Connective Prediction for Implicit Discourse Relation Recognition via Knowledge Distillation

Hongyi Wu[1], **Hao Zhou**[1], **Man Lan**[1,2,3,*], **Yuanbin Wu**[1] **and Yadong Zhang**[1]
[1]School of Computer Science and Technology, East China Normal University, Shanghai, China
[2]Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[3]Lingang Laboratory, Shanghai, China
{hongyiwu,hzhou,yadongzhang}@stu.ecnu.edu.cn
{mlan,ybwu}@cs.ecnu.edu.cn

Abstract

Implicit discourse relation recognition (IDRR) remains a challenging task in discourse analysis due to the absence of connectives. Most existing methods utilize one-hot labels as the sole optimization target, ignoring the internal association among connectives. Besides, these approaches spend lots of effort on template construction, negatively affecting the generalization capability. To address these problems, we propose a novel **C**onnective **P**rediction via **K**nowledge **D**istillation (**CP-KD**) approach to instruct large-scale pre-trained language models (PLMs) mining the latent correlations between connectives and discourse relations, which is meaningful for IDRR. Experimental results on the PDTB 2.0/3.0 and CoNLL 2016 datasets show that our method significantly outperforms the state-of-the-art models on coarse-grained and fine-grained discourse relations. Moreover, our approach can be transferred to explicit discourse relation recognition (EDRR) and achieve acceptable performance. Our code is released in https://github.com/cubenlp/CP_KD-for-IDRR.

1 Introduction

Discourse relation recognition (DRR) aims at detecting semantic relations between two arguments (sentences or clauses, they are denoted as *Arg1* and *Arg2*, respectively). As illustrated in Figure 1, the discourse relation `Contingency` (denoted as *sense*) is held between *Arg1* and *Arg2*, and the explicit connective *so* is drawn from the raw text while the implicit connective *because* is manually inserted by annotators. DRR is significant to many natural language processing (NLP) downstream tasks such as causal reasoning (Staliunaite et al., 2021) and question answering (Huang et al., 2021). However, compared with explicit discourse relation recognition (EDRR), implicit discourse relation recognition (IDRR) is still less accurate and practical due to the lack of connectives, which is a major challenge in current discourse analysis research.

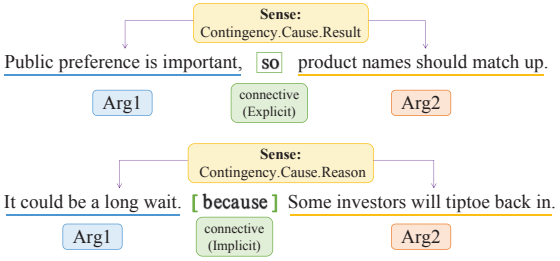

Figure 1: Examples of discourse annotation with explicit and implicit connectives in the PDTB 3.0 corpus.

The connectives (e.g., *because*, *so*, etc.) are critical linguistic cues for identifying discourse relations. On the one hand, with the aid of explicit connectives, a simple frequency-based mapping is sufficient to achieve over 85% classification accuracy on EDRR (Xue et al., 2016). On the other hand, human annotators utilized connectives to aid relation annotation in the most popular PDTB benchmark datasets (Prasad et al., 2008; Webber et al., 2019). For instance, annotators first manually inserted a connective expression, and then determined the abstract relation in consideration of both the implicit connective and argument pairs. Therefore, several studies recognize implicit discourse relations by incorporating connective information.

Several studies incorporate connective information to recognize implicit discourse relations. One method uses the probability distribution of connectives among sense labels in the corpus (Asr and Demberg, 2020), but this requires a consistent label distribution, which is not always the case. For instance, the connective *since* is more likely to represent the relation `contingency` in the training data but `temporal` in the test data. Other methods predict implicit connectives before recognizing relations (Zhou et al., 2010), or project connectives and relations into the same latent space and transfer knowledge (Nguyen et al., 2019). However, these methods perform poorly because of introducing additional parameters that require training with large

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5908–5923

amounts of labeled data.

Inspired by Schick and Schütze (2021), several studies exploited the advantage of prompt learning (Liu et al., 2023) to guide PLMs to predict connectives between argument pairs and then map them to corresponding discourse relations (Xiang et al., 2022; Zhou et al., 2022). However, this paradigm predicted connectives by fitting the outputs of models to one-hot hard labels, regardless of the internal association among connectives. As we all know, a discourse relation corresponds to multiple connectives, but previous studies only selected one of them as the positive sample, while other connectives with similar meanings under the same sense labels were treated as negative samples. Besides, the correlation between connectives and discourse relations utilized in these studies is a direct mapping, which is vulnerable and inaccurate. Finally, both of them spend lots of effort on template construction, which negatively affects the generalization capability.

