# Consistent Inference for Dialogue Relation Extraction

**Xinwei Long**[1,2] , **Shuzi Niu**[1] , **Yucheng Li**[1]
[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
longxinwei19@mails.ucas.ac.cn, {shuzi, yucheng}@iscas.ac.com,

## Abstract

Relation Extraction is key to many downstream tasks. Dialogue relation extraction aims at discovering entity relations from multi-turn dialogue scenario. There exist utterance, topic and relation discrepancy mainly due to multi-speakers, utterances, and relations. In this paper, we propose a consistent learning and inference method to minimize possible contradictions from those distinctions. First, we design mask mechanisms to refine utterance-aware and speaker-aware representations respectively from the global dialogue representation for the utterance distinction. Then a gate mechanism is proposed to aggregate such bi-grained representations. Next, mutual attention mechanism is introduced to obtain the entity representation for various relation specific topic structures. Finally, the relational inference is performed through first order logic constraints over the labeled data to decrease logically contradictory predicted relations. Experimental results on two benchmark datasets show that the F1 performance improvement of the proposed method is at least 3.3% compared with SOTA.

## 1 Introduction

As a fundamental information extraction task, Relation Extraction (RE) is widely used in various downstream tasks, such as knowledge base completion, question answering, and dialogue generation. It predicts relations among entities given a piece of text [Miwa and Bansal, 2016]. More and more studies [Zeng *et al.*, 2020] put their attention on extracting relations in long text, such as a document and dialogue.

Dialogue relation extraction discovers relations from a piece of dialogue text [Yu *et al.*, 2020]. Figure 1 demonstrates a snippet of dialogue from four different speakers about family life. It focuses on relations between speakers, which is important for a dialogue generation system [Choi and Chen, 2018]. Speaker relations in this dialogue example are expressed as a graph in Figure 1. There are two kinds of entities, speakers and objects they are talking about. Though dialogue relation extraction is similar to document level relation extraction, multiple speakers, various topic structures and relations all make it a more challenging task.

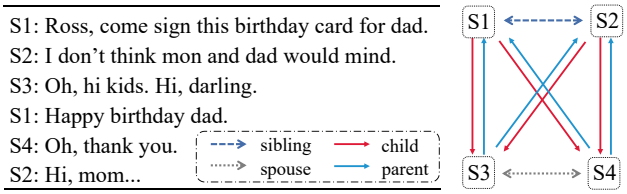


Figure 1: Relations from a dialogue about family life.

**Utterance Discrepancy.** Sentences in the multi-turn dialogue text are expressed in terms of different speakers. The given text leads to the ambiguity of coreference and transition of opinions. Speaker 1 calls speaker 2 "Ross" in the text of Figure 1, and speaker 2 calls himself/herself "I". If two sentences from different speakers are treated as a whole directly, it will confuse the learner who is "Ross" and "I" respectively.

**Topic Structure Discrepancy.** The topic structure in the dialogue text is more flexible due to frequently changing speakers and topics. For document level relation extraction, it needs multiple sentences to predict the relation of a certain entity pair. But sentences are usually located in a certain part of the document. For dialogue relation extraction, these evidence sentences are usually scattered in the whole text. For example in Figure 1, "Dad" is mentioned in the first sentence and the last sentence but two, and at the end of the dialogue Speaker 4 "Dad" appears. Therefore this multi-scale dependency is important to dialogue relation extraction.

**Relation Discrepancy.** Relation extractions are often reduced to multi-class or multi-label classification problem assuming that relations are independent. In fact, all relations in the same dialogue text, speaker-speaker, speaker-object, are connected. The independence assumption easily tends to predict logically contradictory results over entity pairs in the same dialogue text. For example, the predicted label of $Parent(S3, S1)$ may be unequal to that of $Child(S1, S3)$, which is logically inconsistent.

To tackle these challenges, we propose a **Co**nsistent **In**ference network, referred to as **CoIn**. We differentiate context representations between utterances and speakers by corresponding mask strategies to refine consistent representations for both grains. Then we introduce a gate mechanism to aggregate these bi-grained representations and obtain the entity representation from widely distributed men-

tions for each relation specific topic by mutual attention. Finally, we perform 2-hop relational inference based on First Order Logic over labeled data and add feasible 2-hop paths as soft probabilistic rules [Bach *et al.*, 2017] to a differentiable framework by penalizing predicted relations' violated rules. In terms of FOL, $R(a, b)$ means relation $R$ exists between entity $a$ and $b$. In the example graph of Figure 1, starting from $S1$, there are 5 feasible 2-hop reasoning paths, summarized as inverse (e.g. $Parent(S3, S1) \Rightarrow Child(S1, S3)$), symmetric (e.g. $Sibling(S1, S2) \Rightarrow Sibling(S2, S1)$) and transitive rules (e.g. $Spouse(S4, S3) \wedge Parent(S3, S1) \Rightarrow Parent(S4, S1)$).

