Not Just Classification: Recognizing Implicit Discourse Relation on Joint Modeling of Classification and Generation

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li*, **Qiaoming Zhu**
Soochow University
{fjiang,yxfansupery}@stu.suda.edu.cn,
{xmchu,pfli,qmzhu}@suda.edu.cn

Abstract

Implicit discourse relation recognition (IDRR) is a critical task in discourse analysis. Previous studies only regard it as a classification task and lack an in-depth understanding of the semantics of different relations. Therefore, we first view IDRR as a generation task and further propose a method joint modeling of the classification and generation. Specifically, we propose a joint model, CG-T5, to recognize the relation label and generate the target sentence containing the meaning of relations simultaneously. Furthermore, we design three target sentence forms, including the question form, for the generation model to incorporate prior knowledge. To address the issue that large discourse units are hardly embedded into the target sentence, we also propose a target sentence construction mechanism that automatically extracts core sentences from those large discourse units. Experimental results both on Chinese MCDTB and English PDTB datasets show that our model CG-T5 achieves the best performance against several state-of-the-art systems.

1 Introduction

Discourse relation describes the logical connection between two discourse units (e.g., clauses, sentences, or paragraphs). As an essential discourse analysis task, discourse relation recognition is to recover what rhetorical relation exists between discourse units (DUs). Due to the absence of explicit connectives, implicit discourse relation recognition (IDRR) is still a challenging task and research hotspot. Moreover, IDRR is beneficial to many downstream natural language processing (NLP) applications, such as machine translation (Webber et al., 2017), text generation (Bosselut et al., 2018), and text summarization (Xu et al., 2020).

With the success of representation learning in discourse analysis, most existing methods of IDRR focus on three aspects: enhancing discourse units representation (Ji and Eisenstein, 2015; Qin et al., 2016; Liu and Li, 2016), enhancing semantic interaction (Guo et al., 2018; Ruan et al., 2020; Guo et al., 2020), and joint learning with other tasks (Bai and Zhao, 2018; Nguyen et al., 2019; He et al., 2020). They all regard IDRR as a classification task and lack a deeper understanding of the relation semantics; even recent work (Nguyen et al., 2019; He et al., 2020) with labeling embedding cannot directly introduce the prior knowledge of the discourse relation semantics into their models.

DU_1	Ningbo Free Trade Zone ... has achieved fruitful results after three years of construction. ... the development level is among the best...
DU_2	... the Ningbo Free Trade Zone had completed a total of US$812 million in import and ... At the same time, the bonded zone has ...
Relation	Elaboration
Target Sentence	DU_2 is a detailed description of DU_1.

Table 1: The example of implicit discourse relation between two DUs where DU_1 and DU_2 are paragraphs that contain several sentences. Also, there is no explicit hint in DU_1 and DU_2 for the relation. The full sentences of these two DUs are shown in Appendix A.

In the stage of implicit discourse relation annotation, annotators usually not only give the relation type but also provide a description or basis for the relation. Therefore, we hope the model, like a human, gives a target sentence instead of a simple label index for understanding the relation deeper. The target sentence should describe the core information of two DUs and their relation through natural language. As an example in Table 1, the $Elaboration$ relation can be transformed by definition into the target sentence: *"DU_2 is a detailed description of DU_1"*. The model can more explicitly learn the semantics of the $Elaboration$ relation through such a form of the learning goal.

The Question-Answering (QA) method can incorporate prior knowledge into the model using the generation instead of the classification. It has

[*] Corresponding author

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2418–2431
November 7–11, 2021. ©2021 Association for Computational Linguistics

achieved success in a few fine-grained tasks, such as named entity recognition (Li et al., 2019) and coreference resolution (Wu et al., 2020). However, it is a challenge to directly apply the traditional Question-Answering method to IDRR due to the following two issues. First, unlike the above fine-grained tasks where the answers or the clues exist in the input context, the discourse relation in IDRR is implicit between two DUs and does not appear explicitly in the context. It makes the IDRR model unable to extract the answer from the input directly. Second, since DU usually is large and contains several sentences, the target sentence which contains two DUs as shown in Table 1, is too long to encode. Therefore, it is essential to extract the core information from DU in a short form before embedding them into the target sentence.

Besides, the classification model and the generation model have their complementary advantages. The former usually has better performance on major classes due to searching in a limited space, while the latter can introduce prior knowledge to capture the semantics of minor classes better, and its result is a natural language expression with stronger interpretability. Therefore, how to combine the advantages of the classification model and the generation model is another challenge.

Different from previous work, we first regard IDRR as a text generation task and design three forms of the target sentence to represent prior knowledge. In particular, inspired by the annotation work (Pyatkin et al., 2020), we use questions instead of answers to describe the discourse relation between two DUs as the target sentence. Therefore, our model can understand discourse relation deeper by generating a target sentence that describes the relation meaning instead of an index of the relation type. Moreover, we design a method to automatically extract the core information from these large DUs by semantic role labeling and then compress the DU into a short form.

To address the second challenge, we propose a CG-T5 model that combines the classification and generation model to leverage their complementary advantages. Specifically, inspired by pre-trained models (e.g. BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019)), we first extract hidden states of the encoder in T5 (Raffel et al., 2020) and feed them to the span representation layer for the classification task and then use the T5 decoder for the generation task. Finally, we combine these two models with a jointly learning mechanism. Our CG-T5 can integrate the advantage of two different models: the classification model constrains the generation model, while the generation model explains and supports the classification model. In summary, the main contributions of this paper are fourfold:

- We regard IDRR as a generation task to generate the target sentence containing the meaning of relations, which can introduce prior knowledge to understand discourse relations deeper.

- We propose a joint learning model CG-T5 to integrate the classification task and generation task.

- We design three forms of the target sentence, including the question form, and propose a construction method to extract the core information from the large DUs automatically.

- The experimental results both on MCDTB and PDTB datasets show that our CG-T5 outperforms the SOTA baselines.

2 Related Work

We first briefly introduce relative discourse corpora, then summarize the existing methods of IDRR, and finally introduce the success of the question-answering method in fine-grained tasks.

