Global and Local Hierarchical Prompt Tuning Framework for Multi-level Implicit Discourse Relation Recognition

Lei Zeng[1,2], Ruifang He[1,2*], Haowen Sun[1,2], Jing Xu[1,2], Chang Liu[1,2] and Bo Wang[1,2]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China
{leizeng,rfhe,haowensun123,jingxu,changliu,bowang}@tju.edu.cn

Abstract

Multi-level implicit discourse relation recognition (MIDRR) is a challenging task to recognize the hierarchical discourse relations between the arguments with the absence of connectives. Recent methods tend to incorporate the static hierarchical structure containing all senses (defined as global hierarchy) into prompt tuning through a path prompt template or hierarchical label refining. However, hierarchical modeling is independent of the verbalizer, resulting in a failure to effectively utilize the output probability distribution information of verbalizer. Besides, they ignore the utilization of the dynamic hierarchical label sequence for each instance (defined as local hierarchy) in prompt tuning. In this paper, we propose a global and local hierarchical prompt tuning (GLHPT) framework, which utilize prior knowledge of PLMs while better incorporating hierarchical information from two aspects. We leverage bottom-up propagated probability as the global hierarchy to inject it into multi-level verbalizer (MLV). Furthermore, we design a local hierarchy-driven contrastive learning (LHCL) to improve the probability distribution of MLV. Finally, our model achieves competitive results on two benchmacks.

Keywords: Prompt Learning, Hierarchy, Implicit Discourse Relation Recognition

1. Introduction

Implicit discourse relation recognition (IDRR) (Pitler et al., 2009) aims to discover discourse relations between a pair of text segments (named arguments) without the guidance of explicit connectives (e.g., *because*, *but*) (Xiang and Wang, 2023). It is vital for textual coherence and is considered as the essential step for many downstream tasks involving more context, such as question answering (Liakata et al., 2013), text summarization (Li et al., 2020b) and event relation extraction (Tang et al., 2021). Meanwhile, the sense labels for each discourse relation follow a hierarchical classification structure in the annotation process. Figure 1 shows a discourse instance with multi-level labels, it consists of two arguments (i.e., Arg1 and Arg2) and is annotated with relation senses. The label of the top-level is *Comparison*, while the sub-label *Contrast* is the fine-grained semantic expression of *Comparison*. Besides, when annotating the implicit relations, annotators simulate adding connectives (e.g., *however*) to help better understand the semantics of labels.

With the widespread application of pre-trained language models(PLMs), IDRR has also achieved considerable improvement (Lei et al., 2017; Bai and Zhao, 2018; Ruan et al., 2020; Xiang et al., 2022a; Wu et al., 2022; Jiang et al., 2023). Under the pre-train and fine-tuning paradigm, researchers encode the representation of argument pairs by designing sophisticated neural networks for relation classification. Although these task-specific neural networks can effectively learn a kind of contextual semantics of arguments, they introduce some additional parameters that relies on a large scale of data to train. Moreover, some studies suggest that the task objective is often inconsistent with that of PLMs, which restrains the finetuned models to take full advantage of knowledge in PLMs (Liu et al., 2023).

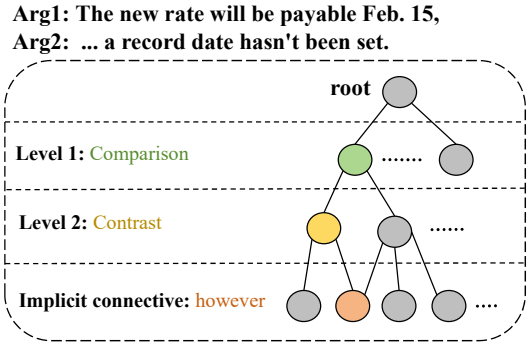


Figure 1: An Instance for multi-level IDRR. The implicit connective is not present in the original discourse context but is assigned by annotators. All senses are organized in a three-layer hierarchical structure, and the implicit connectives is the most fine-grained senses.

Inspired by prompt tuning (PT) (Schick and Schütze, 2021b), several studies reformulate the IDRR task as cloze questions to bridge the gap between the masked language model (MLM) and downstream IDRR tasks (Zhou et al., 2022; Xiang et al., 2022c, 2023). Although remarkable perfor-

*The Corresponding author

LREC-COLING 2024, pages 7760–7773

mances have been achieved via prompt tuning on the single-level IDRR, the inherent discourse label hierarchy is ignored.

To solve this issue, the latest research attempts to inject label hierarchy knowledge into prompt tuning. DiscoPrompt (Chan et al., 2023) transforms the hierarchy in Figure 1 to "*Comparison -> Contrast -> however;* $\cdots$ *; Temporal -> Synchrony -> when*" as the path prompt and adds it as the prefix of arguments to be classified. But their manners of selecting a few special connective nodes form path prompt template, which lacks the integrity of hierarchical label modeling. PEMI (Zhao et al., 2023) propose a hierarchical label refining method for the prompt verbalizer to deeply integrate hierarchical guidance into the prompt tuning.

However, existing methods still have two limitations. **1)** Though they exploit static dependencies among labels (global hierarchy), hierarchical modeling is independent of the verbalizer, resulting in a failure to effectively utilize the output probability distribution information of verbalizer. **2)** Dynamic label structure information (local hierarchy) corresponding to each sample has shown its importance in text classification under the pre-training and fine-tuning paradigm (Jiang et al., 2022). Hierarchy-aware prompt tuning (Wang et al., 2022b; Chan et al., 2023; Zhao et al., 2023) only consider the graph of the entire static label hierarchy and ignore the dynamic label.

Based on the above observations, we propose a novel **G**lobal and **L**ocal **H**ierarchical **P**rompt Tuning (**GLHPT**) framework, to fully exploit the hierarchical knowledge of labels and adapt prior knowledge in PLMs to downstream MIDRR tasks. Specifically, following previous definition(Jiang et al., 2022), we define global hierarchy as the whole hierarchical structure containing all senses, which is static and irrelevant to each instance. While local hierarchy is defined as a dynamic hierarchical sense label sequence corresponding to each input instance. **1)** For the aforementioned limitation 1, we design a **G**lobal **H**ierarchical **B**ottom-**U**p **P**ropagated **P**robability(**GHBUPP**) method utilizes the output probability distribution information of multi-level verbalizer for hierarchical modeling, which injects global hierarchical constraints into verbalizer and adapt prior knowledge in PLMs to downstream MIDRR task. **2)** For the aforementioned limitation 2, we design a **L**ocal **H**ierarchy-driven **C**ontrastive **L**earning (**LHCL**) to learn the hierarchy-aware matching relationship to guide the distance between `<MASK>` token representations, which can further utilize the local labels structure information to improve the probability distribution of multi-level verbalizer. Finally, our multi-task framework carries out joint learning at all levels.