We conduct comprehensive experiments on two benchmark datasets, DialogRE [Yu *et al.*, 2020] and MPDD [Chen *et al.*, 2020b], and CoIn shows the $3.3\%$ and $6.2\%$ improvement in terms of F1 (DialogRE) and accuracy (MPDD) than state-of-the-art models. Ablation studies prove the effectiveness of each module. Especially, comparison results between models with and without relational reasoning regularization ($R^3$) suggest that it enhances the performance significantly.

## 2 Related Work

**Intra- and Inter- Sentence Relation Extraction**, which aims at predicting relations between entities within a sentence or a document has been widely explored [Miwa and Bansal, 2016; Zhang *et al.*, 2018]. Early work focuses on predicting relations of two given entities in a piece of text, ignoring the interactions across entities and sentences. Recent studies [Nan *et al.*, 2020; Zeng *et al.*, 2020] expand the extraction scope from a single sentence to an entire document, considering both the intra- and inter-sentences relations. [Yao *et al.*, 2019] proposes a large-scale document level RE dataset, which requires reasoning ability to infer the complex relations across a long distance. [Wang *et al.*, 2020; Zhou *et al.*, 2020] leverage external tools or manual features to construct a document graph, and deploy graph network to learn the contextual representation. GAIN [Zeng *et al.*, 2020] aggregates features from a mention-level graph and conducts reasoning over an entity-level graph.

**Relation Extraction over Dialogue** is a newly defined task by DialogRE [Yu *et al.*, 2020], which focuses on extracting relations between speakers and arguments in a dialogue. DialogRE is an English dialogue relation extraction dataset, consisting of 1788 dialogues and 36 relations. MPDD [Chen *et al.*, 2020b] is a similar dialogue corpus with 24 relations, built on five Chinese TV series. Both DialogRE and MPDD deal with multi-turn, multi-participants and multi-domain dialogue RE challenges. Similar to some document level RE methods, a heterogeneous document graph is constructed over a dialogue [Chen *et al.*, 2020a]. The latent multi-view graph is proposed to capture critical features from long texts [Xue *et al.*, 2020b]. A more efficient model relation refinement mechanism reports the state-of-the-art performance [Xue *et al.*, 2020a]. Ignoring the distinction between documents and dialogues, all of them treat dialogues as flat long texts without the discrimination of utterances and speakers. Moreover, they do not consider the implicit connection between triplets, causing the complex inter-speaker relation extraction bottleneck.

**Symbolic knowledge** is well-known to be effective at commonsense reasoning tasks, but purely symbolic approaches are proved insufficient to deal with the uncertainty of natural language [Rocktäschel *et al.*, 2015]. Probabilistic soft logic [Bach *et al.*, 2017] is introduced to integrate logic rules into neural models, which can be optimized in a differentiable framework. Besides, several teacher-student models [Hu *et al.*, 2016; Zhang *et al.*, 2020] are proposed to distill logic knowledge for student models. In terms of neural models, some methods [Li and Srikumar, 2019; Maji *et al.*, 2020] compile logical statements into computation graphs through auxiliary named neuron, while other methods [Asai and Hajishirzi, 2020; Wang and Pan, 2020] penalize logical violations as a posterior regularization.

## 3 Methodology

The dialogue is a semi-structured piece of text and has complex relations compared with a flat document, which leads to utterance, structure and relation discrepancies. To learn a unified model from these discrepancies, we propose a consistent learning and inference framework. Here we first formalize the dialogue relation extraction task and then introduce our proposed framework in detail.

### 3.1 Problem Formulation

Given a dialogue relation extraction dataset $D = \{(U_i, S_i, G_i)\}_{i=1}^{N}$, there are $N$ dialogues and the set of all possible relations in $D$ is denoted as $R$. For each triplet $(U, S, G) \in D$, a dialogue with $|U|$ utterances $U = \{u_1, \ldots, u_{|U|}\}$ involves a set of speakers $S$, and $G = \{(e_i, R_{ij}, e_j)\}$ is a triplet set. An utterance in a dialogue is identified by a pause or a change of speaker. Thus each utterance $u_t$ contains a sequence of words and usually belongs to one speaker. Each triplet $(e_i, R_{ij}, e_j)$ represents there exist a relation set $R_{ij} \subset R$ between entity $e_i$ and $e_j$. Each triplet set $G$ is naturally a small knowledge base. Our goal is to learn a relation classifier $f(\cdot, \cdot)$ based on $D$. At the inference stage, given any new dialogue and a group of entity pairs, we use this classifier $f(\cdot, \cdot)$ to predict relations for each entity pair.