To address above-mentioned problems, we propose a novel **C**onnective **P**rediction via **K**nowledge **D**istillation (**CP-KD**) approach for identifying implicit discourse relations. As suggested in Hinton et al. (2015), knowledge distillation is a popular technique for training the student model to emulate the well-informed teacher model. Specifically, we **first** design a knowledgeable teacher model to generate meaningful soft labels that capture more associations among connectives than one-hot hard labels to guide the optimization of the student model. **Secondly**, we add answer hints representing the relations of arguments as input to the teacher model, which exploits the implicit knowledge between connectives and sense labels, rather than using the direct mapping relationships in the previous studies. This approach mitigates issues of connective ambiguity and the possibility of multiple similar connectives mapping to the same discourse relation. **Finally**, we design a simple but effective template matching the pattern of implicit discourse data, and demonstrate that simple templates can achieve acceptable performance as well. In addition, the method we propose alleviates the dependence of prompt learning on templates and has good generalization across different templates. Extensive experiments show that our proposed model outperforms prior state-of-the-art systems on the PDTB dataset by around 3%.

Our contributions are summarized as follows:

- We propose a novel Connective Prediction via Knowledge Distillation (CP-KD) approach for the IDRR task, which achieves the SOTA performance on the PDTB 2.0/3.0 datasets and CoNLL-2016 Shared Task as well.
- Our proposed method performs label softening via knowledge distillation to capture the implicit correlations between connectives and sense labels, which previous methods ignored.
- Our method can be easily transferred from IDRR to EDRR, and experiments demonstrate that our method still performs well for EDRR.

2 Related work

2.1 Implicit Discourse Relation Recognition

Previous studies focused on the feature engineering of linear classifiers to classify implicit discourse relations. For example, Lin et al. (2009) was the first to consider fine-grained classification, and they further used four different feature types to characterize context and component resolution trees.

Along with the booming development of deep learning, most work designs neural networks for IDRR. For instance, Liu et al. (2021) proposed combining the context representation module and bilateral multi-perspective matching module to understand different relational semantics deeply. In addition, Wu et al. (2022) designed a label-focused encoder to learn a global representation of input instances and their level-specific context. It also uses a label-sequence decoder to output predicted labels in a top-down manner. Moreover, several methods have recognized implicit discourse relations with the aid of annotated connectives. Specifically, Kishimoto et al. (2020) proposed to introduce the auxiliary task of connectives prediction in the pre-training process and use explicit discourse relationship data for data enhancement. Kurfalı and Östling (2021) performed implicit discourse relation classification without relying on any labeled implicit relation and sidestepped the lack of data through the explicitation of implicit relations. However, these methods contradicted the original pre-training task and performed poorly on fine-grained discourse relations.

Inspired by Schick and Schütze (2021), several studies exploited the advantage of prompt learning (Liu et al., 2023) to predict connectives between argument pairs to better utilize the knowledge embedded in the PLMs. Specifically, Zhou et al. (2022) manually designed different templates that meet

the task goal and follow natural language patterns. However, this method requires a lot of effort to find a suitable template to achieve better performance. Xiang et al. (2022) developed a multi-prompt ensemble to fuse predictions from different prompting results. However, both of them predict the connective by fitting the outputs of models to hard labels (i.e., one-hot vectors), regardless of the rich semantic correlations among relations.

Another related work is Jiang et al. (2022), which uses a multi-data multi-task teacher model with explicit and implicit discourse data to optimize a single-data single-task student model. Unlike their work, which leverages knowledge distillation to transfer explicit discourse data to the student model, our work captures the intrinsic association of discourse connectives through softened category label distributions from the teacher model, thus guiding the student model.

2.2 Knowledge Distillation

Knowledge distillation has three prominent roles in conventional tasks: model compression, label softening, and domain migration. The principle of model compression is to transfer knowledge from one large-scale model to another lightweight model, thus enabling the model lighter without losing performance. For example, Yang et al. (2019) combined the knowledge of multiple teachers to perform question-and-answer matching. Li et al. (2020) proposed an idea to speed up Transformer model training and reasoning: training a larger model first and then compressing the model.

In knowledge distillation, the predictions from the teacher model are called soft labels, and the student model improves performance through dark knowledge, including inter-class similarity carried by the soft labels. For instance, Tang et al. (2016) found that soft labels from teacher models provide significant regularization for student models. And Cheng et al. (2020) verified mathematically that the soft label gives the student model higher learning speed and better performance than the optimization learning from the original data.

The principle of domain migration is to transfer knowledge from the teacher model to the student model in different domains. Specifically, Fang et al. (2021) found that samples from various fields shared a typical local pattern and obtained this local information for domain migration through knowledge distillation. Choi et al. (2022) extracted domain knowledge from the existing domain pretrained models and transferred it to other PLMs through knowledge distillation.

3 Method

In this section, we introduce our proposed Connective Prediction via Knowledge Distillation (CP-KD) method in detail. We first present the prompt-guided connective prediction model in Section 3.1 and then describe the overall framework of our CP-KD approach in Section 3.2.

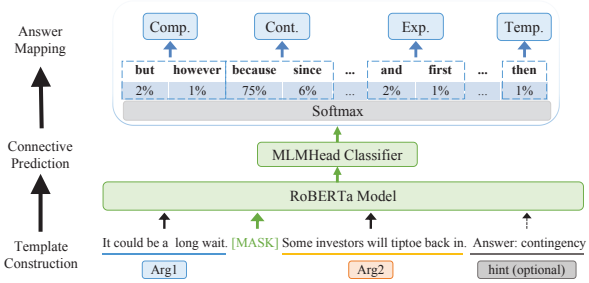

Figure 2: Illustration of Connective Prediction model.

