Using Soft Prompt Learning for Implicit Discourse Relation Recognition

Jie Yu

Abstract—**Implicit Discourse Relation Recognition (IDRR) involves inferring discourse relations between sentences or paragraphs based on contextual information in the absence of explicit connectives, which is a challenging task in discourse parsing. With the advancement of pre-trained language models (PLMs), recent studies have focused on using prompt-based learning methods for IDRR, where specially designed prompt templates are manually crafted to enhance the performance of IDRR tasks. However, manually designed templates have limited expressive power and demands significant manual search effort and time investment. To address these challenges, we propose a soft prompt learning method for IDRR in this paper. It introduces a set of trainable embedding vectors and inserts them to the input arguments to form a learnable prompt template. This approach eliminates the need for extensive manual search effort and provides stronger expressive capabilities compared to discrete templates. Experimental results on the PDTB 2.0 dataset demonstrate that our method achieves superior performance compared to state-of-the-art models. The code for our method is available at https://github.com/L1ngYi/SPLM-IDRR.**

Index Terms—**Implicit Discourse Relation Recognition; Soft Prompting; Pre-trained Language Model; Connective Prediction; Cloze-Prompt Template; Discourse Parsing**

I. Introduction

Implicit Discourse Relation Recognition aims to infer discourse relations between sentences or paragraphs that lack explicit connectives, which is a core task in discourse parsing. This task has extensive applications in various downstream natural language processing (NLP) tasks, such as question answering [1], text summarization [2], and event relation extraction [3].

In recent years, significant progress has been achieved in IDRR research with the advancement of pre-trained language models such as BERT [4] and RoBERTa [5]. These models leverage contextual representations from large-scale corpora to substantially enhance the ability of capturing discourse relations. However, the inherent absence of explicit connectives in IDRR tasks continues to pose challenges for models in handling the complexity of discourse relations.

To improve IDRR performance, researchers have begun exploring prompt-based learning methods. The core idea of prompt-based learning is to reformulate specific tasks as cloze-style questions, thereby bridging the gap between traditional masked language models (MLMs) and downstream tasks. For instance, the PCP framework employed manually crafted templates tailored to discourse relation recognition tasks [6], while the PLSE method integrated mutual information maximization into cloze-style templates to enhance the capture of global logical-semantic information, compensating for the limitations of pre-trained language models in this regard [7]. Although these methods have improved the performance of IDRR tasks, manually designed prompts face significant limitations. First, crafting appropriate templates often requires extensive manual search effort and domain expertise, as well as intuitive insights. Second, manual templates have limited expressive power, making them less adaptable to IDRR tasks and insufficient for capturing nuanced semantics and patterns.

To address the limitations of manually designed prompt templates, soft prompts currently have emerged as a popular method for adapting pre-trained language models to downstream tasks. Unlike traditional handcrafted prompts (also called hard prompts), which use fixed natural language phrases as cues, soft prompts directly operate in the embedding space through parameterized vectors. This approach allows for greater flexibility and efficiency in adapting to various tasks. Studies like P-Tuning have demonstrated that using continuous, learnable vectors as prompts and inserting them into input texts can effectively improve model performance, especially in low-resource scenarios [8]. Therefore, compared to current prompt-based learning methods, soft prompting, due to its learnable and flexible nature, shows good promise for enhancing IDRR performance.

Inspired by the P-Tuning method, this study introduces a soft prompt learning method for IDRR. Building upon the pre-training language model used in PLSE[7], we insert a set of learnable prompt vectors into discrete prompt templates. To ensure flexibility in experimental adjustments and accommodate the learning requirements of different parameter types, we decouple the learning rates of model parameters and template parameters, thereby improving the efficiency and performance of the model. We evaluate our enhanced framework on the PDTB 2.0 dataset [9] which has been widely used for IDRR tasks. Experimental results demonstrate the effectiveness of our proposed method when comparing it with state-of-the-art approaches.

The remainder of this paper is structured as follows. Section 2 briefly introduces the related work. Section 3 details the proposed method. Section 4 describes the datasets, the design and the results of the experiments. Finally, the conclusions are in section 5.

II. Related Work

A. Implicit Discourse Relation Recognition

Dai et al. highlighted that connectives serve as critical cues for predicting discourse relations, significantly improving accuracy [10]. However, unlike explicit discourse relation

Manuscript received April 19, 2025; revised August 12, 2025. This work was supported in part by Project of Philosophy and Social Sciences for a Powerful Education Province in Henan under grant No. 2025JYQS0213, and Educational Reform Project of Henan University under grant No. HDXJJG2024-009.

Jie Yu is a lecturer of the School of College English Teaching and Research, Henan university, Zhengzhou, 450046 China (e-mail: 10310101@vip.henu.edu.cn).

recognition (EDRR), which focuses on logical relations explicitly signaled by connectives such as "because" or "however", implicit discourse relation recognition deals with scenarios where such explicit markers are absent. The lack of direct linguistic identifiers makes IDRR a more challenging task, requiring models to infer hidden connections between sentences or paragraphs.