2.1 Relative Discourse Corpora

In English, one of the most popular discourse corpora is Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) . It annotates about 2.3K Wall Street Journal articles with three-level discourse relations (4 classes, 16 types, and 23 sub-types), including 18.4K explicit relations and 16k implicit relations.

In Chinese, two popular discourse corpora are Chinese Discourse TreeBank (CDTB) (Li et al., 2014) and Macro Chinese Discourse TreeBank (MCDTB) (Jiang et al., 2018). CDTB contains 500 articles with two-level discourse relations (4 classes and 17 types) between clauses and sentences. Following the PDTB-style annotation, it annotated both explicit discourse relation and implicit discourse relation. MCDTB annotated 720 articles from Xinhua News with 3 classes and 15 types relations between paragraphs. Since there are few connectives between paragraphs, it annotated all discourse relations as implicit relations in MCDTB.

2.2 Implicit Discourse Relation Recognition

Most previous studies on IDRR can be divided into the following three categories: enhancing DU representation, enhancing the interaction between DUs, and joint learning of IDRR and other tasks.

In English, early work explored the methods of enhancing DU representation via the shallow convolutional neural network (Zhang et al., 2015), recursive neural network (Ji and Eisenstein, 2015), collaborative gated neural network (Qin et al., 2016), or attention mechanism (Liu and Li, 2016). To enhance the interaction between DUs, Guo et al. (2018) and Ruan et al. (2020) proposed various interactive attention mechanisms for IDRR. Guo et al. (2020) proposed a knowledge-enhanced attention neural network to introduce external knowledge to enhance the interaction. Besides, a few studies combined IDRR with other tasks for joint learning, e.g., explicit relation recognition (Lan et al., 2017), connective prediction (Bai and Zhao, 2018; Shi and Demberg, 2019), and label embedding learning (Nguyen et al., 2019; He et al., 2020).

In Chinese, Zhou et al. (2019) used the macro-structural features and macro-semantic representations to enhance DU representation. Sun et al. (2020) established a large heterogeneous discourse graph on the entire corpus and used the GCN-based model to enhance the interaction between DUs. Xu et al. (2019) and Jiang et al. (2019) jointly learned the relation recognition with the topic modeling and the nuclearity recognition, respectively.

2.3 Formalizing Fine-grained Tasks as QA

A few fine-grained tasks can be formalized as QA and have achieved success due to introducing prior knowledge to their tasks, such as relation extraction (Li et al., 2019), named entity recognition (Li et al., 2020), and co-reference resolution (Wu et al., 2020). It is worth noting that the above studies all take questions as the input and extract the answers from the context as the output.

3 IDRR as Text Generation

To deeper understand the semantics of discourse relation, we regard IDRR as a text generation task, as shown in Figure 1. Unlike previous work, we use a generation model instead of the classification model to generate the target sentence representing the semantics of discourse relation. Then, we obtain the relation type by mapping it into the corresponding discourse relation. In this section, we mainly describe our solution for the first challenge of obtaining the target sentence: its different forms and its corresponding construction method.

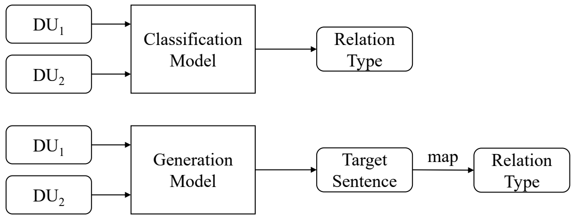

Figure 1: The classification task (upper) and the generation task (lower) for the IDRR.

3.1 Forms of Target Sentence

Unlike other fine-grained tasks that can extract the target sentence (answer) from the context, we can not directly transform their methods to IDRR without manual annotations. In IDRR, the relation label is implicit between two DUs, and there is no hint of them in the context. To alleviate this issue, we design the following three forms of the target sentence to map the relation sense to the templates according to the relation definition: the name, explanation and question of relation.

Form	**Target Sentence Template**
Name	It's Elaboration. 这是解说关系。
Exp-Rel	Because CI_2 is a detailed description of CI_1, it's an Elaboration. 因为CI_2是对CI_1的详细说明，所以是解说关系。
Rel-Exp	It's an Elaboration, because CI_2 is a detailed description of CI_1. 是解说关系，因为CI_2是对CI_1的详细说明。
Q_1	Can you explain in detail about CI_1? 对于CI_1这件事可以详细的解释一下吗?
Q_2	What does CI_2 explain in detail? CI_2是对什么事情的详细解释?

Table 2: The example of the *Elaboration* in MCDTB maps to our designed three forms of templates, where CI_1 and CI_2 are the core information of DU_1 and DU_2, respectively. The complete target sentence templates are shown in the Appendix B.

Relation Name As an intuitive choice, we use relation name (Name) as the target sentence, as shown in Table 9. Using the relation name can introduce prior knowledge by itself, and there is no need to extract external information from context.

Relation Explanation Furthermore, we believe that using only the relation name is not enough, and we design the target sentence as an explanation of relation, as shown in Table 9. It has two variants: the explanation is before the relation name (Exp-Rel) and after the relation name (Rel-Exp).

Although this method contains the relation semantics more comprehensively, it is necessary to extract the core information (CI_1 and CI_2) from two DUs to form the target sentence.

Relation Question Inspired by intra-sentence discourse relation annotation (Pyatkin et al., 2020), we use question instead of declarative sentence as the target sentence to capture discourse relation better. On the one hand, this form extracting core information (CI_1 or CI_2) from only one DU can reduce cascading errors. On the other hand, the question sentence integrating prior knowledge can better connect the semantics of two DUs more naturally.[1] This form also has two variants: the question is guided by the core information of the first DU (Q_1) and guided by that of the second DU (Q_2), as shown in Table 9.

3.2 Constructing Target Sentence

Although we build three forms of the target sentence, the latter two forms need to integrate the core information of the DU into the template. However, it is a challenge to extract the DU's core information without manual annotations. The DU usually contains lots of tokens and we directly insert them into the target sentence will make the target sentence too long to represent the semantics of the corresponding relation. Therefore, we design a core information extraction method to compress DU into a short form, which contains three steps: extracting, filtering, and selecting.