3.1 Prompt-Guided Connective Prediction

The prompt-guided connective prediction method aims to predict the most probable connective between arguments and then map it to the corresponding sense label. As illustrated in Figure 2, it has three main processes, including template construction, connective prediction, and answer mapping.

Template Construction: In this module, we give different inputs for the teacher and student models. For the student model, given a pair of arguments, we transfer them to $\boldsymbol{x}_{\text{prompt-s}}$ with the template:

$$\boldsymbol{x}_{\text{prompt-s}} = \mathbf{T}(\boldsymbol{x}_{\text{Arg1}}, \boldsymbol{x}_{\text{Arg2}}), \quad (1)$$

where $\boldsymbol{x}_{\text{Arg1}}$ and $\boldsymbol{x}_{\text{Arg2}}$ correspond to two arguments, respectively (as shown in Figure 1) and $\mathbf{T}$ represents template function. In the PDTB corpus, almost all implicit discourse data satisfy the "*Arg1* connective *Arg2*" sequence order, where the connective is manually inserted by annotators. Therefore, we design a simple but effective template "*Arg1* [MASK] *Arg2*" for our main experiment, where the symbol [MASK] represents the masked token in place of the predictable connective.

For the teacher model, we add the answer hint as input and combine it with the given argument pairs to $\boldsymbol{x}_{\text{prompt-t}}$ with a new template:

$$\boldsymbol{x}_{\text{prompt-t}} = \mathbf{T}(\boldsymbol{x}_{\text{Arg1}}, \boldsymbol{x}_{\text{Arg2}}, \boldsymbol{x}_{\text{hint}}), \quad (2)$$

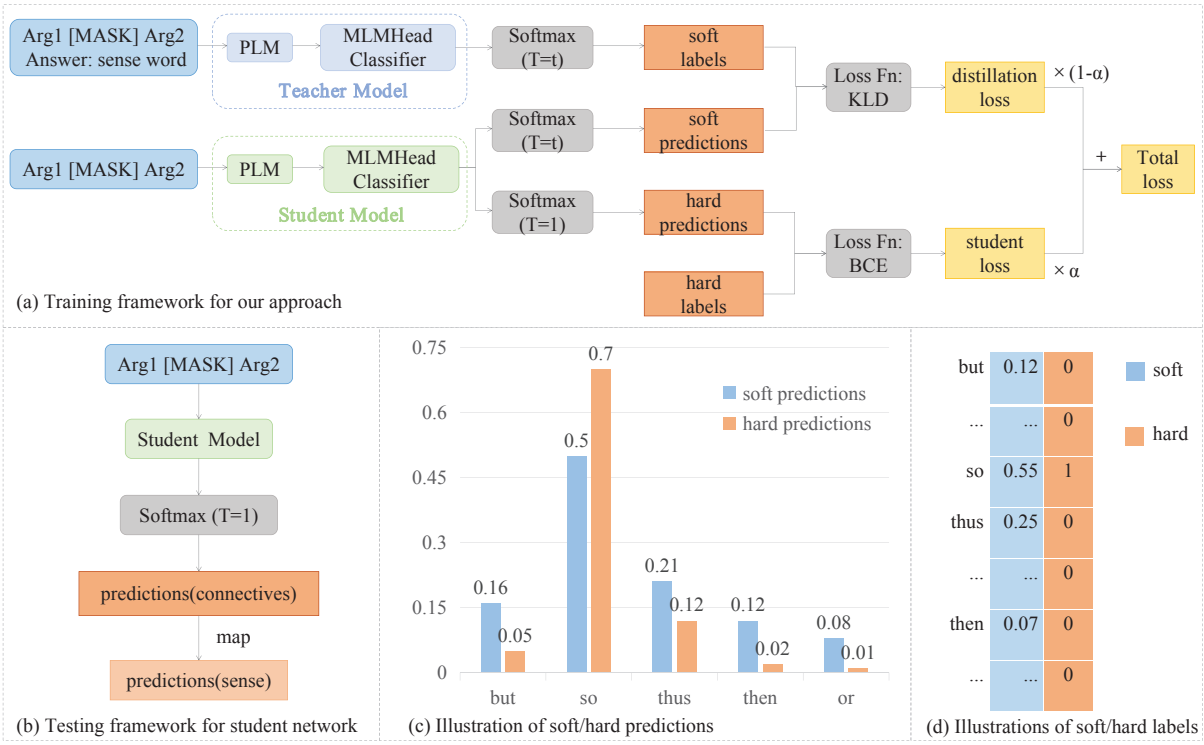

Figure 3: Illustration of our proposed Connective Prediction via Knowledge Distillation(CP-KD) framework. Among them, (a) and (b) represent training and testing framework for our approach respectively, and (c) and (d) indicate the illustration of predictions and labels respectively.

where $\boldsymbol{x}_{\text{hint}}$ represents the specific sense label, such as `Contingency`. For simplicity and clarity, we use "*Arg1* [MASK] *Arg2* Answer: *sense*" as the teacher template.