As a vital task in natural language processing, IDRR has garnered significant attention in recent years. Traditional approaches to IDRR primarily relied on rule-based and handcrafted feature methods [11, 12]. Although these methods laid the groundwork for understanding discourse relations, they often struggled to handle the complexity of linguistic phenomena effectively. With the advent of deep learning, a shift toward neural network-based methods for automatic IDRR emerged. Techniques such as convolution neural networks (CNN) [13] and long short-term memory networks (LSTM) [14] were employed to model implicit discourse relations. These neural approaches offered better performance by leveraging richer feature representations.

Subsequently, methods based on pre-trained language models like BERT [4] and RoBERTa [5] marked a significant milestone in IDRR research. These models, trained on large-scale corpora, can self-learn rich semantic information from context, enhancing the accuracy of implicit discourse relation recognition [4, 5]. The availability of annotated datasets such as the Princeton Discourse Treebank (PDTB) [9] has further propelled research in this domain, providing ample resources for model training and evaluation.

Despite these advancements, several challenges remain in IDRR research. Effectively addressing ambiguity, handling complex contexts, and ensuring robust performance in cross-lingual applications are ongoing hurdles. Developing methods capable of overcoming these challenges are still required for further progress in this field.

B. Prompt-based Learning

With the rapid development of large-scale pre-trained language models such as BERT [4], RoBERTa [5], and GPT-3 [15], prompt-based learning methods have emerged as a popular research direction in the NLP field. The essence of prompt-based learning is to reformulate specific tasks into a cloze-style format, thereby narrowing the gap between traditional masked language models and downstream tasks.

Currently, some studies [16, 17] have proposed manually crafted prompt methods to enhance the performance of information-driven tasks such as the IDRR tasks. However, these approaches face significant limitations, particularly the need for extensive experimentation to identify effective templates that yield satisfactory performance. In the IDRR domain, many prompt-based learning methods have been introduced. For example, the PCP framework [6] reformulated IDRR as an explicit connective prediction task and designed prompt templates to guide pre-trained language models in outputting appropriate connectives. Similarly, PLSE [7] employed ClozePrompt templates to transform IDRR into an masked language model task. These approaches leverage the capabilities of pre-trained language model while utilizing prompts to align the task with the model's inherent structure, showcasing the potential of prompt-based learning in addressing the challenges of IDRR.

C. Prompt Tuning

With the rapid development of prompt-based learning methods, traditional manual design of task-specific templates is no longer sufficient to meet the demands of language model prompting. Manual prompts rely on researchers' experience and experimentation to craft natural language templates, a process that is not only inefficient but also fails to fully exploit the potential capabilities of pre-trained language models. To address these limitations, soft templates—also known as Prompt Tuning—have emerged as a promising approach. Soft templates optimize prompt representations through parameterization and have become an important paradigm for adapting pre-trained language models to various downstream tasks.

Unlike manually crafted hard templates, which use fixed natural language phrases, soft templates are represented as continuous embedding vectors that can be dynamically adjusted through training to suit specific tasks. This method reduces the complexity of manual template design while leveraging the semantic knowledge of pre-trained language models, embedding task requirements implicitly into the model's input. The advent of soft templates marks a pivotal shift in prompt learning, outperforming traditional methods in task adaptation performance while offering effective solutions for low-resource and few-shot learning scenarios. Recently, several techniques have been proposed for prompt tuning by mining training corpora [18], using gradient-based search methods [19], or employing pre-trained generative models [20]. Among them, P-Tuning proposed by Liu et al. [8] demonstrates the potential of soft templates in bridging the gap between fine-tuning and task-specific performance.

In addition, recent research in prompt tuning has focused on enhancing stability, parameter efficiency, and convergence speed. For example, Residual Prompt Tuning augments soft prompt embeddings with residual connections, yielding substantial performance gains over baseline [32]. Furthermore, SuperPos-Prompt accelerates convergence by superposing multiple pretrained token embeddings, achieving superior results compared to residual reparameterization on standard NLP benchmarks such as GLUE and SuperGLUE [31].

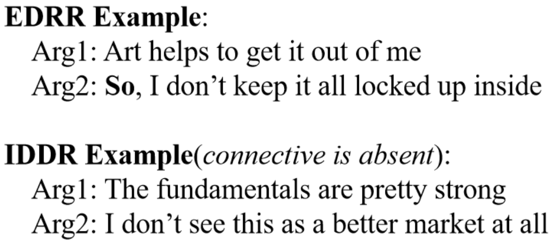
Meanwhile, in multilingual settings, recent studies have shown that freezing pre-trained language model parameters and tuning only soft prompts is sufficient to maintain strong cross-lingual transfer capabilities. This suggests that soft prompts provide a compact and effective means for knowledge adaptation across languages, making them particularly valuable in low-resource or zero-shot multilingual scenarios [30].

III. Methodology

In this section, we state the problem of implicit discourse relation recognition, and detail the proposed method.

