A Joint Model for Dropped Pronoun Recovery and Conversational Discourse Parsing in Chinese Conversational Speech

Jingxuan Yang[1], **Kerui Xu**[1], **Jun Xu**[2,3*], **Si Li**[1], **Sheng Gao**[1], **Jun Guo**[1], **Nianwen Xue**[4] **and Ji-Rong Wen**[2,3]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Gaoling School of Artificial Intelligence, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
[4]Department of Computer Science, Brandeis University

{yjx, xukerui, lisi, gaosheng, guojun}@bupt.edu.cn
junxu@ruc.edu.cn, jirong.wen@gmail.com, xuen@brandeis.edu

Abstract

In this paper, we present a neural model for joint dropped pronoun recovery (DPR) and conversational discourse parsing (CDP) in Chinese conversational speech. We show that DPR and CDP are closely related, and a joint model benefits both tasks. We refer to our model as DiscProReco, and it first encodes the tokens in each utterance in a conversation with a directed Graph Convolutional Network (GCN). The token states for an utterance are then aggregated to produce a single state for each utterance. The utterance states are then fed into a biaffine classifier to construct a conversational discourse graph. A second (multi-relational) GCN is then applied to the utterance states to produce a discourse relation-augmented representation for the utterances, which are then fused together with token states in each utterance as input to a dropped pronoun recovery layer. The joint model is trained and evaluated on a new Structure Parsing-enhanced Dropped Pronoun Recovery (SPDPR) dataset that we annotated with both two types of information. Experimental results on the SPDPR dataset and other benchmarks show that DiscProReco significantly outperforms the state-of-the-art baselines of both tasks.

1 Introduction

Pronouns are often dropped in Chinese conversations as the identity of the pronoun can be inferred from the context (Kim, 2000; Yang et al., 2015) without causing the sentence to be incomprehensible. The task of dropped pronoun recovery (DPR) aims to locate the position of the dropped pronoun and identify its type. Conversational discourse parsing (CDP) is another important task that aims to analyze the discourse relations among utterances

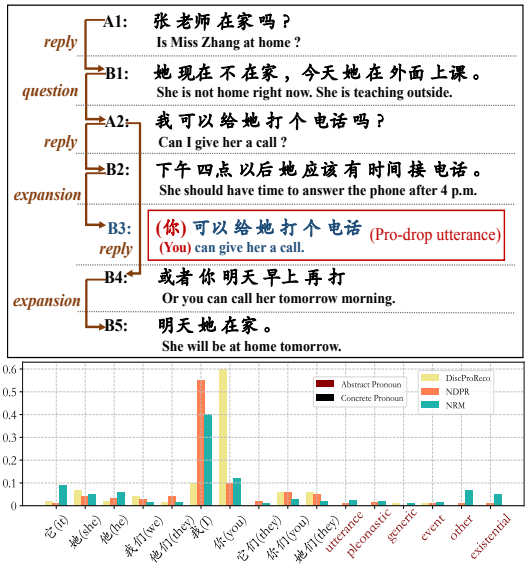

Figure 1: **Top:** A conversation snippet in which the dropped pronoun is shown in bracket. **Bottom:** Pronoun recovery results by two baselines and the proposed DiscProReco. Baselines which ignore the relation "(B3 expands B2) replies A2" mistakenly recover the dropped pronoun 你(you) as 我(I) since the utterance B_3 is considered semantically similar to A_2.

in a conversation, and plays a vital role in understanding multi-turn conversations.

Existing work regards DPR and CDP as two independent tasks and tackles them separately. As an early attempt of DPR, Yang et al. (2015) employ a Maximum Entropy classifier to predict the position and type of dropped pronouns. Zhang et al. (2019) and Yang et al. (2019) attempt to recover the dropped pronouns by modeling the referents with deep neural networks. More recently, Yang et al. (2020) attempt to jointly predict all dropped pronouns in a conversation snippet by modeling dependencies between pronouns with general conditional random fields. A major shortcoming of these DPR methods is that they overlook the discourse relation *(e.g., reply, question)* between conversa-

* Corresponding author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1752–1763
August 1–6, 2021.

tional utterances when exploiting the context of the dropped pronoun. At the same time, previous CDP methods (Li et al., 2014; Afantenos et al., 2015; Shi and Huang, 2019) first predict the relation for each utterance pair and then construct the discourse structure for the conversation with a decoding algorithm. The effectiveness of these methods are compromised since the utterances might be incomplete when they have dropped pronouns.

To overcome these shortcomings, we propose a novel neural model called DiscProReco to perform DPR and CDP jointly. Figure 1 is a Chinese conversation snippet between two speakers A and B that illustrates the advantages of such a joint approach. In this example, a pronoun "你 (you)" is dropped in utterance B_3. It is critical for the DPR model to know that both utterances B_2 and B_3 are in *reply* to the utterance A_2, when recovering this dropped pronoun. Methods which ignore the structure ("(B3 expands B2) replies A2") will more likely consider the utterance B_3 to be semantically similar to A_2, and wrongly recover the pronoun as "我 (I)".