Our contributions are summarized as follows:

- Our work attempts to simultaneously incorporate global and local hierarchical information into prompt tuning to comprehensively explore hierarchical knowledge.
- We propose a global and local hierarchical prompt tuning framework for MIDRR. It utilizes the output of multi-level verbalizer for hierarchical modeling, to inject hierarchical constraints and adapt prior knowledge in PLMs to downstream MIDRR task.
- Experimental results demonstrate that our model achieves the competitive results at multiple levels on two popular datasets, and excels handling label imbalance and few-shot discourse relation.

2. Method

2.1. Overview

Given a pair of arguments instance x = (x_{arg_1},x_{arg_2}) and the set of total labels $\mathbb{L}$ = (L^1,..,L^d,..,L^D), where L^d is the level-d label set. The target label sequence is $\boldsymbol{y}$ =(y^1,..,y^d,..,y^D), where $y^d \in L^d$ is the prediction of level d. Figure 2 shows the overall architecture of our model.

2.2. Multi-Level Verbalizer

Prompt tuning is a universal approach to stimulate the potential of PLMs for most downstream tasks, it is important to find the appropriate template that matches the target task. For a pair of arguments $\boldsymbol{x}$ = (x_{arg_1},x_{arg_2}), we transfer them to x_{prompt} with the template:

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

where T represents template function. We discuss our method of constructing templates with a simple hard template "`<s>` x_{arg_1} `<mask>` x_{arg_2} `</s>`". An argument pair x is reformulated into a prompt template T by concatenating two arguments and inserting some PLM-specific tokens such as `<mask>`, `<s>` and `</s>`, as the input of a PLM. The `<MASK>` token is added for multi-level label predictions, while the `<s>` and `</s>` tokens are used to indicate the beginning and end of an input word sequence, respectively. Note that some PLMs use other tokens like `[MASK]`, `[CLS]`, `[SEP]`, but they have the same meaning as described above.

Then we feed x_{prompt} to the $RoBERTa$ (Liu et al., 2019) model to obtain the hidden states $h_{1:n}$:

$$h_{1:n} = RoBERTa(x_{prompt}), \tag{2}$$

where $h_{1:n} \in \mathbb{R}^{n \times r}$ and n is the length of x_{prompt} and r is the hidden state dimension of RoBERTa.

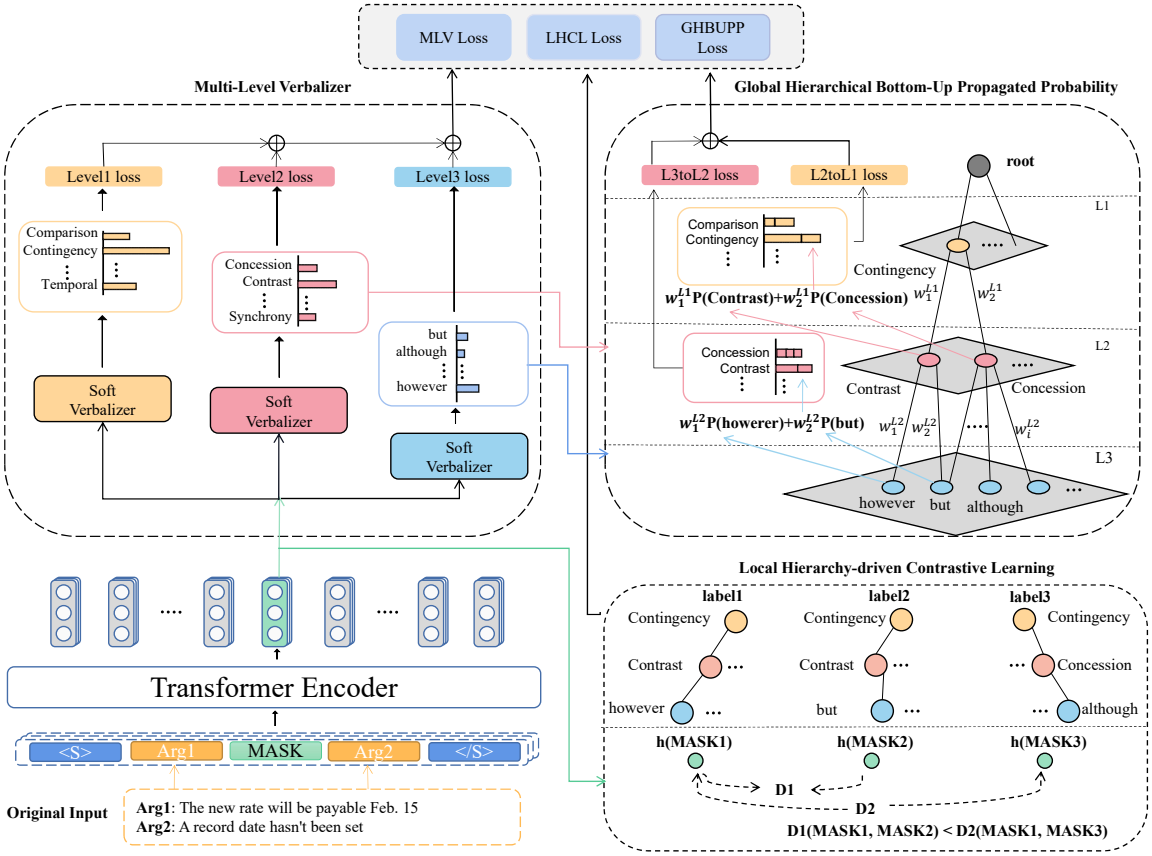


Figure 2: The overall architecture of our framework, which mainly consists of Multi-Level Verbalizer (MLV), Global Hierarchical Bottom-Up Propagated Probability (GHBUPP), and Local Hierarchy-driven Contrastive Learning (LHCL).