We first extract all candidate tuples from the given DU through the Semantic Role Labeling (SRL) tool[2]. Then we use the following three rules to filter out those redundant tuples: (1) **Streamlining core semantics**. We remove unimportant elements except for arguments and predicates from the extracted candidate tuples. (2) **Ensuring semantic integrity**. We remove the semantically incomplete tuples (i.e., the tuple does not contain both A0 and A1 in SRL) from the candidate tuples. (3) **Reducing semantic overlap**. We remove those small tuples contained in the larger candidate tuples due to semantic overlap. Finally, considering that important information is usually in front of the DUs, we extract the first tuple containing complete semantics and reproduce them in the original order as the CI. In particular, to ensure the form of "Subject-Verb-Object", we place the predicate in the second position.

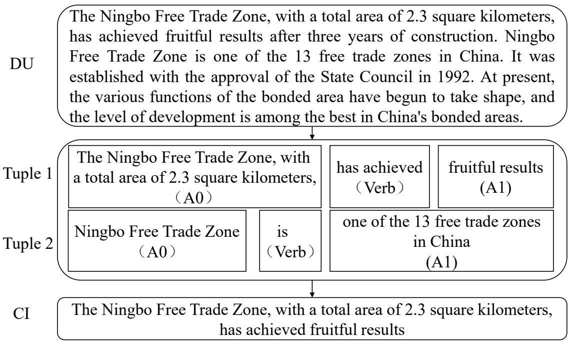
DU
The Ningbo Free Trade Zone, with a total area of 2.3 square kilometers, has achieved fruitful results after three years of construction. Ningbo Free Trade Zone is one of the 13 free trade zones in China. It was established with the approval of the State Council in 1992. At present, the various functions of the bonded area have begun to take shape, and the level of development is among the best in China's bonded areas.
Tuple 1
The Ningbo Free Trade Zone, with a total area of 2.3 square kilometers, （A0）
has achieved （Verb）
fruitful results (A1)
Tuple 2
Ningbo Free Trade Zone （A0）
is （Verb）
one of the 13 free trade zones in China (A1)
CI
The Ningbo Free Trade Zone, with a total area of 2.3 square kilometers, has achieved fruitful results

Figure 2: The example of extracting the core information of discourse unit.

We use the example in Figure 2 to illustrate the process. There is a DU that contains four sentences, and we extract all candidate SRL tuples from it. Then we filter them by three rules we proposed to obtain two simplified tuples. Finally, we select the sentence that contains the first tuple as CI. Besides, if there is more than one tuple in a sentence, we combine them into one CI by the comma.

4 CG-T5 Model

To integrate the advantages of the classification model and the generation model in IDRR, we propose the Classification and Generation T5 (CG-T5) model that recognizes relation class and generates the target sentence simultaneously, as shown in Figure 3. Thanks to the excellent performance of T5 (Raffel et al., 2020) on many NLP tasks, we choose it as the backbone of CG-T5. Better than BERT (Devlin et al., 2019), an encoder architecture for classification, and GPT-2 (Radford et al., 2019), a decoder architecture for generation, T5 has an encoder-decoder architecture that allows us to naturally joint learning classification and generation[3].

CG-T5 comprises three parts: the classification module based on the encoder, the generation module based on the decoder, and the joint learning module. Therefore, when given two DUs as the input, CG-T5 has two outputs: the class label from the classification module and the target sentence from the generation module.

[1] We have also tried to use the question as input and the second DU as output following the question-answering method in the fine-grained task. However, this method did not achieve good performance due to the unlimited generation space.

[2] LTP (http://ltp.ai/index.html) for Chinese and AllenNLP (https://allennlp.org) for English.

[3] We have also tried the general encoder-decoder model (Rothe et al., 2020) with a similar architecture, but the effect is far worse than the pre-trained T5.

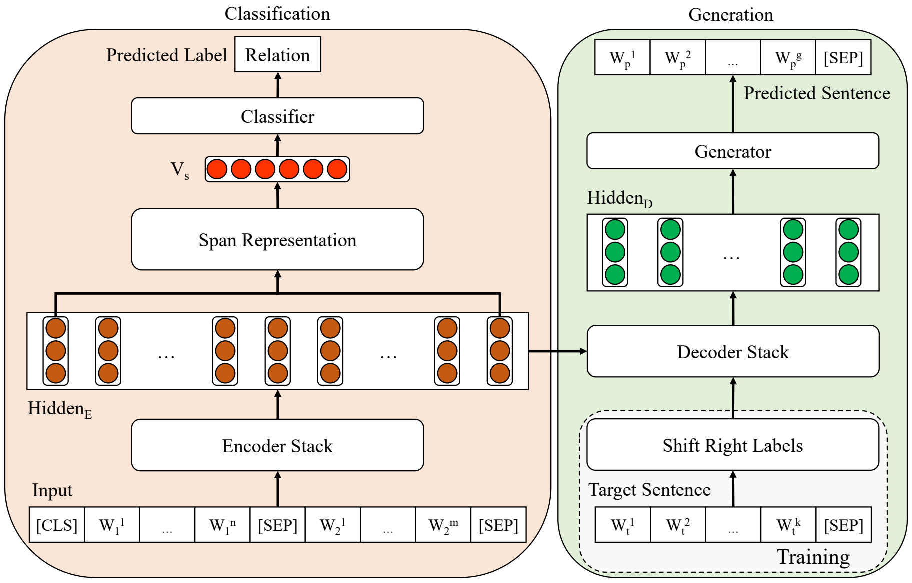

Figure 3: The architecture of CG-T5 model we proposed.

4.1 Classification Module

In the encoding layer, consistent with the input of the traditional classification model like BERT, we first encode two discourse units (DU$_1$ and DU$_2$) into a single string as follows.

$$S = [CLS]W_1[SEP]W_2[SEP] \tag{1}$$

where $W_1 = \{w_1^1, w_1^2, ..., w_1^n\}$ and $W_2 = \{w_2^1, w_2^2, ..., w_2^m\}$ represent the the token sequences of DU$_1$ and DU$_2$, respectively.