Given a pro-drop utterance and its context, DiscProReco parses the discourse structure of the conversation and recovers the dropped pronouns in the utterance in four steps: (i) Each utterance is parsed into its dependency structure and fed into a directed GCN to output the syntactic token states. The utterance state is then obtained by aggregating the token states in the utterance. (ii) The utterance states of a conversation are fed into a biaffine classifier to predict the discourse relation between each utterance pair and the discourse structure of the conversation is constructed. (iii) Taking the discourse structure as input, another (multi-relational) GCN updates the utterance states and fuses them into the token states for each utterance to produce discourse-aware token representations. (iv) Based on the discourse structure-aware context representation, a pronoun recovery module is designed to recover the dropped pronouns in the utterances. When training this model, all components are jointly optimized by parameter sharing so that CDP and DPR can benefit each other. As there is no public dataset annotated with both dropped pronouns and conversational discourse structures, we also construct Structure Parsing-enhanced Dropped Pronoun Recovery (SPDPR) corpus, which is the first corpus annotated with both types of information. Experimental results show that DiscProReco outperforms all baselines of CDP and DPR.

Contributions: This work makes the following contributions: (i) We propose a unified framework DiscProReco to jointly perform CDP and DPR, and show that these two tasks can benefit each other. (ii) We construct a new large-scale dataset SPDPR (Section 4) which supports fair comparison across different methods and facilitates future research on both DPR and CDP. (iii) We present experimental results which show that DiscProReco with its joint learning mechanism realizes knowledge sharing between its CDP and DPR components and results in improvements for both tasks (Section 5). The code and SPDPR dataset is available at https://github.com/ningningyang/DiscProReco.

2 Problem Formulation

We first introduce the problem formulation of these two tasks. Following the practices in (Yang et al., 2015, 2019, 2020), we formulate DPR as a sequence labeling problem. DPR aims to recover the dropped pronouns in an utterance by assigning one of 17 labels to each token that indicates the type of pronoun that is dropped before the token (Yang et al., 2015). CDP is the task of constructing the conversational discourse structure by predicting the discourse relation (Xue et al., 2016) among utterances. The discourse relations may characterize one utterance as agreeing with, responding to, or indicate understanding of another utterance in the conversational context.

Let us denote an input pro-drop utterance of n tokens as $\mathrm{X} = (w_1, w_2, \cdots, w_n)$, and its contextual utterances as $\mathrm{C} = (\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_m)$ where the i-th contextual utterance X_i is a sequence of l_i tokens: $\mathrm{X}_i = (w_{i,1}, \cdots, w_{i,l_i})$. Our task aims to (1) model the distribution $\mathrm{P}(\mathrm{X_j}|\mathrm{X_i}, \mathrm{C})$ to predict the relation between each pair of utterances (i.e., $(\mathrm{X}_i, \mathrm{X}_j)$) for CDP, and (2) model $\hat{\mathrm{Y}} = \arg\max_{\mathrm{Y}} \mathrm{P}(\mathrm{Y}|\mathrm{X}, \mathrm{C})$ to predict the recovered pronoun sequence $\hat{\mathrm{Y}}$ for the input utterance X. Each element of $\hat{\mathrm{Y}}$ is chosen from one of the T possible labels from $\mathcal{Y} = \{y_1, \cdots, y_{\mathrm{T}-1}\} \cup \{\mathrm{None}\}$ to indicate whether a pronoun is dropped before the corresponding token in utterance X and the type of the dropped pronoun. The label "None" means no pronoun is dropped before this token.

3 The DiscProReco Framework

3.1 Model Overview

The architecture of DiscProReco is illustrated in Figure 2. Given a pro-drop utterance X and its

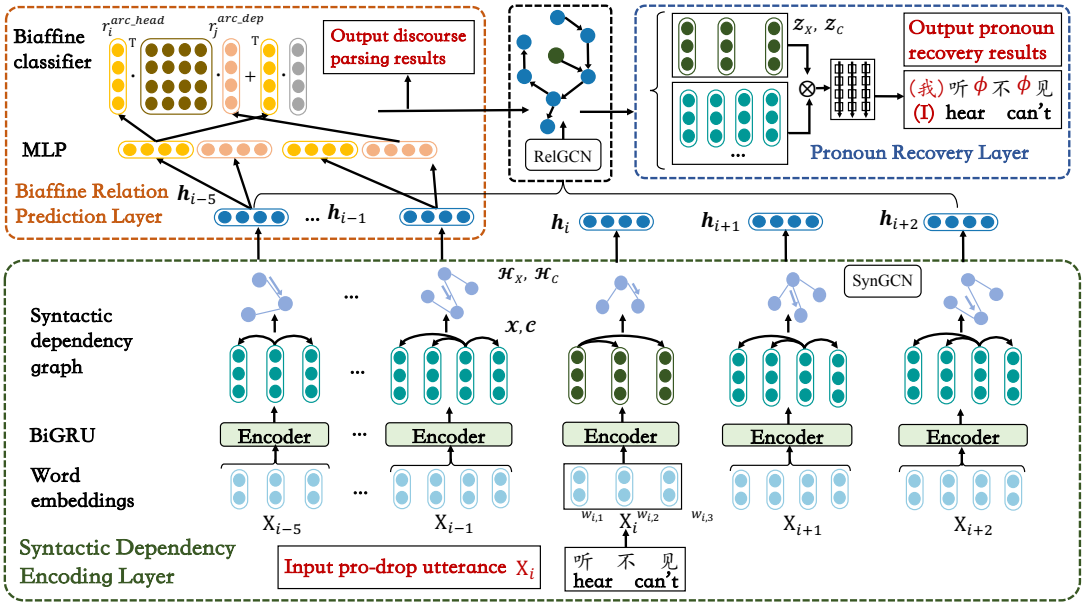

Figure 2: Overview of DiscProReco, which explores conversational discourse structures to learn effective referent representations that are used to recover dropped pronouns. DiscProReco consists of four components, and the details are introduced in Section 3.

context C, DiscProReco first represents tokens of these utterances as d-dimensional pre-trained word embeddings (Li et al., 2018), and then feed them into a BiGRU (Chung et al., 2014) network, to represent sequential token states $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{\mathcal{C}} \in \mathbb{R}^{m \times l_m \times d}$ as the concatenation of forward and backward hidden states outputted from BiGRU. The syntactic dependency encoding layer then revises the sequential token states by exploiting the syntactic dependencies between tokens in the same utterance using a directed GCN and generates utterance representations. After that, the biaffine relation prediction layer predicts the relation between each pair of utterances. The discourse structure then is constructed based on the utterance nodes and the predicted relations. The discourse structure encoding layer further encodes the inter-utterance discourse structures with a multi-relational GCN, and employs the discourse-based utterance representations to revise the syntactic token states. Finally, the pronoun recovery layer explores the referent semantics from the context C and predicts the dropped pronouns in each utterance.