A. Problem Statement

Discourse relation recognition (DRR) aims to identify the existence of logical relationships between adjacent discourse units within the same text. The discourse units can be clauses, sentences, and paragraphs. As shown in Figure 1, there is usually one argument pairs (Arg1, Arg2) given in the DRR tasks where Arg1 and Arg2 are the discourse units, and the tasks are to predict the discourse relations between the argument pairs. The DRR tasks can be divided into explicit discourse relationship recognition, i.e., EDRR, and implicit discourse relation recognition, i.e., IDRR, based on the presence of significant connective words, which is like the word “so” in the EDRR example shown in Figure 1. Because connective words can provide linguistic information clues in DRR tasks, EDDR usually achieve high accuracy rate. However, due to the lack of connective words, IDRR can only rely on semantic information of discourse units, making this task quite challenging. In this paper, we focus on using soft prompt learning technique to address IDRR based on current pre-trained language models like BERT.



EDRR Example:
Arg1: Art helps to get it out of me
Arg2: **So**, I don’t keep it all locked up inside

IDDR Example(*connective is absent*):
Arg1: The fundamentals are pretty strong
Arg2: I don’t see this as a better market at all

Fig. 1. The examples showing the EDRR and IDRR [7]. In the EDRR example, the explicit connective word is “so”, while the connective word is absent in the IDRR example.

B. Pipeline of the Proposed Method

Figure 2 illustrates the framework of our method. Given one argument pairs (Arg1, Arg2), the input will be first templated by filling some placeholders or special markers. Specifically, besides the two arguments, there are a mask token, i.e., `[mask]`, and several soft markers, i.e., `[P_1][P_2][P_3]`, as shown in Figure 2. The mask token indicates the placeholder substituting for the connectives to be predicted, while the soft markers are the placeholders representing the soft prompts which can be adjusted and learned during training. Then, the soft prompts will pass through a prompt encoder which is implemented as a long short-term memory network. With the prompt encoder, the inter-relationships between the soft prompts will be encoded and the embedding vectors of these soft prompts are generated. By concatenating the embedding of these soft prompts with the embedding of the other tokens in the input, all the input are mapped into the high-dimensional vector space.

Next, the mapped embedding vectors are input to the masked language model to predict the absent connectives as shown in Figure 2. Finally, according to the mapping between the connectives and the discourse relations, the discourse relation implied by the input argument pair is inferred and the loss is derived during training. With the loss, not only the model but also the learnable vectors corresponding to these soft prompts are updated. During testing, the argument pairs are templated in the same way as the training, and using the learned vectors to substitute for each soft prompt, the discourse relations between the arguments can be derived.

In the following section, we detail the proposed method including the prompt template and the prompt turning.

C. Prompt Template

In our method, instead of directly using raw text as input, we construct a template to integrate the input text, thereby transforming the IDRR task into a masked language model task. In previous studies, such templates were often handcrafted, for example, the template `Arg1:Arg1. Arg2:Arg2. The connective between Arg1 and Arg2 is [MASK].` Although these approaches appear intuitive and straightforward, crafting a high-performance handcrafted prompt template heavily depends on experience and intuition. It often requires considerable time for manual search and may limit the model’s capabilities due to the constraints of the template’s expressive power. Therefore, we employ soft prompt templates to address this issue. Soft prompt templates can not only help the model understand the task but also automatically optimize and adjust the template during training, enabling the model to better adapt to IDRR tasks.

Specially, we use the soft prompt template like `[P_1][P_2][P_3]Arg1[mask]Arg2` in this paper. Here, `Arg1` and `Arg2` are still the two arguments in the input representing the discourse units, and the symbol `[mask]` also represents a placeholder for the connective. Differently, there are several `[P_i]` tokens which represent the learnable soft prompts.

D. Prompt Tuning

The masked language model used in this paper is the RoBERTa model which is further pre-trained by Wang et al. [7] with a logic-semantic enhancement approach. By leveraging handcrafted templates to construct connective prediction tasks, Wang et al. [7] have employed a multi-head cross-attention module alongside a mutual information maximization objective to train RoBERTa model and perform global logic-semantic learning. Since IDRR task is not directly signaled by explicit connectives (e.g., “because” or “but”) but instead relies on contextual understanding across discourse units, it needs to capture dependencies spanning sentences and even paragraphs in long texts. Therefore, we employ LSTM network which effectively captures long-distance dependencies to learn the inter-relationships between soft prompts, which is similar with the research [8].

While using the masked large language model to predict the absent connectives, we employed a verbalizer that maps connectives to implicit discourse relation labels. This verbalizer defines a discrete answer space for IDRR, which is a subset of the pre-trained language model vocabulary. For PDTB dataset, the verbalizer is shown in Table I, which consists of a set of high-frequency and low-ambiguity connectives manually selected by Wang et al. [7] to represent the corresponding discourse relations.