### 3.2 Architecture

To alleviate the semantic ambiguity between speakers, we design mask mechanisms to refine utterance-aware and speaker-aware representations. From the utterance perspective, we employ intra-utterance and window-limited inter-utterance word relations separately to derive two complementary utterance level representations. From the speaker perspective, we use inter-speaker and intra-speaker word relations separately to obtain two complementary speaker level representations. Then we propose a consistency and discrepancy fusion gate to aggregate two complementary representations in terms of utterances and speakers respectively, which is supposed to learn flexible representations for different topic structures. Next, mutual attention mechanism is proposed to obtain discriminative utterance-aware and speaker-aware representations for various kinds of relations. Finally, this bi-grained representation is used for relation classification. For optimization, we

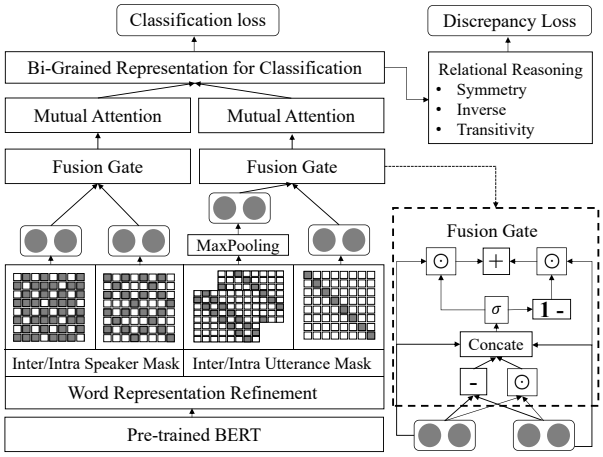


Figure 2: Architecture of CoIn

perform relation reasoning over the ground truth triplet set $G$, an minimize the discrepancy between network and logical outputs. The whole architecture is shown in Figure 2.

### 3.3 Word Representation Refinement

Given a dialogue $U$ with $|U|$ utterances, we concatenate all the utterances as a long input sequence with special tokens like"$[cls], [s_0], x_{00}, \ldots, [s_i], x_{i0}, \ldots, x_{ij}, [sep]$". $[cls]$ and $[sep]$ are the start and end token of a dialogue respectively. $[s_i]$ is the speaker index of the utterance $u_i$, and $x_{ij}$ represents the j-th token in $u_i$. Through a pretrained BERT, we obtain word representations $H \in \mathbb{R}^{n_t \times d_{model}}$, where $d_{model}$ is the dimension of learned word representations and $n_t$ denotes the input sequence length. Considering utterance orders and entity types, we utilize sinusoid position embeddings to encode the utterance id, and randomly initialized vectors for type encoding. For each word, we add the utterance and type representation denoted as $E_u$ and $E_t$ to $H$ as $\widetilde{H} = H + E_u + E_t$.

Rich features, such as domain and emotion, is encoded in $\widetilde{H}$. Some are essential to predict scene-specific relations. Others involve much noises from the ambiguity of coreference and transition of speakers. To refine information from the mixed global representation, we deploy self-attention module with mask strategies to focus on information in a certain scope. The multi-head self-attention mechanism [Vaswani *et al.*, 2017] with different masks can be formalize as:

$$H_q = \text{Concat}(\text{head}_1^q, ..., \text{head}_h^q), \tag{1}$$

$$\text{head}_i^q = \text{softmax}(\frac{\widetilde{H}W_i^Q \cdot (\widetilde{H}W_i^K)^T}{\sqrt{d_k}} + M_q)\widetilde{H}, \tag{2}$$

where $W^Q \in \mathbb{R}^{d \times d_q}, W^K \in \mathbb{R}^{d \times d_k}$ are trainable parameters. $M_q, q \in \{u, w, s, o\}$ denotes the mask.

**Utterance-aware Representation Refinement**

To capture useful information within a local scope, we design intra-utterance mask in Equation 3 and window-limited inter-utterance masks in Equation 4. The former focuses on word relations within each utterance. The latter attends word relations from its left and right $k$-th utterances to handle the transitivity of coreference and opinions. We design inter-utterance masks $M_w^k$with different window size $k \in \{1, \ldots, K\}$, $K$ is a hyper-parameter.

$$M_u[i,j] = \begin{cases} 0 & \text{uid}(i) = \text{uid}(j) \\ -\infty & otherwise \end{cases} \tag{3}$$

$$M_w^k[i,j] = \begin{cases} 0 & \text{uid}(i) = \text{uid}(j) \pm k \\ -\infty & otherwise \end{cases} \tag{4}$$

where $i$ and $j$ are token positions, $\text{uid}(i)$ is the utterance index of $i$-th token. $M_u$ and $\{M_w^k\}_{k=1}^K$ are fed into Equation 1 separately. Then we obtain the intra-utterance representation $H_u$ and window limited inter-utterance representation $\{H_w^k\}_{k=1}^K$. Moreover, we deploy max-pooling to aggregate representations from windows of different sizes through $H_w = \max(H_w^1, ..., H_w^K)$.