3.2 Syntactic Dependency Encoding Layer

As the sequential token states overlook long-distance dependencies among tokens in a utterance, this layer takes in the sequential token states $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{C}}$, and revises them as syntactic token states as $\boldsymbol{\mathcal{H}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{C}}$ by exploring the syntactic dependencies between the tokens based on a directed GCN.

Specifically, for each input utterance in X and C, we first extract syntactic dependencies between the tokens with Stanford's Stanza dependency parser (Qi et al., 2020). Using the output of the dependency parser, we construct a syntactic dependency graph for each utterance in which the nodes represents the tokens and the edges correspond to the extracted syntactic dependencies between the tokens. Following the practices of (Marcheggiani and Titov, 2017; Vashishth et al., 2018), three types of edges are defined in the graph. The node states are initialized by the sequential token states $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{C}}$, and then message passing is performed over the constructed graph using the directed GCN (Kipf and Welling, 2017), referred to as **SynGCN**. The syntactic dependency representation of token $w_{i,n}$ after $(k+1)$-th GCN layer is defined as:

$$\boldsymbol{h}_{w_{i,n}}^{k+1} = \mathrm{ReLU}\left(\sum_{u \in \mathcal{N}_+(w_{i,n})} g_e^k \cdot \left(\mathbf{W}_e^k \boldsymbol{h}_u^k + \boldsymbol{b}_e^k\right)\right),$$

where $\mathbf{W}_e^k \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_e^k \in \mathbb{R}^d$ are the edge-specific parameters, $\mathcal{N}_+(w_{i,n}) = \mathcal{N}(w_{i,n}) \cup \{w_{i,n}\}$ is the set of $w_{i,n}$'s neighbors including itself, and $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$ is the Rectified Linear Unit. g_e^k is an edge-wise gating mechanism which incorporates the edge importance as:

$$g_e^k = \sigma\left(\hat{\boldsymbol{w}}_e^k \boldsymbol{h}_u^k + \hat{b}_e^k\right),$$

where $\hat{\boldsymbol{w}}_e^k \in \mathbb{R}^{1 \times d}$ and $\hat{b}_e^k \in \mathbb{R}$ are independent

trainable parameters for each layer, and $\sigma(\cdot)$ is the sigmoid function. The revised syntactic token states $\boldsymbol{\mathcal{H}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{C}}$ of the pro-drop utterance and context are outputted for subsequent discourse structure prediction and pronoun recovery.

3.3 Biaffine Relation Prediction Layer

For conversational discourse parsing, we jointly predict the arc $s_{i,j}^{(\mathrm{arc})}$ and relation $\mathbf{s}_{i,j}^{(\mathrm{rel})}$ between each pair of utterances utilizing the biaffine attention mechanism proposed in (Dozat and Manning, 2017). Given the syntactic token states $\boldsymbol{\mathcal{H}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{C}}$, we make an average aggregation on these token states of each utterance X_i to obtain the syntactic utterance representation $\boldsymbol{h}_{\mathrm{X}_i}$.

For a pair of utterances $(\mathrm{X}_i, \mathrm{X}_j)$ in the conversation snippet, we feed the representations of these two utterances into a biaffine function to predict the probability of an arc from X_i to X_j as:

$$\boldsymbol{r}_i^{(\mathrm{arc\_head})} = \mathrm{MLP}^{(\mathrm{arc\_head})}(\boldsymbol{h}_{\mathrm{X}_i}),$$

$$\boldsymbol{r}_j^{(\mathrm{arc\_dep})} = \mathrm{MLP}^{(\mathrm{arc\_dep})}(\boldsymbol{h}_{\mathrm{X}_j}),$$

$$s_{i,j}^{(\mathrm{arc})} = \boldsymbol{r}_i^{(\mathrm{arc\_head})}\mathbf{U}^{(\mathrm{arc})}\boldsymbol{r}_j^{(\mathrm{arc\_dep})} + \boldsymbol{r}_i^{(\mathrm{arc\_head})^T}\boldsymbol{u}^{(\mathrm{arc})},$$

where MLP is the multi-layer perceptron that transforms the original utterance representation $\boldsymbol{h}_{\mathrm{X}_i}$ and $\boldsymbol{h}_{\mathrm{X}_j}$ into head or dependent-specific utterance states $\boldsymbol{r}_i^{(\mathrm{arc\_head})}$ and $\boldsymbol{r}_j^{(\mathrm{arc\_dep})}$. $\mathbf{U}^{(\mathrm{arc})}$ and $\mathbf{u}^{(\mathrm{arc})}$ are weight matrix and bias term used to determine the probability of a arc.

