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

Anonymous ACL submission

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 dealing with IDRR. However, prior work primarily transform IDRR into a connective-cloze task based on the masked language model (MLM), which limits the predicted word to one single token. Besides, these methods use hand-crafted verbalizers which are time-consuming and less convincing. In this paper, we propose NCPrompt, an **N**SP-based prompt learning and **C**ontrastive learning method for IDRR. Specifically, we automatically search the optimal verbalizer for IDRR based on the statistical and expressive features of connectives. Furthermore, we transform the IDRR task into a next sentence prediction (NSP) task and introduce contrastive learning by constructing augmentation views. In this way, the answer words of multiple tokens can convey more precise meaning and contrastive learning can help to generate more informative embeddings, expected to boost the model performance. To our knowledge, we are the first to apply NSP to handle the IDRR task. Experiments on the PDTB 3.0 corpus have demonstrated the effectiveness and superiority of our proposed 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 downstream 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 rather challenging task which depends on understanding the semantics of natural language text.

The challenge and key point to the IDRR task is to learn high-quality semantic features of argument pairs. Leveraging the powerful ability of 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. Prompt learning can bridge the gap between pre-training and downstream task objective by reformulating the downstream tasks to match the pre-training tasks of PLMs and can thus fully exploit the semantic knowledge embedded in PLMs. Combining the design considerations of prompt learning and the specific characteristics of the IDRR task, how to develop a prompt learning-based solution to the IDRR task still exists challenges.

On the one hand, how to design prompt templates to transform the IDRR task into pre-training tasks of PLMs is the critical step and core challenge of prompt learning methods. 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. The MLM can only predict one single token for the masked slot, and thus only individual connectives can be selected as answer words, which extremely limits the construction of verbalizers. Meanwhile, it is obvious that phrases can convey more precise meaning than individual words. Therefore, we propose to transform the IDRR task into the next sentence prediction (NSP) task of PLMs by inserting the various-length connectives between argument pairs and predicting their logical relationship in order to expand the construction of verbalizers.

On the other hand, the construction of verbalizers is also a great challenge due to the characteristics of the IDRR task. Specifically, the relation hierarchy in the Penn Discourse TreeBank (PDTB) 3.0 corpus (Webber et al., 2019) is complicated,

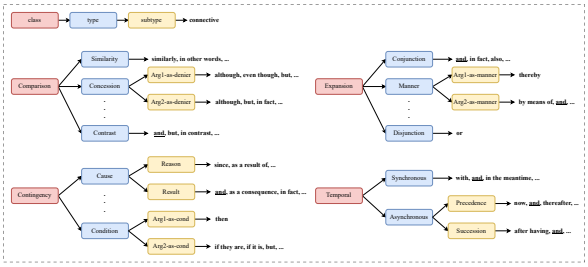

Figure 1: The annotation hierarchy of discourse relation senses in the PDTB 3.0 corpus (**and** can be assigned to various relation senses by annotators).

and the annotated implicit connectives are numerous as well as ambiguous according to Figure 1. Obviously, to select the least ambiguous and most representative connectives as answer words from massive candidates brings too much trouble. Existing models (Xiang et al., 2022b; Zhou et al., 2022; Xiang et al., 2023; Wu et al., 2023) manually construct verbalizers, consuming lots of human efforts. Meanwhile, it's also hard to ensure the superiority of such manually selected verbalizers which rely on domain expertise. Suboptimal verbalizers may negatively impact the performance of prompt learning methods (Gao et al., 2021). Therefore, we propose to automatically construct the verbalizer for IDRR based on the statistical and expressive features of candidate connectives, in order to reduce manual efforts and obtain optimal verbalizer.

Also, inspired by ConnPrompt (Xiang et al., 2022b) utilizing a multi-prompt ensemble strategy, we notice that multiple prompts create a kind of augmentation views which naturally constitute the data augmentation process of contrastive learning (Chen et al., 2020). By discrimination among positive and negative samples in the representation space, contrastive learning can generate more informative embeddings (Dehghan and Amasyali, 2022). Therefore, we propose to combine the contrastive learning loss with the classification loss specially for NSP-based prompt learning methods in order to capture critical semantic information among embeddings and further boost the model performance.