Connective Prediction: Then we feed $\boldsymbol{x}_{\text{prompt}}$ to the RoBERTa (Liu et al., 2019) model to obtain the representation of [MASK] token $\boldsymbol{h}_{\text{mask}}$, and input the token into MLMHead model to acquire scores $\boldsymbol{e}_{\text{mask}}$ of each word in its vocabulary $\boldsymbol{V}$.

$$\boldsymbol{e}_{\text{mask}} = \textbf{MLMHead}(\boldsymbol{h}_{\text{mask}}). \tag{3}$$

According to the hierarchy sense labels and implicit connectives, we manually select a discrete answer space $\boldsymbol{V}_{\text{a}}$, which is a subset of PLM's vocabulary $\boldsymbol{V}$. During the training, a softmax layer is applied on $\boldsymbol{e}_{\text{mask}}$ to normalize it into probabilities:

$$P_i = \frac{\exp\left(\boldsymbol{e}_i\right)}{\sum_{k=1}^{|\boldsymbol{V}_{\text{a}}|} \exp\left(\boldsymbol{e}_k\right)}, \quad v_i \in \boldsymbol{V}_{\text{a}}, \tag{4}$$

where $|\boldsymbol{V}_{\text{a}}|$ is the size of vocabulary $\boldsymbol{V}_{\text{a}}$. Afterwards we use cross-entropy to calculate the loss between the model prediction and the selected golden connective:

$$\mathcal{L}_{\text{MLM}}(\boldsymbol{x}) = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} y_i \log P_i, \tag{5}$$

where $\mathcal{M}$ denotes the set of masked tokens and y_i represents the golden label.

Answer Mapping: Finally, we map the predicted connective (e.g., *because*) to the corresponding sense label (e.g., *Cause*). For implicit discourse relation data, each sample has been annotated with the connective appropriate to it on PDTB and CoNLL16 datasets (detailed in Section 4.1). However, the number of connectives marked in the original samples is large, and the ambiguity is high. As a result, we select the **most frequent** and **less ambiguous** connectives as the answer words. At the same time, we only select those tokenized connectives with a single token as answer words since most masked PLMs predict a single word. We present the final answer sets we select on the PDTB 2.0/3.0 and CoNLL16 datasets in Appendix B.

3.2 Overall Framework of CP-KD

As illustrated in Figure 3, our proposed CP-KD approach consists of two branches: a teacher model T, which aims to combine soft type constraints between connectives and sense labels with prompt-guided connective prediction model to instruct the optimization of the student model, and a student(distilled) model S, which is forced to produce

vectorized outputs that are similar to the results of the teacher model.

In the training stage, the optimization goal of the teacher model is to correctly predict the golden connective when adding sense words as answer hints. To enable the teacher model to predict connectives without relying on "reciting answers," we select a fraction of the random samples to add sense words as hints (detailed in Section 4.6.1).

Meanwhile, the student model requires to serve testing scenarios where extra sense labels are missing. Therefore, the student is expected to tap the deep semantic relationships of argument pairs with the guidance of a knowledgeable teacher. As shown in Figure 3(a), in the training stage, the student model S is required to match not only the ground-truth one-hot labels but also the probability outputs of the teacher model T:

$$\mathcal{L}_s = \alpha \mathcal{L}_{GT}^S + (1-\alpha)\tau^2 \mathcal{L}_{KD}, \tag{6}$$

where α is the coefficient to trade off such two terms and τ is the temperature rate parameter used to alleviate category imbalance. In addition, $\mathcal{L}_{GT}^S$ is the ground-truth loss using one-hot labels to predict connectives, and $\mathcal{L}_{KD}$ is the knowledge distillation loss utilizing the Kullback-Leibler divergence(Hershey and Olsen, 2007) to quantify the difference of output distribution from student's soft predictions to teacher's soft labels:

$$\mathcal{L}_{GT}^S = -\frac{1}{|K|}\sum_{i=1}^{K} y_i \log \frac{\exp\left(\boldsymbol{e}_i\right)}{\sum_{k=1}^{|\mathcal{V}_a|}\exp\left(\boldsymbol{e}_k\right)}, \tag{7}$$

$$\mathcal{L}_{KD} = \sum_{i=1}^{K} \tilde{P}_T(i) \log\left(\tilde{P}_T(i)/\tilde{P}_S(i)\right), \tag{8}$$

where y_i is the golden label, K is the size of instance, $\tilde{P} = \mathrm{softmax}\left(\tilde{Z}/\tau\right)$, and $\tilde{Z}$ is the pre-softmax logits output by the model.

As shown in Figure 3(b), in the inference stage, the well-trained student model aims to predict connectives between a pair of arguments and then map it to corresponding discourse relations.

It is worth mentioning that the inclusion of temperature rate τ in the $\mathrm{softmax}$ layer contributes to flattening the distribution, narrowing the gap between two models and making the distillation focus on whole logits, as illustrated in Figure 3(c). Furthermore, as seen in Figure 3(d), soft labels output by the teacher model carry more information among connectives than one-hot labels. For example, the connective *so* is semantically similar to *thus*, yet hard labels do not carry such information.