Once the discourse relations are predicted, the cross-entropy loss calculated by Eq. (1) will be derived to tune the model.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} -\log P(l_i = V(c_i) \mid T(x_i)) \tag{1}$$

In this equation, N represents the number of training examples, T denotes the prompt template used to transform the input argument pair x, V is the verbalizer that maps the

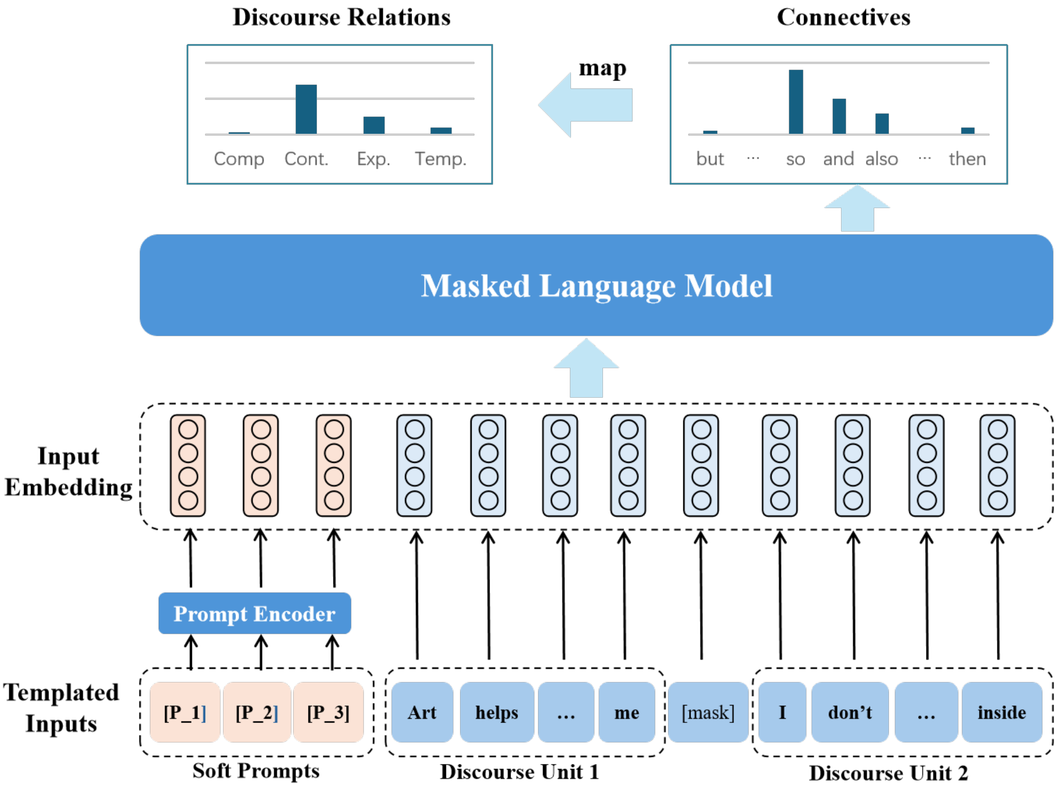


Fig. 2. The pipeline of the proposed method. The P_1, P_2 and P_3 are the soft markers representing the soft prompts to be learned during training, and the [mask] indicates the masked token substituting for absent connectives.

TABLE I
THE VERBALIZER MAPPING CONNECTIVES BETWEEN IMPLICIT DISCOURSE RELATION LABELS AND CONNECTIVES ON PDTB 2.0 DATASET, CONSISTING OF FOUR **TOP-LEVEL** AND 11 **SECOND-LEVEL** CLASSES.

Top-level	**Second-level**	**Connectives**
Comparison	Concession Contrast	however, although, though but
Contingency	Cause Pragmatic cause	because, so, thus, consequently, therefore as, since
Expansion	Alternative Conjunction Instantiation List Restatement	instead, rather and, also, fact, furthermore instance, example finally specifically, indeed, particular
Temporal	Asynchronous Synchrony	then, after, before meanwhile, when

connective c to the implicit discourse relation label l, and P estimates the probability of the gold semantic label l.

While tuning the model with the losses, a differentiated learning rate strategy is adopted, considering that the pre-trained masked language model already contains significant useful information and there are a large number of parameters. That is, we assign a specific learning rate to LSTM network to keep it separate from that of the masked large language model. By carefully balancing the learning rates, this strategy will enable the soft templates to adapt dynamically without over-fitting and destabilizing the pre-trained masked language model.

IV. Experiments

In order to validate the effectiveness of the proposed method, extensive experiments have been done. In this section, we describe the experimental setup and results.

A. Dataset

We used the Penn Discourse Treebank 2.0 (PDTB 2.0) for the evaluation in this paper. It is a widely-used corpus comprising 2,312 articles from the Wall Street Journal (WSJ), annotated with discourse relations through a lexically-driven approach [9]. The discourse relations are organized into a three-level hierarchy: classes, types, and sub-types. Following established practices [21], the dataset is split into training (sections 2-20), validation (sections 0-1), and test sets (sections 21-22). Consistent with prior research [5, 22], our evaluation focuses on the four top-level implicit discourse relation classes and 11 prominent second-level types, which can be seen in Table I.

B. Implementation Details

Our implementation was based on the open-source framework OpenPrompt [23], which facilitated the construction of prompt-learning experiments. We used the PLSE method [7] to pre-train the masked large language model of RoBERTa [5] and employed AdamW [24] as the optimizer. All experiments were conducted on a Tesla T4 GPU.