Then we pick out the vectors corresponding to the <mask> tokens h_m and construct hierarchies multi-level verbalizer framework.

Instead of picking up verbalizer through handcraft or rules, we create each verbalizer as a virtual vector $W_d \in R^{r\times|L^d|}$ and initialize it by corresponding label embeddings, where r is the hidden state dimension of RoBERTa and $|L^d|$ is the number of labels of d-th level. We can get a verbalizers $\boldsymbol{V} = \{V_1, ..V_d, ..V_D\}$. The <mask> token representations h_m is connected to all vebalizers and the d-th verbalizer predict the d-th level label. The probabilistic predictions of the d-th level is:

$$\hat{\boldsymbol{y}}^d = \{\hat{y}_i^d\}_{i=1}^{|L^d|} = softmax(h_m W_d + b_d), \quad (3)$$

there are different operations for each PLM.

We train this model through cross entropy loss to approximate the real distribution of d-th level as follows:

$$\mathcal{L}^d = -\sum_{i=1}^{|L^d|} y_i^d log(\hat{y}_i^d), \quad (4)$$

where $\boldsymbol{y}^d = \{y_i^d\}_{i=1}^{|L^d|}$ is the one-hot representation of ground-truth relation. Finally, for each input instance, we can calculate the loss of the multi-level verbalizer framework as:

$$\mathcal{L}_{MLV} = \sum_{d=1}^{D} \mathcal{L}^d = -\sum_{d=1}^{D}\sum_{i=1}^{|L^d|} y_i^d log(\hat{y}_i^d). \quad (5)$$

2.3. Global Hierarchical Bottom-Up Propagated Probability

Although single-level IDRR has achieved success, there is still a lack of guidance from hierarchical labels between verbalizers. In order to further exploit the global hierarchy information and narrow the gap between the prompt tuning and hierarchical objectives on IDRR task. Inspired by (Chan et al., 2023; Ji et al., 2023; Zhao et al., 2023), we propose a Global Hierarchical Bottom-Up Propagated Probability (GHBUPP) method utilizes the output probability distribution information of multi-level verbalizer (MLV) for hierarchical modeling, which can inject global hierarchical constraints into MLV and adapt prior knowledge in PLMs to downstream MIDRR task.

Specifically, due to the arguments are annotated by different semantic granularity in the process of labeling for MIDRR, all the labels can form a graph

G with D levels. According to the global label dependency graph G, for a particular label i of d-th level y_i^d, its relevant child-labels in level d+1 can be denoted as:

$$Y_i^{d+1} = \{y^{d+1} | Parent(y_j^{d+1}) = y_i^d\}, \quad (6)$$

where $Parent(\cdot)$ means the parent node of it.

In addition, the weight of influence of each child node on the parent node depends on numerous factors, e.g., the label distribution of datasets, the semantic importance of the parent label, polysemy and so on. Hence we apply several learnable weight units in the process of GHBUPP to balance the influence of multiple factors, which is equal to adding weights to the edges in G. All the weights are acquired through the iteration of prompt tuning. For a particular label y_i^d, its relevant child labels normalized weight list can be denoted as $\boldsymbol{w}_i^d = [w_{i1}^d, w_{i2}^d, .., w_{in}^d]$, where n is the number of set Y_i^{d+1}, which is the number of child labels of y_i^d.

For argument pair x, the bottom-up propagated probability on label y_i^d can be calculated through:

$$p_i^d = P(y_i^d|x) = \sum_{j=1}^{n} w_{ij}^d P(y^{d+1}|x) y^{d+1} \in Y_i^{d+1}. \quad (7)$$

Thus, the loss of propagation probability of the d-th level is:

$$\mathcal{L}_P^d = -\sum_{i=1}^{|L^d|} y_i^d log(\hat{p}_i^d), \quad (8)$$

where $\hat{p}_i^d$ is the normalized result of p_i^d. The propagation probability is bottom up in our method, the loss of the GHBUPP can be obtained as follows:

$$\mathcal{L}_{GHBUPP} = \sum_{d=1}^{D-1} \mathcal{L}_P^d = -\sum_{d=1}^{D-1} \sum_{i=1}^{|L^d|} y_i^d log(\hat{p}_i^d). \quad (9)$$

2.4. Local Hierarchy-driven Contrastive Learning

Despite global hierarchies have been effectively utilized, MLV suffers from the absence of local hierarchical label guidance. Besides, existed hierarchical infusion methods (Chen et al., 2021; Wang et al., 2022b; Chan et al., 2023; Zhao et al., 2023) overlook the use of local hierarchies in prompt tuning.

Therefore, we design a Local Hierarchy-driven Contrastive Learning(LHCL) to learn the hierarchy-aware matching relationship to guide the distance between <MASK> token representations, which can further improve the hierarchical recognition of multi-level verbalizer. As shown in the lower right part of figure 2, label sequences of different instances share different numbers of label nodes, *label1* and *label2* share two labels in the hierarchy, while *label1* and *label3* only share one. Naturally, the distance between *label1* and *label2* should be closer than that between *label1* and *label3*.

Following SimCSE(Gao et al., 2021b; Wang et al., 2022a), denote $B = \{(X_n, Y_n)\}$ as one batch where $Y_n = \{y^1, .., y^d, ..y^D\}_n$, $n \in N$, where N denotes the batch size and D denotes the depth of the label hierarchy. we duplicate a batch of training instances B as $B+$ and feed B as well as $B+$ through Encoder E with diverse dropout augmentations to obtain $2N$ sets of hidden vectors for all corresponding `[MASK]` tokens $Z = \{z \in \{h^d\} \cup \{h_+^d\}\}$. In addition, labels at different levels have different levels of importance, so we add a coefficient λ^d for each layer to control the degree of labels importance.Thus we calculate the score of shared local labels between instance i and j:

$$l_{ij} = \sum_{d=1}^{D} \lambda^d \mathbb{I}_{y_i^d, y_j^d}, \quad (10)$$

where $\mathbb{I}_{y_i^d, y_j^d}$ equals 1 when y_i^d and y_j^d are equal, and 0 otherwise. In fact, when the weight coefficients λ for different layers are equal to 1, l_{ij} represents the number of shared nodes.