In this paper, we propose NCPrompt, an **N**SP-based prompt learning and **C**ontrastive learning method for IDRR. First, we automatically construct the verbalizer for IDRR based on *statistics refinement* and *relevance refinement* process. Second, we reformulate the IDRR task into an NSP task by inserting the searched connectives between argument pair and predicting the coherence score. Third, we construct positive and negative samples and define contrastive loss combined with the classification loss to boost the model performance. Experiments on PDTB 3.0 corpus have demonstrated the superiority of our proposed NCPrompt over other competitive baselines. More importantly, our NCPrompt validates the potential of NSP and provides inspiration for other NLP tasks.

2 Related Work

2.1 Implicit Discourse Relation Recognition

IDRR is a major challenge in NLP research whose difficulty lies in learning informative representations of argument pairs. With the emergence of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and other powerful PLMs, *pre-train and fine-tune* paradigm has been applied in IDRR (Chen et al., 2016; Liu and Li, 2016; Ruan et al., 2020; Li et al., 2020; Liu et al., 2020; Xiang et al., 2022a) which transfer the pre-trained representations to downstream tasks to encode argument pairs into semantic embeddings. For example, IPAL (Ruan et al., 2020) designs a cross-coupled network to combine self-attention and interactive-attention mechanisms integrated with BERT. However, such paradigm may result in poor utilization of PLM knowledge due to the inconsistency between the pre-training and downstream task objective.

Recently, the prompt learning paradigm is proposed to bridge the gap between pre-training and downstream task objective and successfully employed for IDRR. ConnPrompt (Xiang et al., 2022b) and PCP (Zhou et al., 2022) first apply prompt learning to IDRR by simply transforming IDRR into a connective-cloze task. Subsequently, the CP-KD (Wu et al., 2023) and AdaptPrompt (Wang et al., 2023) 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. Also, DiscoPrompt (Chan et al., 2023b) injects discourse label structure information into prompts. However, the use of NSP task and automatic construction of verbalizers have hardly been explored in current methods.

2.2 Next Sentence Prediction

Prompt learning is playing a dominative role in NLP, however, most work (Schick et al., 2020; Hu et al., 2022b; Zhang and Wang, 2023) design cloze-format prompts based on MLM, limiting the an-