For top-level classes on PDTB2.0, we trained for 10 epochs and selected the model that performed best on the validation set. Due to GPU resource constraints, we used a batch size of 32 and a learning rate of 1×10^{-5}. For second-level classes on PDTB2.0, considering the increased complexity of the task, we trained for 15 epochs and selected the best-performing model on the validation set. After conducting a detailed hyperparameter search, we set the batch size for the training data to 16, and for the validation and test sets to 32. We set the learning rate for the masked large language model to 1×10^{-5} and for the prompt encoder to 1×10^{-3}.

C. Baselines

To validate the effectiveness of our method, we compared it with a set of state-of-the-art methods.

- **CG-T5 [25]**: By viewing IDRR as a generation task, it proposes a method joint modeling of the discourse relation recognition and generation.
- **LDSGM [26]**: Consider multi-level IDRR as a conditional label sequence generation task, it proposes a label dependence-aware sequence generation model for IDRR.
- **PCP [6]**: It instructs large-scale pre-trained models to use knowledge relevant to discourse relation and utilizes the strong correlation between connectives and discourse relations to help the model recognize implicit discourse relations.
- **GOLF [27]**: It proposes a novel global and local hierarchy-aware contrastive framework to sufficiently exploit global and local hierarchies of classes to learn better discourse relation representations.
- **DiscoPrompt [28]**: Considering that it is more effective to predict the paths inside the hierarchical tree rather than flat labels or connectives, it proposes a prompt-based path prediction method to utilize the interactive information and intrinsic senses among the hierarchy in IDRR.
- **ChatGPT [29]**: It is an improved dialogue generation approach that strengthens interactivity and reliability in dialogue systems.
- **PLSE [7]**: It seamlessly injects knowledge relevant to discourse relation into pre-trained language models through prompt-based connective prediction, while designing a novel self-supervised learning objective based on mutual information maximization to derive enhanced representations of logical semantics for IDRR.

D. Overall Performance Comparison

To evaluate the effectiveness of the proposed soft prompt learning framework, we compare its performance with these selected methods on the PDTB 2.0 dataset. The comparison is carried out on both the top-level (4-class) and second-level (11-class) implicit discourse relation recognition tasks. And the template of `[P_1][P_2][P_3]Arg1[mask]Arg2` was adopted for the proposed method during comparison.

Two commonly used evaluation metrics were adopted: Macro-F1 and Accuracy. The macro-F1 is defined as the unweighted mean of class-wise F1 scores:

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

where P_i and R_i denote the precision and recall for class i, and C is the total number of classes.

The accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

TABLE II
MACRO-F1 SCORES (%) AND ACCURACY (%) EVALUATED ON THE PDTB 2.0 DATASET. **BOLD NUMBERS** REPRESENT THE BEST RESULTS.

Method	PDTB-Top		PDTB-Second	
	F1	Acc.	F1	Acc.
CG-T5 [25]	*57.18*	*65.54*	*37.76*	*53.13*
LDSGM [26]	*63.73*	*71.18*	*40.49*	*60.33*
PCP [6]	*64.95*	*70.84*	*41.55*	*60.54*
GOLF [27]	*65.76*	*72.52*	*41.74*	*61.16*
DiscoPrompt [28]	*65.79*	*71.70*	*43.68*	*61.02*
ChatGPT [29]	*36.11*	*44.18*	*16.20*	*24.54*
PLSE [7]	68.12	73.23	47.22	62.85
Ours	**69.32**	**73.90**	**48.91**	**62.95**

Table II shows the comparison results. From the results, we can see that our method achieves the best performance on the PDTB2.0 dataset. For the top-level discourse relation classes, the macro-F1 increases by 1.2% and the accuracy improves by 0.7%. For the second-level discourse relation classes, the macro-F1 rises by 1.7%, and the accuracy improves by 0.1%. Compared to latest method of PLSE, our method still demonstrates certain performance improvements. These results indicate the effectiveness of our work, and show the superiority of continuous templates using soft prompts over traditional handcrafted templates.

To provide more detailed insights, we further break down the macro-F1 scores across the four top-level and the eleven second-level discourse relation classes. Table III shows the results over four top-level discourse relation classes: *Comparison*, *Contingency*, *Expansion* and *Temporal*. These results indicate that our model outperforms these baselines in each class, especially in the *Comparison* and *Temporal* classes, with respective gains of 2.98% and 0.86% over PLSE method. Table IV reports the fine-grained macro-F1 performance for the eleven second-level discourse relation classes. The results show that our model achieves the best macro-F1 scores in six classes, including *Contrast*, *Cause*, *Restatement*, *Asynchronous*, and *Synchrony*. Notably, in the *Synchrony* class, our model achieves 66.67% in terms of macro-F1, a substantial improvement over PLSE (33.33%) and other baselines (0%), reflecting its strength in handling rare and challenging discourse categories.

E. Evaluation under Few-shot Settings

To assess the robustness of our model under limited supervision, we conducted a series of few-shot experiments using only 30%, 50%, and 70% of the full training data. This setting simulates real-world scenarios where annotated discourse relations are scarce, particularly in fine-grained classification.

