Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition

Hao Zhou[1], Man Lan[1,2,*], Yuanbin Wu[1], Yuefeng Chen[3] and Meirong Ma[3]

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[3]Shanghai Transsion Co., Ltd., Shanghai, China

`hzhou@stu.ecnu.edu.cn,{mlan,ybwu}@cs.ecnu.edu.cn`
`{yuefeng.chen,meirong.ma}@transsion.com`

Abstract

Due to the absence of connectives, implicit discourse relation recognition (IDRR) is still a challenging and crucial task in discourse analysis. Most of the current work adopted multitask learning to aid IDRR through explicit discourse relation recognition (EDRR) or utilized dependencies between discourse relation labels to constrain model predictions. But these methods still performed poorly on fine-grained IDRR and even utterly misidentified on most of the few-shot discourse relation classes. To address these problems, we propose a novel **P**rompt-based **C**onnective **P**rediction (**PCP**) method for IDRR. Our method instructs large-scale pre-trained models to use knowledge relevant to discourse relation and utilizes the strong correlation between connectives and discourse relation to help the model recognize implicit discourse relations. Experimental results show that our method surpasses the current state-of-the-art model and achieves significant improvements on those fine-grained few-shot discourse relation. Moreover, our approach is able to be transferred to EDRR and obtain acceptable results. Our code is released in `https://github.com/zh-i9/PCP-for-IDRR`.

1 Introduction

Discourse relations recognition (DRR) aims to identify the semantic relation between two discourse units (e.g., sentences or clauses, they are denoted as *Arg1* and *Arg2* respectively). DRR is essential to many natural language processing (NLP) downstream tasks involving more context, such as document-level machine translation (Xiong et al., 2019) and machine reading comprehension (Mihaylov and Frank, 2019), and text summarization (Xu et al., 2020).

Compared with explicit discourse relation recognition (EDRR), due to the absence of explicit connectives, implicit discourse relation recognition (IDRR) is more challenging and attractive to researchers. Figure 1 shows an IDRR sample in the

---

Arg1: There will be a lot of malice.

Arg2: Men who have played hard all their lives aren't about to change their habits.

Relation: Contingency.Cause.Reason

Connective: because

Figure 1: An IDRR instance in PDTB 2.0. For IDRR, the *Connective* are not present in the original discourse context, but are assigned by the annotator according to semantic relations between *Arg1* and *Arg2*.

Penn Discourse Treebank 2.0 (PDTB 2.0) (Prasad et al., 2008).

Traditional machine learning based methods primarily relied on human-designed and shallow linguistic features (Lin et al., 2009; Zhou et al., 2010). With the rapid development of deep learning, a large number of methods prefer to utilize deep neural networks to recognize discourse relation between two arguments (Liu et al., 2016; Liu and Li, 2016; Dai and Huang, 2018). Recently, pre-trained language models (PLMs), such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), play a dominative role in many NLP tasks through contextual representation learning. Leveraging the success of PLMs, extensive research work design a wide variety of post-processing neural networks (Liu et al., 2020; Wu et al., 2021) to extract the semantic information by fine-tuning them on the task-specific dataset.

On the other side, previous work showed the accuracy of explicit discourse relation recognition achieved more than 93% using the discourse connective alone (Pitler and Nenkova, 2009), which reveals strong correlations between connectives and discourse relations. Therefore, more and more studies focus on implicit discourse relation recognition through connective prediction. Kishimoto et al. (2020) applied an auxiliary pre-training task, which

3848

utilized the representation of the [CLS] token to predict explicit connectives, to BERT using domain text containing explicit connectives. **Since** the idea of their method is contradictory with the original next sentence prediction (NSP) task of BERT and does not make use of the knowledge that BERT had learned through the masked language model (MLM) task. The experimental results showed that their method does not perform well on coarse-grained and fine-grained discourse relations. Kurfalı and Östling (2021) came up with a pipeline approach, where a connective selected from candidates and two implicit arguments were first integrated as *"Arg1 Connective Arg2"* sequence. Then this integrated sequence was input into an explicit discourse relation classifier trained in explicit data to identify implicit discourse relations in a distant supervision manner. **Since** there are a lot of connective candidates (about 100 in PDTB 2.0), and it requires a large number of model inferences for each sample. Obviously, the simple sequence does not make full use of the semantic knowledge embedded in the PLMs. Therefore, how to make better connective prediction becomes a crucial step for implicit discourse relation recognition.