Then, we send the input (S) to the Encoder Stack (i.e., the encoder of T5) to obtain the encoder hidden states ($Hidden_E$) as follows.

$$Hidden_E = Encoder_{CG-T5}(S) \tag{2}$$

Since T5 does not use the state at the position of **[CLS]** for pre-training tasks like BERT, we use the Endpoint[4] of the span that combines the hidden states of the head (H_0) and tail (H_{-1}) for representing two DUs, as shown in Equation 3. Then we feed the output of the span representation layer V_s into a linear layer with softmax to classify the implicit discourse relation (r) between two DUs, as shown in Equation 4 and 5 .

$$V_s = Con(H_0, H_{-1}) \tag{3}$$

$$P = SoftMax(Linear(V_s)) \tag{4}$$

$$r = argmax(P) \tag{5}$$

4.2 Generation Module

There are two inputs in the decoding layer when training: the hidden state of the encoding layer ($Hidden_E$) and the golden target sentence (T). We take the T as the same style in the encoding layer as follows.

$$T = W_t[SEP] \tag{6}$$

where $W_t = \{w_t^1, w_t^2, ..., w_t^k\}$ represents the token sequence of the target sentence. Then we feed them into the decoding layer to get the decoder hidden states ($Hidden_D$) as follows.

$$Hidden_D = Decoder_{CG-T5}(T, Hidden_E) \tag{7}$$

Finally, we use a linear layer generator with softmax to produce the predicted target sentence. When testing, our model generates the predicted sentence $W_p = \{w_p^1, w_p^2, ..., w_p^g\}$ according to the input of two DUs.

4.3 Joint Learning Module

We joint learning the above two modules. The loss function for the classifier ($Loss_{cla}$) and generator ($Loss_{gen}$) is cross-entropy loss, and the total loss ($Loss$) is the sum of the two losses as follows.

$$Loss = Loss_{cla} + Loss_{gen} \tag{8}$$

[4] We have evaluated our model using various span representations (Toshniwal et al., 2020), including Average Pooling, Attention Pooling, Endpoint, Coherent, etc., and the Endpoint achieves the best performance. The reason is that this representation not only obtains the representation of [CLS] (the head) commonly used in the pre-training model to represent the Span for classification, but also considers the last hidden state information (the tail) that is closest to the generation module.

5 Experimentation

In this section, we first introduce the datasets and experimental settings. Then, we evaluate our model CG-T5 on MCDTB and PDTB.

5.1 Datasets and Experimental Settings

We mainly evaluate our model on two popular discourse relation datasets Chinese MCDTB and English PDTB. First of all, considering the macro discourse unit is longer, and the connection between discourse units is more obscure in Chinese, we first conduct the experiments on MCDTB. Then, we conduct the experiments on PDTB, one of the most popular discourse relation corpus in English, to verify the generality of our model. Besides, we also conduct the experiments on another Chinese dataset CDTB and the results are shown in Appendix D.3.

MCDTB: following previous work (Jiang et al., 2019; Sun et al., 2020), we use the same dataset division and five-fold cross-validation for the experiments.

PDTB: following previous work (Ji and Eisenstein, 2015; Kim et al., 2020), we adopt the most-used dataset splitting PDTB-Ji that takes the sections 2-20 as the training set, 0-1 as the development set, and 21-22 for testing.

We use Pytorch and Huggingface (Wolf et al., 2020)[5] as a deep learning framework, and the key parameter settings of our model are described in Appendix C. Since there is no official Chinese T5 model, we use the parameter weights provided by the third party[6]. It is a T5 (base) model with 12-layer encoders and 12-layer decoders trained by automatic summarization task on about a 30G corpus. In English, we use the official parameter weight[7] of T5 (base) for our CG-T5 model. Besides, we use AllenNLP instead of LTP tools to extract the core information from discourse units in English.

Following previous work, we use Micro-F1 and Macro-F1 to evaluate the IDRR models on the top-level and second-level class. In MCDTB, there are 3 classes on the top-level and 15 types on the second-level. In PDTB, there are 4 classes on the top-level and 11 types on the second-level[8].

Model	**Second-Level**		**Top-Level**	
	Micro-F1	**Macro-F1**	**Micro-F1**	**Macro-F1**
MSRN	54.18	14.92	65.62	50.10
STGSN	56.71	15.68	67.63	57.87
DAGGNN	-	-	70.01	55.38
BERT (base)	61.67	25.12	70.25	64.21
GPT-2 (Q_1)	51.81	17.29	60.86	49.81
Name (cla)	60.98	31.09	69.78	63.66
Name (gen)	61.69	31.50	70.67	64.27
Rel-Exp (cla)	61.13	30.48	70.76	64.66
Rel-Exp (gen)	60.02	28.53	70.62	64.59
Exp-Rel (cla)	60.44	27.42	70.73	64.53
Exp-Rel (gen)	61.02	27.70	70.29	64.30
Q_2 (cla)	61.04	29.65	70.21	64.02
Q_2 (gen)	60.47	27.20	69.59	63.24
Q_1 (cla)	**63.61**	**31.94**	**72.57**	**66.43**
Q_1 (gen)	63.12	31.80	72.43	66.24

Table 3: The performance comparison on Chinese MCDTB (five-fold cross-validation). Note that the performance of our model reported in the table is that of the output of classification (cla) and generation (gen) with the joint modeling.

5.2 Experimentation on MCDTB

To exhibit the effectiveness of our CG-T5 model, we select the following strong baselines: 1) **MSRN** (Zhou et al., 2019): it uses Support Vector Machine (SVM) as a classifier and uses word vectors and structural features to enhance the discourse unit representation. 2) **STGSN** (Jiang et al., 2019): it proposes a structure and topic gated semantic network for enhancing the discourse representation. 3) **DAGGNN** (Sun et al., 2020): it is a GCN-based neural network on the discourse pair graph to enhance the interaction between DUs. 4) **BERT** (Devlin et al., 2019): we view discourse relation recognition as text pair classification and choose BERT (base) at the same scale as our model for fair comparison and real-world application. 5) **GPT-2** (Radford et al., 2019): we also add the representative generation model GPT-2 into baselines.

