NCPrompt: NSP-Based Prompt Learning and Contrastive Learning for Implicit Discourse Relation Recognition

Yuetong Rong, Yijun Mo*
School of Computer Science and Technology,
Huazhong University of Science and Technology, Wuhan, China
{m202273670,moyj}@hust.edu.cn

Abstract

Implicit Discourse Relation Recognition (IDRR) is an important task to classify the discourse relation sense between argument pairs without an explicit connective. Recently, prompt learning methods have demonstrated success in IDRR. However, prior work primarily transform IDRR into a connective-cloze task based on the masked language model (MLM), which limits the predicted connective to one single token. Also, they fail to fully exploit critical semantic features shared among various forms of templates. In this paper, we propose NCPrompt, an **N**SP-based prompt learning and **C**ontrastive learning method for IDRR. Specifically, we transform the IDRR task into a next sentence prediction (NSP) task, which can allow various-length answer connectives and enlarge the construction of the verbalizer for prompt-learning methods. Also, we notice that various prompt templates naturally constitute positive samples applied for self-supervised contrastive learning. And the usage of NSP naturally creates hard negative samples by introducing different candidate connectives between the same example. To our knowledge, we are the first to combine self-supervised contrastive learning with prompt learning to obtain high-quality semantic representations. Experiments on the PDTB 3.0 corpus have demonstrated the effectiveness and superiority of our model.

1 Introduction

Implicit Discourse Relation Recognition (IDRR) aims at classifying the relation sense between a pair of text segments (called arguments) without an explicit connective (Xiang and Wang, 2023). IDRR provides essential information for many Natural Language Processing (NLP) tasks, such as question answering (Jansen et al., 2014) and machine translation (Li et al., 2014). Without explicit connectives as triggers, IDRR is a challenging task that heavily depends on understanding the semantics of natural language text and even performs poorly with ChatGPT (Chan et al., 2023b,a).

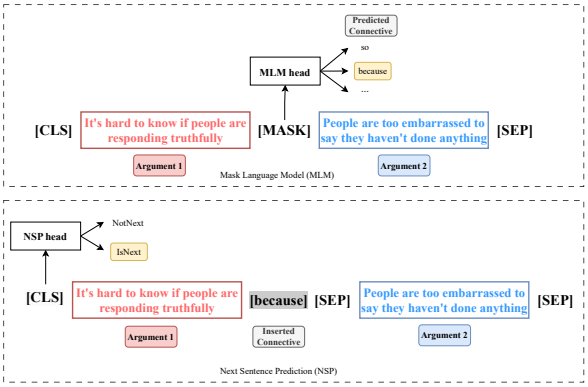


Figure 1: Examples of comparison between MLM-based and NSP-based prompt learning for IDRR in the PDTB 3.0 corpus.

The challenge and key point of the IDRR task is to learn high-quality semantic features of argument pairs. Leveraging the powerful ability of the Pre-trained Language Model (PLM) in representation learning, the *pre-train, prompt, and predict* paradigm, also known as prompt learning (Liu et al., 2023), has replaced the *pre-train and fine-tune* paradigm as the mainstream solution for IDRR. Most existing models (Xiang et al., 2022b; Zhou et al., 2022; Xiang et al., 2023; Wu et al., 2023) reformulate the IDRR task into a connective-cloze task, consistent with the masked language model (MLM) task of PLMs, and map the predicted connective to a relation sense label through the defined verbalizers. Despite their success, these prompt-learning methods have some limitations.

On the one hand, the MLM task can only predict one single token for the masked slot. Thus, only individual connectives instead of phrases can be selected as answer words, limiting the construction

This work is supported by Hubei intelligent edge computing research institute, Hubei science and technology talent service project (2024DJC078).

Findings of the Association for Computational Linguistics: EACL 2024, pages 1159–1169

of verbalizers. Obviously, phrases can convey more precise and detailed meaning than individual words, expected to improve the model performance. In fact, besides the commonly-used MLM task, there also exists a sentence-level pre-training task, next sentence prediction (NSP), in BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019), which is trained to distinguish whether two sentences appear consecutively within a document (Qiu et al., 2020). We believe it is appropriate to reformulate the IDRR task to match the NSP task in prompt-learning methods since both tasks characterize the relationship between sentences. As illustrated in Figure 1, different from the MLM task requiring the single-token answer words, various-length connectives can be inserted between argument pairs, and NSP can tell which one connects the sentences most reasonably, which allows multi-token answer words and expands the construction of verbalizers.

On the other hand, mutual and critical semantic features shared between various forms of prompt templates have not been fully explored in current methods. Most prompt-learning methods for IDRR (Xiang et al., 2023; Wu et al., 2023; Chan et al., 2023b) elaborately design one single type of template, PCP (Zhou et al., 2022) searches the one that achieves the best performance from all candidate templates, and ConnPrompt (Xiang et al., 2022b) employs a simple majority voting decision to fuse different prompt predictions. In fact, we notice that different prompt templates wrapping the same example can convey critical shared information and naturally constitute multiple augmentation views applied for self-supervised contrastive learning (Jaiswal et al., 2020). Self-supervised contrastive learning creates augmentation views and unmatching views of the same example as positive and negative samples, and captures more informative semantic features by pulling positives together and pushing negatives apart. Self-supervised contrastive learning has been widely applied in computer vision tasks (Tian et al., 2020; Chen et al., 2020; Ciga et al., 2022), but barely observed in NLP since defining multiple views for text representations is less intuitive than it is for images.