Inspired by Schick and Schütze (2021b), we propose a novel **P**rompt-based **C**onnective **P**rediction (**PCP**) method for implicit discourse relation recognition. We exploit the advantage of prompt learning (Liu et al., 2021) to bridge the gap between connective prediction in the pre-training and fine-tuning stage and make better use of the knowledge of PLMs. Specifically, we **first** manually design templates matching natural language patterns, which elicit large-scale pre-trained language models for connective prediction using specific knowledge (in Section 3.2). **Secondly**, we select the less ambiguous connectives corresponding to each discourse relation based on the frequency of the connectives in the dataset to implement answer mapping (mapping predicted connectives to discourse relaiton labels) (in Section 3.3). Experimental results prove that the simple but proper template can outperform the current state-of-the-art (SOTA) model based on fine-tuning and achieve zero breakthroughs on few-shot fine-grained discourse relation. Furthermore, our approach based on connective frequency effectively avoids the trouble of manually selecting answer words and takes full advantage of the relevance between discourse relations and connectives.

Our contributions are summarized as follows:

- We propose a Prompt-based Connective Prediction (PCP) method for implicit discourse relation recognition, which achieves SOTA performance on the PDTB 2.0 dataset and CoNLL-2016 Shared Task as well.

- The method we proposed breaks the bottleneck of previous work on the few-shot fine-grained discourse relation of PDTB 2.0.

- Our approach can be easily transferred from IDRR to EDRR, and we have experimentally demonstrated that our approach still performs well for EDRR.

2 Related Work

2.1 Implicit Discourse Relation Recognition

Previous methods based on machine learning primarily relied on manually-designed linguistic features (Lin et al., 2009; Zhou et al., 2010). With the rapid development of deep learning, most of methods prefer to utilize deep neural networks, such as RNN (Liu et al., 2016), CNN (Varia et al., 2019), or LSTM (Liu and Li, 2016; Lan et al., 2017; Dai and Huang, 2018), to recognize discourse relation between two arguments.

Along with the booming development of pre-trained language models, most work designs various post-processing neural networks for information interaction to improve overall performance by fine-tuning their models (Van Ngo et al., 2019; Wu et al., 2020). For example, Liu et al. (2020) combines different levels of representation learning modules to address implicit discourse relation recognition. Wu et al. (2021) designs a label attentive encoder to learn the global representation of an input instance and its level-specific context and employs a label sequence decoder to output the predicted labels in a top-down manner.

Recently, several methods have emerged for implicit discourse relation recognition through connective prediction (Kishimoto et al., 2020; Kurfalı and Östling, 2021). They use efficient pre-trained language models and additional explicit data, but their performance is less than ideal. We think that their methods are contradictory to the original pre-training task, and there is a distribution difference between the explicit and implicit data.

2.2 Prompt-based Models

Prompt-based methods have received considerable attention with the emergence of GPT-3 (Brown

et al., 2020). A series of research work (Schick and Schütze, 2021b; Chen et al., 2022b) have demonstrated that prompt learning can effectively stimulate knowledge from PLMs compared with conventional fine-tuning. Therefore, many prompt-based models have been proposed and they achieved outstanding performance in widespread NLP tasks, such as text classification (Schick and Schütze, 2021a; Hu et al., 2021), text matching (Jiang et al., 2022), named entity recognition (Cui et al., 2021; Chen et al., 2021), and relation extraction (Han et al., 2021; Chen et al., 2022a), particularly on few-shot and zero-shot settings. The main idea of their approaches is to convert the downstream task into a cloze-style task that is closer to the pre-training step of PLMs with prompts.

As far as we know, we are the first to apply prompt learning (Schick and Schütze, 2021b) for implicit and explicit discourse relation recognition task.

3 The Prompt-based Connective Prediction Approach

In this section, we will introduce our prompt-based connective prediction (PCP) approach to recognize implicit discourse relations in detail. Figure 2 shows the overall architecture of our model.