Dataset	Top-level Senses	Train	Dev.	Test
PDTB 2.0	Comparison (Comp.)	1,894	191	146
	Contingency (Cont.)	3,281	287	276
	Expansion (Exp.)	6,792	651	556
	Temporal (Temp.)	665	54	68
	Total	12,632	1,183	1,046
PDTB 3.0	Comparison (Comp.)	1,937	190	154
	Contingency (Cont.)	5,916	579	529
	Expansion (Exp.)	8,645	748	643
	Temporal (Temp.)	1,447	136	148
	Total	17,945	1,653	1,474

Table 1: Statistics of top-level senses in PDTB datasets.

4 Experiment

4.1 Dataset

The Penn Discourse Treebank (PDTB 2.0/3.0) PDTB corpora are annotated with information related to discourse semantic relation. Among them, PDTB 2.0 (Prasad et al., 2008) contains 2312 Wall Street Journal (WSJ) articles, while PDTB 3.0 (Webber et al., 2019) has made a series of modifications based on Version 2, including annotation of 13,000 additional tokens and incorporation of new senses. 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. We evaluate our model on both coarse-grained and fine-grained discourse relations. Table 1 shows the statistics of the top-level senses. We introduce the CoNLL16 dataset in Appendix A.

4.2 Baselines

To validate the effectiveness of our method, we compare our approach with the advanced models in recent years. First of all, we select some strong baselines based on the neural network, including ESDP (Wang and Lan, 2016), MANN (Lan et al., 2017), and RWP-CNN (Varia et al., 2019). Their work mainly focused on the top-level senses of PDTB 2.0 and CoNLL16 cross-level senses. Secondly, we compare our method with competitive baselines based on PLMs, such as HierMTN-CRF (Wu et al., 2020), BERT-FT (Kishimoto et al., 2020), BMGF-RoBERTa (Liu et al., 2021) and LDSGM (Wu et al., 2022). These methods achieve impressive performance at the fine-grained second-level senses with the help of large-scale PLMs. Finally, we compare our approach with the latest

Model	PDTB2-Top		PDTB2-Second		CoNLL	Blind	PDTB3-Top		PDTB3-Second	
	F1	Acc.	F1	Acc.	Acc.	Acc.	F1	Acc.	F1	Acc.
ESDP	-	-	-	-	40.91	34.20	-	-	-	-
MANN	47.80	57.39	-	-	39.40	40.12	-	-	-	-
RWP-CNN	50.20	59.13	-	-	39.39	39.36	-	-	-	-
HierMTN-CRF	55.72	65.26	33.91	52.34	-	-	-	-	-	-
BERT-FT	58.48	65.26	-	54.32	-	-	-	-	-	-
BMGF-RoBERTa	63.39	69.06	37.95	58.13	57.26	55.19	66.92*	71.98*	41.28*	61.87*
LDSGM	63.73	71.18	40.49	60.33	-	-	68.89*	73.47*	37.44*	60.06*
PCP$_{base}$	64.95	70.84	41.55	60.54	60.98	57.31	69.82*	73.81*	49.87*	63.36*
ConnPrompt	64.26*	71.61*	39.16*	61.02*	59.14*	53.44*	69.92	74.36	41.88*	57.19*
Our CP-KD$_{base}$	**68.86**	**75.43**	**44.77**	**64.00**	**62.79**	57.24	**72.07**	**77.00**	**50.12**	**66.21**
PCP$_{large}$	67.79	73.80	44.04	61.41	63.36	58.51	71.95*	75.17*	49.00*	66.42*
Our CP-KD$_{large}$	**71.88**	**76.77**	**47.78**	**66.41**	**67.23**	**59.86**	**75.52**	**78.56**	**52.16**	**67.84**

Table 2: Experimental results on PDTB 2.0/3.0 and CoNLL16 datasets. The best results of each part are underlined. Models in the third part of the table use RoBERTa-base as PLMs, while the last part uses RoBERTa-large as PLMs.

work PCP (Zhou et al., 2022) and ConnPrompt (Xiang et al., 2022). Both utilize the strategy of prompt learning to predict connectives and achieve state-of-the-art performance on PDTB 2.0 and PDTB 3.0 datasets, respectively. Since almost all previous methods were not experimented on PDTB 2.0/3.0 and CoNLL16 datasets at the same time, to comprehensively evaluate the performance, we choose several competitive models in the last three years (including BMGF-RoBERTa, LDSGM, PCP, and ConnPrompt) to re-implement on three datasets.[1]

4.3 Implementation Details

In this work, we use *RobertaForMaskedLM* [2] as the backbone of our method, where *RobertaEncoder* is to obtain context representation of inputs and *RobertaLMHead* is to acquire each vocabulary token prediction score for [MASK] token position.

We adopt AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate of $1e^{-5}$ to update the model parameters and set batch size as 16 and accumulated gradients as 2 for training and validation. Since the knowledge distillation method is sensitive to hyperparameters, we use the optimization algorithm of grid search to explore the practical effect under different parameters, where α takes value from 0.3 to 0.7 and τ from 1 to 5. All our experiments are performed on one RTX 3090. All other parameters are initialized with the default values in PyTorch Lightning[3], and our model is all implemented by Transformers[4].