**Speaker-aware Representation Refinement**

Different speakers tend to have different opinions in the same thing or the same opinion with different words. Thus we refine speaker level representations by dividing word relations in the self-attention matrix into intra-speaker and inter-speaker two classes. On the one hand, both kinds of relations are easy to model the long range dependency between words, which are useful for relations across sentences within a long distance. On the other hand, we distinguish intra-speaker and inter-speaker relations to model the consistency and discrepancy with different weights to learn flexible representations for different scenarios. Thus, we define the intra-speaker mask [Liu *et al.*, 2020] in Equation 5 and its complementary mask $M_o = -\infty - M_s$.

$$M_s[i,j] = \begin{cases} 0 & \text{sid}(i) = \text{sid}(j) \\ -\infty & otherwise \end{cases} \tag{5}$$

$\text{sid}(i)$ is the speaker index of $U$'s $i$-th token. We feed $M_s$ and $M_o$ to Equation 1 separately and obtain speaker aware representations $H_s$ and $H_o$.

### 3.4 Consistency and Discrepancy Fusion Gate

We decompose word relations into intra-utterance and inter-utterance within a window, and obtain two complementary views of word representations $H_u$ and $H_w$. To capture the similarity and difference between two views, we introduce a consistency and discrepancy fusion gate to derive the utterance aware representation. Similar to [Mou *et al.*, 2016], we design heuristic features, such as subtraction and element-wise multiplication in Equation 6, to capture the discrepancy and consistency signal. These heuristic features are fed into a gate in Equation 7 to control which (intra- or inter-utterance representation) is important for a relation as Equation 8.

$$I_{u,w} = [H_u; H_w; H_u - H_w; H_u \odot H_w] \tag{6}$$

$$G_{u,w} = \sigma(WI_{u,w} + b) \tag{7}$$

$$H_L = G_{u,w} \odot H_u + (1 - G_{u,w}) \odot H_w \tag{8}$$

Similarly to capture the consistency and discrepancy between intra-speaker and inter-speaker representations, we feed $H_s$ and $H_o$ to the fusion gate like Equation 6∼8 to get the speaker-aware representations $H_S$.

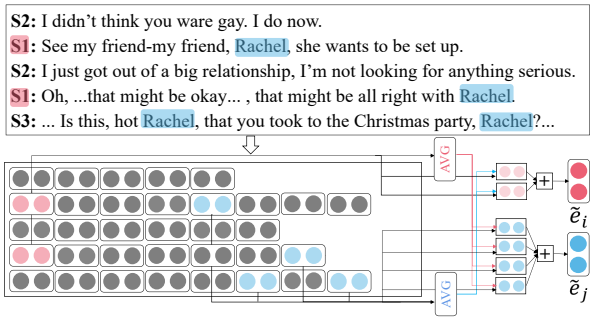


Figure 3: Mutual Attention for Entity Pair (Speaker1,Rachel)

### 3.5 Mutual Attention

To learn discriminative representations for relations, we propose a mutual attention mechanism to derive an entity pair representation. Given an entity pair $(e_i, e_j)$ and dialogue $U$, there are $n_i$ mentions of $e_i$ and $n_j$ mentions of $e_j$ in $U$. Each mention of $e_i$ is represented with the average word representations from $H_L$ and $H_S$ denoted as $m^l_{i,k}, l \in \{L, S\}$ respectively. The coarse entity representation is computed as the averaged representation of all mentions $\bar{e}^l_i = \frac{1}{n_i}\sum_k m^l_{i,k}$.

However, $e_i$ do not only exist a relation with $e_j$, it could be overlapped by other triples. As Figure 3 shown, each speaker entity involves several relations, where entities should attend different mentions when involving different relations. Therefore, we propose a mutual attention module to obtain a fine entity representation from head or tail related mention representations in Equation 9 and 10.

$$s^l_{j,k} = \bar{e}^l_j W_1 m^l_{i,k} + W_2 m^l_{i,k} + b \tag{9}$$

$$\tilde{e}^l_i = \sum_{k=0}^{n_i} a^l_{j,k} \cdot m^l_{i,k}, \text{ where } a^l_{j,k} = \frac{exp(s^l_{j,k})}{\sum_{w=0}^{n_i} exp(s^l_{j,w})} \tag{10}$$

For each entity pair, we obtain $e_i$'s fine representation $\tilde{e}^l_i$ based on how $e_j$ attends to all $e_i$'s mentions and obtain $e_j$'s fine representation $\tilde{e}^l_j$ based on how $e_i$ attends to all $e_j$'s mentions. This attention mechanism helps learn the mutual information between $\bar{e}^l_i$ and $\bar{e}^l_j$, referred to as Mutual attention. We represent each pair as $p^l_{ij} = [\tilde{e}^l_i; \tilde{e}^l_j], l \in \{L, S\}$.