3.1 Connective Prediction with Prompt

For convenience and clarity, we discuss our method with a simple template "*Arg1* <mask> *Arg2*.". In this template, *Arg1* and *Arg2* correspond to two arguments in the original texts respectively (as shown in Figure 1), and the symbol <mask> represents the masked token in place of their connective. Given a pair of arguments x_{arg1} and x_{arg2}, we transfer them to x_{prompt} with the template:

$$x_{prompt} = \mathbf{T}(x_{arg1}, x_{arg2}) \quad (1)$$

where $\mathbf{T}$ represents template function. Then we feed x_{prompt} to the RoBERTa (Liu et al., 2019) model to obtain the representation of <mask> token $\boldsymbol{h}_{mask}$:

$$\boldsymbol{h}_{mask} = \mathbf{RoBERTa}(x_{prompt}) \quad (2)$$

Same as MLM task in pre-training step, we input $\boldsymbol{h}_{mask}$ into MLMHead model, which predicts scores of the language modeling head (scores for each vocabulary token before $SoftMax$), and acquire the output $\boldsymbol{e}_{mask}$:

$$\boldsymbol{e}_{mask} = \mathbf{MLMHead}(\boldsymbol{h}_{mask}) \quad (3)$$

We select the token with the highest score in $\boldsymbol{e}_{mask}$ as the connective predicted by the model. During the training, we use cross-entropy to calculate the loss between the model prediction and the golden connective:

$$Loss = CE(target, SoftMax(\boldsymbol{e}_{mask})) \quad (4)$$

where CE is cross-entropy loss function and $target$ is golden connective (mapped from golden discourse relation label) index in the vocabulary. Finally, we are mapping the predicted connective (e.g., *because*) to the corresponding discourse relation label (e.g., *Cause*).

Then, we will detail the two above-mentioned crucial parts in our methods, **prompt construction** (Section 3.2) and **answer search** (Section 3.3).

3.2 Prompt Construction

For prompt-based methods, one key challenge is to find an appropriate template matching the target task. In this work, we transfer implicit discourse relation recognition to a connective prediction task and utilize manual search to identify suitable templates for the connective prediction task. Moreover, we use the accuracy of the top-level senses of the PDTB 2.0 development set as the main metric to evaluate different templates. Table 1 shows our designed templates based on the frequent position of connectives and natural language instructions.

In PDTB 2.0 dataset, for all implicit discourse relation samples, their connective position is between *Arg1* and *Arg2*. In other words, they all satisfy the "*Arg1* connective *Arg2*" sequence order. Therefore, we designed the first two templates in Table 1 based on the position of connectives. Perhaps because there are some connectives that may not be appropriate for the sentence form "That's *connective ...*" (e.g., *but*), the accuracy value of the second template is lower than the first.

Inspired by (Schick and Schütze, 2021b), we design templates conforming to natural language patterns to instruct PLMs to output the content we want to obtain. In this work, our target is connective prediction using PLMs and then mapping connective to the final discourse relation label. Therefore, we use the template "The connective (or conjunction) between Arg1 and Arg2 is <mask>." to direct PLMs for connective prediction. Meanwhile, we also attempt to place segment token </s> between two arguments and instruction sentences, which is to allow the model to take the coherence of overall
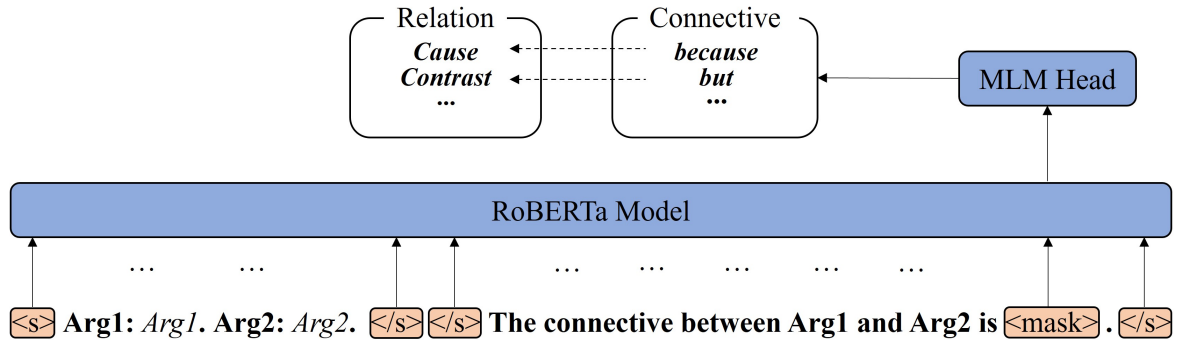
Figure 2: The model architecture of our Prompt-based Connective Prediction method.