TABLE III
MACRO-F1 SCORES (%) FOR TOP-LEVEL DISCOURSE RELATION CLASSES ON PDTB 2.0.

Model	Comp.	Cont.	Exp.	Temp.
CG-T5	55.40	57.04	74.76	41.54
GOLF	67.71	62.90	79.41	54.55
DiscoPrompt	62.55	64.45	78.77	57.41
PLSE	65.02	64.49	**80.60**	62.39
Ours	**68.00**	**65.55**	80.49	**63.25**

TABLE IV
MACRO-F1 SCORES (%) FOR SECOND-LEVEL DISCOURSE RELATION CLASSES ON PDTB 2.0.

Second-level	GOLF	DiscoPrompt	PLSE	Ours
Comp.Concession	0.00	9.09	0.00	0.00
Comp.Contrast	61.95	59.26	61.92	**64.98**
Cont.Cause	63.35	63.83	64.94	**67.85**
Cont.Pragmatic Cause	0.00	0.00	0.00	0.00
Exp.Alternative	63.49	**72.73**	51.85	56.00
Exp.Conjunction	60.28	61.08	**60.58**	56.85
Exp.Instantiation	73.36	69.96	**76.50**	73.49
Exp.List	27.78	37.50	**45.45**	26.09
Exp.Restatement	59.84	60.00	61.39	**61.26**
Temp.Asynchronous	63.82	57.69	63.49	**64.86**
Temp.Synchrony	0.00	0.00	33.33	**66.67**

As shown in Figure 3, the model exhibits stable and consistent performance across all training proportions. In the PDTB-TOP (4-class) setting, the macro-F1 and the accuracy both remain relatively steady, showing only modest increases as more data becomes available. Notably, even at 30% supervision, the model retains competitive performance, indicating its ability to learn meaningful patterns under data-constrained conditions. In the more complex PDTB-Second (11-class) setting, the performance also progresses gradually. Although both macro-F1 and accuracy are lower compared to the top-level task—as expected given the increased difficulty—the model avoids sharp degradation and maintains a consistent trend across different levels of supervision.

These observations highlight the model's generalization capability in few-shot scenarios. Rather than relying solely on large amounts of data, the hybrid prompt framework demonstrates resilience and stability, making it a promising approach for discourse relation recognition in few-shot settings.

F. Robustness to Template Perturbation

Prompt-based models are often sensitive to the structure and semantics of their input templates. To evaluate how well the proposed method copes with structural deviations, we conducted a series of controlled template perturbation experiments. Each variation alters the original template in a specific way to simulate potential design noise or deployment inconsistencies.

We considered three types of perturbations applied to the base soft template: (1) shifting the `[MASK]` token to the end of the sequence, thus decoupling it from its natural contextual position; (2) inserting an unrelated token (`"XX"`) immediately before the `[MASK]`; and (3) inserting a misleading discourse cue (`"WHY"`) before the `[MASK]` to introduce semantic ambiguity. All experiments were conducted on both the top-level classes and second-level classes of PDTB dataset.

The results are shown in Table V. ID 1 represents the original template structure which we have adopted in the comparison experiments, while IDs 2–4 introduce position shifts or irrelevant token insertions. Across both tasks, the original template achieves the highest macro-F$_1$ and accuracy scores, as expected. However, performance under perturbation remains stable. In the top-level discourse relation recognition task, macro-F$_1$ varies from 69.32% (original) to 66.30%(inserting the `"WHY"` before the `[MASK]`), 63.08% (with `"XX"`) and 59.23% (`[MASK]` is moved to the end), reflecting a controlled decline without abrupt collapse. A similar pattern is observed in the second-level discourse relation recognition setting, where macro-F$_1$ decreases from 46.82% to 45.55%, 43.89% and 43.53% under analogous conditions. Accuracy metrics mirror this trend, suggesting that the model retains strong prediction capacity despite significant template modification. These findings indicate that our soft prompt design exhibits a desirable level of structural robustness: it maintains a consistent performance profile under syntactic and semantic disturbances.

G. Parameter Sensitivity Analysis

To further understand how the position and the number of soft prompt tokens affect model performance, we conducted a series of controlled experiments, with results shown in Table VI. Each row corresponds to a different variant of the input template, either modifying the position of the `[P_i]` soft prompt tokens or the position of the `[mask]` token. Our goal is to identify optimal template structures for both top-level and second-level discourse relation recognition tasks.

For instance, the #1 template is `[P_1][P_2][P_3] Arg1 [mask] Arg2`, which places all soft tokens at the beginning and inserts the `[mask]` token between the two arguments. In #2 template, i.e., `Arg1 [mask] Arg2 [P_1][P_2][P_3]`, we move the soft prompt tokens to the end. #3 template places the soft prompt tokens behind the `[mask]` token. #4 template places the soft prompt tokens before the `[mask]` token but behind the argument `Arg1`. It can be seen that the position of the soft prompt tokens indeed affect the performance of the proposed method. And among these variant templates, the #1 temple yields the best performance overall, which has been selected for comparison with other methods in our experiments.