As we have mentioned, various prompt templates naturally constitute positive samples (Jian et al., 2022). Also, using NSP can naturally create hard negative samples by introducing different candidate connectives between the same example. Hard negative samples that have different labels from the anchor but that are embedded nearby are likely to provide the most useful gradient information during training (Robinson et al., 2020). In IDRR, the same argument pair connected with improper connectives can highlight connective differences, guiding the model to capture critical semantic features of embeddings. Built on these motivations, we propose NCPrompt, an **N**SP-based prompt learning and **C**ontrastive learning method for IDRR. To our knowledge, our work is the first to combine self-supervised contrastive learning with prompt learning benefitting from the NSP task. Our main contributions are as follows:

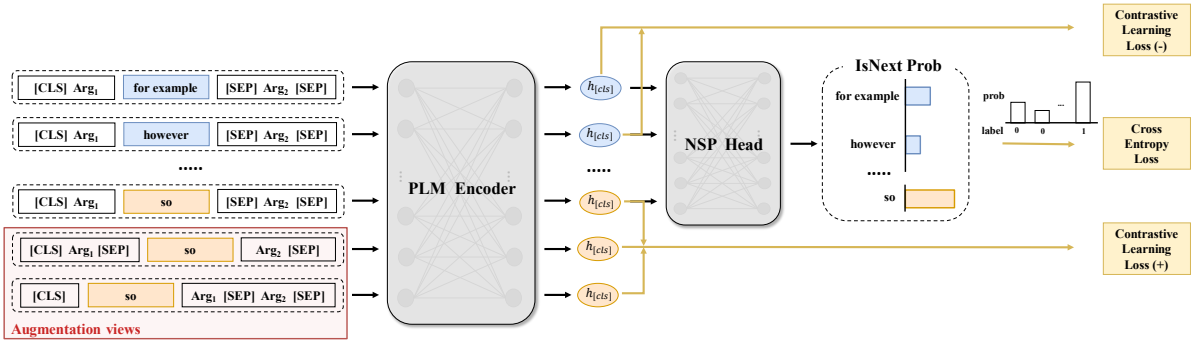
- We reformulate the IDRR task into an NSP task that allows various-length answer words and enlarges the construction of the verbalizer for prompt-learning methods.
- We are the first to combine self-supervised contrastive learning with prompt learning, which can capture critical semantic representation features.
- Experiments on PDTB 3.0 corpus have demonstrated the superiority of our proposed NCPrompt over competitive baselines.

2 Related Work

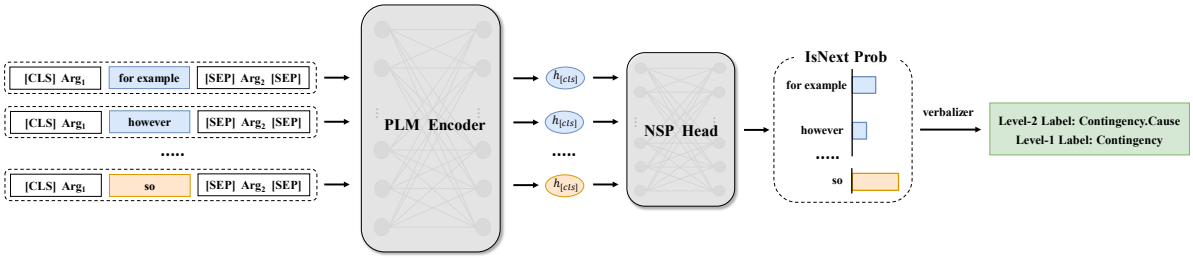
2.1 Implicit Discourse Relation Recognition

IDRR is a major challenge in NLP whose difficulty lies in learning informative representations of argument pairs. With the emergence of powerful PLMs like BERT (Devlin et al., 2019), *pre-train and fine-tune* paradigm has been applied in IDRR (Ruan et al., 2020; Li et al., 2020; Liu et al., 2020; Xiang et al., 2022a; Liu and Strube, 2023) which transfer the pre-trained representations to downstream tasks to encode argument pairs into embeddings. However, such paradigm may result in poor utilization of PLM knowledge.

Recently, the prompt learning paradigm is proposed to bridge the gap between pre-training and downstream task objectives and successfully employed for IDRR. After the first trial of prompt learning for IDRR of ConnPrompt (Xiang et al., 2022b) and PCP (Zhou et al., 2022), the CP-KD (Wu et al., 2023) and AdaptPrompt (Wang et al., 2023a) model combine knowledge distillation with prompt learning, and TEPrompt (Xiang et al., 2023) introduces auxiliary tasks to represent the intrinsic correlation between connectives and relations. DiscoPrompt (Chan et al., 2023b) injects discourse la-



(a) Training process of NCPrompt.



(b) Testing process of NCPrompt.