Template	PDTB Top-level Acc.
Arg1 <mask> *Arg2*.	69.23
Arg1. That's <mask> *Arg2*.	67.37
Arg1: *Arg1*. Arg2: *Arg2*. The connective between Arg1 and Arg2 is <mask>.	71.09
Arg1: *Arg1*. Arg2: *Arg2*. The conjunction between Arg1 and Arg2 is <mask>.	71.43
Arg1: *Arg1*. Arg2: *Arg2*.</s></s>The connective between Arg1 and Arg2 is <mask>.	71.68
Arg1: *Arg1*. Arg2: *Arg2*.</s></s>The conjunction between Arg1 and Arg2 is <mask>.	72.27

Table 1: Searching templates on *RoBERTa-large*. The symbol </s> and <mask> represents segment and mask token in RoBERTa tokenizer.

inputs into account. We can see that the last template achieves the best performance on the accuracy value, and our final experimental results are also based on this template (as shown in Figure 2).

3.3 Answer Search

For other prompt-based methods, like Schick and Schütze (2021a), they regarded "terrible", "bad", "okay", "good", and "great" as the answer word to the rating that a customer gave to a restaurant on a 1- to 5-star scale based on their review's text. They artificially selected answer words that appear to fit the golden label, but these words may not fit well with most of the samples in the dataset.

For implicit discourse relation data, each sample has been annotated with the connective appropriate to this sample on PDTB 2.0 and CoNLL16 dataset (detailed in Section 4.1). As a result, we select the **most frequent** and **less ambiguous** connectives as the answer words of the corresponding discourse relations based on the frequency of connectives in each discourse relation and the repetition rate of some connectives in different discourse relations. At the same time, to unify the model's input and output and facilitate the mapping between answer words and discourse relation labels, we select those connectives that are tokenized as a single token by

Top-level	Second-level	Answer Set
Comparison	Concession	although, nevertheless
	Contrast	but, however
Contingency	Cause	because, as, so, consequently, thus
	Pragmatic cause	since
Expansion	Alternative	instead, rather, or
	Conjunction	and, also, furthermore
	Instantiation	instance, example
	List	first
	Restatement	indeed, specifically
Temporal	Asynchronous	then, subsequently, previously, earlier, after
	Synchrony	meanwhile

Table 2: Mapping between implicit discourse relation labels and connectives on PDTB 2.0 dataset, which has four top-level and 11 second-level senses. The answer set of top-level senses is a union set of second-level.

the RoBERTa tokenizer.

Table 2 and 3 show the final answer sets we selected based on the above conditions on PDTB 2.0 and CoNLL16 dataset. It is worth mentioning that some second-level senses have subtypes (also be called the third-level senses). For examples, the second-level sense *Cause* has two subtypes *Reason* and *Result*. We select the connectives for each subtype and merge them into the second-level senses. Moreover, for those samples whose connectives are

Cross-level Senses	Answer Set
Comp.Concession	although, nevertheless
Comp.Contrast	but, however
Cont.Cause.Reason	because, as
Cont.Cause.Result	so, thus, consequently
Cont.Condition	if
Exp.Alternative	unless, or
Exp.Alternative.Chosen alternative	instead
Exp.Conjunction	and, also, furthermore
Exp.Exception	rather
Exp.Instantiation	instance, example
Exp.Restatement	specifically
Temp.Asynchronous.Precedence	then, subsequently
Temp.Asynchronous.Succession	previously, earlier, after
Temp.Synchrony	meanwhile
EntRel	none

Table 3: Mapping between implicit discourse relation labels and connectives on CoNLL16 dataset which has 15 cross-level senses. The *EntRel* sense represents that there is no discourse relation between two arguments, but there is a relation between the entities in them.

not in the corresponding answer set, we select the first connective of the corresponding answer set as the golden connective of this sample.

4 Experiments

4.1 Dataset

The Penn Discourse Treebank 2.0 (PDTB 2.0)
PDTB 2.0 is a large scale corpus annotated with information related to discourse relation, containing 2,312 Wall Street Journal (WSJ) articles (Prasad et al., 2008). PDTB 2.0 has three senses levels (i.e., classes, types, and sub-types). We follow Ji and Eisenstein (2015) to take the sections 2-20 as the training set, 0-1 as the development set, and 21-22 as the testing set. Meanwhile, we evaluate our model on the four top-level implicit classes and the 11 major second-level implicit types (Varia et al., 2019; Liu et al., 2020; Wu et al., 2021). Table 4 and 5 show the detailed statistics of the top-level and second-level senses respectively[1].