One distinctive characteristics of conversational discourse parsing is that the head of each dependent utterance must be chosen from the utterances before the dependent utterance. Thus we add an upper triangular mask operation on the results of arc prediction to regularize the predicted arc head:

$$\mathbf{s}^{(\mathrm{arc})} = \mathrm{mask}(\mathbf{s}^{(\mathrm{arc})}).$$

We minimize the cross-entropy of gold head-dependent pair of utterances as:

$$\mathrm{loss}_{\mathrm{arc}} = -\sum_{j=1}^{m} \delta(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C}) \log(\mathrm{P}_{\mathrm{arc}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C})),$$

$$\mathrm{P}_{\mathrm{arc}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C}) = \mathrm{softmax}(\mathbf{s}_i^{(\mathrm{arc})}).$$

After obtaining the predicted directed unlabeled arc between each utterance pair, we calculate the score distribution $\mathbf{s}_{i,j}^{(\mathrm{rel})} \in \mathbb{R}^k$ of each arc $\mathrm{X}_i \rightarrow \mathrm{X}_j$, in which the t-th element indicates the score of the t-th relation as the arc label prediction function in (Dozat and Manning, 2017). In the training phase, we also minimize the cross-entropy between gold relation labels and the predicted relations between utterances as:

$$\mathrm{loss}_{\mathrm{rel}} = -\sum_{j=1}^{n} \delta(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C}) \log(\mathrm{P}_{\mathrm{rel}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C})),$$

$$\mathrm{P}_{\mathrm{rel}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C}) = \mathrm{softmax}(\mathbf{s}_{i,j}^{(\mathrm{rel})}).$$

3.4 Discourse Structure Encoding Layer

After the relations are predicted, we construct the discourse structure as a multi-relational graph in which each node indicates an utterance, and each edge represents the relation between a pair of utterances. In order to utilize the discourse information in dropped pronoun recovery process, we first encode the discourse structure, and then utilize the discourse information-based utterance representations to improve token states which are used to model the pronoun referent.

Specifically, we apply a multiple relational GCN (Vashishth et al., 2020), referred to as **Rel-GCN**, over the graph to encode the discourse structure based utterance representations $\boldsymbol{\mathcal{R}}$ and utilize the updated representations to further revise syntactic token states $\boldsymbol{\mathcal{H}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{C}}$ for outputting discourse structure based token states $\boldsymbol{\mathcal{Z}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{Z}}_{\mathrm{C}}$. The node states of the graph are initialized as the average aggregation of token states of corresponding utterances. The representation of utterance X_i in the $(k+1)$-th layer is updated by incorporating the discourse relation state $\boldsymbol{h}_{rel}^k$ as:

$$\boldsymbol{r}_i^{k+1} = f\left(\sum_{(j,rel)\in\mathcal{N}(\mathrm{X}_i)} \mathrm{P}_{\mathrm{rel}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C})\mathbf{W}_{\lambda(\mathrm{rel})}^k \phi\left(\boldsymbol{r}_j^k, \boldsymbol{h}_{rel}^k\right)\right),$$

where $\boldsymbol{r}_j^k$ and $\boldsymbol{h}_{rel}^k$ denote the updated representation of utterance j and relation rel after the k-th GCN layers, and $\mathbf{W}_{\lambda(rel)}^k \in \mathbb{R}^{d\times d}$ is a relation-type specific parameter. Following the practice of (Vashishth et al., 2020), we take the composition operator ϕ as multiplication in this work. Please note that we take in the label distribution $\mathrm{P}_{\mathrm{rel}}(\mathrm{X}_j|\mathrm{X}_i, \mathrm{C})$ from the relation prediction layer and compute the weighted sum of each kind of relation to update the utterance representation, rather than taking the hard predicted relation by applying an argmax operation over the distribution.

After encoding the constructed discourse structure with a message passing process, we obtain the discourse relation-augmented utterance representations $\boldsymbol{\mathcal{R}}$, and then utilize the updated utterance

representations to revise the syntactic token states with a linear feed-forward network:

$$\boldsymbol{z}_{w_{i,n}} = \mathbf{W}_1 \cdot \left[\boldsymbol{h}_{w_{i,n}}^{k+1}; \boldsymbol{r}_i^{k+1}\right] + \boldsymbol{b}_1,$$

where $\boldsymbol{h}_{w_{i,n}}^{k+1}$ refers to the token state of $w_{i,n}$ outputted from the $(k+1)$-th layer of SynGCN, $\boldsymbol{r}_i^{k+1}$ refers to the state of the corresponding utterance that the token belongs to, outputted from the $(k+1)$-th layer of RelGCN. The operation thus augments syntactic token states $\boldsymbol{\mathcal{H}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{H}}_{\mathrm{C}}$ with discourse information-based utterance representation to obtain discourse context-based token states $\boldsymbol{\mathcal{Z}}_{\mathrm{X}} = (\boldsymbol{z}_{w_1}, \ldots, \boldsymbol{z}_{w_n})$ and $\boldsymbol{\mathcal{Z}}_{\mathrm{C}} = (\boldsymbol{z}_{w_{1,i}}, \ldots, \boldsymbol{z}_{w_{i,l_i}})$, which will be used to model the referent semantics of the dropped pronoun in the dropper pronoun recovery layer.