Figure 2: Overview of the proposed NCPrompt model

bel structure information into prompts. Also, PLSE (Wang et al., 2023b) designs a Cloze-Prompt template based on explicit connective prediction to inject knowledge from unannotated explicit data into the pre-training phase, and implicit connective prediction to bridge the gap between the pre-training and the downstream task. However, most of these work design cloze-format prompts based on MLM, limiting the answer words to one single token. Therefore, we first propose to transform IDRR into NSP by inserting various-length connectives between argument pairs and predicting whether they come consecutively.

2.2 Contrastive Learning

Contrastive learning constructs more informative representations by pulling similar examples (called positives) together and pushing dissimilar examples (called negatives) apart. Some existing methods (Jiang et al., 2023; Long and Webber, 2022; Yi-heng et al., 2024) apply **S**upervised **C**ontrastive **L**earning (Khosla et al., 2020) for IDRR, which contrast examples from the same relation label as positives against the negatives from different labels. For example, Long and Webber (2022) leverages the sense hierarchy to get contrastive learning representation, making examples from the same types at level-2 or level-3 stay close to each other while sister types far apart.

Differently, **S**elf-supervised **C**ontrastive **L**earning (Jaiswal et al., 2020) focuses on one single example, which contrasts augmentation views of the example as positives against the negatives from unmatching views. On the foundation of NSP-based prompt learning, we notice the applicability of self-supervised contrastive learning in NLP besides computer vision, and we first propose to apply it for IDRR.

3 The Proposed NCPrompt Model

Figure 2 presents the overview of our NCPrompt model. Overall, we first design prompts to reformulate the IDRR task into an NSP task. The PLM outputs the coherence score of the argument pair connected by each connective from the answer space. During training, we design positives and negatives for self-supervised contrastive learning, and combine the contrastive loss with the connective classification loss. During testing, the connective with the highest coherence score is mapped into the answer relation sense label through the verbalizer.

3.1 Prompt Template

The argument pair is transformed to the format of NSP-input through the prompt template $T(Arg_1, Arg_2) = T(x)$. Specifically, every an-

Level-1 labels	Level-2 labels	Connective words
Comparison	Concession	*although, however*
	Contrast	*in contrast*
Contingency	Cause	*because, so*
	Cause+Belief	*in fact*
	Condition	*if*
	Purpose	*for*
Expansion	Conjunction	*and*
	Equivalence	*in other words*
	Instantiation	*for example*
	Level-of-detail	*specifically*
	Manner	*thereby*
	Substitution	*instead*
Temporal	Asynchronous	*then, previously*
	Synchronous	*simultaneously*

Table 1: Answer space of our NCPrompt and mapping to the level-2 and level-1 implicit discourse relation sense labels in the PDTB 3.0 corpus.

swer connective v_i is inserted between an argument pair x to form a particular prompt template $p_i = T(x, v_i)$, denoted as:

$$\boldsymbol{T(x, v_i)} = [\boldsymbol{CLS}] + \boldsymbol{Arg_1} + \boldsymbol{v_i} + [\boldsymbol{SEP}] + \boldsymbol{Arg_2} + [\boldsymbol{SEP}]$$

where v_i refers to every candidate connective in the answer space $V = \{v_1, v_2, ..., v_k\}$ and $|V| = k$ is the total number of connectives.

According to the annotation statistics of the PDTB 3.0 corpus, we select the least ambiguous and most representative connectives as answer space, and map each of them to a specific level-2 relation sense label and then to a level-1 relation label. Our verbalizer is shown in Table 1.

Specifically, we can obtain the manually-annotated labels of all three levels of relation senses as its ground-truth relation labels for every argument pair. Accordingly, we can obtain the *ground-truth connective label* which is not the manually-annotated one but the specific connective mapped to the ground-truth level-3 relation sense label through our verbalizer.

3.2 Self-supervised Contrastive Learning

In NCPrompt, due to NSP-based prompt learning, we develop a method to construct positive and negative samples and introduce self-supervised contrastive learning to obtain informative embeddings.

Specifically, there is one ground-truth connective denoted as $\boldsymbol{v_g}$ for argument pair x, leading to prompt $p_g = T(x, v_g)$. For positives, we design two auxiliary prompts as the augmentation views of the anchor p_g, as below:

$$\boldsymbol{T_1(x, v_g)} = [\boldsymbol{CLS}] + \boldsymbol{Arg_1} + [\boldsymbol{SEP}] + \boldsymbol{v_g} + \boldsymbol{Arg_2} + [\boldsymbol{SEP}]$$
$$\boldsymbol{T_2(x, v_g)} = [\boldsymbol{CLS}] + \boldsymbol{v_g} + \boldsymbol{Arg_1} + [\boldsymbol{SEP}] + \boldsymbol{Arg_2} + [\boldsymbol{SEP}]$$

We denote the above two prompts as p_g^1 and p_g^2 respectively. Evidently, we construct our augmentation prompts by altering the order of the connectives and the two arguments. We emphasize the appropriate combination of x and v_g to be mapped to the ground-truth relation labels, while neglecting the potential grammatical logic inconsistencies arising from different permutations of connective order where the connective is situated either between or at the beginning of the arguments. These various forms of prompt templates wrapping the ground-truth sample (x, v_g) can convey critical mutual information from multiple views.

Meanwhile, the argument pair x connected with improper connectives $v_{i(i \neq g)}$ naturally constitute negative samples, denoted as $p_{i(i \neq g)} = T(x, v_{i(i \neq g)})$. Actually, these negatives only different from the anchor in connectives serve as hard negative samples, which have different labels from the anchor but are embedded nearby, expected to provide substantial connective guidance.