Table 3 shows the performance of our CG-T5 (the bottom ten lines using different forms of the target sentence) and other baselines on MCDTB. The pre-trained BERT performs better than the other classification-based baselines. However, there is still a significant gap between the performance of the traditional generation model GPT-2 (Q_1) and that of the classification model. It indicates that the traditional generation model is not suitable to recognize implicit discourse relation.

In addition, CG-T5 using the relation name (Name) and the explanation of relation (Rel-Exp and Exp-Rel) as the target sentence achieve similar performance. In particular, Table 3 shows that our

[5] https://huggingface.co/transformers/

[6] https://github.com/renmada/t5-pegasus-pytorch/

[7] https://huggingface.co/t5-base/tree/main

[8] We follow (Ji and Eisenstein, 2015) to exclude 5 minor second-level classes in our experiments because none of these classes appear in the test or dev sets.

generation model and classification model with relation name (Name) and joint learning reach 31.09 and 31.50 at Second-Level in Macro-F1, significantly gain 5.97 and 6.3 improvement, respectively, in comparison with BERT, which demonstrates this form recognizes classes with fewer samples better.

Our CG-T5 using the question sentence guided by the first DU (Q_1) achieves the best performance and improves the fine-grained (15 types) IDRR up to 1.94 and 6.82 in Micro-F1 and Macro-F1, respectively, in comparison with the best baseline BERT. There are two following reasons, as mentioned in Section 3.1. First, using the question sentence extracting core information from only one DU to construct the target sentence can reduce cascading errors. Second, the question sentence highlights the meaning of discourse relation and difference between various types, which enables the model to understand the discourse relation better than the other two forms. Besides, according to our statistics, the average length of Q1's target sentence is 45.44 words, and that of Exp-Rel's target sentence is 87.73 words. It is more difficult for the model to learn the relation from the Exp-Rel because the relation description takes up a smaller proportion in the target sentence. Compared with generating only relation name (Name) that lacks core information of DUs and generating explanation and relation (Exp-Rel) that doesn't pay enough attention to the relation description, using questions as the target sentence (Q1) can balance learning discourse relation description and the core information of discourse unit, achieving the best performance.

In addition, we also notice the performance of the generation model (gen) is slightly lower than that of the classification model (cla). The reason may be that the pre-trained T5 in Chinese is different from the vanilla T5 in English.

However, the performance of CG-T5 using the question sentence guided by the second DU (Q_2) is not good as that guided by the first DU (Q_1). We believe that the reason is the uneven distribution of the nuclearity that the first DU is usually the nucleus. In MCDTB, 63.94% of the first DU is the nucleus, 2.52% of the second DU is the nucleus, and 33.54% of the discourse relation pairs are equally important. Therefore, the model (Q_1) using the question sentence guided by the first DU composing of more important information can better grasp the connection between the two DUs to recognize discourse relation better.

5.3 Experimentation on PDTB

To evaluate the model generality, we also conduct experiments on another dataset PDTB and select six strong baselines for fair comparison as follows: 1) **Bai2018** (Bai and Zhao, 2018): it uses different grained text representations on ELMO to enhance DU representation. 2) **Bai2019** (Bai et al., 2019): it adds the memorizing mechanism to their previous work (Bai and Zhao, 2018). 3) **Nguyen2019** (Nguyen et al., 2019): it uses multi-task learning via label embedding. 4) **Guo2020** (Guo et al., 2020): it is a knowledge-enhanced attention neural network that enhances the interaction between discourses by introducing external knowledge. 5) **He2020** (He et al., 2020): it translates the discourse relations in low-dimensional embedding space and propose a joint learning framework with the semantic features of arguments. In addition, we also reproduce 6) **BERT (base)**[9] at the same scale as our model for fair comparison.

Model	**Second-Level**		**Top-Level**	
	Micro-F1	**Macro-F1**	**Micro-F1**	**Macro-F1**
Bai2018	48.22	-	-	51.06
Bai2019	49.15	-	60.69	52.19
Nguyen2019	49.95	-	-	53.00
Guo2020	-	-	57.25	47.90
He2020	-	-	59.94	51.24
BERT (base)	51.88	36.10	63.91	55.13
Name (cla)	48.03	34.60	61.12	53.21
Name (gen)	48.22	35.21	61.21	53.57
Rel-Exp (cla)	51.20	37.43	64.00	56.29
Rel-Exp (gen)	51.20	35.86	63.14	54.78
Exp-Rel (cla)	51.68	36.17	63.52	55.70
Exp-Rel (gen)	51.40	34.38	62.95	55.56
Q_2 (cla)	51.68	37.49	63.91	56.10
Q_2 (gen)	51.49	36.76	64.20	55.69
Q_1 (cla)	52.17	37.53	63.23	55.35
Q_1 (gen)*	**53.13**	**37.76**	**65.54**	**57.18**

Table 4: The performance comparison on PDTB. Note that the performance of our model reported in the table is that of the output of classification (cla) and generation (gen) with the joint modeling. Superscript * indicates the model is significantly superior to the BERT model with a p-value < 0.01.

Table 4 shows the performance of our model on PDTB-Ji at the top-level and second-level classes. Consistent with Kim et al. (2020)'s conclusion, BERT is indeed the best baseline, achieves 51.88 and 36.10 in Micro-F1 and Macro-F1 at the second-level and 63.91 and 55.13 in Micro-F1 and Macro-F1 at the top-level, respectively.

Similar to the performance on MCDTB, our CG-T5 with Q_1 achieves the best performance and almost all its F1-measures integrating classification

[9] We use the same parameter settings in Kim et al. (2020).

or generation mechanism is better than the strong baseline BERT. In particular, compared with BERT, our generation model (gen) improve the Micro-F1 and Macro-F1 in 11-way classification by 1.25 and 1.66, and in 4-way classification by 1.63 and 2.05, respectively. These results indicate that our model can achieve the best performance under the same order of magnitude of model parameters, proving the effectiveness of our model on English IDRR.