The CoNLL 2016 Shared Task (CoNLL16)
The CoNLL 2016 shared task (Xue et al., 2016)

---

[1]We found that there are some data samples with two senses. In our data statistics and experiments process, we uniformly considered the first sense of these samples as their golden label for avoiding ambiguity. The same operation was performed on CoNLL16 dataset.

Top-level Senses	Train	Dev.	Test
Comparison (Comp.)	1894	191	146
Contingency (Cont.)	3281	287	276
Expansion (Exp.)	6792	651	556
Temporal (Temp.)	665	54	68
Total	12632	1183	1046

Table 4: The implicit data statistics of top-level senses in PDTB 2.0.

Second-level Senses	Train	Dev.	Test
Comp.Concession	180	15	17
Comp.Contrast	1566	166	128
Cont.Cause	3227	281	269
Cont.Pragmatic cause	51	6	7
Exp.Alternative	146	10	9
Exp.Conjunction	2805	258	200
Exp.Instantiation	1061	106	118
Exp.List	330	9	12
Exp.Restatement	2376	260	211
Temp.Asynchronous	517	46	54
Temp.Synchrony	147	8	14
Total	12406	1165	1039

Table 5: The implicit data statistics of second-level senses in PDTB 2.0.

provides more abundant annotation than PDTB for shadow discourse parsing. The PDTB section 23 and Wikinews texts following the PDTB annotation guidelines were organized as the test sets. CoNLL16 merges several labels of PDTB. For example, *Contingency.Pragmatic cause* is merged into *Contingency.Cause.Reason* to remove the former type with very few samples. Finally, there is a flat list of 15 sense classes to be classified, detailed senses as shown in Table 3 the first column.

4.2 Baselines

To validate the effectiveness of our method, We compare our method with previous state-of-the-art methods. First of all, we select some strong baselines based on neural network including NNMA (Liu and Li, 2016), ESDP (Wang and Lan, 2016), MANN (Lan et al., 2017), PDRR (Dai and Huang, 2018) and RWP-CNN (Varia et al., 2019). Their work mainly focused on the top-level senses of PDTB 2.0 and CoNLL16 cross-level senses. Secondly, we compare our method with competitive baselines based on PLMs, such as DER (Bai and

| Model | PDTB-Top | | PDTB-Second | | CoNLL-Test | CoNLL-Blind |
	Macro-F1	Acc.	Macro-F1	Acc.	Acc.	Acc.
NNMA (Liu and Li, 2016)	46.29	57.57	-	-	-	-
ESDP (Wang and Lan, 2016)	-	-	-	-	**40.91**	34.20
MANN (Lan et al., 2017)	47.80	57.39	-	-	39.40	**40.12**
PDRR (Dai and Huang, 2018)	48.82	57.44	-	-	-	-
RWP-CNN (Varia et al., 2019)	**50.20**	**59.13**	-	-	39.39	39.36
DER (Bai and Zhao, 2018)	51.06	-	-	48.22	-	-
ELMo-C&E (Dai and Huang, 2019)	52.89	59.66	-	48.23	-	-
MTL-MLoss (Van Ngo et al., 2019)	53.00	-	-	49.95	-	-
HierMTN-CRF (Wu et al., 2020)	55.72	65.26	33.91	52.34	-	-
BERT-FT (Kishimoto et al., 2020)	58.48	65.26	-	54.32	-	-
DS-CP (Kurfalı and Östling, 2021)	59.24	-	39.33	55.42	-	-
BMGF-RoBERTa (Liu et al., 2020)	63.39	69.06	37.95	58.13	<u>57.26</u>	<u>55.19</u>
LDSGM (Wu et al., 2021)	<u>63.73</u>	<u>71.18</u>	<u>40.49</u>	<u>60.33</u>	-	-
PCP w/ RoBERTa-base (PCP-base)	64.95	70.84	41.55	60.54	60.98	57.31
PCP w/ RoBERTa-large (PCP-large)	**67.79**	**73.80**	**44.04**	**61.41**	**63.36**	**58.51**

Table 6: Experimental results on PDTB 2.0 and CoNLL16. The best results of previous baselines are underlined. Models in the first part of table are based on neural networks and others in the second part (including our method) are based on different PLMs. We have bolded the best performance results in each of the two parts.