Although supervised contrastive learning has been applied for IDRR (Jiang et al., 2023; Long and Webber, 2022; Yi-heng et al., 2024), self-supervised contrastive learning is hardly observed. To further clarify the difference, we illustrate the comparison in Figure 3.

3.3 Model Prediction

Feeding the prompt templates into the PLM encoder M, we can obtain the hidden state vector of $[CLS]$ token denoted as $h_{[CLS]}$ for every input. Then, the NSP head outputs the prediction scores of the relationship between input sentences, denoted as $q_M(n|input)$:

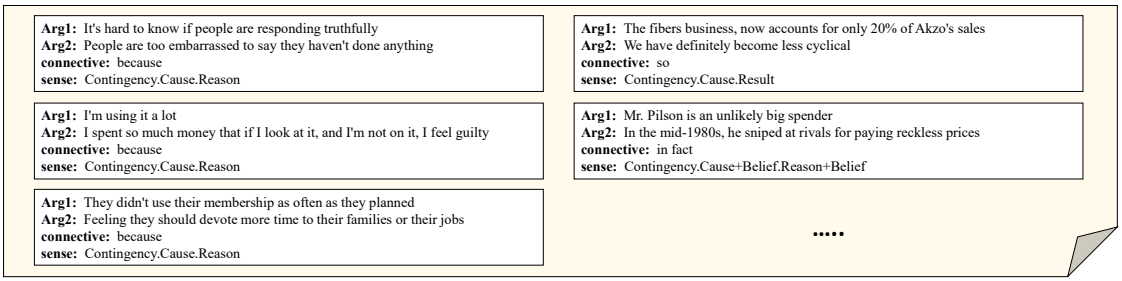
$$q_M(n|input) = W_{nsp} h_{[CLS]} + b_{nsp}, \quad (1)$$

where $n \in \{IsNext, NotNext\}$, W_{nsp} and b_{nsp} are learnable parameters. We take the $IsNext$ score of NSP head as the output logit of the current connective v_i towards prompt p_i, which is:

$$p(v_i|x) = q_M(n = IsNext|x; p_i) \quad (2)$$

Then, a softmax layer is applied to the output logits of all candidate connectives V to normalize them into output probabilities:

$$P(v_i|x) = \frac{exp\, p(v_i|x)}{\sum_{j=1}^{k} exp\, p(v_j|x)} \quad (3)$$



	Self-supervised Contrastive Learning	Supervised Contrastive Learning
Anchor	[CLS] It's hard to know if people are responding truthfully **because** [SEP] People are too embarrassed to say they haven't done anything [SEP]	[CLS] It's hard to know if people are responding truthfully [SEP] People are too embarrassed to say they haven't done anything [SEP]
Positives	[CLS] It's hard to know if people are responding truthfully [SEP] **because** People are too embarrassed to say they haven't done anything [SEP]	[CLS] I'm using it a lot [SEP] I spent so much money that if I look at it, and I'm not on it, I feel guilty [SEP]
	[CLS] **because** It's hard to know if people are responding truthfully [SEP] People are too embarrassed to say they haven't done anything [SEP]	[CLS] They didn't use their membership as often as they planned [SEP] Feeling they should devote more time to their families or their jobs [SEP]
Negatives	[CLS] It's hard to know if people are responding truthfully **so** [SEP] People are too embarrassed to say they haven't done anything [SEP]	[CLS] The fibers business, now accounts for only 20% of Akzo's sales [SEP] We have definitely become less cyclical [SEP]
	[CLS] It's hard to know if people are responding truthfully **and** [SEP] People are too embarrassed to say they haven't done anything [SEP]	[CLS] Mr. Pilson is an unlikely big spender [SEP] In the mid-1980s, he sniped at rivals for paying reckless prices [SEP]

Figure 3: Examples of comparison between **S**elf-supervised **C**ontrastive **L**earning and **S**upervised **C**ontrastive **L**earning for IDRR in the PDTB 3.0 corpus. For our self-supervised contrastive learning, various connection permutations of the proper answer word and the specific argument pair serve as positives, ignoring any potential logic inconsistencies of these sequences.

During training, the output probability distributions of connectives are utilized to compute classification loss with the *ground-truth connective label*, combined with the contrastive loss. During testing, we choose the connective with the highest output probability as the answer connective $\widehat{v}$ of the current argument pair:

$$\widehat{v} = v_{\underset{i}{argmax}\ P(v_i|x)} \tag{4}$$

Then, the answer connective is mapped into the level-2 and level-1 answer relation senses through the verbalizer in Table 1 as our prediction results.

3.4 Training Strategies

Our overall training goal consists of both cross entropy loss L_{CE} for connective classification and contrastive learning loss L_{CL} for bringing positive samples closer and pushing negative samples away.

Cross Entropy Loss: We define the cross entropy loss as follow:

$$L_{CE} = -\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{v}^{(m)}\ log(\widehat{\boldsymbol{v}}^{(m)}), \tag{5}$$

where $\boldsymbol{v}^{(m)}$ and $\widehat{\boldsymbol{v}}^{(m)}$ are the ground-truth and predicted connective labels of the m-th training argument pair respectively and M is the batchsize.

