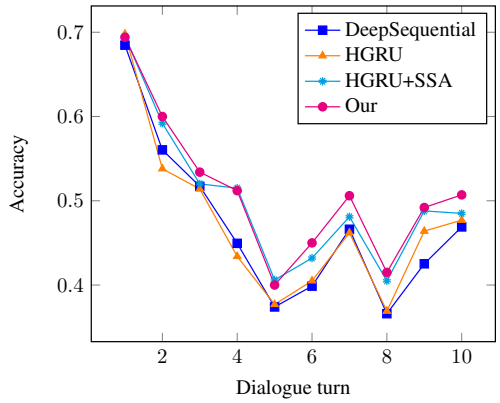


Figure 6: Comparison of prediction accuracy between typical models at different dialogue turns.

find that assigning distinct hidden states for opposite directions (e.g. r_{ij} and r_{ji}) yields better performance than using the same states for both directions. For this result, we speculate that the former learns more flexible representations, and thus it can better fit the asymmetric fact of each utterance pair on this task.

5.4 Case Study

As shown in Figure 5, given a dialogue example from three speakers, different discourse structures are predicted by these models, where we have the following observations. **First**, *DeepSequential* is based on RNN, therefore, it is difficult to correctly predict the long-distance dependencies such as $u_1 \rightarrow u_6$. Besides, it tends to be confronted with the problem of error propagation, which results in errors on u_7, u_8 given $u_4 \rightarrow u_6$. **Second**, *HGRU+SSA* allows more efficiently semantic interaction and does not suffer from error propagation. As a result, it correctly predicts u_6, but still fails on u_7 and u_8. It indicates that this task is still very challenging. **Third**, with both auxillary losses, our model is able to correctly predict both u_7 and u_8 due to its better representation learning for intermediate layers.

Surprisingly, all models fail to predict the correct relation type of $u_4 \rightarrow u_5$. This is because that u_5 seems to be confusing and not related to the given context. In fact, this phenomenon is common in the actual conversation scenario. Therefore, the application of conversational discourse parsing still faces challenges.

5.5 Accuracy at Different Dialogue Turns

Since the prediction of each EDU may depend on any preceding one, the accuracy may decrease as the dialogue turn increases. Hence, we investigate the accuracy of our model and other system at different dialogue turns, which is shown in Figure 6. Overall, all models have a similar downward trend. *DeepSequential* exhibits the worst performance especially after the 6th dialogue turn because of error propagation. In comparison to *HGRU*, *HGRU+SSA* achieves better performance in most cases, demonstrating the advantage of directly learning edge-specific vectors for EDU pairs and utilizing important dialogue features. Particularly, when introducing the two auxiliary loss terms, the accuracy of our model is further improved for later dialogue turns.

6 Conclusions

In this paper, we propose a Structure Self-Aware model for conversational discourse parsing. Particularly, it adopts an edge-centric GNN to directly learn the implicit structural information between each EDU pair. Besides, we explore two effective auxiliary losses for relation recognition and structure distillation to enhance representation learning. Compared with previous models, ours avoids the serious defect of error propagation, but also makes better use of the structural information in training data. In the future, we plan to continuously refine our model by enhancing its robustness for domain transfer.

Acknowledgements

The project was supported by National Natural Science Foundation of China (No. 62036004, No. 61672440), Natural Science Foundation of Fujian Province of China (No. 2020J06001), Youth Innovation Fund of Xiamen (Grant No. 3502Z20206059), and the Fundamental Research Funds for the Central Universities (Grant No. ZK20720200077)

References

[Afantenos *et al.*, 2015] Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. Discourse parsing for multi-party chat dialogues. In *EMNLP*, 2015.

[Beck *et al.*, 2018] Daniel Beck, Gholamreza Haffari, and Trevor Cohn. Graph-to-sequence learning using gated graph neural networks. In *ACL*, 2018.

[Braud *et al.*, 2017] Chloé Braud, Maximin Coavoux, and Anders Søgaard. Cross-lingual rst discourse parsing. In *EACL*, 2017.

[Cai and Lam, 2020] Deng Cai and Wai Lam. Graph transformer for graph-to-sequence learning. In *AAAI*, 2020.

[Clark *et al.*, 2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020.

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

[Hinton *et al.*, 2014] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2014.

[Holmer, 2008] Torsten Holmer. Discourse structure analysis of chat communication. *Language@ Internet*, 2008.

[Joty *et al.*, 2015] Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 2015.

[Kipf and Welling, 2017] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[Li *et al.*, 2014] Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. Text-level discourse dependency parsing. In *ACL*, 2014.

[Li *et al.*, 2016] Qi Li, Tianshi Li, and Baobao Chang. Discourse parsing with attention-based hierarchical neural networks. In *EMNLP*, 2016.

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

[Liu and Lapata, 2017] Yang Liu and Mirella Lapata. Learning contextually informed representations for linear-time discourse parsing. In *EMNLP*, 2017.

[Lowe *et al.*, 2015] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 2015.

[Mann and Thompson, 1988] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 1988.

[Marcheggiani and Titov, 2017] Diego Marcheggiani and Ivan Titov. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP*, 2017.

[Muller *et al.*, 2012] Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. Constrained decoding for text-level discourse parsing. In *COLING*, 2012.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *EMNLP*, 2014.

[Perret *et al.*, 2016] Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. Integer linear programming for discourse parsing. In *NAACL-HLT*, 2016.

[Prasad *et al.*, 2008] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. The penn discourse treebank 2.0. In *LREC*, 2008.

[Romero *et al.*, 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2014.

[Shi and Huang, 2019] Zhouxing Shi and Minlie Huang. A deep sequential model for discourse parsing on multi-party dialogues. In *AAAI*, 2019.

[Song *et al.*, 2019] Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. Semantic neural machine translation using amr. *TACL*, 2019.

[Velickovic *et al.*, 2018] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[Wang *et al.*, 2017] Yizhong Wang, Sujian Li, and Houfeng Wang. A two-stage parsing method for text-level discourse analysis. In *ACL*, 2017.

[Wang *et al.*, 2020] Tianming Wang, Xiaojun Wan, and Hanqi Jin. AMR-to-text generation with graph transformer. *TACL*, 2020.

[Yang and Li, 2018] An Yang and Sujian Li. Scidtb: Discourse dependency treebank for scientific abstracts. In *ACL*, 2018.

[Yin *et al.*, 2019] Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. Graph-based neural sentence ordering. In *IJCAI*, 2019.

[Yu *et al.*, 2018] Nan Yu, Meishan Zhang, and Guohong Fu. Transition-based neural rst parsing with implicit syntax features. In *COLING*, 2018.

[Zhang *et al.*, 2019] Biao Zhang, Deyi Xiong, Jinsong Su, and Jiebo Luo. Future-aware knowledge distillation for neural machine translation. *TASLP*, 2019.

[Zhu *et al.*, 2019] Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. Modeling graph structure in transformer for better AMR-to-text generation. In *EMNLP-IJCNLP*, 2019.