#5–#7 templates explore the impact of the number of soft prompt tokens, ranging from 2 to 5. Although five soft prompt tokens slightly improve the accuracy of the top-level discourse relation recognition, performance on second-level classification degrades, suggesting that more soft tokens may not always yield better generalization, especially for fine-grained tasks.

V. CONCLUSION

This paper proposes a soft prompt learning method for implicit discourse relation recognition. Different from existing methods, we employ soft prompt template instead of handcrafted prompt template when using pre-trained language model. Besides eliminating the reliance on manual experience and intuition, it can not only help the model understand the task but also automatically optimizes and adjusts the template during training, enabling the model to better adapt to discourse relation recognition task.

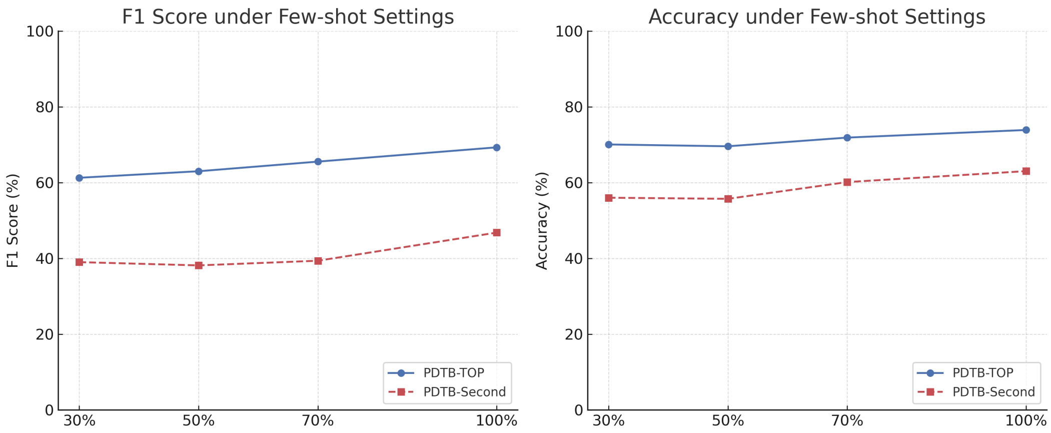


Fig. 3. Few-shot evaluation results of the hybrid prompt model on PDTB-TOP and PDTB-Second tasks. Left: F1 scores across four levels of training supervision (30%, 50%, 70%, and 100%). Right: Corresponding accuracy values. The curves indicate that the model maintains stable performance under reduced training data, with no sharp degradation observed.

TABLE V
Performance of the our soft prompt method under different template structures on PDTB-Top and PDTB-Second tasks. Each row corresponds to a controlled perturbation of the original template.

ID	Template	PDTB-Top		PDTB-Second	
		F1	Acc	F1	Acc
1	[P_1] [P_2] [P_3] Arg1 [mask] Arg2	69.32	73.90	46.82	63.04
2	[P_1] [P_2] [P_3] Arg1 Arg2 [mask]	59.23	68.93	43.89	58.61
3	[P_1] [P_2] [P_3] Arg1 XX [mask] Arg2	63.08	71.22	45.55	62.85
4	[P_1] [P_2] [P_3] Arg1 WHY [mask] Arg2	66.30	72.28	43.53	60.92

TABLE VI
Macro-F1 scores (%) and accuracy (%) evaluated on the PDTB2.0 dataset using different templates. Here, [P_i] refers to soft prompt markers, and [mask] refers to masked tokens.

ID	Template	PDTB-Top		PDTB-Second	
		F1	Acc	F1	Acc
1	[P_1] [P_2] [P_3] Arg1 [mask] Arg2	67.14	73.14	48.91	62.95
2	Arg1 [mask] Arg2 [P_1] [P_2] [P_3]	68.12	72.75	45.49	61.21
3	Arg1 [mask] [P_1] [P_2] [P_3] Arg2	67.11	73.22	42.00	59.29
4	Arg1 [P_1] [P_2] [P_3] [mask] Arg2	68.07	73.18	42.00	61.41
5	[P_1] [P_2] [P_3] [P_4] Arg1 [mask] Arg2	66.69	72.37	42.37	59.58
6	[P_1] [P_2] [P_3] [P_4] [P_5] Arg1 [mask] Arg2	67.34	73.23	43.34	59.67
7	[P_1] [P_2] Arg1 [mask] Arg2	67.09	72.75	43.67	60.54

Through a series of controlled and comparative experiments, we have validated the effectiveness and robustness of the proposed soft prompt tuning framework. Our method achieves competitive performance compared to state-of-the-art baselines, while also demonstrating notable stability under few-shot training conditions. Furthermore, it maintains consistent performance in the face of prompt template perturbations, indicating structural resilience. The parameter sensitivity analysis further reveals that the position and number of soft tokens are critical design factors that influence model effectiveness. Overall, these findings support the conclusion that soft prompting is both a practical and theoretically grounded approach for implicit discourse relation recognition.