Model	Comp.	Cont.	Exp.	Temp.
Bai2018	47.85	47.85	70.60	36.87
Bai2019	47.15	55.24	70.82	38.20
Nguyen2019	48.44	56.84	73.66	38.60
Guo2020	43.92	57.67	73.45	36.33
He2020	47.98	55.62	69.37	38.94
BERT	47.19	**59.20**	72.63	41.51
Q_1 (gen)	**55.40**	57.04	**74.76**	**41.54**

Table 5: The results of different relations on PDTB (top-level multi-class classification).

We further analyze the performance of the 4-way classification on PDTB at the top-level on different classes, shown in Table 5. We notice that our improvement mainly comes from the relation $Comparison$, where there is a significant increase of 8.21%. The reason is that the two words "negate" and "opposite" in the target sentence (question) of the $Concession$ and $Contrast$ relations can more accurately represent their meanings, which helps the model better recognize the two relations.

6 Analysis

To further demonstrate the effectiveness of our CG-T5, we choose the 1st-fold data set of MCDTB as an example to further analyze the following two parts: the ablation experiments of joint modeling and the text generation quality assessment.

6.1 Ablation Study

We conduct ablation studies to evaluate CG-T5 with joint modeling of classification and generation, as shown in Table 6. We can find that both the only-generation (only-gen) model (i.e., vanilla T5) and the only-classification (only-cla) model achieve better performance than GPT-2 and BERT, respectively. It is worth noting that our CG-T5 further improves the performance of the classification model (cla) and generation model (gen) simultaneously by joint modeling, which proves that our CG-T5 model can integrate the advantages of these two models, and achieve better performance. In addition, we find that the improvement of the generation model is more significant than the classification model in joint modeling architecture.

Model	Second-Level		Top-Level	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
Q_1 (cla)	**62.75**	31.25	**71.79**	**65.59**
Q_1 (gen)	62.36	**31.80**	71.41	65.22
only-cla	62.20	29.47	70.48	63.83
vanilla T5	58.73	23.27	67.70	60.55
BERT	60.59	24.62	69.32	62.47
GPT-2	51.55	16.19	60.59	48.87

Table 6: The ablation experiments of our best model (Q_1) on MCDTB (the 1st-fold data set).

Since we notice that the performance of the generation model is similar to that of the classification model in Table 3, we further analyze the difference between the two models with Q_1 in the joint framework, as shown in Table 7. Although the performance gap between the two outputs is not significant and the agreement rate is 91.34%, its oracle value (as long as one of two outputs is correct, the final result is correct) has a further improvement. Specifically, we find the generation model performs better for the minor classes with fewer samples, while the classification model performs better for the major classes. It demonstrates that our model can effectively integrate the classification model and the generation model to complement each other's advantages.

Joint Model	Second-Level	Top-Level
Q_1 (gen)	62.36	71.41
Q_1 (cla)	62.75	71.79
Oracle	65.00	74.19
Agreement	91.34%	93.51%

Table 7: The Micro-F1 comparison between the classification and generation of our best model (Q_1) on MCDTB (the 1st-fold data set).

6.2 Text Generation Quality Assessment

We select the Rouge scores[10] commonly used in the generation task to evaluate the generation quality of our best model (Q_1), as shown in Table 8. Due to the more advanced generation model architecture, our model achieves better performance than the traditional GPT-2. Moreover, since joint modeling with the classification can constrain the generation, our model CG-T5 outperforms the vanilla

[10] https://github.com/JialeGuo/py_rouge_zh

T5 model and achieves excellent performances of 83.99, 79.69, 77.22, and 82.97 on Rouge-1/2/3/L. It proves that our model can effectively extract the core information of discourse units and generate the question sentence based on the corresponding discourse relation.

Q_1	Rouge-1	Rouge-2	Rouge-3	Rouge-L
GPT-2	75.83	69.48	65.86	74.33
vanilla T5	83.34	78.70	76.16	82.18
CG-T5 (gen)	**83.99**	**79.69**	**77.22**	**82.97**

Table 8: The generated quality of our model (Q_1) on MCDTB (the 1st-fold data set).

7 Conclusion

In this paper, we regard IDRR as a generation task to generate the target sentence, which can introduce prior knowledge to understand discourse relations deeper. Moreover, we propose a joint learning model, CG-T5, to integrate the classification model and the generation model and design three forms of the target sentence and the construction method to automatically extract core content from the large DUs. The experimental results both on MCDTB and PDTB datasets show that our CG-T5 outperforms the SOTA baselines. In future work, we will focus on how to construct more robust and automatic target sentences and how to integrate the question generation and the answer generation to recognize discourse relations better.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (No. 61836007, 61772354 and 61773276.), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 571–583.

Hongxiao Bai, Hai Zhao, and Junhan Zhao. 2019. Memorizing all for implicit discourse relation recognition.

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 173–184.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 933–941.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACL-HLT)*, pages 4171–4186.

Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 7822–7829.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 547–558.

Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. 2020. TransS-driven joint learning architecture for implicit discourse relation recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 139–148.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics (TACL)*, 3:329–344.

Feng Jiang, Peifeng Li, and Qiaoming Zhu. 2019. Joint modeling of recognizing macro Chinese discourse nuclearity and relation based on structure and topic gated semantic network. In *Natural Language Processing and Chinese Computing (NLPCC)*, pages 276–286.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A macro-level Chinese discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3493–3504.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5404–5414. Association for Computational Linguistics.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1299–1308.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5849–5859. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1340–1350.

Yancui Li, Wenhe Feng, Jing Sun, Fang Kong, and Guodong Zhou. 2014. Building Chinese discourse corpus with connective-driven dependency tree structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2105–2114. Association for Computational Linguistics.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1233.

Linh The Nguyen, Linh Van Ngo, Khoat Than, and Thien Huu Nguyen. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4201–4207.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 2961–2968. European Language Resources Association.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2263–2270. The Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics (TACL)*, 8:264–280.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-propagative attention learning for implicit discourse relation recognition. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. International Committee on Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics (IWCS)*, pages 188–199.

Zhenhua Sun, Feng Jiang, Peifeng Li, and Qiaoming Zhu. 2020. Macro discourse relation recognition via discourse argument pair graph. In *Natural Language Processing and Chinese Computing (NLPCC)*, pages 108–119.