3.5 Pronoun Recovery Layer

This layer takes in the revised token representations $\boldsymbol{\mathcal{Z}}_{\mathrm{X}}$ and $\boldsymbol{\mathcal{Z}}_{\mathrm{C}}$, and attempts to find tokens in context C that describe the referent of the dropped pronoun in the pro-drop utterance X with an attention mechanism. The referent representation is then captured as the weighted sum of discourse context-based token states as:

$$aw_{i,i',n'} = \mathrm{softmax}(\mathbf{W}_2 \left(\boldsymbol{z}_{w_i} \odot \boldsymbol{z}_{w_{i',n'}}\right) + b_2),$$

$$\boldsymbol{r}_{w_i} = \sum_{i'=1}^{m} \sum_{n'=1}^{l_{i'}} aw_{i,i',n'} \cdot \boldsymbol{z}_{w_{i',n'}}.$$

Then we concatenate the referent representation $\boldsymbol{r}_{w_i}$ with the syntactic token representation $\boldsymbol{h}_{w_i}^{k+1}$ to predict the dropped pronoun category as follows:

$$\boldsymbol{hr}_{w_i} = \tanh\left(\mathbf{W}_3 \cdot \left[\boldsymbol{h}_{w_i}^{k+1}; \boldsymbol{r}_{w_i}\right] + \boldsymbol{b}_3\right),$$

$$P\left(y_i | w_i, C\right) = \mathrm{softmax}\left(\mathbf{W}_4 \cdot \boldsymbol{hr}_{w_i} + \boldsymbol{b}_4\right).$$

The objective of dropped pronoun recovery aims to minimize cross-entropy between the predicted label distributions and the annotated labels for all sentences as:

$$\mathrm{loss}_{\mathrm{dp}} = -\sum_{q \in Q} \sum_{i=1}^{l_i} \delta\left(y_i | w_i, \mathrm{C}\right) \log\left(P\left(y_i | w_i, \mathrm{C}\right)\right),$$

where Q represents all training instances, l_i represents the number of words in pro-drop utterance; $\delta\left(y_i | w_i, \mathrm{C}\right)$ represents the annotated label of w_i.

3.6 Training Objective

We train our DiscProReco by jointly optimizing the objective of both discourse relation prediction and dropped pronoun recovery. The total training objective is defined as:

$$\mathrm{loss} = \alpha \cdot (\mathrm{loss}_{\mathrm{arc}} + \mathrm{loss}_{\mathrm{label}}) + \beta \cdot \mathrm{loss}_{\mathrm{dp}}, \quad (1)$$

where α and β are weights of CDP objective function and DPR objective function respectively.

4 The SPDPR Dataset

To verify the effectiveness of DiscProReco, we need a conversational corpus containing the annotation of both dropped pronouns and discourse relations. To our knowledge, there is no such a public available corpus. Therefore, we constructed the first Structure Parsing-enhanced Dropped Pronoun Recovery (SPDPR) dataset by annotating the discourse structure information on a popular dropped pronoun recovery dataset (i.e., Chinese SMS).

The Chinese SMS/Chat dataset consists of 684 multi-party chat files and is a popular benchmark for dropped pronoun recovery (Yang et al., 2015). In this study, we set the size of the context snippet to be 8 utterances which include the current pro-drop utterance plus 5 utterances before and 2 utterances after. When performing discourse relation annotation we ask three linguistic experts to independently choose a head utterance for the current utterance from its context and annotate the discourse relation between them according to a set of 8 pre-defined relations (see Appendix A). The inter-annotator agreement for discourse relation annotation is 0.8362, as measured by Fleiss's Kappa. The resulting SPDPR dataset consists of 292,455 tokens and 40,280 utterances, averaging 4,949 utterance pairs per relation, with a minimum of 540 pairs for the least frequent relation and a maximum of 12,252 for the most frequent relation. The SPDPR dataset also annotates 31,591 dropped pronouns (except the "None" category).

5 Experiments

5.1 Experimental Settings

In this work, 300-dimensional pre-trained embeddings (Li et al., 2018) were input to the BiGRU encoder, and 500-dimensional hidden states were uitilized. For SynGCN and RelGCN, we set the number of GCN layers as 1 and 3 respectively, and augment them with a dropout rate of 0.5. The

Model	SPDPR			TC of OntoNotes			BaiduZhidao		
	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
MEPR	37.27	45.57	38.76	-	-	-	-	-	-
NRM	37.11	44.07	39.03	23.12	26.09	22.80	26.87	49.44	34.54
BiGRU	40.18	45.32	42.67	25.64	36.82	30.93	29.35	42.38	35.83
NDPR	49.39	44.89	46.39	39.63	43.09	39.77	41.04	46.55	42.94
XLM-RoBERTa-NDPR	54.03	50.18	52.46	43.14	46.37	45.13	46.04	49.12	47.54
Transformer-GCRF	52.51	48.12	49.81	40.48	44.64	42.45	43.30	46.54	43.92
DiscProReco	59.58	53.68	57.37	-	-	-	-	-	-
DiscProReco(XLM-R-w/o RelGCN)	56.32	52.28	55.67	**44.62**	**47.14**	**46.98**	**47.31**	**50.43**	**48.19**
DiscProReco(XLM-R)	**61.13**	**54.26**	**59.47**	-	-	-	-	-	-

Table 1: Experimental results produced by the baseline models, the proposed model DiscProReco and two variants of DiscProReco on all three conversation datasets in terms of precision, recall and F-score.

Stanza dependency parser (Qi et al., 2020) returns 41 kinds of dependency edges. We remove 13 types of them which connects the punctuation with other tokens, and irrelevant to referent description. During training, we utilized Adam optimizer (Kingma and Ba, 2015) with a 0.005 learning rate and trained our model for 30 epochs. The model performed best on the validation set is used to make predictions on the test set. We repeat each experiment 10 times and records the average results.