According to one (i.e. MPDD) or multiple relations (i.e. DialogRE) per pair in the ground truth triplet set $G$, relation extraction is solved by multi-class or multi-label classification method. For multi-class classifier, utterance-aware and speaker-aware representations $p^L_{ij}$ and $p^S_{ij}$ is fed into a feed forward neural network layer (FFNN) followed by softmax function as Equation 11. For multi-label classification method, we adopt a FFNN for each relation $r$ and use sigmoid function following the previous work [Chen *et al.*, 2020a] to judge whether relation $r$ exists in this pair as Equation 12.

$$P(\hat{R}_{ij}|p^L_{ij}, p^S_{ij}) = \text{softmax}(\text{FFNN}([p^L_{ij}; p^S_{ij}])) \tag{11}$$

$$P(r|p^L_{ij}, p^S_{ij}) = \sigma(\text{FFNN}_r([p^L_{ij}; p^S_{ij}])) \tag{12}$$

### 3.6 Relational Reasoning Regularization

For each dialogue $U$, its ground truth triplet set $G$ is naturally a small and incomplete knowledge base. Traditional relation classification loss functions focus on minimizing the discrepancy between the predicted relation label and its ground truth label in $G$, which ignores the possible logical constraints between relations. Thus logical constraints existing between ground truth relations may not be satisfied by traditional methods. To alleviate the possible logical contraction problem, we attempt to perform 2-hop relational reasoning over $G$ and minimize the discrepancy between 2-hop relation paths and predicted relation labels.

Each triplet $(e_i, r, e_j) \in G$ is transformed into $r(e_i, e_j)$ with First Order Logic (FOL). 2-hop reasoning paths over $G$ include three relations of relations: inverse, symmetry and transitivity. We formalize these relations of relations as first order logic rules as Equation 13.

$$\begin{aligned} \forall e_i, e_j, r(e_i, e_j) &\rightarrow \bar{r}(e_j, e_i) \\ \forall e_i, e_j, r(e_i, e_j) &\rightarrow r(e_j, e_i) \\ \forall e_i, e_j, \exists e_k, r_1(e_i, e_k) \wedge r_2(e_k, e_j) &\rightarrow r_3(e_i, e_j) \end{aligned} \tag{13}$$

To define the discrepancy between the network output like Equation 12 and logical reasoning result like Equation 14 in a differentiable framework, we utilize a probabilistic mapping [Bach *et al.*, 2017; Wang and Pan, 2020] from logic rules in $\{0, 1\}$ to the interval $[0, 1]$. Without loss of generality, we define each rule body and head as $\phi_b$ and $\phi_h$, where $\phi_b$ or $\phi_h$ could be an atomic proposition, such as $r(e_i, e_j)$, conjunctive or disjunctive normal form, such as $r_1(e_i, e_k) \wedge r_2(e_k, e_j)$. The probability that $\phi \in \{\phi_b, \phi_h\}$ holds is denoted as $P(\phi)$. Equation 14 shows the probability derived from network outputs with multi-label classifier in Equation 12.

$$\begin{aligned} P(r(e_i, e_j)) &= P(r|p^L_{ij}, p^S_{ij}) \\ P(\neg r(e_i, e_j)) &= 1 - P(r|p^L_{ij}, p^S_{ij}) \\ P(\wedge_{n=1}^N r_n(e_{n_1}, e_{n_2})) &= \frac{1}{N}\sum_n P(r_n|p^L_{n_1 n_2}, p^S_{n_1 n_2}) \end{aligned} \tag{14}$$

For each rule $\phi_b \rightarrow \phi_h$ derived from $G$, denoted as positive instances, we measure its discrepancy by $d(P(\phi_b), P(\phi_h))$. Those $\phi_b$ and $\phi_h$ cannot be derived directly or indirectly from $G$, such as $\neg\phi_b \rightarrow \neg\phi_h$, are negative instances. Due to the huge number difference between positive and negative instances, we compute each loss with Equation 15 and 16.

$$loss^P_r(U, G, \Theta) = \frac{1}{S_P}\sum_i d(P(\phi^i_b), P(\phi^i_h)), \tag{15}$$

$$loss^N_r(U, G, \Theta) = \frac{1}{S_N}\sum_i d(1 - P(\phi^i_b), 1 - P(\phi^i_h)), \tag{16}$$

where $d(\cdot, \cdot)$ is the distance metric, and We use mean squared error (MSE) for better convergence. $S_P$ and $S_N$ is the number of positive and negative rules for $U$. $\Theta$ is model parameters.

We train both the neural module and logical regularization in a multi-task framework as $\omega \cdot loss_{class} + (1-\omega) \cdot (\lambda_1 loss^P_r + \lambda_2 loss^N_r)$. $loss_{class}$ is the cross entropy loss based on probabilities in Equation 12 or Equation 11. $loss^P_r$ and $loss^N_r$ are defined as Equation 15 and 16 respectively. $\omega$, $\lambda_1$, and $\lambda_2$ are the hyper-parameters fine-tuning in the validation.

| Dataset | DialogRE | MPDD |
|---|---|---|
| Dialog Num. | 1073 / 358 / 357 | 1482 / 400 / 400 |
| Relation Num. | 4992 / 1597 / 1529 | 5225 / 1370 / 1307 |
| A/M of Length △ | 225.8 / 678 | 199.7 / 775 |
| A/M of Utterance | 21.8 / 42 | 8.9 / 20 |
| A/M of Speakers | 3.3 / 9 | 2.5 / 8 |

Table 1: Dataset Statistics. △: The length of tokenized sub-word sequences. *A/M* represents the average/max value of each item.