Contrastive Learning Loss: As illustrated in Figure 2, for the anchor p_g, we obtain 2 positive samples and $k-1$ negative samples in total, which are input into the PLM in the same batch. For a positive pair of examples (i, j), the contrastive learning loss is defined as:

$$l(i,j) = -log\frac{exp(sim(z_i, z_j)/\tau)}{\sum_{l=1}^{k+2} \mathbb{1}_{[l\neq i]} exp(sim(z_i, z_l)/\tau)} \tag{6}$$

where $\mathbb{1}_{[l\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $l \neq i$, $sim(\cdot)$ is the standard cosine similarity and τ is a temperature hyper-parameter. And in our NCPrompt, z is consistent with $h_{[CLS]}$ for every input sample. The contrastive learning loss is computed across all positive pairs in a batch:

$$L_{CL} = \frac{1}{M}\sum_{m=1}^{M}[\sum_{\substack{i,j\in\{p_g,p_g^1,p_g^2\}\\ i\neq j}} l(i,j)] \tag{7}$$

Our total loss is a weighted average of L_{CE} and L_{CL}, which is:

$$L = (1-\lambda)\cdot L_{CE} + \lambda \cdot L_{CL} \tag{8}$$

where λ is a scalar weighting hyper-parameter for the contrastive loss.

Models	PLMs	Level-1		Level-2	
		Acc	F1	Acc	F1
DAGRN	Word2Vec	57.33	45.11	-	-
NNMA	Glove	57.67	46.13	-	-
IPAL	BERT	57.33	51.69	-	-
PLR	BERT	63.84	55.74	-	-
MANF	BERT	64.04	56.63	-	-
ContrastiveIDRR	BERT	68.14	61.84	57.64	42.62
ConnPrompt	BERT	67.46	61.23	-	-
PCP	BERT	67.26	61.53	-	-
TEPrompt	BERT	68.83	63.22	-	-
Ours	BERT	68.62	63.17	58.05	48.62
ConnPrompt	ERNIE	68.35	63.79	-	-
Ours	ERNIE	**70.26**	**66.73**	**60.64**	**51.29**
ChatGPT	gpt-3.5-turbo	32.97	28.79	-	-

Table 2: Acc(%) and F1 score(%) of our NCPrompt and baselines for IDRR on the PDTB 3.0 corpus. The boldface is the best result among all models and the underline is the best result among models in a group.

Models	Comp.	Cont.	Expa.	Temp.
DAGRN	27.34	62.56	64.71	38.91
NNMA	29.15	63.33	65.10	41.03
IPAL	37.31	66.40	66.86	41.25
PLR	35.16	66.97	69.33	43.40
MANF	35.83	66.77	70.00	40.22
ContrastiveIDRR	50.94	72.07	71.07	53.28
ConnPrompt$_{BERT}$	49.64	68.84	72.71	53.73
PCP	48.63	71.84	70.02	55.64
TEPrompt	52.80	71.19	**73.25**	55.64
Ours$_{BERT}$	52.15	72.47	71.41	56.67
ConnPrompt$_{ERNIE}$	54.86	70.52	72.64	57.14
Ours$_{ERNIE}$	**61.34**	**72.73**	72.85	**60.00**

Table 3: F1 score(%) of binary classification for level-1 relation labels on the PDTB 3.0 corpus.

4 Experiment Settings

4.1 Dataset

We conduct our experiments on the PDTB 3.0 corpus, which includes more than one million words of English texts from Wall Street Journal. Also, we follow the conventional data splitting (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. Our experiments focus on the recognition of 4 level-1 implicit discourse relation sense labels and 14 level-2 relation labels. With regard to those instances with multiple annotated labels, we treat them as separate instances following early works (Xiang et al., 2022b, 2023) during training. During testing, the prediction matching the first one of the ground-truth relation labels is regarded as the correct answer. Appendix Table 7 presents the data statistics of implicit discourse relation instances in the dataset.

4.2 Baselines

To validate the effectiveness of our proposed NCPrompt, we compare our method with the recent advanced models. First, we select baselines based on the *pre-train and fine-tune* paradigm:

- **DAGRN** (Chen et al., 2016) uses a gated relevance network to capture semantic interaction.
- **NNMA** (Liu and Li, 2016) represents arguments with the neural networks with multi-level attention.
- **IPAL** (Ruan et al., 2020) uses a cross-coupled network to propagate attention.
- **PLR** (Li et al., 2020) proposes a penalty-based loss re-estimation method to regulate the attention learning.
- **MANF** (Xiang et al., 2022a) fuses semantic connection and linguistic evidence for relation recognition.
- **ContrastiveIDRR** (Long and Webber, 2022) leverages the sense hierarchy to apply supervised contrastive learning for IDRR.

Second, we select some models based on the *pre-train, prompt, and predict* paradigm:

- **ConnPrompt** (Xiang et al., 2022b) transforms the IDRR task as a connective-cloze prediction task based on BERT and other PLMs.
- **PCP** (Zhou et al., 2022) proposes a prompt-based connective prediction method based on RoBERTa.
- **TEPrompt** (Xiang et al., 2023) designs a task enlightenment prompt learning model to fuse learned features from related tasks for IDRR.