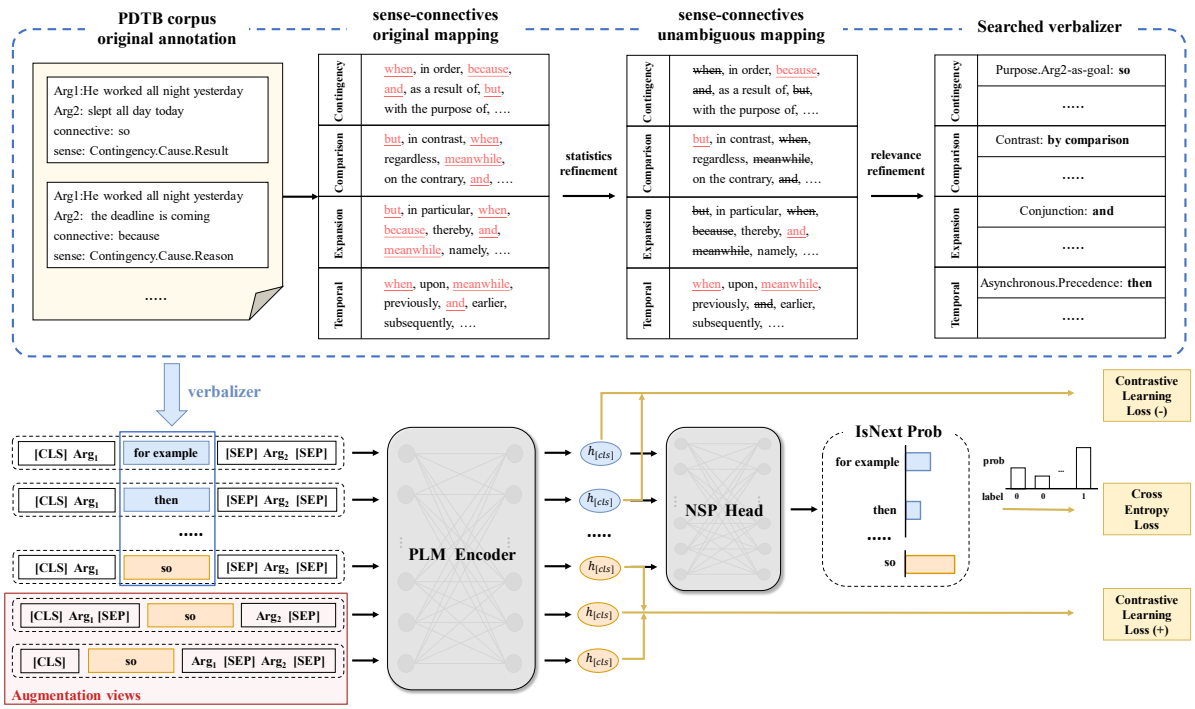

Figure 2: Overview of the proposed NCPrompt model (**so** is the gold connective of current argument pair).

swer words to one single token. In fact, besides the common MLM task, there also exists a sentence-level pre-training task, NSP, in BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019), which set no restriction to the length of answer words. NSP trains the PLM to identify whether two input sentences are continuous segments from the training corpus. In the past, the necessity of the NSP task has been questioned by RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020), so NSP is hardly utilized in prompt learning until the NSP-BERT (Sun et al., 2021) model proves its abilities.

NSP-BERT transforms the single-sentence classification tasks into NSP and achieves performance comparable to the MLM-based PET (Schick and Schuetze, 2021) model, proving the potential of NSP in prompt learning. However, NSP-BERT only conducts experiments on fundamental NLP tasks instead of focusing on challenging tasks like IDRR. Therefore, we first propose to transform IDRR into NSP by inserting the answer connectives between argument pairs and predicting whether the two arguments come consecutively.

3 The Proposed NCPrompt Model

Figure 2 presents the overview of our proposed NCPrompt model. Overall, we first design prompt 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 combine the contrastive loss on multiple views of the argument pair with the connective classification loss. During testing, the connective with the highest coherence score is mapped into the answer relation sense through our automatically constructed verbalizer.

3.1 Prompt Template

The argument pair is transformed to the format of PLM-input through the prompt template $T(Arg_1, Arg_2) = T(x)$. The comparison between MLM-based and NSP-based prompt learning for IDRR is illustrated in Figure 3. In MLM-based prompt learning models, $T(x)$ usually consists of the input, the $[MASK]$ token and sometimes task-specific texts. In ConnPrompt (Xiang et al., 2022b), the authors design a kind of connective-cloze prompt template, denoted as $T_{conn}(x)$:

$$T_{conn}(x) = [CLS] + Arg_1 + [MASK] + Arg_2 + [SEP]$$

The PLM estimates the probability of each word in its vocabulary for the $[MASK]$ token to predict a connective-bearing answer word, but only single-token words can be predicted. In NSP-based prompt learning models, however, there is no such restriction. In our NCPrompt, connectives of single

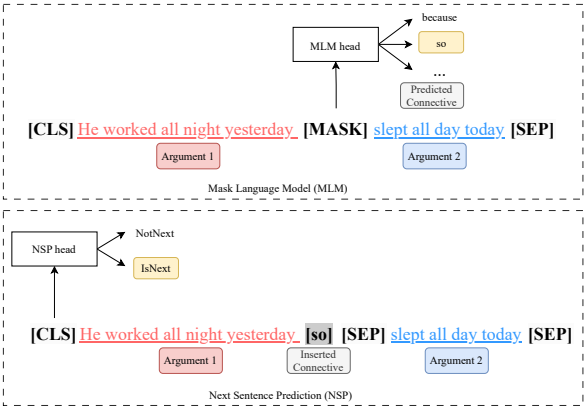

Figure 3: Examples of comparison between MLM-based and NSP-based prompt learning for IDRR.