# 4 Experiments

## 4.1 Experimental Settings

**Datasets.** (1) **DialogRE** [Yu *et al.*, 2020]. We follow the standard settings offered by the original paper, and deploy F1 score as the metric. (2) **MPDD** [Chen *et al.*, 2020b]. It is not a specialized RE dataset, but a comprehensive corpus supporting several dialogue related tasks. We adapted the corpus from the fine-grained relation classification sub-task. Besides, there are many speakers with names appearing both in train and test splits. To avoid information leakage, we follow the settings of DialogRE, which anonymizes speakers' name and filter out the duplicate instances. We deploy accuracy as the metric, in line with the original paper. More details of DialogRE and processed MPDD can be found in Table 1.

**Implementation Details.** We adopt BERT-base architecture with the fine-tuning learning rate of $2e-5$. We insert special tokens for speaker and utterance indices among utterances and obtain a input sequence. We then split the above sequence into sub-word pieces. For those tokenized sequences with length longer than 512, we split the sequences into two overlapped sub-sequences.

**Hyper-parameters.** Experiments are conducted on a sever with a GeForce GTX 1080Ti GPU, 64G memory. Our model was implemented by Pytorch with CUDA 11.0. We use a self-attention layer with dropout $0.2$ and learning rate $5e-4$. The number of windows $K$ is set to 2 from $\{i\}_{i=1}^{4}$. We use AdamW [Loshchilov and Hutter, 2019] as optimizer with Cosine Annealing scheduler [Loshchilov and Hutter, 2017]. The threshold $\tau$ of multi-label classifier, Trade-off parameters $\lambda_1$ and $\lambda_2$ are set to $0.5$[1].

**Rules.** For DialogRE, we define 24 inverse/symmetric rules for the 24 relation types based on the schema offered by original paper. For MPDD, we define 24 inverse/symmetric rule for 24 relation types and 6 transitive rules for 4 types, refering to type definitions.

## 4.2 Baseline Models

Baselines include two benchmark models offered by dataset DialogRE (BERT and BERTs) [Yu *et al.*, 2020], two advanced document-level RE models (GCGCN [Zhou *et al.*, 2020] and GAIN [Zeng *et al.*, 2020]), and state-of-the-art methods [Chen *et al.*, 2020a; Xue *et al.*, 2020a; Xue *et al.*, 2020b] in the DialogRE challenge. Except that HGAT [Chen *et al.*, 2020a] encodes tokens with LSTM, other baselines deploy BERT-base model [Devlin *et al.*, 2019] as encoder.

[1]Source codes and pre-processed data are released in https://github.com/xinwei96/CoIn_dialogRE

| Model | F1-DialogRE | | Acc-MPDD | |
|---|---|---|---|---|
| | Dev | Test | Dev | Test |
| BERT [Yu *et al.*, 2020] | 60.6 | 58.5 | 31.0* | 31.6* |
| $BERT_s$ [Yu *et al.*, 2020] | 63.0 | 61.2 | 37.2* | 36.9* |
| HGAT [Chen *et al.*, 2020a] | 57.7 | 56.1 | - | - |
| GDPNet [Xue *et al.*, 2020b] | 67.1 | 64.9 | - | - |
| SimRE [Xue *et al.*, 2020a] | - | 66.7 | 36.9* | 37.9* |
| GCGCN [Zhou *et al.*, 2020] | 66.9* | 67.6* | 42.5* | 39.1* |
| GAIN [Zeng *et al.*, 2020] | 69.8* | 69.0* | 42.2* | 43.6* |
| CoIn (Our Model) | **71.1** | **72.3** | **46.5** | **49.8** |

Table 2: Performance Comparison on DialogRE and MPDD.

## 4.3 Accuracy Analysis

Obviously our propose method CoIn outperforms the best state-of-the-art baseline by about $3.3\%$ and $6.2\%$ in terms of F1 and accuracy on DialogRE and MPDD respectively, where the best baseline on two corresponding sets is GAIN and GCGCN respectively[2]. CoIn is more suitable for dialogue RE task for the following two possible reasons. One is that CoIn extracts speaker aware information which is scattered in the whole dialogue text, while document level methods locate evidences that is continuously distributed in a part of a document, such as several paragraphs. The other is that utterance aware information refinement in CoIn helps alleviate the semantic ambiguity due to frequently topic switching, which falls outside of document level methods. The next ablation study will show which plays a more important role.

Compared with advanced dialogue RE methods, document level RE methods achieve better performances on both sets. State-of-the-art Document level methods tend to take advantage of reasoning ability over entity graphs to model complex interaction among entity pairs and relations. This is not considered in existing dialogue models because they treat this task as multi-label or multi-class classification task under the relation independence assumption. Distinguished from the reasoning ability over entity graphs from evidences in document level method, CoIn performs relational reasoning on a ground truth triplet set, which avoid possible topic divergence. Thus existing dialogue RE methods do not work.