Then the local hierarchical contrastive loss for each instance pair (i, j) can be calculated as:

$$\mathcal{L}_l^{ij} = -\alpha_{ij} \log \frac{\exp(sim(h_i, h_j)/\tau)}{\sum_{j \in B_+} \exp(sim(h_i, h_j)/\tau)}, \quad (11)$$

$$\alpha_{ij} = \frac{l_{ij}}{\sum_{k \in B_+} l_{ik}}, \quad (12)$$

where α_{ij} is the normalization of l_{ij}, τ is the temperature of contrastive learning, h_i is the hidden states of `<MASK>` for instance i. Thus, the local hierarchy-driven contrastive loss is:

$$\mathcal{L}_{LHCL} = -\frac{1}{N} \sum_{i \in B} \sum_{j \in B_+} \mathcal{L}_l^{ij}. \quad (13)$$

With this contrastive loss, for an instance pair (i, j), the more sub labels they share, the larger the weight α_{ij} will become, thus increasing the value of their loss term $\mathcal{L}_l^{ij}$. In consequence, the distance between h_i and h_j will become closer. On the contrary, if they share fewer sub labels in the label sequence, their distance between h_i and h_j will be optimized relatively farther. Thus, the local hierarchy contrastive loss utilizes the similarity scores α_{ij} to guide the distance between between `<MASK>` representations of different instances.

The overall training objective is the combination of hierarchies multi-level verbalizer loss, bottom-up propagated probability loss, and local hierarchical contrastive loss:

$$\mathcal{L} = \mathcal{L}_{MLV} + \lambda_1 \mathcal{L}_{GHBUPP} + \lambda_2 \mathcal{L}_{LHCL}, \quad (14)$$

where λ_1 and λ_2 are the hyperparameters controlling the weights of corresponding loss.

3. Experiment Settings

3.1. Dataset

The experiments are conducted on two datasets, the PDTB 2.0 (Prasad et al., 2008) and the PDTB 3.0 (Webber et al., 2019), to validate the performance of our method. PDTB corpora are annotated with information related to discourse semantic relation. Following previous work (Wu et al., 2020, 2022; Zhao et al., 2023), we regard the connectives as the third level for MIDRR. The PDTB 2.0 contains 4 (Top Level), 11 (Second Level) and 102 (Connectives) categories for each level. The PDTB 3.0 contains 4 (Top Level), 14 (Second Level) and 186 (Connectives) categories for each level. For data partitioning, 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.More details of the PDTB-Ji splitting are shown in Appendix A.

3.2. Implementation Details

Our work uses Pytorch and Huggingface libraries for development, and also verifies the effectiveness of our model on MindSpore library. For better comparison with recent models, We use the pretrained RoBERTa-base (Liu et al., 2019) as our Transformer encoder. We adopt AdamW optimizer with a learning rate of 1e-6 and a batch size of 8 to update the model parameters for 10 epochs. The evaluation step is set to 400 and chooses models based on the best result on the development set. All experiments are conducted with a single RTX 3090 GPU and one training process takes about 80 minutes. Finally, We choose the macro-F1 and accuracy as our validation metrics and report the mean performance over 5 random seeds.

3.3. Baselines

In this section, we select some baselines for PDTB 2.0 and 3.0 separately and introduce them briefly.
Baselines for PDTB 2.0:

1) **RoBERTa-FT**: (Liu et al., 2019) improves the BERT by removing the NSP task. We conduct experiments for each level separately.

2) **HierMTN-CRF** (He et al., 2020): a model that firstly deals with multi-level IDRR simultaneously and chooses the label sequence based on a CRF layer.

3) **BMGF** (Liu et al., 2020): a model that proposes a bilateral multi-perspective matching encoder to enhance the arguments interaction on both text span and sentence level.

4) **LGSGM** (Wu et al., 2022): a label sequence generation model that leverages the label dependencies between discourse relations through GCN.

5) **PCP** (Zhou et al., 2022): a model that firstly constructs manual template to mine the strong correlation between connectives and discourse relation.

6) **ChatGPT** (Chan et al., 2023): a ChatGPT based method equipped with an in-context learning prompt template.

7) **GOLF** (Jiang et al., 2023): a global and local hierarchy-aware contrastive framework, to model and capture the information from these two kinds of hierarchies with the aid of contrastive learning.

8) **DiscoPrompt** (Chan et al., 2023): a prompt-based path prediction model to utilize the interactive information and intrinsic senses among the hierarchy in IDRR.

9) **PEMI** (Zhao et al., 2023): a model that leverages parameter efficient prompt tuning to drive the arguments to match the pre-trained space and realize the approximation with few parameters .

Baselines for PDTB 3.0:

1) **RoBERTa-FT**: (Liu et al., 2019) We conduct experiments for each level separately on PDTB3.0.

2) **MANF** (Xiang et al., 2022b): a dual attention model to encodes word-pairs offsets to enhance semantic interaction.

3) **ConnPrompt** (Xiang et al., 2022c): a model that transforms the relation prediction task as a connective-cloze prediction task.

4) **TEprompt** (Xiang et al., 2023): a dual attention model to encodes word-pairs offsets to enhance semantic interaction.

5) **GOLF** (Jiang et al., 2023): a global and local hierarchy-aware contrastive framework.

6) **PEMI** (Zhao et al., 2023): a parameter efficient prompt tuning framework.