5.2 Dropped Pronoun Recovery

Datasets and Evaluation Metrics We tested the performance of DiscProReco for DPR on three datasets: (1) TC section of OntoNotes Release 5.0, which is a transcription of Chinese telephone conversations, and is released in the CoNLL 2012 Shared Task. (2) BaiduZhidao, which is a question answering corpus (Zhang et al., 2019). Ten types of concrete pronouns were annotated according to the pre-defined guidelines. These two benchmarks do not contain the discourse structure information and are mainly used to evaluate the effectiveness of our model for DPR task. (3) The SPDPR dataset, which contains 684 conversation files annotated with dropped pronouns and discourse relations. Following practice in (Yang et al., 2015, 2019), we reserve the same 16.7% of the training instances as the development set, and a separate test set was used to evaluate the models. The statistics of the three datasets are shown in Appendix B.

Same as existing efforts (Yang et al., 2015, 2019), we use Precision(P), Recall(R) and F-score(F) as metrics when evaluating the performance of dropped pronoun models.

Baselines We compared DiscProReco against existing baselines, including: (1) MEPR (Yang et al., 2015), which leverages a Maximum Entropy classifier to predict the type of dropped pronoun before each token; (2) NRM (Zhang et al., 2019), which employs two MLPs to predict the position and type of a dropped pronoun separately; (3) BiGRU, which utilizes a bidirectional GRU to encode each token in a pro-drop sentence and then makes prediction; (4) NDPR (Yang et al., 2019), which models the referents of dropped pronouns from a large context with a structured attention mechanism; (5) Transformer-GCRF (Yang et al., 2020), which jointly recovers the dropped pronouns in a conversational snippet with general conditional random fields; (6) XLM-RoBERTa-NDPR, which utilizes the pre-trained multilingual masked language model (Conneau et al., 2020) to encode the pro-drop utterance and its context, and then employs the attention mechanism in NDPR to model the referent semantics.

We also compare two variants of DiscProReco: (1) DiscProReco (XLM-R-w/o RelGCN), which replaces the BiGRU encoder with the pre-trained XLM-RoBERTa model, removes the RelGCN layer, and only utilizes SynGCN to encode syntactic token representations for predicting the dropped pronouns. (2) DiscProReco(XLM-R) which uses the pre-trained XLM-RoBERTa model as an encoder to replace the BiGRU network in our proposed model.

Experimental Results Table 1 reports the results of DiscProReco and the baseline methods on DPR. Please note that for the baseline methods, we directly used the numbers originally reported in the corresponding papers. From the results, we observed that our variant model DiscProReco(XLM-

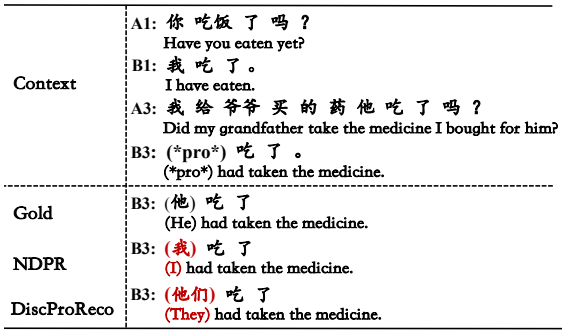

Figure 3: Results of different DPR models.

	STAC		SPDPR	
Model	**Arc**	**Rel**	**Arc**	**Rel**
MST	68.8	50.4	-	-
ILP	68.6	52.1	-	-
Deep+MST	69.6	52.1	81.06	40.93
Deep+ILP	69.0	53.1	80.53	41.38
Deep+Greedy	69.3	51.9	81.32	42.38
Deep Sequential	73.2	55.7	83.00	43.45
DiscProReco(w/o DPR)	**74.1**	**57.0**	84.51	51.34
DiscProReco	-	-	**87.97**	**53.07**

Table 2: Micro-averaged F-score (%) of conversational discourse parsing on two standard benchmarks.

R-w/o RelGCN) outperforms existing baselines on three datasets by all evaluation metrics, which prove the effectiveness of our system as a standalone model for recovering dropped pronouns. We attribute this to the ability of our model to consider long-distance syntactic dependencies between tokens in the same utterance. Note that the results for feature-based baseline MEPR (Yang et al., 2015) on OntoNotes, and BaiduZhidao are not available because several essential features cannot been obtained. However, our proposed DiscProReco still significantly outperforms DiscProReco (XLM-R-w/o RelGCN) as it achieved 3.26%, 1.40%, and 1.70% absolute improvements in terms of precision, recall and F-score respectively on SPDPR corpus. This shows that discourse relations between utterances are crucially important for modeling the referent of dropped pronouns and achieving better performance in dropped pronoun recovery. This is consistent with the observation in (Ghosal et al., 2019). The best results are achieved when our model uses uses the pretrained XLM-RoBERTa (i.e., DiscProReco(XLM-R)). Note that discourse relations are not available for Ontonotes and BaiduZhidao datasets and thus we do not have joint learning results for these two data sets.

Error Analysis We further investigated some typical mistakes made by our DiscProReco for DPR. Resolving DPR involves effectively modeling the referent of each dropped pronoun from the context to recover the dropped pronoun. As illustrate in Figure 3, both DiscProReco and NDPR model the referent from the context. The former outperforms the latter since it considers the conversation structure that the utterance B3 is a reply to A3 but not an expansion to the utterance B1. However, just modeling the referent from the context is insufficient. In Figure 3, the referent of the dropped pronoun was correctly identified but the dropped pronoun is mistakenly identified as "(他们/they)". This indicates that the model needs to be augmented with some additional knowledge, such as the difference between singular and plural pronouns.