Performances on DialogRE are consistently better than those on MDPP. This result suggests relation extraction in MDPP is harder. The dialogue text is shorter in terms of the number of utterance list in Table 1, which are not easy especially for speaker relation extraction. The higher performance improvement on MDPP suggests CoIn is fit for this scenario.

## 4.4 Ablation Studies

We conduct a series of ablation studies from the top to bottom layer including Relational Reasoning Regularization ($R^3$), Mutual Attention, Fusion Gate and Mask Mechanisms for Speaker-Aware and Utterance-aware Representation Refinement, to explore their roles in CoIn. Results are in Table 3.

Precision and Recall of CoIn without $R^3$ drop sharply by $0.8\%$, $1.9\%$ respectively. It suggests $R^3$ is critical to CoIn. Obviously the recall reduction is more than that of precision.

[2]$*$ represents the results produced by running author released codes.

| Model | P | R | F1 |
|---|---|---|---|
| Full Model | 74.7 | 70.0 | 72.3 |
| w.o. Relational Reasoning Reg. ($R^3$) | 73.9 | 68.1 | 70.8 |
| w.o. $R^3$ & Mutual Attention | 75.1 | 65.8 | 70.2 |
| w.o. $R^3$ & Gate fushion | 72.6 | 66.7 | 69.5 |
| w.o. $R^3$ & Speaker aware | 74.4 | 66.7 | 70.3 |
| w.o. $R^3$ & Utterance aware | 71.8 | 63.9 | 67.7 |

Table 3: Ablation studies on DialogRE. where *w.o. X* denotes the model without *X* modules.

Possible reasons will be explored in next subsection. Moreover, the Performance for CoIn without $R^3$ in Table 3 is still higher than that of the best baseline in Table 2 on DialogRE. Therefore we further study roles of other modules and use CoIn without $R^3$ as baseline to compare performances between it with and without the other three modules separately.

In the ablation study of Mutual Attention, we replace the mention-aware representation $\widetilde{e}_i^l$ with the coarse-grained entity representation $\bar{e}_i^l$ for CoIn without $R^3$. Without $\widetilde{e}_i^l$, the recall and F1 score decreases by $2.3\%$ and $0.6\%$ respectively. Mutual attention learns discriminative representations by different mention distributions. Without the fusion gate, F1 score sharply declines by $1.3\%$, proving that fusion gate is essential for the decoupled information.

Among all modules, CoIn without $R^3$ and Utterance aware mask mechanism achieves the lowest performance. In other words, Utterance aware mask mechanism, which focuses on utterances within a local window, are most important for CoIn. This agrees with the commonsense that words in a local context are closely related. Long range dependency on the global dialogue history through speaker aware mask mechanism further improves the performance as shown in Table 3.

### 4.5 Discussion

To explore why relational reasoning regularization ($R^3$) prompts the recall performance, we analyze its effect on the prediction performance for different relations. First we investigate its effect on two kinds of relation prediction performance of DialogRE: relations **N**ot constrained in $\mathbf{R}^3$ (denoted as $NR$) and **R**elations constrained in $\mathbf{R}^3$ (denoted as $RR$).

The ablation study of $R^3$ on $NR$ and $RR$ is shown in Table 4. (1) The recall performance improvement on both sets benefits from 2-hop reasoning over the the ground truth triplet set $G$ in terms of $R^3$. Traditional classification loss only focus on one step reasoning over $G$. $R^3$ helps each instance look at one more step ahead over $G$, which leads to more involved entities. (2) The performance of $NR$ is surprisingly higher than that of $RR$. It indicates that relations in $RR$ are more complex to predict than those in $NR$. (3) The performance

| | $RR$ | | | $NR$ | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| with $R^3$ | 63.7 | 61.7 | 62.7 | 86.2 | 78.2 | 82.0 |
| without $R^3$ | 61.4 | 58.6 | 60.0 | 87.0 | 77.4 | 81.9 |
| | +2.3 | +3.1 | +2.7 | -0.8 | +0.8 | +0.1 |

Table 4: Performance Comparison on $RR$ and $NR$ of DialogRE.

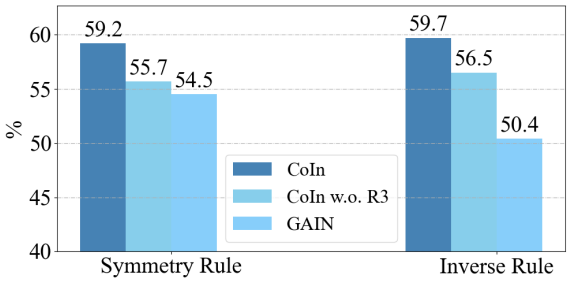


Figure 4: Recall of Consistent triplet pairs on DialogRE.

improvement is higher on $RR$. Complex relations usually deal with multiple speakers and utterances, which needs $R^3$'s help. Simple relations may only need local information, e.g. in an utterance without $R^3$'s influence.