token or multiple tokens can become part of the prompt template, denoted as $T_{NC}(x)$:

$$T_{NC}(x) = [CLS] + Arg_1 + \boldsymbol{v} + [SEP] + Arg_2 + [SEP]$$

where $\boldsymbol{v}$ refers to every candidate connective in the constructed answer space $V = \{v_1, v_2, ..., v_k\}$ and $|V| = k$ is the total number of connectives in the answer space. Specifically, every connective v_i constitutes a specific prompt $T_{NC}(x, v_i) = p_i$. And there is one gold connective denoted as v_g for a argument pair, leading to prompt p_g. Inspired by ConnPrompt (Xiang et al., 2022b), we also design two auxiliary prompts as the augmentation views of the positive sample p_g, as below:

$$T^{a_1}_{NC}(x, v_g) = [CLS]{+}Arg_1{+}[SEP]{+}v_g{+}Arg_2{+}[SEP]$$

$$T^{a_2}_{NC}(x, v_g) = [CLS]{+}v_g{+}Arg_1{+}[SEP]{+}Arg_2{+}[SEP]$$

We denote the above two prompts as $p^{a_1}_g$ and $p^{a_2}_g$ respectively. For contrastive learning, prompts p_g, $p^{a_1}_g$ and $p^{a_2}_g$ constitute the positive samples together, while $p_{i(i\neq g)}$ constitute the negative samples.

3.2 Model Prediction

After 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 $q_M(n|x)$:

$$q_M(n|x) = W_{nsp}h_{[CLS]} + b_{nsp},$$

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:

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

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

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

During training, the output probability distributions of connectives are utilized to compute classification loss with the gold connective label of the current argument pair, 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:

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

Then, the answer connective is mapped into the answer relation sense through the verbalizer.

3.3 Verbalizer Construction

In prompt learning methods for IDRR, verbalizer is an essential component to select the least ambiguous and most representative connectives as answer space and then map each of them to a specific relation sense. Motivated by LM-BFF (Gao et al., 2021) and KPT (Hu et al., 2022a), we propose *statistics refinement* and *relevance refinement* process in order to develop a pipeline of automatic verbalizer construction.

Statistics Refinement: In the original mapping between annotated connectives and relation senses, some connectives can be annotated to multiple top-level relation senses as shown in Figure 2. Therefore, referring to the classical TF-IDF (Sparck Jones, 1972), we measure the importance of ambiguous connectives u to each top-level relation sense y_i, where $i \in \{1, .., n\}$ and $n = 4$ represents the total number. The computation formula of Term Frequency (TF) is as follow:

$$TF(u, y_i) = \frac{count(u, y_i)}{\sum_j count(u_j, y_i)}$$

where $count(u, y_i)$ denotes the times that connective u is annotated to the relation sense y_i. TF can represent the annotation frequency and expressive

ability of ambiguous connectives towards each top-level relation sense. The computation formula of Inverse Category Frequency (ICF) is as follow:

$$ICF(u) = log(1 + \frac{n}{|j : u \in y_j|})$$

$$TF.ICF(u, y_i) = TF(u, y_i) \times ICF(u)$$

where $|j : u \in y_j|$ denotes the number of top-level relation senses containing connective u. ICF can measure the ability of connectives to distinguish different top-level relation senses. Combining TF and ICF, we obtain the TF.ICF indicator to perform statistics refinement on ambiguous connectives by classifying each of them into the relation sense with the highest TF.ICF score:

$$sense(u) = y_{\underset{i}{argmax}\{TF.ICF(u,y_i)\}}$$

Accordingly, after handling the ambiguous connectives, we can obtain the one-to-one mapping between candidate connective set and each top-level discourse relation sense, denoted as C_i.