3.4. Results and Analysis

In this section, we display the main results of three levels on PDTB 2.0 and PDTB 3.0 (Table 1) and the label-wise F1 of level 2 (Table 2). We can obtain the following observations from these results:

- In table1, our model achieves comparable performance with strong baselines and achieves state-of-the-art performance at both top and second-level classes in PDTB 2.0. It demonstrates that our model can more accurately identify discourse relations at different levels, demonstrating improved classification and generalization capabilities, and effectively handling complex hierarchical relationships. This further validates the model's capabilities and effectiveness. Specifically, our method gains a considerable improvement of 2.07% top-level F1 score, 1.37% top-level accuracy, 1 .99% second-level F1 score, 0.17% second-level accuracy and 0.97% connectives F1 score over the existing state-of-the-art model on PDTB

Model	Embedding	Top Level		Second Level		Connectives	
		F_1	ACC	F_1	ACC	F_1	ACC
PDTB2.0							
HierMTN-CRF (Wu et al., 2020)	BERT	55.72	65.26	33.91	53.34	10.37	30.00
BMGF (Liu et al., 2020)	RoBERTa-base	63.39	69.06	-	58.13	-	-
RoBERTa-FT(Fine-tuning)	RoBERTa-base	64.18	70.34	43.24	59.63	10.03	30.35
LDSGM (Wu et al., 2022)	RoBERTa-base	63.73	71.18	40.49	60.33	10.68	32.20
PCP (Zhou et al., 2022)	RoBERTa-base	64.95	70.84	41.55	60.54	-	-
ChatGPT (Chan et al., 2023)	-	44.09	50.24	19.88	31.95	-	-
GOLF (Jiang et al., 2023)	RoBERTa-base	65.76	72.52	41.74	61.16	11.79	32.85
DiscoPrompt (Chan et al., 2023)	T5-base	60.66	70.63	45.99	60.84	-	-
PEMI (Zhao et al., 2023)	RoBERTa-base	64.05	71.13	41.31	60.66	10.87	**35.32**
Ours	RoBERTa-base	**67.83**	**73.89**	**46.14**	**61.33**	**12.76**	33.22
PDTB3.0							
RoBERTa-FT(Fine-tuning)	RoBERTa-base	68.45	72.31	52.04	61.73	13.32	39.56
MANF(Xiang et al., 2022b)	BERT	56.63	64.04	-	-	-	-
ConnPrompt (Xiang et al., 2022c)	RoBERTa-base	69.51	73.84	-	-	-	-
TEprompt (Xiang et al., 2023)	RoBERTa-base	**72.26**	75.51	-	-	-	-
GOLF (Jiang et al., 2023)	RoBERTa-base	70.88	75.03	55.30	63.57		
PEMI (Zhao et al., 2023)	RoBERTa-base	69.06	73.27	52.73	63.09	10.52	39.92
Ours	RoBERTa-base	71.59	**75.53**	**56.50**	**64.87**	**16.64**	**42.10**

Table 1: We report the mean Macro-F1 score (%), Accuracy (%) over 5 random seeds on PDTB 2.0 and PDTB3.0. Bold: best results. Underlined: second highest.

Second-level	Label-wise F1(%)		
	PEMI	GOLF	Ours
Comp.Concession (2%)	**8.11**	0	7.90
Comp.Contrast (12%)	60.20	61.95	**62.74**
Cont.Cause (26%)	61.82	65.35	**66.35**
Cont.Pragmatic cause (1%)	0	0	0
Expa.Alternative (1%)	60.54	**63.49**	51.63
Expa.Conjunction (19%)	50.71	**60.28**	49.39
Expa.Instantiation (12%)	73.81	75.36	**75.81**
Expa.List (1%)	30.55	27.78	**33.87**
Expa.Restatement (20%)	55.60	59.84	**60.19**
Temp.Asynchronous (5%)	53.04	**63.82**	57.17
Temp.Synchrony (1%)	0	0	**36.62**

Table 2: The second-level label-wise F1 scores(%) on PDTB 2.0. The proportion of each sense is listed behind its name.

Model	Macro-F1		
	Top	Sec	Conn
Ours	**67.83**	**46.14**	**12.76**
w.o. GHBUPP	66.62	45.52	11.74
w.o. LHCL	67.37	45.31	12.21
w.o. GHBUPP+LHCL	65.71	45.07	12.02
w.o. MLV+GHBUPP+LHCL	64.18	43.24	10.03

Table 3: Ablation study on PDTB 2.0. "*w/o*" stands for "without"; **GHBUPP** means Global Hierarchical Bottom-Up Propagated Probability; **LHCL** mean Local Hierarchy-driven Contrastive Learning and **MLV** mean Multi-Level Verbalizer.

2.0. Moreover, in the case of PDTB 3.0, our model also also achieved the best results on multiple metrics. Note that, the number of connectives in PDTB3 is 150 types after deduplication, GOLF(Jiang et al., 2023) did not remove duplicate categories, so we did not compare it with it in the Connectives layer.

- Table 2 showcases the label-wise F1 comparison for the second-level senses, results show that our framework enhances the F1 performance of most second-level senses, with a notable increase in *Temp.Synchrony(1%)* from 0% to 36.62%, *Expa.List(1%)* also has a significant increase. Our model breaks the bottleneck of previous work in two few-shot second-level senses.

- Despite large language models (LLMs) such as ChatGPT (OpenAI, 2022) achieving good performance on various few-shot and and zero-shot tasks for various understanding and reasoning tasks (Bang et al., 2023), they are far behind us in all metrics in PDTB 2.0 as shown in table 1, which indicates that chatGPT may struggle to comprehend the abstract sense of discourse relation. Therefore, IDRR remains a challenging and crucial task for the NLP community, requiring further exploration.

3.5. Ablation Study and Analysis

We conduct the ablation study on PDTB 2.0 to deeply analyze the impact of individual modules in our framework. The main parts of our work are