Since RoBERTa (Liu et al., 2019) has abandoned the NSP task, our NCPrompt model is based on BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019). Considering the fairness of the comparative experiments, we use the same PLM for our baselines and exclude RoBERTa (Liu et al.,

Level-2 labels	ContrastiveIDRR	Ours$_{BERT}$	Ours$_{ERNIE}$
Comparison.Concession	40.25	46.08	**57.56**
Comparison.Contrast	39.53	**44.44**	42.59
Contingency.Cause	66.32	**66.82**	66.51
Contingency.Cause+Belief	0	0	0
Contingency.Condition	50.00	**78.57**	76.92
Contingency.Purpose	89.84	92.55	**92.97**
Expansion.Conjunction	56.02	58.75	**60.74**
Expansion.Equivalence	0	13.33	**15.79**
Expansion.Instantiation	61.08	59.71	**63.30**
Expansion.Level-of-detail	48.11	42.11	**50.00**
Expansion.Manner	34.48	35.90	**40.00**
Expansion.Substitution	37.21	50.98	**51.52**
Temporal.Asynchronous	61.78	**65.95**	64.71
Temporal.Synchronous	12.00	25.45	**35.48**

Table 4: F1 score(%) of binary classification for level-2 relation labels on the PDTB 3.0 corpus.

2019). Specifically, we re-implement ContrastiveIDRR (Long and Webber, 2022) and PCP (Zhou et al., 2022) on BERT (Devlin et al., 2019).

Moreover, as ChatGPT has demonstrated strong capabilities in contextual understanding and interactive dialogue, we also propose to try ChatGPT on zero-shot IDRR task by designing appropriate template like:

"Choose the most appropriate connective between Arg1 and Arg2 from one of the given connectives: " + Answer space + "Arg1: " + Arg1 + "Arg2: " + Arg2 + "Connective: " + ***ChatGPT output***

4.3 Parameter Settings

In NCPrompt, we conduct our experiments on two PLMs with the NSP task: BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019). Specifically, we adopt the *bert-base-uncased* model and *ernie-2.0-en* model implemented in PyTorch by HuggingFace transformers, and run with CUDA on RTX 3090. We set the batchsize to 4 and learning rates to 1e-5 and hyper-parameters τ, λ to $0.4, 0.5$.

5 Result and Analysis

5.1 Overall Result

We implement a 4-way classification on the level-1 discourse relation labels and a 14-way classification on the level-2 relation labels of the PDTB 3.0 corpus. We adopt the commonly used macro $F1$ score and accuracy (Acc) as evaluation metrics. Table 2 compares the overall performance between our NCPrompt and baselines. In the table, models in the first group all use *pre-train and fine-tune* paradigm. The second and third groups represent BERT-based and ERNIE-based prompt learning methods for IDRR respectively while the last group is the latest ChatGPT solution.

Models	PLMs	Level-1		Level-2	
		Acc	F1	Acc	F1
Ours$_{multi-token}$	BERT	**68.62**	**63.17**	**58.05**	**48.62**
w single-token	BERT	67.94	62.22	56.07	46.79
Ours$_{multi-token}$	ERNIE	**70.26**	**66.73**	**60.64**	**51.29**
w single-token	ERNIE	69.85	64.99	56.34	47.60

Table 5: Acc(%) and F1 score(%) of ablation study on the answer space of our NCPrompt.

Models	PLMs	Level-1		Level-2	
		Acc	F1	Acc	F1
Ours$_{CL}$	BERT	**68.62**	**63.17**	**58.05**	**48.62**
w prompt p_g^1	BERT	67.80	61.93	55.66	46.77
w prompt p_g^2	BERT	66.64	61.11	56.48	48.17
w/o CL	BERT	66.44	60.87	55.05	44.79
Ours$_{CL}$	ERNIE	**70.26**	**66.73**	**60.64**	**51.29**
w prompt p_g^1	ERNIE	68.55	63.73	56.48	50.06
w prompt p_g^2	ERNIE	68.62	64.02	58.66	47.45
w/o CL	ERNIE	68.01	63.60	57.84	46.21

Table 6: Acc(%) and F1 score(%) of ablation study on the contrastive learning of our NCPrompt.

In the first group, models using BERT generally outperform NNMA(Liu and Li, 2016) and DAGRN(Chen et al., 2016). This can be attributed to their utilization of Transformer-based PLM, which can provide dynamic and contextual embeddings. However, Glove and Word2Vec language models transfer English words into static word embeddings. We also notice that ContrastiveIDRR (Long and Webber, 2022) achieves quite competitive performance, which can be attributed to their good selection of positive and negative samples for supervised contrastive learning based on sense hierarchy.