Relevance Refinement: Since some connectives may be more relevant to the corresponding relation sense than others, we propose a relevance refinement method to search the most representative connectives towards each second-level relation sense. We first narrow down the candidate connective set C_i based on their conditional likelihood over training data using the initial PLM without fine-tuned. Specifically, we construct the pruned connective set by selecting top a (hyper-parameter) connectives that achieve the highest output logit on training data for each top-level relation sense, denoted as:

$$C_i^p = \underset{v \in C_i}{Top - a}\{\sum_{x \in D_{train}^i} p(v|x)\}$$

where D_{train}^i is the training data of each top-level relation sense y_i. Then, for each second-level relation sense y_i^j, we continue to search top b (hyper-parameter) connectives, denoted as:

$$C_i^j = \underset{v \in C_i^p}{Top - b}\{\sum_{x \in D_{train}^{i,j}} p(v|x)\}$$

By performing permutations on the connective set C_i^j, we can get all candidate verbalizers that contain the most representative connective mapped to each second-level relation sense. To achieve our final verbalizer, we first select a subset of c (hyper-parameter) verbalizers that maximize zero-shot $F1$ score on training data and fine-tune the selected verbalizers to find the only one that maximizes $F1$ score on development data.

Relation sense	Connective words
Comparison	*in contrast, by comparison, however, but*
Contingency	*so, in order, as a result, therefore, consequently, since*
Expansion	*for instance, for example, in fact, and, thereby*
Temporal	*then, previously*

Table 1: Answer space of our NCPrompt and connection to the top-level discourse relation senses in the PDTB corpus.

In this way, we automatically construct the verbalizer for IDRR that every connective is directly mapped to a second-level relation sense and then mapped to the top-level relation sense. Our final verbalizer is shown in Table 1.

3.4 Training Strategies

Inspired by Jian et al. (2022) who combine a contrastive learning loss with the standard MLM loss in prompt-based few-shot learners, we first propose to introduce contrastive learning to NSP-based prompt learning methods. Actually, the NSP task applied in prompt learning naturally creates hard negative samples (Robinson et al., 2020) which differ from the positive sample p_g only in connective and thus provide significant connective guidance for contrastive learning, expected to capture critical semantic features of embeddings. Therefore, 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:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} g_i \; log \; P(x_i),$$

where g_i is the gold connective label of the i-th training argument pair and N is the batchsize. The gold connective label is not the manually annotated implicit connective by annotators but the specific connective mapping to the gold relation sense label through our constructed verbalizer.

Contrastive Learning Loss: As illustrated in Figure 2, after creating augmentation views of p_g, we obtain 3 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),

Relation Sense	Train	Dev.	test
Comparison	1937	190	154
Contingency	5916	579	529
Expansion	8645	748	643
Temporal	1447	136	148
Total	17945	1653	1474

Table 2: Statistics of implicit relation instances in the PDTB 3.0 corpus with top-level relation senses.

the contrastive learning loss is defined as:

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

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:

$$L_{CL} = \frac{1}{N}\sum_{i=1}^{N}[\sum_{\substack{l,j\in\{p_g, p_g^{a_1}, p_g^{a_2}\}\\ l\neq j}} l(j,l)]$$

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

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

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

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 four top-level discourse relation senses, namely *{Comparison, Contingency, Expansion, Temporal}*. Table 2 presents the statistics of implicit discourse relation instances in dataset.

4.2 Baselines

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

- **DAGRN** (Chen et al., 2016) adopts a gated relevance network to capture the 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.
- **BMGF** (Liu et al., 2020) combines representation, matching, and fusion modules for implicit discourse analysis.
- **MANF** (Xiang et al., 2022a) fuses semantic connection and linguistic evidence for relation recognition.

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, and achieves state-of-the-art performance on the PDTB 3.0 corpus.
- **PCP** (Zhou et al., 2022) proposes a prompt-based connective prediction method based on RoBERTa, and achieves state-of-the-art performance on the PDTB 2.0 corpus.

For fair comparisons, we only select the models which simply apply prompt learning and ignore models further combined with other strategies like TEPrompt (Xiang et al., 2023). Since RoBERTa (Liu et al., 2019) and some PLMs have abandoned the NSP task, we re-implement PCP based on BERT (Devlin et al., 2019) and ERNIE (Zhang et al., 2019) on the PDTB 3.0 corpus using the verbalizer of ConnPrompt.