Zhao, 2018), ELMo-C&E (Dai and Huang, 2019), MTL-MLoss (Van Ngo et al., 2019), HierMTN-CRF (Wu et al., 2020), BERT-FT (Kishimoto et al., 2020), DS-CP (Kurfalı and Östling, 2021), BMGF-RoBERTa (Liu et al., 2020) and LDSGM (Wu et al., 2021). These methods achieve impressive performance at the fine-grained second-level senses with the help of large-scale PLMs.

4.3 Implementation Details

In this work, we use *RobertaForMaskedLM* as the backbone of our method, in which *RobertaEncoder* is used for obtaining context representation of inputs and *RobertaLMHead* is to acquire each vocabulary token prediction score for <mask> token position. About prompt settings, we select the last template in Table 1 and answer set in Table 2 and 3 to perform our expreiments on PDTB 2.0 and CoNLL16 dataset.

We adopt AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate of $1e^{-5}$ and weigh decay coefficient of $1e^{-4}$ to update the model parameters and set batch size as 4 for training and validation. At the same time, we add label smoothing with a coefficient of 0.05 into the cross-entry loss function to alleviate overfitting. Based on the above settings, we can usually get the best performance results in the first 3 epochs. All our experiments were performed on one RTX 3090. All other parameters are initialized with the default

values in PyTorch Lightning[2] and our model are all implemented by transformers[3].

4.4 Experimental Results and Analysis

We first evaluate our model on the four coarse-grained top-level and 11 fine-grained second-level senses (denoted as PDTB-Top and PDTB-Second) of PDTB 2.0 with Macro-F1 score and accuracy value. Then we conduct 15-class classification on the CoNLL16 dataset and consider accuracy as the main metric, denoted as CoNLL-Test and CoNLL-Blind for the test and blind-test set.

Table 6 shows the main experimental results of our method and other baselines. Intuitively, our method achieves the new SOTA performance with substantial improvements for coarse-grained and fine-grained implicit discourse relation recognition, which demonstrates that prompt-based connective prediction can effectively mine specific knowledge about connectives and discourse relations in the large-scale PLMs and improve the model's ability to recognize implicit discourse relations.

To better evaluate the performance of our method on fine-grained implicit discourse relation recognition, we compare it with two previous competitive models at each second-level sense of PDTB 2.0.

[2] https://github.com/Lightning-AI/lightning
[3] https://github.com/huggingface/transformers

Second-level Senses	BMGF-RoBERTa	LDSGM	PCP-large	PCP-large w/o </s>
Comp.Concession	0.0	0.0	8.00	**14.81**
Comp.Contrast	59.75	<u>63.52</u>	**63.88**	63.49
Cont.Cause	59.60	<u>64.36</u>	**65.64**	64.90
Cont.Pragmatic cause	0.0	0.0	0.0	0.0
Expa.Alternative	60.0	<u>63.46</u>	66.67	**70.00**
Expa.Conjunction	<u>60.17</u>	57.91	57.78	**58.35**
Expa.Instantiation	67.96	<u>72.60</u>	**74.01**	73.39
Expa.List	0.0	<u>8.98</u>	29.63	**37.50**
Expa.Restatement	53.83	<u>58.06</u>	**61.00**	55.32
Temp.Asynchronous	56.18	<u>56.47</u>	57.81	**60.00**
Temp.Synchrony	0.0	0.0	0.0	**12.50**
Macro-F1	37.95	<u>40.49</u>	44.04	**46.38**
Acc	58.13	<u>60.33</u>	**61.41**	60.64

Table 7: Experimental results on PDTB 2.0 second-level senses. The best results of previous baselines are underlined and the best performance of our method is bolded. We also show changes in the results of our approach (PCP-large and PCP-large without segment token </s>) compared to previous best model.

As shown in Table 7, our PCP approach surpasses the previous state-of-the-art model in almost all second-level senses, except for *Expa.Conjunction*. Noticeable improvements were also achieved by our method on two few-shot second-level senses (i.e. *Comp.Concession* and *Expa.List*).