Prompt-learning methods in the second and third groups generally outperform or rival models based on the *pre-train and fine-tune* paradigm. This proves that prompt learning can better utilize the semantic knowledge embedded in PLMs than the traditional fine-tune paradigm by reformulating downstream tasks into the pre-training tasks of PLMs. Also, ERNIE-based methods outperform the BERT-based ones. Although they both employ Transformer-based PLMs, ERNIE uses knowledgeable masking strategies to optimize the pre-training processes. Therefore, it can be seen that the improvements in the pre-training process are expected

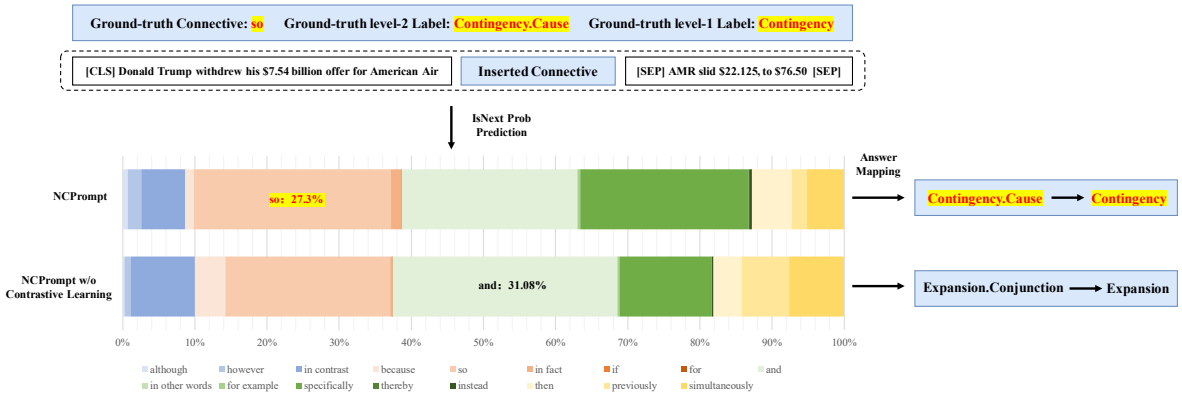


Figure 4: Case study of the predicted connectives for the NCPrompt and NCPrompt w/o contrastive learning.

to benefit the prompt-learning model performance.

We notice that our NCPrompt offers distinctive advantages in prompt-learning methods for IDRR. Similar results can also be observed in the binary classification for level-1 relation labels in Table 3. Obviously, our NCPrompt$_{BERT}$ achieves better performance than BERT-based ConnPrompt (Xiang et al., 2022b) and PCP (Zhou et al., 2022) which simply transform IDRR into a connective-cloze task based on MLM. Moreover, our model achieves comparable performance to TEPrompt (Xiang et al., 2023) which introduces auxiliary tasks for enlightenment prompt learning. This validates the effectiveness and superiority of our method. Although the NSP task has long been questioned by RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020), we validate its potential in prompt learning, and when combined with self-supervised contrastive learning, it can rival or beat advanced MLM-based prompt learning methods.

Moreover, only ContrastiveIDRR (Long and Webber, 2022) and our NCPrompt can classify level-2 relation labels for IDRR, and our model always performs better. Similar results can also be observed in the binary classification for level-2 relation labels in Table 4. This clearly demonstrates the capabilities of self-supervised contrastive learning in NLP. Various forms of prompt templates constitute positives, and the same example connected with false answer words constitute hard negatives, creating informative representation features together.

Finally, the ChatGPT-based model performs the worst among all on zero-shot IDRR task. This result reveals that IDRR is still a challenging and tricky task for ChatGPT, consistent with the results in Chan et al. (2023b,a). Although ChatGPT has exhibited powerful abilities in NLP, there still exist various tasks that cannot be easily solved at the current state, motivating us to design unique and innovative methods for specific research.

5.2 Ablation study

Answer Space: We claim that connectives with multiple tokens can convey more precise meaning than single-token words, like *"in contrast"* can highlight the contrasting relationship more clearly than *"but"*. To prove that, we replace the multi-token connectives in Table 1 with single-token ones and re-design the verbalizer in Appendix Table 8.

Table 5 shows the ablation study results on the answer space of our NCPrompt. Compared with NCPrompt with multi-token connectives, using single-token ones degrades the model performance for both level-1 and level-2 relation classification. This result indicates that multi-token connectives are indeed more expressive and effective in prompt learning for IDRR. Therefore, it is promising to reformulate the IDRR task into an NSP task and introduce various-length answer words.

Self-supervised Contrastive Learning: Table 6 shows the ablation study results on the contrastive learning part of our NCPrompt. *w prompt* p_g^1 and *w prompt* p_g^2 only introduce one form of augmentation prompt template respectively. *w/o CL* only trains the PLM parameters with cross entropy loss.

Removing the contrastive learning loss results in a distinct reduction of our model performance, with $F1$ score decreasing by about 3% for level-1 classification and decreasing by about 4% for level-2 classification. This proves that contrastive learning can indeed boost relation recognition performance by capturing significant connective information. Meanwhile, we can observe that the

model performance also degrades if there is only one form of augmentation view. This suggests that the model can learn more representation features with increasing numbers of positive samples for self-supervised contrastive learning.

5.3 Case Study

We use a case study to compare our NCPrompt and NCPrompt w/o contrastive learning. Figure 4 illustrates the $IsNext$ probability distribution of all candidate words connecting the current arguments. Each connective constitutes a specific prompt input into the PLM, and the PLM outputs the coherence score to be normalized into probabilities. The most probable connective is mapped into the answer relation label, as our prediction result.