5.3 Conversational Discourse Parsing

Datasets and Evaluation Metrics We evaluated the effectiveness of our DiscProReco framework for CDP task on two datasets as: (1) STAC, which is a standard benchmark for discourse parsing on multi-party dialogue (Asher and Lascarides, 2005). The dataset contains 1,173 dialogues, 12,867 EDUs and 12,476 relations. Same as existing studies, we set aside 10% of the training dialogues as the validation data. (2) SPDPR, which is constructed in our work containing 684 dialogues and 39,596 annotated relations. Following (Shi and Huang, 2019), we also utilized micro-averaged F-score as the evaluation metric.

Baselines We compared our DiscProReco with existing baseline methods: (1) MST (Afantenos et al., 2015): A approach that uses local information in two utterances to predict the discourse relation, and uses the Maximum Spanning Tree (MST) to construct the discourse structure; (2) ILP (Perret et al., 2016): Same as MST except that the MST algorithm is replaced with Integer Linear Programming (ILP); (3) Deep+MST: A neural network that encodes the discourse representations with GRU, and then uses MST to construct the discourse structure; (4) Deep+ILP: Same as Deep+MST except that the MST algorithm is replaced with Integer Linear Programming (ILP); (5) Deep+Greedy: Similar to Deep+MST and Deep+ILP except that this model uses a greedy decoding algorithm to select the parent for each utterance; (6) Deep Sequential (Shi and Huang, 2019): A deep sequential neural net-

work which predicts the discourse relation utilizing both local and global context.

In order to explore the effectiveness of joint learning scheme, we also make a comparison of our DiscProReco with its variant, referred to as DiscProReco(w/o DPR), which predict the discourse relation independently, without recovering the dropped pronouns.

Experimental Results We list the experimental results of our approach and the baselines in Table 2. For the STAC dataset, we also reported the original results of the STAC benchmark from an existing paper (Shi and Huang, 2019), and apply our DiscProReco to this corpus. For the SPDPR dataset, we ran the baseline methods with the same parameter settings. From the results we can see that the variant of our approach DiscProReco (w/o DPR) outperforms the baselines of discourse parsing. We attribute this to the effectiveness of the biaffine attention mechanism for dependency parsing task (Yan et al., 2020; Ji et al., 2019). However, our approach DiscProReco still significantly outperforms all the compared models. We attribute this to the joint training of the CDP task and the DPR task. The parameter sharing mechanism makes these two tasks benefits each other. Note that the results for the joint model is not available for STAC as STAC is not annotated with dropped pronouns.

5.4 Interaction between DPR and CDP

We also conducted experiments on SPDPR to study the quantitative interaction between DPR and CDP. Firstly, during the training process, we optimize our DiscProReco model utilizing the objective function in Eq. 1 until the CDP task achieves a specific F-score (i.e., gradually increases from 30.64 to 50.38). Then we fix the CDP components and continue to optimize the components of DPR task. We conduct this experiment to explore the influence of CDP task on the DPR task. Secondly, we set the ratio between α and β in Eq. 1 varies from 0.25 to 1.25 and record the F-score of DPR and CDP respectively. We conduct this experiment to study the interanction between these two tasks by modifying their weights in the objective function.

Results of these two experiments are shown in Figure 4. According to Figure 4 (a), the performance of DPR is increased in terms of all evaluation metrics as the F-score of CDP increases, which indicates that exploring the discourse relations between utterances benefits dropped pronoun recovery. Moreover, Figure 4 (b) illustrate the performance of DPR and CDP when the ratio between α to β varies gradually. Results show that the performance of CDP remains stable, while the performance of DPR increases at beginning and then decrease sharply as the ratio increases, indicating that DiscProReco framework should pay more attention to DPR during the optimizing process.

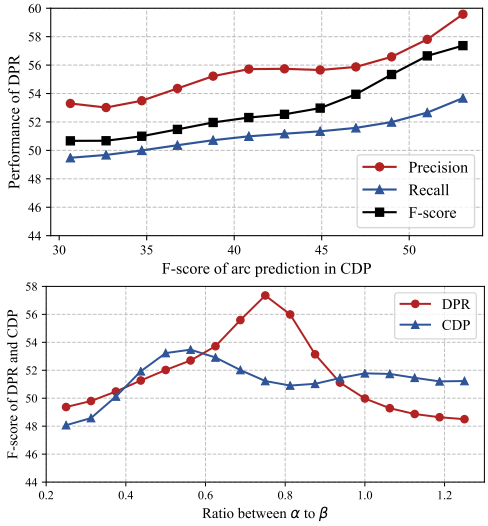

Figure 4: Exploratory results. (a) Interaction between DPR and CDP; (b) Effects of parameters (i.e., α and β)

6 Related Work

Dropped pronoun recovery is a critical technique that can benefit many downstream applications (Wang et al., 2016, 2018; Su et al., 2019). Yang et al. (2015) for the first time proposed this task, and utilized a Maximum Entropy classifier to recover the dropped pronouns in text messages. Giannella et al. (2017) further employed a linear-chain CRF to jointly predict the position and type of the dropped pronouns in a single utterance using hand-crafted features. Due to the powerful semantic modeling capability of deep learning, Zhang et al. (2019); Yang et al. (2019) introduced neural network methods to recover the dropped pronoun by modeling its semantics from the context. All these methods represent the utterances without considering the relationship between utterances, which is important to identify the referents. Zero pronoun resolution is also a closely related line of research to DPR (Chen and Ng, 2016; Yin et al., 2017, 2018). The main difference between DPR and zero pronoun resolution task is that DPR considers both anaphoric and non-anaphoric pronouns, and doesn't attempt to resolve it to a referent.