In addition, in order to further explore our model performance on some few-shot senses, we removed the segment token </s> from the inputs to make it more consistent with the pre-training step of the RoBERTa (corresponding to the forth template in Table 1). The new experimental results are shown in the last part of Table 7. The results indicate that our model without segment token breaks the bottleneck of previous work in three few-shot second-level senses (i.e. *Comp.Concession*, *Expa.List* and *Temp.Synchrony*) of PDTB 2.0. We achieved significant improvements in these few-shot senses.

4.5 Case Study

For the fine-grained sense *Cont.Pragmatic cause*, our method and previous state-of-the-art model predict incorrectly for all samples. Therefore, we check our model prediction results for all seven samples of this sense in test set, and the final results statistics are summarized in Table 8.

We can see that our method tends to predict the connectvies of these samples as high-frequency connectvies, like *because* and *and*, that belong to other discourse relations. So after the answer mapping, the model predicts them as other second-

Mapped Senses	Predicted Connectvies	Count
Exp.Restatement	indeed	1
Exp.Conjunction	and	2
Cont.Cause	because	3
	as	1

Table 8: Statistics on the prediction results of our method for the second-level sense *Pragmatic cause*.

level senses. However, after successively examining these samples, we found that the results predicted by our method seemed to be closer to its actual discourse relation. In other words, the actual discourse relation of these samples that annotated as *Cont.Pragmatic cause* is more similar to the *Cont.Cause* sense. This is probably the reason for merging *Cont.Pragmatic cause* into *Cont.Cause* in CoNLL16 dataset.

5 Discussion

5.1 Why not Predict Discourse Relations Directly

Our approach to implicit discourse relation recognition is prompt-based connective prediction. What happens if we use prompt directly to predict implicit discourse relations. To figure out this question, we construct a new template and search new answer set for prompt-based implicit discourse relation prediction (**PIDRP**). Specifically, the new template is shown below:

- Arg1: *Arg1*. Arg2: *Arg2*. The discourse relation between Arg1 and Arg2 is <mask>.

and we directly use the top-level senses as their answer sets (i.e. mapping the sense *Comparison* to answer set [comparison]).

We use the same pre-trained model and data processing on top-level senses of PDTB to compare the performance gap between the two methods. As shown in Table 9, the PIDRP method performs worse than the PCP method on all four top-level senses of the PDTB, especially on the *Temporal* sense. We think that the main reason of poor performance is that connective prediction is closer to the natural language patterns when the model is in pre-training stage than direct implicit discourse relation prediction.

Top-level Senses	PCP	PIDRP
Comparison	70.38	65.26
Contingency	64.18	64.02
Expansion	80.17	79.80
Temporal	56.41	39.13
Macro-F1	67.79	62.05

Table 9: Performance comparison between PCP and PIDRP on four top-level senses of PDTB 2.0.

5.2 Generalization to Explicit Discourse Relation Recognition

Inspired by the attempt of section 5.1, we transfer our PCP method to explicit discourse relation recognition. We want to explore how well the prompt-based method predicts discourse relation when given connectives.

Similarly, we have designed a simple template in line with natural language conventions for the prompt-based explicit discourse relation recognition (**PEDRR**). The new template is as follows:

- Arg1: *Arg1*. Arg2: *Arg2*. The connective between Arg1 and Arg2 is *Connective*. In summary, the discourse relation between Arg1 and Arg2 is <mask>.

where the *Connective* represents connectives that appear in the original text but ont in *Arg1* or *Arg2*.

We are equal to regard the second-level sense as the corresponding answer set (detailed in Table 10). But for some special senses in which their output of RoBERTa tokenizer is not a single token, we find

Top-level	Second-level	Answer Set
Comparison	Concession	concession
	Contrast	contrast
Contingency	Cause	cause
	Pragmatic cause	justification
Expansion	Alternative	alternative
	Conjunction	conjunction
	Instantiation	instance
	List	list
	Restatement	repetition
Temporal	Asynchronous	asynchronous
	Synchrony	simultaneous

Table 10: Mapping between explicit discourse relation labels and connectives on PDTB 2.0 dataset.

synonyms for them (e.g. *Instantiation*, *Restatement* and *Synchrony*) or use the corresponding subtype as the answer word (e.g. *Pragmatic cause*) to ensure that the result after tokenization is a single token.