### 4.6 Consistency Analysis

Logical constraints as $R^3$ are directly added for the training stage. Whether logical constraints will take effect on the inference stage remains unknown. To measure the logical consistency of predicted labels, we propose a consistent triplet recall measure. It is defined as the proportion of the number of correctly predicted triplet pairs satisfying a symmetric, inverse, or transitive rule to the ground truth number of triplet pairs satisfying a corresponding rule. Only symmetric and inverse rules exist on DialogRE, so recall of consistent triplet pairs are computed in Figure 4.

CoIn's consistency is significantly higher than the best baseline GAIN for both rules. It suggests that CoIn is better at consistent inference. The Consistency performance of CoIn without $R^3$ is still higher than GAIN though lower than CoIn. This indicates both $R^3$ and representation learning modules do good to consistent inference. Their roles in the consistency improvement are different for two rules. For inverse rules, discriminative representation learning modules are more important because two entity pairs belongs to different relations. For symmetric rules, the representation distinction between entity pairs is not so important as logical constraints, because they belong to the same relation.

## 5 Conclusion

To solve utterance, structure and relation discrepancy problems, we propose a Consistent Inference networks with masking strategies refining multi-grained word representation and mutual attention aggregating entity representation. Finally, a relational reasoning regularization is introduced to minimize the discrepancy between network and logical outputs. Experimental results demonstrate a significant improvement compared with SOTA and prove the effectiveness of each module through ablation studies. In the future, few-shot and overlapped relation detection will be studied in dialogue RE tasks.

## Acknowledgments

This research work was funded by the National Natural Science Foundation of China under Grant No.62072447.

## References


[Asai and Hajishirzi, 2020] Akari Asai and Hannaneh Hajishirzi. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the ACL*, pages 5642–5650, 2020.

[Bach *et al.*, 2017] Stephen H. Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *J. Mach. Learn. Res.*, 18:109:1–109:67, 2017.

[Chen *et al.*, 2020a] Hui Chen, Pengfei Hong, Wei Han, Navonil Majumder, and Soujanya Poria. Dialogue relation extraction with document-level heterogeneous graph attention networks. *CoRR*, abs/2009.05092, 2020.

[Chen *et al.*, 2020b] Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of The LREC*, pages 610–614, 2020.

[Choi and Chen, 2018] Jinho D. Choi and Henry Y. Chen. Semeval 2018 task 4: Character identification on multiparty dialogues. In *Proceedings of The SemEval@NAACL-HLT*, pages 57–64, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019.

[Hu *et al.*, 2016] Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard H. Hovy, and Eric P. Xing. Harnessing deep neural networks with logic rules. In *ACL*, 2016.

[Li and Srikumar, 2019] Tao Li and Vivek Srikumar. Augmenting neural networks with first-order logic. In *ACL*, pages 292–302, 2019.

[Liu *et al.*, 2020] Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. *CoRR*, abs/2009.06504, 2020.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[Maji *et al.*, 2020] Subhadeep Maji, Rohan Kumar, Manish Bansal, Kalyani Roy, and Pawan Goyal. Logic constrained pointer networks for interpretable textual similarity. In *Proceedings of the 29th IJCAI*, pages 2405–2411, 2020.

[Miwa and Bansal, 2016] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the ACL, ACL 2016*, 2016.

[Mou *et al.*, 2016] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the ACL*, pages 130–136, 2016.

[Nan *et al.*, 2020] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. Reasoning with latent structure refinement for document-level relation extraction. In *ACL*, pages 1546–1557, 2020.

[Rocktäschel *et al.*, 2015] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. Injecting logical background knowledge into embeddings for relation extraction. In *NAACL-HLT*, pages 1119–1129, 2015.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[Wang and Pan, 2020] Wenya Wang and Sinno Jialin Pan. Integrating deep learning with logic fusion for information extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 9225–9232. AAAI Press, 2020.

[Wang *et al.*, 2020] Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. Global-to-local neural networks for document-level relation extraction. In *Proceedings of the 2020 Conference on EMNLP*, pages 3711–3721, 2020.

[Xue *et al.*, 2020a] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. An embarrassingly simple model for dialogue relation extraction. *CoRR*, 2020.

[Xue *et al.*, 2020b] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. Gdpnet: Refining latent multi-view graph for relation extraction. *CoRR*, 2020.

[Yao *et al.*, 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. In *ACL*, pages 764–777, 2019.

[Yu *et al.*, 2020] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 4927–4940, 2020.

[Zeng *et al.*, 2020] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on EMNLP*, pages 1630–1640, 2020.

[Zhang *et al.*, 2018] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on EMNLP*, pages 2205–2215, 2018.

[Zhang *et al.*, 2020] Zhenyu Zhang, Xiaobo Shu, Bowen Yu, Tingwen Liu, Jiapeng Zhao, Quangang Li, and Li Guo. Distilling knowledge from well-informed soft labels for neural relation extraction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 9620–9627, 2020.

[Zhou *et al.*, 2020] Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. Global context-enhanced graph convolutional networks for document-level relation extraction. In *COLING*, 2020.