Existing discourse parsing methods first predicted the probability of discourse relation, and then applied a decoding algorithm to construct the discourse structure (Muller et al., 2012; Li et al., 2014; Afantenos et al., 2015; Perret et al., 2016). A deep sequential model (Shi and Huang, 2019) was further presented to predict the discourse dependencies utilizing both local information of two utterances and the global information of existing constructed discourse structure. All these methods consider how to do relation prediction independently. However, in this work, we explore the connection between the CDP and DPR, and attempt to make these two tasks mutually enhance each other.

7 Conclusion

This paper presents that dropped pronoun recovery and conversational discourse parsing are two strongly related tasks. To make them benefit from each other, we devise a novel framework called DiscProReco to tackle these two tasks simultaneously. The framework is trained in a joint learning paradigm, and the parameters for the two tasks are jointly optimized. To facilitate the study of the problem, we created a large-scale dataset called SPDPR which contains the annotations of both dropped pronouns and discourse relations. Experimental results demonstrated that DiscProReco outperformed all baselines on both tasks.

Acknowledgments

This work was supported by the National Key R&D Program of China (2019YFE0198200), the National Natural Science Foundation of China (No. 61872338, No. 61832017), Beijing Academy of Artificial Intelligence (BAAI2019ZD0305), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098 and BUPT Excellent Ph.D. Students Foundation (No.CX2020305).

References

Stergos D. 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, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 928–937. The Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2005. *Logics of Conversation*. Studies in natural language processing. Cambridge University Press.

Chen Chen and Vincent Ng. 2016. Chinese zero pronoun resolution with deep neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Chris Giannella, Ransom K. Winder, and Stacy Petersen. 2017. Dropped personal pronoun recovery in chinese SMS. *Nat. Lang. Eng.*, 23(6):905–927.

Tao Ji, Yuanbin Wu, and Man Lan. 2019. Graph-based dependency parsing with graph neural networks. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2475–2485. Association for Computational Linguistics.

Young Joo Kim. 2000. Subject/object drop in the acquisition of korean: A cross-linguistic comparison. *Journal of East Asian Linguistics*, 9(4):325–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France,*

April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Jiwei Li, Rumeng Li, and Eduard H. Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 2061–2069. ACL.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 138–143. Association for Computational Linguistics.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1506–1515. Association for Computational Linguistics.

Philippe Muller, Stergos D. Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 1883–1900. Indian Institute of Technology Bombay.

Jérémy Perret, Stergos D. Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 99–109. The Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 101–108. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance rewriter. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 22–31. Association for Computational Linguistics.

Shikhar Vashishth, Shib Sankar Dasgupta, Swayambhu Nath Ray, and Partha P. Talukdar. 2018. Dating documents using graph convolution networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1605–1615. Association for Computational Linguistics.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018. Translating pro-drop languages with reconstruction models. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4937–4945. AAAI Press.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 983–993. The Association for Computational Linguistics.

Nianwen Xue, Qishen Su, and Sooyoung Jeong. 2016. Annotating the discourse and dialogue structure of SMS message conversations. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016, LAW@ACL 2016, August 11, 2016, Berlin, Germany*. The Association for Computer Linguistics.

Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. A graph-based model for joint chinese word segmentation and dependency parsing. *Trans. Assoc. Comput. Linguistics*, 8:78–92.

Jingxuan Yang, Jianzhuo Tong, Si Li, Sheng Gao, Jun Guo, and Nianwen Xue. 2019. Recovering dropped pronouns in chinese conversations via modeling their referents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 892–901. Association for Computational Linguistics.

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Ji-Rong Wen, and Nianwen Xue. 2020. Transformer-gcrf: Recovering chinese dropped pronouns with general conditional random fields. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 137–147. Association for Computational Linguistics.

Yaqin Yang, Yalin Liu, and Nianwen Xue. 2015. Recovering dropped pronouns from chinese text messages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 309–313. The Association for Computer Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, and Ting Liu. 2017. Chinese zero pronoun resolution with deep memory network. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1309–1318. Association for Computational Linguistics.

Qingyu Yin, Yu Zhang, Weinan Zhang, Ting Liu, and William Yang Wang. 2018. Deep reinforcement learning for chinese zero pronoun resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 569–578. Association for Computational Linguistics.

Weinan Zhang, Ting Liu, Qingyu Yin, and Yu Zhang. 2019. Neural recovery machine for chinese dropped pronoun. *Frontiers Comput. Sci.*, 13(5):1023–1033.

Relation	Description
Different Participant	
Agreement	a participant provides a response to a previous request or suggestion
Understanding	a participant indicates understanding of a previous utterance
Directive	a participant asks another one to do something
Question	a general request for another participant
Answer	a participant provides the information requested by another participant
Feedback	a participant responds to another speaker's utterance
Same Participant	
Expansion	a participant provides an elaboration of a previous utterance
Contingency	a participant continues to say something else

Table 3: Explanation of discourse relations.

	Training		Test	
	#Sentences	#DPs	#Sentences	#DPs
SPDPR	35,933	28,052	4,346	3,539
TC	6,734	5,090	1,122	774
Zhidao	7,970	5,097	1,406	786

Table 4: Statistics of training and test sets on three conversational datasets.

A Discourse Relations

The discourse relation describes a participant may speak a utterance to agree with, respond to, or indicate understanding of another utterance in the conversational context. According to (Xue et al., 2016) , each utterance is assumed only related to one previous utterance. All relations are summarized as 6 types between *same-participant* utterance pairs, and 2 types between *different-participant* utterance pairs, as summarized in Table 3.

B Statistics of DPR Datasets

The statistics of three dropped pronoun recovery benchmarks (i.e., SPDPR, TC section of OntoNotes and BaiduZhidao) are shown in Table 4.