As shown in Table 11, the variant of our method PEDRR is also able to achieve results close to those of previous SOTA models on the top-level senses of PDTB 2.0, which effectively demonstrates the generalizability of our approach to EDRR.

Model	Acc.	F1
(1)Connective Only (Pitler and Nenkova, 2009)	93.67	-
(1)+Syntax+Conn-Syn (Pitler and Nenkova, 2009)	94.15	-
(2)ELMo-C&E (Dai and Huang, 2019)	95.39	94.84
(3)RWP-CNN (Varia et al., 2019)	**96.20**	**95.48**
PEDRR (Ours)	94.78	93.59

Table 11: Experimental results of our PEDRR method and other strong baselines on PDTB 2.0 top-level senses for EDRR.

6 Conclusion

In this paper, we propose a novel prompt-based connective prediction method for coarse-grained and fine-grained implicit discourse relation recognition. Experimental results demonstrate that our method achieves state-of-the-art performance on the PDTB 2.0 dataset and the CoNLL-2016 Shared Task. Furthermore, our proposed method breaks the bottleneck of previous work in the few-shot fine-grained discourse relation of PDTB 2.0. Finally, we experimentally prove that our approach can be transferred from IDRR to EDRR and still performs well for EDRR. We will later explore the applicability of our approach to some Chinese discourse relations datasets for fine-grained DRR.

Limitations

In this section, we will point out the limitations of our work, which can be summarized in the following two aspects.

Firstly, in the step of prompt construction (Section 3.2), we manually design templates that meet the task goal of connective prediction and follow natural language patterns. But we do not make more attempts at other templates. We think there may be a new template more suitable for this task goal, which could achieve better performance. Furthermore, we have not explored how well template ensemble works. Therefore, this is a crucial research area in our future work.

Secondly, the data statistics in Table 5 show a category imbalance issue in the second-level senses of PDTB 2.0. The experimental results in Table 7 also show that our method does not solve this problem well, as it markedly improves the model performance on few-shot classes while there is a slight decrease on "more-shot" classes. In the future, we will explore new methods that can significantly improve the model performance in fine-grained implicit discourse relation recognition on few-shot classes while also improving the model performance on "more-shot" classes.

Acknowledgements

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*, pages 571–583, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiang Chen, Ningyu Zhang, Lei Li, Xin Xie, Shumin Deng, Chuanqi Tan, Fei Huang, Luo Si, and Hua-

jun Chen. 2021. Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner. *arXiv preprint arXiv:2109.00720*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022b. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using bart. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2976–2987.

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

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. Ptr: Prompt tuning with rules for text classification. *ArXiv*, abs/2105.11259.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

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*, 3:329–344.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *ArXiv preprint*, abs/2201.04337.

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*, pages 5404–5414, Online. Association for Computational Linguistics.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. 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*, pages 1299–1308.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv preprint*, abs/2107.13586.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3830–3836. ijcai.org.

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*, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification

via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Todor Mihaylov and Anette Frank. 2019. Discourse-aware semantic self-attention for narrative reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2541–2552, Hong Kong, China. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

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'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Linh Van Ngo, Khoat Than, Thien Huu Nguyen, et al. 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*, pages 4201–4207.

Siddharth Varia, Christopher Hidey, and Tuhin Chakrabarty. 2019. Discourse relation prediction: Revisiting word pairs with convolutional networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 442–452, Stockholm, Sweden. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2021. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. *ArXiv preprint*, abs/2112.11740.

Changxing Wu, Chaowen Hu, Ruochen Li, Hongyu Lin, and Jinsong Su. 2020. Hierarchical multi-task learning with crf for implicit discourse relation recognition. *Knowledge-Based Systems*, 195:105637.

Wei Xiang and Bang Wang. 2022. A survey of implicit discourse relation recognition. *ArXiv preprint*, abs/2203.02982.

Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Modeling coherence for discourse neural machine translation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7338–7345. AAAI Press.

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*, pages 5021–5031, Online. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Zhi Min Zhou, Man Lan, Zheng Yu Niu, Yu Xu, and Jian Su. 2010. The effects of discourse connectives prediction on implicit discourse relation recognition. In *Proceedings of the SIGDIAL 2010 Conference*, pages 139–146, Tokyo, Japan. Association for Computational Linguistics.