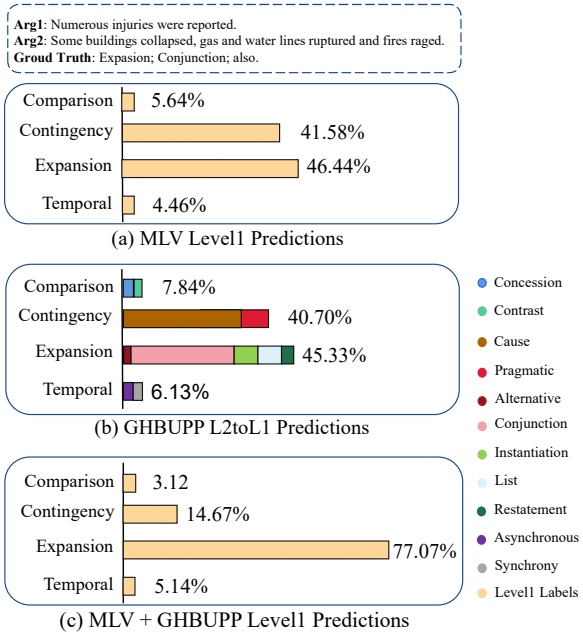
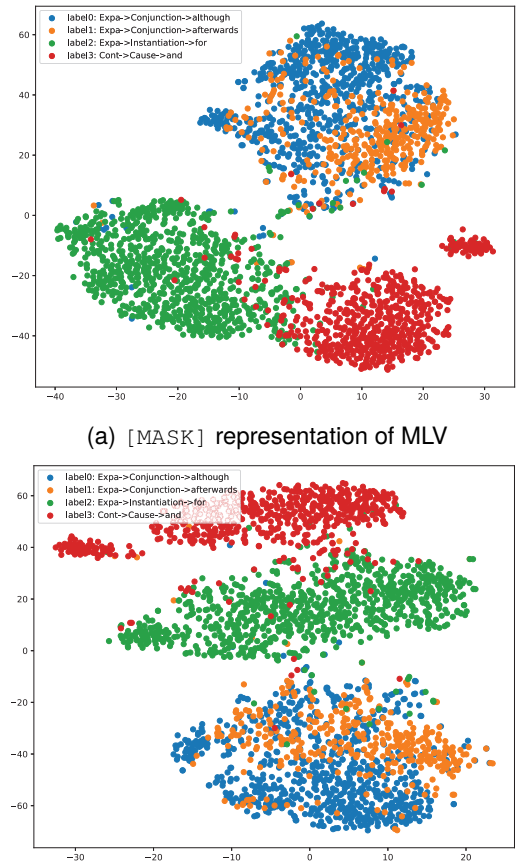


Figure 3: An Instance for GHBUPP analysis.



(a) [MASK] representation of MLV

(b) [MASK] representation of MLV+LHCL

Figure 4: Visualization of LHCL method

the multi-level verbalizer (**MLV**), global hierarchical bottom-up propagated probability (**GHBUPP**) and local hierarchical contrastive learning (**LHCL**), we test our model by gradually removing each component of our model. For fairness, when we remove MLV, we choose fine-tuned RoBERTa MLM with a learnable verbalizer for testing. It treats IDRR of different levels as an individual classification but shares the parameters of the encoder.

From Table 3, we can observe that: **1)** Eliminating any modules would hurt the performance across all three levels, MLV contributes mostly, which demonstrate the effectiveness of each module. **2)** Simultaneously removing both GHBUPP and LHCL has a worse performance compared to removing either one alone. This suggest that GHBUPP and LHCL have a mutually reinforcing effect, LHCL can optimize the representation of <MASK> by introducing local hierarchy to enhance the recognition of discourse relationships. **3)** While removing MLV, GHBUPP and LHCL together results in the worst F1 score. This suggests that MLV is more beneficial for MIDRR compared to individual veberlizer, GLBUPP and LHCL successfully integrated global and local hierarchical information to improve MLV recognition performance.

3.6. GHBUPP Method Analysis

As shown in Figure 3, we use an instance to explain the impact of our GHBUPP method. Figure 3a shows the predicted probability distribution of level1 soft verbalizer of MLV, Figure 3b shows the predicted probability distribution of GHBUPP on level1, which is obtained by multiplying the probability distribution of level2 soft verbalizer by the weight. We can clearly observe a high degree of similarity in their probability distributions, which indicates that the upper layer's verbalizer prediction can be obtained by the lower layer verbalizer based on the global hierarchical structure. There is global hierarchical information in MLV. Futhermore, the probability predictions of Figure 3c exhibit higher discriminative power, indicating that GHBUPP has a strengthening effect on MLV.

3.7. LHCL Visualization Analysis

As shown in Figure 4, we select some local label sequences for visual analysis of LHCL, and we visualize the <MASK> representations of different instances using t-SNE. *label0* and *label1* share two node labels, *label0* and *label2* share one node label, and *label0* and *label3* do not share a node label. Comparing Figure 4a and Figure 4b, we can observe that *label0* is closest to *label1* and farthest from *label3*. Thus, LHCL learns the local hierarchy-aware matching relationship to guide the distance between <MASK> token representations between input instances. Meanwhile, from the ablation study

in Table 3, we demonstrate that LHCL has an enhancing effect on the recognition performance of MLV.

4. Related Work

4.1. Implicit Discourse Relation Recognition

We introduce deep learning methods for the IDRR (Pitler et al., 2009) through two routes.

One route is **pre-train and fine-tuning paradigm**. Conventional pre-train and fine-tuning paradigm usually approaches the IDRR task as a classification problem, the early work (Zhang et al., 2015; Qin et al., 2016; Rutherford et al., 2017; Bai and Zhao, 2018) tends to design a sophisticated downstream neural network for argument representation learning. Besides, other methods (Liu and Li, 2016; Lan et al., 2017; Guo et al., 2018; Ruan et al., 2020) attempt to capture interactions between argments for argument representation learning. Moreover, several methods (Dai and Huang, 2018; Kishimoto et al., 2018; Guo et al., 2020; Kishimoto et al., 2020) achieve more robust representations through data augmentation or knowledge projection. However, these methods overlook the exploration of relation patterns. Along with the booming development of deep learning, some work simultaneously focuses on exploring the representation of argument pairs and discourse relations. For instance, TransS (He et al., 2020) utilizes a triplet loss to establish spatial relationships between arguments and relation representation. (Liu et al., 2021b) proposed combining the context representation module and bilateral multi-perspective matching module to understand different relational semantics deeply. LDSGM (Wu et al., 2022) utilizes the graph convolutional networks to incorporate label dependencies into text representations. While some studies (Long and Webber, 2022; Jiang et al., 2023) leverages the sense hierarchy to obtain contrastive learning representation. Despite the success of the fine-tuning paradigm, these methods may suffer from distinct training strategies in the pretraining and fine-tuning stages, which restrains the finetuned models to take full advantage of knowledge in PLMs.