4.4 Experimental Results and Analysis

We first evaluate our model on the coarse-grained top-level and fine-grained second-level senses of PDTB 2.0/3.0 (denoted as PDTB2-Top, PDTB2-Second, PDTB3-Top, and PDTB3-Second, respectively) with Macro F1 score and accuracy value. Then we conduct cross-level classification on the CoNLL16 dataset and consider accuracy as the primary metric, denoted as CoNLL and Blind for the test and blind-test set.

Table 2 shows the main results, from which we can reach the following conclusions. **First**, our method achieves the new SOTA performance with substantial improvements on almost all implicit discourse recognitions, which proves the superiority and generalization of our approach. Specifically, when considering accuracy, it obtains 3.82%, 2.98%, 1.81%, 2.64% and 2.85% improvements over the best results of previous baselines (Part 3) on PDTB2-Top, PDTB2-Second, CoNLL, PDTB3-Top, and PDTB3-second classifications, respectively. In terms of F1, it also performs consistently better than previous models. **Second**, compared with the latest work PCP and ConnPrompt, **the most significant improvement** of our approach is utilizing knowledge distillation to obtain implicit

[1] Since the LDSGM model utilizes the hierarchical relationship between top-label and second-level, it does not apply to the cross-level recognition of CoNLL16. We mark the data we re-implement with a superscript.

[2] https://huggingface.co/docs/transformers/model_doc/roberta

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

[4] https://github.com/huggingface/transformers

Model	PDTB2-Top		PDTB2-Second		CoNLL	Blind	PDTB3-Top		PDTB3-Second	
	F1	Acc.	F1	Acc.	Acc.	Acc.	F1	Acc.	F1	Acc.
CP-KD$_{base}$	**68.86**	**75.43**	**44.77**	**64.00**	**62.79**	**57.24**	**72.07**	**77.00**	**50.12**	**66.21**
w/o KD	63.78	70.55	39.14	61.31	58.62	53.44	70.00	74.36	46.29	61.74
w/o MLM	66.48	73.42	43.16	62.46	60.97	54.93	70.44	75.10	48.65	63.98
w/o hint	68.49	74.76	43.84	62.75	61.62	54.39	71.64	76.46	49.86	65.67

Table 3: Architecture ablation analysis on PDTB 2.0/3.0 and CoNLL16 dataset.

relationships between connectives and sense labels instead of using direct mapping relationships. The experimental results prove the meaningful soft labels generated by the teacher model contribute to recognizing the implicit relations between argument pairs. **Third**, it can be observed that our CP-KD$_{base}$ approach outperforms the PCP$_{large}$ method on almost all datasets, which proves that knowledge distillation supports the student models to obtain significant performance gains, even over larger models. (See Appendix C for more analysis.)

4.5 Ablation Study

To evaluate the effects of different components, we compare CP-KD with its variants: 1) w/o KD. In this variant, we remove the teacher model and only remain the student model for connective prediction; 2) w/o MLM. In this variant, the teacher model predicts connectives through [CLS] of the PLMs. 3) w/o hint. In this variant, we remove answer hints of the teacher model. We intend to explore whether adding answer hints for the teacher model contributes to learning the deep correlations between connectives and sense labels and thus help implicit discourse relation recognition.

From Part 1 of Table 3, we can observe that our CP-KD model consistently exhibits better performance than their corresponding variants across both coarse-grained and fine-grained labels. Specifically, the knowledge distillation module brings the most significant performance improvements, with about 5% gains in F1 and Acc. metrics on almost all datasets. Moreover, the performance decreases by about 2% when CP-KD w/o MLM as the reference, which proves the prompt-guided method outperforms the conventional pre-train and fine-tuning paradigm model. Finally, the performance improvement on fine-grained classification is more significant than coarse-grained when the teacher model adds answer hints, which demonstrates that answer hints can guide the teacher model to explore the implicit relationships between connectives and sense labels accurately.

4.6 Hyperparameter Tuning

4.6.1 Proportion of Answer Hints

In this section, we explore the appropriate ratio for introducing answer hints. As shown in Table 4, the optimal balance of selected answer hints is 10% for the PDTB 2.0/3.0 datasets. When the ratio is lower, it is difficult for the teacher model to discover the relationship between connectives and sense labels. The teacher model is more inclined to recite the answers when the proportion is higher. We can imagine the teacher model as an experienced professor who teaches the best students when it has seen some samples instead of remembering all the answers. Moreover, the optimal proportion for the CoNLL16 datasets is 40%, which indicates that fine-grained classification requires more cues than coarse-grained to uncover the implicit relationship between connectives and sense labels.

Proportion	PDTB2-Top		PDTB3-Top		CoNLL
	F1	Acc.	F1	Acc.	Acc.
0	68.49	74.76	71.64	76.46	61.62
10%	**68.86**	**75.43**	72.07	**77.00**	62.14
40%	68.13	74.57	**72.35**	76.46	**62.79**
70%	68.06	74.38	72.29	76.59	61.88
100%	67.91	74.38	71.34	76.05	61.88

Table 4: Results of different proportion of answer hints.