We observe that our NCPrompt tends to predict *"so"* with the highest probability 27.3% mapped to *Contingency.Cause* and *Contingency*, which consists with the ground-truth labels. However, removing the contrastive learning part makes it difficult to distinguish between similar connectives and predicts *"and"* with the highest probability 31.08% mistakenly. This result validates that self-supervised contrastive learning provides substantial connective guidance and equips our model with strong abilities to differentiate between similar connectives. Also, we prove that self-supervised contrastive learning can be successfully applied in NLP on the foundation of NSP-based prompt learning.

6 Conclusion

In this paper, we first apply the NSP-based prompt learning method for IDRR and combine the self-supervised contrastive learning. Experiments on the PDTB 3.0 corpus validate the effectiveness of our proposed NCPrompt. We offer a new perspective that NSP-based prompt learning methods can be as remarkable as the commonly-used MLM-based ones, and validate the applicability of self-supervised contrastive learning in NLP.

7 Limitations

Despite achieving good performance, there are still some limitations of our work. First, since our proposed NCPrompt is based on PLMs pre-trained with the NSP task, limited to BERT and ERNIE, we cannot directly implement our model on RoBERTa and ALBERT which abandoned the NSP task. To fix this, we may train an NSP head for RoBERTa from scratch to improve the generalization ability of our proposed method and make fair comparisons with other RoBERTa-based models in our future work. Second, our proposed NSP-based prompt learning requires multiple inputs for the same instance to get the prediction result and this leads to decrease in inference efficiency to some extent. We may figure out how to utilize NSP-based prompt learning more efficiently in our future work. Third, We manually create several discrete prompt templates as positive augmentation views for self-supervised contrastive learning. However, discrete prompt templates may be suboptimal for adequately modeling various types of relations between argument pairs. In the future, we can introduce learnable vectors to form a dynamic soft template and automatically search for optimal templates in an embedding space.

References

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023a. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *arXiv preprint arXiv:2305.03973*.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1735.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Ozan Ciga, Tony Xu, and Anne Louise Martel. 2022. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198.

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

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia

Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 977–986.

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

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive learning for prompt-based few-shot language learners. In *North American Chapter of the Association for Computational Linguistics*, pages 5577–5587.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288.

Xiao Li, Yu Hong, Huibin Ruan, and Zhen Huang. 2020. Using a penalty-based loss re-estimation method to improve implicit discourse relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1513–1518.

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. *ACM Computing Surveys*, 55(9):1–35.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention.

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

Wanqiu Long and Bonnie Webber. 2022. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*.

Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. 2020. Interactively-propagative attention learning for implicit discourse relation recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3168–3178.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer.

Bang Wang, Zhenglin Wang, Wei Xiang, and Yijun Mo. 2023a. Adaptive prompt learning with distilled connective knowledge for implicit discourse relation recognition. *CoRR*, abs/2309.07561.

Chenxu Wang, Ping Jian, and Mu Huang. 2023b. Prompt-based logical semantics enhancement for implicit discourse relation recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 687–699.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Annual Meeting of the Association for Computational Linguistics*.

Wei Xiang, Chao Liang, and Bang Wang. 2023. Teprompt: Task enlightenment prompt learning for implicit discourse relation recognition. In *Annual*

Meeting of the Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. *ACM Computing Surveys*, 55(12):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 *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257.

Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. 2022b. Connprompt: Connective-cloze prompt learning for implicit discourse relation recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911.

WU Yi-heng, LI Jun-hui, and ZHU Mu-hua. 2024. Implicit discourse relation recognition with multi-view contrastive learning. *Computer Engineering & Science/Jisuanji Gongcheng yu Kexue*, 46(4).

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

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 *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858.

A Appendix

In this appendix, we present two tables: Table 7 presents the data statistics of implicit discourse relation instances in the PDTB 3.0 corpus; Table 8 describes our constructed answer space of single-token connectives and mapping to the relation sense labels in the PDTB 3.0 corpus.

Level-2 labels	**Train**	**Dev.**	**Test**
Comparison.Concession	1165	103	98
Comparison.Contrast	742	82	54
Contingency.Cause	4484	450	406
Contingency.Cause+Belief	159	13	15
Contingency.Condition	152	18	15
Contingency.Purpose	1105	97	89
Expansion.Conjunction	3586	298	236
Expansion.Equivalence	254	25	30
Expansion.Instantiation	1163	116	124
Expansion.Level-of-detail	2601	262	208
Expansion.Manner	673	14	17
Expansion.Substitution	342	27	26
Temporal.Asynchronous	1011	102	105
Temporal.Synchronous	436	34	43
Total	17873	1641	1466

Table 7: Data statistics of implicit relation instances in the PDTB 3.0 corpus with level-2 relation senses.

Level-1 labels	**Level-2 labels**	**Connective words**
Comparison	Concession	*although, however*
	Contrast	*but*
Contingency	Cause	*because, so*
	Cause+Belief	*thus*
	Condition	*if*
	Purpose	*for*
Expansion	Conjunction	*and*
	Equivalence	*indeed*
	Instantiation	*generally*
	Level-of-detail	*specifically*
	Manner	*thereby*
	Substitution	*instead*
Temporal	Asynchronous	*then, previously*
	Synchronous	*simultaneously*

Table 8: Answer space of single-token connectives and mapping to the level-2 and level-1 implicit discourse relation sense labels in the PDTB 3.0 corpus.