Another route is **prompt tuning paradigm**. Recently, inspired by (Schick and Schütze, 2021b), several studies (Zhou et al., 2022; Xiang et al., 2022c, 2023) reformulate an single-level IDDR task as cloze questions to bridge the gap between the masked language model (MLM) and downstream NLP tasks. However, the inherent global discourse label hierarchy is ignored, therefor PEMI (Zhao et al., 2023) propose a hierarchical label refining method to deeply integrate global hierarchical guidance into the prompt tuning. Meanwhile, DiscoPrompt (Chan et al., 2023) transform the hierarchy as the path prompt and add it as the prefix of arguments to be classified. Different from previous works, our work introduce both global and local hierarchical information into prompt tuning to fully explore discourse label hierarchy information.

4.2. Prompt Tuning

With some large-scale PLMs have been proposed, such as the BERT (Dai and Huang, 2018) and GPT (Brown et al., 2020), the prompt learning has become a new paradigm for many NLP tasks, which has achieved promising results (Seoh et al., 2021; Liu et al., 2022). The prompt tuning methods can be broadly divided into hard prompt (Gao et al., 2021a; Wang et al., 2021; Schick and Schütze, 2021b) and soft prompt (Hambardzumyan et al., 2021; Qin and Eisner, 2021). The hard prompt methods select template and label words from the vocabulary of PLMs, which require carefully manual designing. Soft prompt methods learn some continuous vectors directly in the feature space of PLMs, which eliminate the need for manually-designed prompts. The design of verbalizers (Schick and Schütze, 2021a; Gao et al., 2021a; Cui et al., 2022) is also an important step in prompt tuning, which aims to reduce the gap between model outputs and label words.

At the same time, there are some efforts to leverage prompts with structural inputs for knowledge customization (Zhong et al., 2022). Injecting hierarchy information into prompts is also promising (Wang et al., 2022b; Ji et al., 2023; Chan et al., 2023; Zhao et al., 2023). For example, using top-level predictions to refine prompts of bottom levels can surpass soft prompts and hard prompts (Wang et al., 2022b). How to employ PLMs to better involve hierarchy knowledge is still worth exploring.

5. Conclusion

In this paper, we propose a Global and Local Hierarchy Prompt Tuning Framework (GLHPT) to utilize the output of multi-level verbalizer for hierarchical modeling, which simultaneously incorporate global and local hierarchical knowledge into prompt tuning. It narrows the gap between hierarchy objective and PLMs, and effectively adapts prior knowledge in PLMs to downstream MIDRR task. Experiment results show that our model achieves highest at multiple levels on PDTB datasets, and can handle the imbalance and low resource situations. In the future, we will further explore the utilization of hierarchical knowledge and the applicability of GLHPT in other NLP tasks.

6. Limitations

Although our model obtains satisfying results, it also exposes some limitations. **First**, since the difficulty of data annotation has leads to few publicly available datasets of IDRR tasks, for a fair comparison to other models, we mainly carry out relevant experiments and analysis on PDTB 2.0. In the future, We plan to comprehensively evaluate our model on more datasets, including datasets from other languages. **Second**, Since the appearance of large pre-trained models such as chatGPT(OpenAI, 2022), many NLP tasks are being accomplished by directly employing large-scale models for in-context learning without fine-tuning. Our method is suitable for situations with insufficient labeled data, but it is difficult to directly extend to a large-scale language model because large language models are hard to fine-tune in many situations. In future work, we plan to enhance the model's capability on this task by leveraging knowledge distillation from large language models.

7. Ethics Statement

In this work, we conformed to recognized privacy practices and rigorously followed the data usage policy. We do not introduce social/ethical bias into the model or amplify any bias from the data. Therefore, we do not see any potential risks.

8. Acknowledgement

Our Work is supported by the National Natural Science Foundation of China (No. 62376192, No. 62376188) and the CAAI-Huawei MindSpore Open Fund.

9. References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *COLING*, pages 571–583.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang, Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *arXiv*.

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. In *NeurIPS*, pages 1877–1901.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *ACL Findings*, pages 35–57.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *ACL*, pages 4370–4379.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *ACL*, pages 7014–7024.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *ACL*, pages 141–151.

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

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. In *ACL*, pages 3816–3830.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *ICLR*.

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 *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 *COLING*, pages 547–558.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. In *ACL*, pages 4921–4933.

Ruifang He, Jian Wang, Fengyu Guo, and Yugui Han. 2020. TransS-driven joint learning architecture for implicit discourse relation recognition. In *ACL*, pages 139–148.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *ACL*, pages 2918–2933.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is 681not enough: Entity-augmented distributed semantics 682for discourse relations. In *TACL*.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *EMNLP*, pages 2418–2431.

Ting Jiang, Deqing Wang, Leilei Sun, Zhongzhi Chen, Fuzhen Zhuang, and Qinghong Yang. 2022. Exploiting global and local hierarchies for hierarchical text classification. In *EMNLP*, pages 4030–4039.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. In *ACL Findings*, pages 8048–8064.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *COLING*, pages 584–595.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting bert to implicit discourse relation classification with a focus on discourse connectives. In *LREC*.

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 *EMNLP*, pages 1299–1308.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *IJCAI*, pages 4026–4032.

Zhenwen Li, Wenhao Wu, and Sujian Li. 2020b. Composing elementary discourse units in abstractive summarization. In *ACL*, pages 6191–6196.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *EMNLP*, pages 747–757.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *ACL*, pages 3154–3169.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, pages 1–35.

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 *IJCAI*, pages 3830–3836.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021a. Exploring discourse structures for argument impact classification. In *ACL*, pages 3958–3969.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021b. On the importance of word and sentence representation learning in implicit discourse relation classification. In *IJCAI*.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *EMNLP*, pages 1224–1233.

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. In *arXiv*.

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *EMNLP*, pages 10704–10716.

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 *ACL*, pages 4201–4207.

TB OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. In *OpenAI*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL*, pages 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. The penn discourse treebank 2.0. In *ICLR*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *NAACL*, pages 5203–5212.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *EMNLP*, pages 2263–2270.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-propagative attention learning for implicit discourse relation recognition. In *COLING*, pages 3168–3178.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *ACL*, pages 281–291.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *ACL*, pages 255–269.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *NAACL*, pages 2339–2352.

Ronald Seoh, Ian Birle, Mrinal Tak, Haw-Shiuan Chang, Brian Pinette, and Alfred Hough. 2021. Open aspect target sentiment classification with natural language prompts. In *EMNLP*, pages 6311–6322.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *ACL*, pages 672–679.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *ACL*, pages 732–742.