4.6.2 Influence of Hyperparameter in KD

As we all know, the knowledge distillation algorithm is sensitive to hyperparameters and random seeds. To explore the effect of hyperparameters, we experiment with ten consecutive random seeds varying α from 0.3 to 0.7 and τ from 1 to 5 on the PDTB 2.0 top-level senses.

As we can observe from Figure 4, the average performance is significantly better when α is smaller, demonstrating that the teacher model's soft labels can carry more information than one-hot

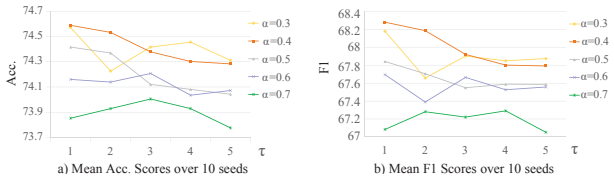

Figure 4: Mean Acc./F1 scores of different hyparameters over 10 seeds on PDTB 2.0 top-level senses.

hard labels. In addition, the average Acc. and F1 scores reach their highest values when both α and τ are small, which proves that when the student model prefers the knowledge of the teacher model, τ needs to be tuned down to prevent the effect of negative labels.[5]

4.7 Case Study

Figure 5 showcases the confusion matrices of both the ConnPrompt (Xiang et al., 2022) and the CP-KD models, tested on the PDTB 2.0 second-level senses. The matrices highlight ConnPrompt's challenge in differentiating between closely related categories, namely `Comp.Contrast` and `Exp.Conjunction`, as well as `Cont.Cause` and `Exp.Restatement`. This confusion emphasizes the criticality of profound semantic comprehension for precise implicit discourse relation recognition.

Contrarily, CP-KD, leveraging the benefits of knowledge distillation, displays superior capabilities in discerning these nuanced differences. This demonstrates that a simplistic reliance on surface-level lexical or syntactic features is inadequate, and a deeper understanding of semantics is necessary. We present this through the following examples:

- **Example 1:** ConnPrompt confuses `Cont.Cause` with `Exp.Restatement`.

  Arg1: He was right.

  Arg2: By midday, the London market was in full retreat.

- **Example 2:** ConnPrompt erroneously identifies `Comp.Contrast` as `Exp.Conjunction`.

  Arg1: Amcore, also a bank holding company, has assets of $1.06 billion.

  Arg2: Central's assets are $240 million.

[5]See the appendix D for details about the results.

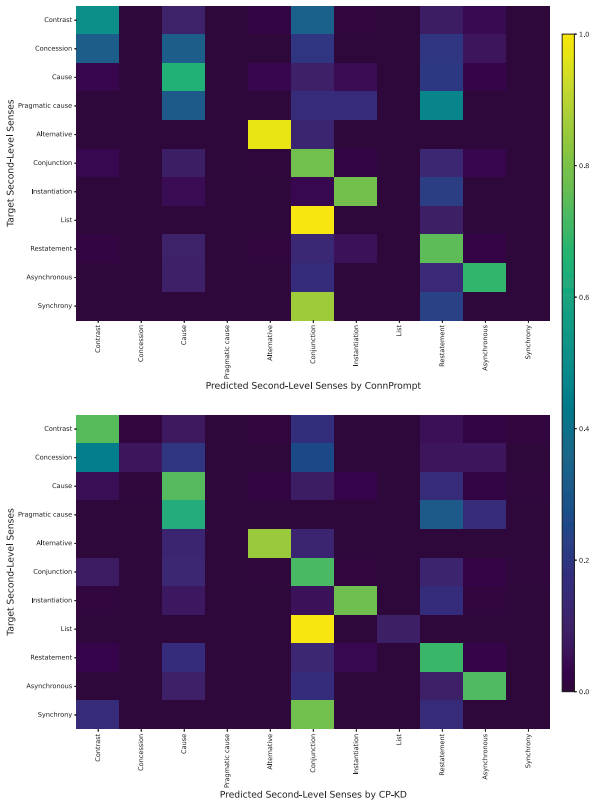

Figure 5: Confusion Matrix for the ConnPrompt and our CP-KD Model on PDTB 2.0 Second-Level Senses

The examples provided above underscore CP-KD's enhanced capability to comprehend the semantic relationships between pairs of arguments. This enhancement can largely be attributed to the integration of knowledge distillation within CP-KD, which fosters a deeper understanding of discourse relations and connectives. Despite the model's praiseworthy performance, we recognize the potential for further optimization and exploration. Specifically, the model requires improvement in handling few-shot categories such as `Exp.List` and `Temp.Synchrony`. To bolster the model's overall predictive precision and robustness, we propose increasing its competency in managing underrepresented senses. This can be achieved by enriching the training set with additional instances of these categories, enhancing the model's familiarity with these senses, thereby augmenting its predictive capabilities.

5 Discussion

5.1 Generalization to Other Prompt Template

Previous studies proved that templates have different impacts on the prediction results of connectives

(Xiang et al., 2022; Zhou et al., 2022). Therefore, in this section, we are tempted to verify the generalization of our method on different templates. Specifically, Zhou et al. (2022) found a relatively best template for connective prediction after abundant experiments. Xiang et al. (2022) designed three prompt templates and made a decision fusion of majority voting as multi-prompt ensembling for final relation sense prediction. For a fair comparison, we replace the templates of our approach to verify the effectiveness of knowledge distillation.