Shubham Toshniwal, Haoyue Shi, Bowen Shi, Lingyu Gao, Karen Livescu, and Kevin Gimpel. 2020. A cross-task analysis of text span representations. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 166–176, Online. Association for Computational Linguistics.

Bonnie Webber, Andrei Popescu-Belis, and Jörg Tiedemann, editors. 2017. *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020.

Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6953–6963. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5021–5031. Association for Computational Linguistics.

Sheng Xu, Peifeng Li, Fang Kong, Qiaoming Zhu, and Guodong Zhou. 2019. Topic tensor network for implicit discourse relation recognition in Chinese. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 608–618. Association for Computational Linguistics.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2230–2235. Association for Computational Linguistics.

Yi Zhou, Xiaomin Chu, Qiaoming Zhu, Feng Jiang, and Peifeng Li. 2019. Macro discourse relation classification based on macro semantics representation. *Journal of Chinese Information Processing*, 33(3):1–7.

Appendices

A Details of Example

DU1: The Ningbo Free Trade Zone, with a total area of 2.3 square kilometers, has achieved fruitful results after three years of construction. Ningbo Free Trade Zone is one of the 13 free trade zones in China. It was established with the approval of the State Council in 1992. At present, the various functions of the bonded area have begun to take shape, and the level of development is among the best in China's bonded areas.

DU2: According to statistics, by the end of last year, the Ningbo Free Trade Zone had completed a total of US$812 million in import and export trade, and the import and export trade volume through the customs of the Free Trade Zone last year alone reached US$365 million. At present, there are ten bonded warehouses in the zone with a storage area of more than 80,000 square meters; last year alone, the zone has stored goods of 2.627 billion yuan. With the adjustment of China's special policies outside the bonded area since April this year, the bonded area's certificate and tax exemption, and the stability advantages of the bonded policy have become more obvious. A large number of domestic and foreign industrial processing projects have successively settled in the area. By the end of December last year, a total of 1,614 enterprises had been established in the zone, with a total investment of 1.2 billion U.S. dollars, of which 260 were foreign-invested enterprises, and the actual utilization of foreign capital was 113 million U.S. dollars. In addition, many domestic enterprises have also connected with the international market through the bonded zone. In order to complement the free trade zone in terms of operating mechanism, the Ningbo Free Trade Zone took the lead in implementing the trial one-stop management system of direct registration of enterprises in accordance with the law in China, which was handled at one time. At the same time, the bonded zone has vigorously promoted the construction of the information expressway network system in the zone to create good supporting conditions for the realization of modern management. (Finish)

B Complete Templates of Target Sentences

Table 9 and Table 10 show the target sentences guided by DU1(Q_1) in MCDTB and PDTB on different relations, respectively.

Relation	Target Sentence Template
Joint	Is there anything similar to CI_1? 和CI_1类似的事情还有吗?
Sequence	What happened after CI_1? 在CI_1之后，发生了什么?
Progression	What goes further than CI_1? 比CI_1更进一步的事情是什么?
Contrast	What is the contrast with CI_1? 和CI_1这件事相对比的是什么?
Supplement	What else do you want to add to CI_1? 对于CI_1这件事还有什么要补充的?
Result-Cause	What is the cause of CI_1? 导致CI_1这件事发生的原因是什么?
Cause-Result	What is the result of CI_1? CI_1这件事所导致的结果是什么?
Background	What's the background of CI_1? CI_1这件事的背景是什么?
Behavior-Purpose	What's the purpose of CI_1? CI_1这件事的目的是什么?
Purpose-Behavior	What behavior was did for CI_1? 为了CI_1这件事，做了什么行为?
Elaboration	What is the detailed explanation for CI_1? 对于CI_1这件事可以详细的解释一下吗?
Summary	What's the summarization for CI_1? 对于CI_1这件事可以总结一下吗?
Evaluation	What's the evaluation on CI_1? 对于CI_1这件事是怎么评价的?
Statement-Illustration	Take an example for CI_1? 对于CI_1这件事，举一个例子?
Illustration-Statement	What does CI_2, the example, mean? 对于CI_2这个例子，它想说明什么?

Table 9: The complete template of Q_1 in MCDTB. CI_1 is the core information of DU_1.

Relation	**Target Sentence Template**
Concession	What event negates part of CI_1?
Contrast	What is the opposite of CI_1?
Cause	What is the cause or result of CI_1?
Pragmatic cause	What is the justification of CI_1?
Conjunction	Is there anything to add about CI_1?
Instantiation	Can you give me an example of CI_1?
Alternative	Can you replace CI_1 with something else?
List	What is the other list member for CI_1?
Restatement	Can you explain CI_1 in detail?
Asynchronous	What happened after CI_1?
Synchrony	What happened in synchrony with CI_1?

Table 10: The complete template of Q_1 in PDTB. CI_1 is the core information of DU_1.

C Details of Experimental Settings

In MCDTB (Jiang et al., 2018), there are 720 documents annotated with 3 classes and 15 types. The distribution of relation classes in MCDTB is shown in Table 11.

PDTB 2.0 (Prasad et al., 2008) annotated 16K implicit relations in over 2K Wall Street Journal (WSJ) articles annotated with 4 classes and 16 types. We only select 11 types following previous

Class (Top-level)	Type (Second-level)
Causality	Result-Cause, Cause-Result, Background, Behavior-Purpose, Purpose-Behavior
Coordination	Joint, Sequence, Progression, Contrast, Supplement
Explanation	Elaboration, Summary, Evaluation, Statement-Illustration, Illustration-Statement

Table 11: The set of two-level discourse relations on MCDTB.

work (Ji and Eisenstein, 2015). The distribution of relation classes in PDTB is shown in Table 12.

Class (Top-level)	Type (Second-level)
Comparison	Concession, Contrast
Contingency	Cause, Pragmatic cause
Expansion	Conjunction, Instantiation, Alternative, List, Restatement
Temporal	Asynchronous, Synchrony

Table 12: The set of two-level discourse relations on PDTB.

The key parameters of our experimental model are shown in Table 13.