Chengyu Wang, Jianing Wang, Minghui Qiu, Jun Huang, and Ming Gao. 2021. TransPrompt: Towards an automatic transferable prompting framework for few-shot text classification. In *EMNLP*, pages 2792–2802.

Ran Wang, Xinyu Dai, et al. 2022a. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *ACL (Short)*, pages 672–679.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *EMNLP*, pages 3740–3751.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. In *Philadelphia, University of Pennsylvania*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. In *AAAI*, pages 11486–11494.

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

Wei Xiang, Chao Liang, and Bang Wang. 2023. TEPrompt: Task enlightenment prompt learning for implicit discourse relation recognition. In *ACL Findings*, pages 12403–12414.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. In *ACM Computing Surveys*, 12, pages 1–34.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022a. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In *ACL Findings*, pages 3247–3257.

Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. 2022b. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In *ACL Findings*, pages 3247–3257.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022c. ConnPrompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *COLING*, pages 902–911.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *EMNLP*, pages 2230–2235.

Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. Context tracking network:

Graph-based context modeling for implicit discourse relation recognition. In *NAACL*, pages 1592–1599.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *ACL*, pages 6477–6492.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. ProQA: Structural prompt-based pre-training for unified question answering. In *NAACL*, pages 4230–4243.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. In *EMNLP Findings*, pages 3848–3858.

10. Appendices

A. Details of PDTB-Ji Splitting

In this section, we provide data statistics for PDTB 2.0 and PDTB 3.0.

Second-level	Sample Size		
	Train	Dev	Test
Comp.Concession	183	15	17
Comp.Contrast	1607	166	128
Cont.Cause	3270	281	269
Cont.Pragmatic cause	64	6	7
Expa.Alternative	147	10	9
Expa.Conjunction	2872	258	200
Expa.Instantiation	1063	106	118
Expa.List	338	9	12
Expa.Restatement	2404	260	211
Temp.Asynchronous	532	46	54
Temp.Synchrony	203	8	14
Total	12683	1165	1039

Table 4: Statistics for relation senses of Level 2 in PDTB 2.0 by PDTB-Ji splitting. Comp, Cont, Exp and Temp represents Comparison, Contingency, Expansion and Temporal separately.

Second-level	Sample Size		
	Train	Dev	Test
Comp.Concession	1164	103	97
Comp.Contrast	741	82	54
Cont.Cause	4475	448	404
Cont.Cause+Belief	159	13	15
Cont.Condition	159	13	15
Cont.Purpose	1092	96	89
Expa.Conjunction	3586	298	236
Expa.Equivalence	254	25	30
Expa.Instantiation	1063	106	118
Expa.Level-of-detail	2601	261	208
Expa.Manner	615	14	17
Expa.Substitution	343	27	26
Temp.Asynchronous	1007	101	105
Temp.Synchrony	435	33	43
Total	17788	1635	1463

Table 5: Statistics for relation senses of Level 2 in PDTB 3.0 by PDTB-Ji splitting.

B. Experimental Results for level-2 senses on PDTB 3.0

Due to the limitation of pages, we provide results of PDTB 3.0 in this section. Table 7 displays the labelwise F1 for level-2 senses on PDTB 3.0

Level	PDTB2.0	PDTB3.0
	Number of categories	
Top Level	4	4
Second Level	11	14
Connectives	102	150

Table 6: The number of categories at different levels in PDTB2.0 and PDTB3.0. Note that the number of connectives in PDTB3 is 150 categories after deduplication.

Second-level	Label-wise F1(%)		
	GOLF	PEMI	Ours
Comp.Concession (7%)	59.09	64.68	**64.25**
Comp.Contrast ((4%))	43.33	**52.94**	48.70
Cont.Cause (26%)	69.47	69.04	**71.33**
Cont.Cause+Belief(1%)	0	0	0
Cont.Condition(1%)	66.67	68.97	**92.31**
Cont.Purpose (4%)	71.60	91.49	93.57
Expa.Conjunction (16%)	**64.09**	58.82	58.32
Expa.Equivalence	**25.39**	0	17.14
Expa.Instantiation (9%)	**75.53**	70.42	68.78
Expa.Level-of-detail(15%)	52.60	54.25	**59.59**
Expa.Manner (4%)	63.53	59.26	**73.91**
Expa.Substitution (2%)	**66.67**	48.98	56.25
Temp.Asynchronous (7%)	68.79	66.67	**68.22**
Temp.Synchrony (2%)	**41.00**	32.73	36.92

Table 7: The second-level label-wise F1 on PDTB 3.0.

C. Template Selection

In the prompt paradigm, using different templates may impact on the task performance. We have designed some different prompt templates in Table 8. Table 9 compares the results of prompt templates. If we choose Prompt4 in Table 8, we input the <mask> representation of each level into the corresponding level's Verbalizer for prediction. From Table 8, it can be seen that the performance of Prompt 1 and Prompt 4 is similar. Therefore, for simplicity, we chose Prompt 1 as our final input template.

Prompt Template	Template Form
Prompt 1	*Arg1*,<MASK>,*Arg2*.
Prompt 2	Arg1:*Arg1*,Arg2:*Arg2*. </s></s>The conjunction between Arg1 and Arg2 is <mask>.
Prompt 3	Arg1:*Arg1*, Arg2:*Arg2*. {"soft":"The connective word between"} Arg1 {"soft":"and"} Arg2 {"soft":"is"} <mask>.
Prompt 4	*Arg1*, *Arg2*. The top level is <mask>; The second level is <mask>;The third conjunction is <mask>.

Table 8: Different prompt template form, {"soft":...} represents the initialized soft prompt template

Prompt Template	Top Level		Second Level		Connectives	
	F_1	ACC	F_1	ACC	F_1	ACC
Prompt 1	**67.83**	**73.89**	**46.14**	**61.14**	12.76	33.22
Prompt 2	65.96	72.37	44.76	60.11	12.02	33.35
Prompt 3	65.75	72.51	44.37	60.46	13.03	33.11
Prompt 4	66.15	73.24	46.00	60.69	**13.27**	**34.09**

Table 9: Different prompt template results.