Moreover, as ChatGPT has demonstrated strong capabilities in contextual understanding and interactive dialogue, we 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***

Model	PLM	Acc	F1
NNMA	Glove	57.67%	46.13%
DAGRN	Word2Vec	57.33%	45.11%
MANF	Word2Vec	60.45%	53.14%
IPAL	BERT	57.33%	51.69%
PLR	BERT	63.84%	55.74%
MANF	BERT	64.04%	56.63%
BMGF	RoBERTa	69.95%	62.31%
ConnPrompt	BERT	68.86%	62.66%
PCP	BERT	66.42%	62.14%
Our Model	BERT	69.13%	63.01%
ConnPrompt	ERNIE	67.98%	63.98%
PCP	ERNIE	70.83%*	65.60%*
Our Model	ERNIE	**71.37%**	**65.73%**
ChatGPT	gpt-3.5-turbo	32.97%	28.79%

Table 3: Overall results of our NCPrompt and baselines for IDRR on the PDTB 3.0 corpus. The boldface and the * are the best and the second best results respectively among all models, the underline is the best result among models in a specific group.

4.3 Experiment 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). BERT is the most representative PLM proposed by Google, while ERNIE is a knowledge-enhanced PLM proposed by Baidu. 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 hyperparameters a, b, c, τ, λ to 50 respectively.

5 Result and Analysis

5.1 Overall Result

We implement a four-way classification on the top-level discourse relation senses of the PDTB 3.0 corpus and adopt the commonly used macro $F1$ score and accuracy (Acc) as evaluation metrics. Table 3 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 group respectively represent methods of BERT-based and ERNIE-based prompt learning for IDRR while the last group is the latest ChatGPT solution.

We first observe that our NCPrompt$_{ERNIE}$ achieves the best Acc and $F1$ score among all models. Also, our NCPrompt$_{BERT}$ offers distinctive advantages in BERT-based prompt learning for IDRR, which validates the effectiveness and superiority of our methods. The usage of NSP enables phrases as answer connectives, which can convey more accurate meaning for PLMs to understand. Based on the NSP task, we introduce automatic verbalizer construction and contrastive learning as well, boosting the model performance together.

In the first group, models using BERT and RoBERTa generally outperform NNMA, DAGRN and MANF$_{Word2Vec}$ using Glove and Word2Vec language model to transfer English words into static word embeddings. This can be attributed to their utilization of Transformer-based PLMs which provide dynamic and contextual embeddings.

Comparing between prompt-learning methods in the second and third groups, we notice that ERNIE-based methods outperform the BERT-based ones. Although they all employ Transformer-based PLMs, ERNIE uses some knowledgeable masking strategies to optimize the pre-training processes. Also, BMGF achieves competitive result with prompt-learning methods due to the usage of RoBERTa pre-trained on a much larger dataset. Therefore, it can be seen that the improvements in the pre-training process are expected to benefit the model performance. In conclusion, we observe that the choice of PLMs indeed has a decisive effect on the results and we should evaluate the performance of models based on the same PLMs to make fair comparisons.

Overall, prompt-learning methods outperform models based on the *pre-train and fine-tune* paradigm in the first group especially when using the same PLM. This result 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.

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 on Verbalizer Construction

Table 4 shows the ablation study results on the verbalizer construction process of our NCPrompt$_{BERT}$. *w manual* replaces our verbalizer

Model	Acc	F1
NCPrompt	**69.13%**	**63.01%**
w manual	68.52%	62.13%
w/o statistics refinement	67.30%	61.58%
w/o relevance refinement	66.82%	60.86%
w/o fine-tune search	64.86%	59.40%

Table 4: Ablation study on the automatic verbalizer construction process of NCPrompt$_{BERT}$.