Parameter	Value
The learning rate	1e-4
The training epoch	10
The batch size	2
The random seed	42
The optimizer	Adam
The hidden size of pre-training model	768
The maximum length of input	512
The maximum length of the target sentence	200

Table 13: The key parameter in experimental settings.

D More Details of Experiments on PDTB and CDTB

D.1 Error Analysis on PDTB

Figure 4 shows the confusion matrix of CG-T5 (Q_1) on 11 relation types. We find that the five major types ($Contrast$, $Cause$, $Conjunction$, $Instantiation$, and $Restatement$), whose samples are greater than 100 in the test set, achieve higher performance (Accuracy: >50). Although the instances of two types ($Pragmatic\ cause$ and $Synchrony$) cannot be recognized due to too few samples, our model with the target sentence still improves the other types (e.g., $Alternative$, $Asynchronous$, and $List$). In addition, we also find that the main errors come from the confusion between the relations $Synchrony$ and $Conjunction$, the relations $Pragmatic\ cause$ and $Cause$, the relations $Pragmatic\ cause$ and $Restatement$, and the relations $List$ and $Conjunction$. They are difficult to distinguish even if our model takes the question as the target sentence due to their similar semantics.

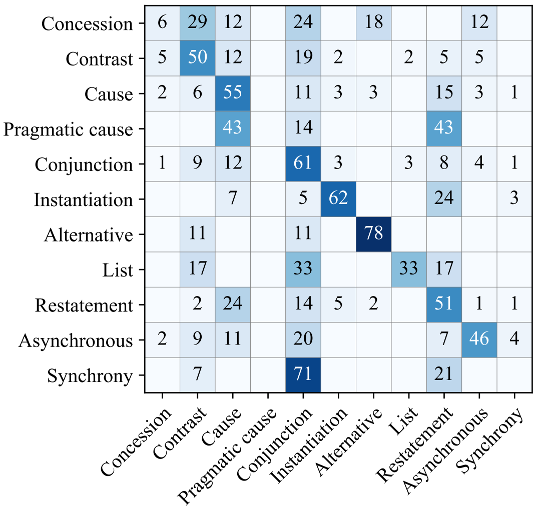

Figure 4: The confusion matrix of CG-T5 model (Q_1) in PDTB. The X-axis is the predicted value, and the Y-axis is the true value. The values are normalized (%).

D.2 Case Study on PDTB

Table 14 shows the two prediction outputs of our joint model CG-T5 (Q_1) and the prediction output of BERT for a sample. It can be seen that BERT can't distinguish the relations $Pragmatic\ cause$ and $Instantiation$ well because the form of the discourse unit in the two relation types is similar. However, our CG-T5 model accurately generates the target sentence and better grasp the difference between the two relations through joint learning. Therefore, both the classification and the generation of our CG-T5 model are correct due to the classification and generation modules can complement each other.

D.3 Experiments on CDTB

In CDTB, following the previous work (Xu et al., 2019)[11], we use 446 articles as the training set and 49 articles as the test set.

we reproduce the following baselines: **Bi-LSTM**, **CNN**, **GCN** (Dauphin et al., 2017) and

[11] We contacted the author and use the latest CDTB V2.0 instead of CDTB V1.0, which corrected some annotated errors and removed five problematic articles (No. 150, 208, 287, 300, and 644).

DU1: TV programmers could let audiences vote on different endings for a movie
DU2: Fox Broadcasting experimented with this concept last year when viewers of "Married ... With Children" voted on whether Al should say "I love you" to Peg on Valentine's Day
True relation: Expansion.Instantiation
True target sentence: Can you give me an example of TV programmers let audiences vote on different endings for a movie?
Relation predicted by BERT: Contingency.Pragmatic cause
Relation predicted by CG-T5: Expansion.Instantiation
The target sentence generated by CG-T5: Can you give me an example of TV programmers let audiences vote on different endings for a movie?

Table 14: The example of the prediction of our CG-T5 model with question sentence (Q_1) and BERT model in PDTB.

Xu19 (Xu et al., 2019). In addition, we also reproduce the **BERT (base)** model for a fairer comparison. The experiment setting parameters are the same as those on MCDTB. To be consistent with the previous work, we use the results of converting 17 types into four classes (top-level).

Model	**Caus.**	**Coor.**	**Elab.**	**Tran.**	**Micro-F1**	**Macro-F1**
Bi-LSTM	24.5	80.9	55.7	-	68.7	40.3
CNN	26.7	81.0	53.6	-	70.3	41.2
GCN	25.5	81.8	50.0	11.8	70.3	43.2
Xu19	30.8	81.5	56.2	15.4	71.0	46.3
BERT	44.1	84.6	67.6	25.0	76.7	55.7
Exp-Rel (gen)	39.5	85.5	**72.1**	18.2	76.6	53.8
Rel-Exp (gen)	39.0	**85.9**	71.0	24.0	76.7	55.0
Name (gen)	42.4	85.4	71.3	27.3	76.7	56.6
Q_1 (gen)	42.5	84.7	68.8	**31.6**	75.8	56.9
Q_2 (gen)	**48.6**	84.9	70.9	30.0	**76.9**	**58.6**

Table 15: Top-level multi-class classification results on CDTB. We report Micro-F1, Macro-F1 and Micro-F1 on each class (Causuality, Coordination, Explanation and Transition).

Table 15 shows that thanks to the large-scale pre-training tasks, the BERT model achieve 76.7 Micro-F1 and 55.7 Macro-F1, which is better than other SOTAs without pre-training tasks. It can be seen that our model with Q_2 achieves the best performance, higher 0.2 and 2.9 on Micro-F1 and Macro-F1 than BERT. It significantly improves 4.5 Macro-F1 scores in the Caus., which benefit from introducing prior knowledge.

Unlike the experimental results on the MCDTB, we notice that the model using Q_2 instead of Q_1 achieves the best performance. The reason may be the semantic difference between types within a class is not significant in the Q_1 form. For example, in the Caus., the difference between $Hypothesis$ and $Conditional$ relationship is more difficult to distinguish by the model using Q_1 than Q_2.