References

[1] A. Rutherford; N. Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 799–808.

[2] A. Cohan; F. Dernoncourt; D.S. Kim; T. Bui; S. Kim; W. Chang; N. Goharian. A discourse-aware attention model for abstractive summarization of long documents. *arXiv:1804.05685,* 2018.

[3] J. Tang; H. Lin; M. Liao; Y. Lu; X. Han; L. Sun; W. Xie; J. Xu. From discourse to narrative: Knowledge projection for event relation extraction. *arXiv:2106.08629,* 2021.

[4] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805,* 2018.

[5] Y. Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692,* 2019.

[6] H. Zhou; M. Lan; Y. Wu; Y. Chen; M. Ma. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. *arXiv:2210.07032,* 2022.

[7] C. Wang; P. Jian; M. Huang. Prompt-based logical semantics enhancement for implicit discourse relation

recognition. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 687–699.

[8] X. Liu; Y. Zheng; Z. Du; M. Ding; Y. Qian; Z. Yang; J. Tang. GPT understands, too. *AI Open,* 2024, *5*, pp. 208–215.

[9] R. Prasad; E. Miltsakaki; N. Dinesh; A. Lee; A. Joshi; L. Robaldo; B. Webber. The penn discourse treebank 2.0 annotation manual. *December,* 2007.

[10] Z. Dai; R. Huang. 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, 2019, pp. 2976–2987.

[11] Z. Lin; M.Y. Kan; H.T. Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 343–351.

[12] Z.M. Zhou; Y. Xu; Z.Y. Niu; M. Lan; J. Su; C.L. Tan. Predicting discourse connectives for implicit discourse relation recognition. In Proceedings of 4th Workshop on Syntax and Structure in Statistical Translation, 23rd International Conference on Computational Linguistics, 2010, pp. 1507–1514.

[13] B. Zhang; J. Su; D. Xiong; Y. Lu; H. Duan; J. Yao. Shallow convolutional neural network for implicit discourse relation recognition. In Proceedings of the 2015 Conference on empirical methods in natural language processing, 2015, pp. 2230–2235.

[14] Y. Liu; S. Li. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. *arXiv:1609.06380,* 2016.

[15] T. Brown; B. Mann; N. Ryder; M. Subbiah; J.D. Kaplan; P. Dhariwal; A. Neelakantan; P. Shyam; G. Sastry; A. Askell; et al. Language models are few-shot learners. *arXiv:2005.14165,* 2020.

[16] W. Xiang; Z. Wang; L. Dai; B. Wang. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 902–911.

[17] H. Zhou; M. Lan; Y. Wu; Y. Chen; M. Ma. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. *arXiv:2210.07032,* 2022.

[18] Z. Jiang; F.F. Xu; J. Araki; G. Neubig. How can we know what language models know? *arXiv:1911.12543,* 2020.

[19] T. Shin; Y. Razeghi; R.L. Logan IV; E. Wallace; S. Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv:2010.15980,* 2020.

[20] T. Gao; A. Fisch; D. Chen. Making pre-trained language models better few-shot learners. *arXiv:2012.15723,* 2020.

[21] Y. Ji; J. Eisenstein. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics,* 2015, *3*, 329–344.

[22] Y. Jiang; L. Zhang; W. Wang. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. *arXiv:2211.13873,* 2022.

[23] N. Ding; S. Hu; W. Zhao; Y. Chen; Z. Liu; H.T. Zheng; M. Sun. Openprompt: An open-source framework for prompt-learning. *arXiv:2111.01998,* 2021.

[24] I. Loshchilov. Decoupled weight decay regularization. *arXiv:1711.05101,* 2017.

[25] F. Jiang; Y. Fan; X. Chu; P. Li; Q. Zhu. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 2418–2431.

[26] C. Wu; L. Cao; Y. Ge; Y. Liu; M. Zhang; J. Su. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 11486–11494.

[27] Y. Jiang; L. Zhang; W. Wang. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. *arXiv:2211.13873,* 2022.

[28] C. Chan; X. Liu; J. Cheng; Z. Li; Y. Song; G.Y. Wong; S. See. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *arXiv:2305.03973,* 2023.

[29] C. Chan; J. Cheng; W. Wang; Y. Jiang; T. Fang; X. Liu; Y. Song. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv:2304.14827,* 2023.

[30] F. Philippy; S. Guo; S. Haddadan; C. Lothritz; J. Klein; T. F. Bissyandé. Soft prompt tuning for cross-lingual transfer: When less is more. *arXiv:2402.03782,* 2024.

[31] M. SadraeiJavaeri; E. Asgari; A. C. McHardy; H. R. Rabiee. SuperPos-Prompt: Enhancing Soft Prompt Tuning of Language Models with Superposition of Multi Token Embeddings. *arXiv:2406.05279,* 2024.

[32] A. Razdaibiedina; Y. Mao; R. Hou; M. Khabsa; M. Lewis; J. Ba; A. Almahairi. Residual prompt tuning: Improving prompt tuning with residual reparameterization. *arXiv:2305.03937,* 2023.