Relation sense	Connective words
Comparison	*similarly, but, however, although*
Contingency	*for, if, because, so*
Expansion	*instead, by, thereby, specifically, and*
Temporal	*simultaneously, previously, then*

Table 5: Answer space of ConnPrompt (Xiang et al., 2022b) and connection to the top-level discourse relation senses in the PDTB corpus.

with the manually constructed one in ConnPrompt (Xiang et al., 2022b) where all answer words are single-token connectives as shown in Table 5. *w/o statistics refinement* doesn't handle the ambiguous connectives. *w/o relevance refinement* constructs the pruned sense-connectives mapping not based on conditional likelihood. *w/o fine-tune search* constructs the final verbalizer without a fine-tune search on development data.

Compared with NCPrompt, using ConnPrompt's manually constructed verbalizer reduces the $F1$ score by almost 1%. This result indicates that multi-token connectives are indeed more expressive and effective in prompt learning for IDRR. Also, the manually selected answer words can be sub-optimal, time-consuming and less convincing.

When removing the statistics refinement module, we first ignore those ambiguous connectives and directly eliminate the verbalizers that connectives are annotated to multiple top-level senses, which downgrades the $F1$ score by more than 1%. This naturally excludes some connectives from their most corresponding relation senses, validating the effectiveness of the statistics refinement process.

The $F1$ score of NCPrompt without the relevance refinement module drops by more than 2%. It means that candidate connectives only decided by annotation statistical information without counting on conditional likelihood logits, will result in fewer representative connectives. Also, the fine-tune search process is proven to be helpful because solely the zero-shot performance can't totally determine the potential of the verbalizers.

Model	Acc	F1
NCPrompt	**69.13%**	**63.01%**
w/o contrastive loss	68.18%	61.76%
w prompt $p_g^{a_1}$	68.59%	62.59%
w prompt $p_g^{a_2}$	68.52%	62.05%

Table 6: Ablation study on the contrastive learning loss of NCPrompt$_{BERT}$.

5.3 Ablation study on Contrastive Learning

Table 6 shows the ablation study results on the contrastive loss of our NCPrompt$_{BERT}$. *w/o contrastive loss* only trains the PLM parameters with cross entropy loss removing the contrastive loss. *w prompt* $p_g^{a_1}$ and *w prompt* $p_g^{a_2}$ only introduce one augmentation positive prompt respectively.

When removing the contrastive learning loss, the $F1$ score decreases by over 1%, which proves that contrastive learning can indeed boost the relation recognition performance by capturing significant connective information. Meanwhile, we can observe that the model performance degrades if there is only one augmentation view of the positive sample p_g, which suggests that the model can learn more representation features with increasing numbers of positive samples for contrastive learning.

6 Conclusion

In this paper, we first apply the NSP-based prompt learning method for IDRR and propose the NCPrompt framework, which further combines automatic verbalizer construction and contrasting learning loss. Experiments on the PDTB 3.0 corpus prove that our NCPrompt can achieve better results than competitive baselines. Also, our successful usage of the NSP task in IDRR validates the potential and capability of this pre-training task and offers a new perspective that NSP-based prompt learning methods can be as remarkable as the commonly-used MLM-based ones by allowing multi-token answer words.

Limitations

In verbalizer construction, we only regard connectives annotated in the PDTB 3.0 corpus as candidates.

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. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Somaiyeh Dehghan and Mehmet Fatih Amasyali. 2022. Supmpn: Supervised multiple positives and negatives contrastive learning model for semantic textual similarity. *APPLIED SCIENCES-BASEL*, 12(19).

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.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021*, pages 3816–3830. Association for Computational Linguistics (ACL).

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. 2022a. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *Computing Research Repository*, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers):2225–2240.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022b. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240.

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.

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.

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*.

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.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578.

Timo Schick and Hinrich Schuetze. 2021. It's not just size that matters: Small language models are also few-shot learners. *Computing Research Repository*, abs/2009.07118:2339–2352.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Yi Sun, Yu Zheng, Chao Hao, and Hangping Qiu. 2021. Nsp-bert: A prompt-based zero-shot learner through an original pre-training task–next sentence prediction. *arXiv e-prints*, pages arXiv–2109.

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

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

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.

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.

Zizhuo Zhang and Bang Wang. 2023. Prompt learning for news recommendation. *arXiv preprint arXiv:2304.05263*.

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.