As shown in Table 5 and 6, our method has successfully generalized different templates. It is worth mentioning that the general template used in this paper is precisely the same as the first template in ConnPrompt (Xiang et al., 2022). When we use multi-template fusion like it, our method achieves better performance on the PDTB 3.0 dataset.

template	method	PDTB2-Top F1	PDTB2-Top Acc.
Arg1: *Arg1*. Arg2: *Arg2*.</s> </s>The conjunction between Arg1 and Arg2 is [MASK].	PCP-base	64.95	70.84
	CP-KD-base	**67.52**	**74.76**
	PCP-large	67.79	73.80
	CP-KD-large	**71.37**	**76.58**

Table 5: Results of CP-KD method on the template on PDTB 2.0 top-level senses.

5.2 Generalization to Explicit Discourse Relation Recognition

Inspired by the attempt of section 5.1, we transfer our method to the EDRR task. Similarly, we design a simple template in line for the explicit discourse relation recognition via knowledge distillation (**KD-EDRR**). The new template is as follows:

- <start> *Connective* <end> *Arg1* [MASK] *Arg2*

where the *Connective* represent connectives that appear in the original text but not in *Arg1* or *Arg2*. In addition, <start> and <end> are marker tokens used to guide the position of connective. Meanwhile, we use the [CLS] token of the masked language model to predict the sense directly, and we introduce the symbol [MASK] to predict connective, which is regarded as an auxiliary task for mining the implicit relationships between connectives and sense labels.

As shown in Table 7, the variant of our method KD-EDRR achieves the new state-of-the-art performance on the top-level senses of PDTB 2.0 for the EDRR task, which effectively demonstrates the generalizability of our approach.

template	ConnPrompt F1	ConnPrompt Acc.	CP-KD-base F1	CP-KD-base Acc.
Arg1 [MASK] Arg2	69.91	74.36	72.07	77.00
Arg1 </s> [MASK] Arg2	69.63	73.61	71.84	76.32
[MASK] Arg1 </s> Arg2	69.00	73.54	71.80	76.53
Multi-Prompt	70.88	75.17	**72.89**	**77.54**

Table 6: Results of CP-KD method on the single template and multi-prompt ensembling on PDTB 3.0 top-level senses.

Model	**Acc.**	**F1**
(1)Connective Only (Pitler and Nenkova, 2009)	93.67	-
(1)+Syntax+Conn-Syn (Pitler and Nenkova, 2009)	94.15	-
(2)ELMo-C&E (Dai and Huang, 2019)	95.39	94.84
(3)RWP-CNN (Varia et al., 2019)	96.20	95.48
(4)PEDRR (Zhou et al., 2022)	94.78	93.59
KD-EDRR (Ours)	**96.39**	**95.59**

Table 7: Experimental results of our KD-EDRR method on PDTB 2.0 top-level senses for EDRR.

6 Conclusion

In this paper, we propose a novel connective prediction via knowledge distillation approach for coarse-grained and fine-grained implicit discourse relation recognition. Experimental results demonstrate that our method achieves state-of-the-art performance on the PDTB 2.0/3.0 datasets and the CoNLL-2016 Shared Task. Furthermore, our proposed method fully uses the correlation between connectives and sense labels and achieves good generalization on different templates. Finally, we experimentally prove that our approach can be transferred from IDRR to EDRR and still performs well for EDRR. We will later explore the applicability of our approach to some Chinese discourse relations datasets for coarse-grained and fine-grained DRR.

Limitations

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

Firstly, in the step of answer mapping (Section 3.1), we only select those connectives that are tokenized with a single token as answer words, since most masked PLMs predict only a single word. Therefore, those connectives tokenized with multiple tokens will be replaced by the most frequent answer word with the same subtype-level sense tags. We believe that this approach will filter out several meaningful connectives as answer words. In the future, we will utilize the generative

model to predict the connectives between argument pairs, which can decode multiple tokens at a single mask position.

Secondly, in section 5.1, we can observe that multi-prompt ensembling is effective for fusing multiple single-prompts for implicit discourse relation recognition. In the future, we will explore multi-teacher knowledge distillation method for the IDRR task, here teacher models are trained with different templates. In this way, we can take advantage of the different prompt templates.

Acknowledgement

We appreciate the support from National Natural Science Foundation of China with the Main Research Project on Machine Behavior and Human-Machine Collaborated Decision Making Methodology (72192820 & 72192824), Pudong New Area Science & Technology Development Fund (PKX2021-R05), Science and Technology Commission of Shanghai Municipality (22DZ2229004), Shanghai Trusted Industry Internet Software Collaborative Innovation Center and East China Normal University International Conference Grant Programme.

References

Fatemeh Torabi Asr and Vera Demberg. 2020. Interpretation of discourse connectives is probabilistic: Evidence from the study of but and although. *Discourse Processes*, 57(4):376–399.

Xu Cheng, Zhefan Rao, Yilan Chen, and Quanshi Zhang. 2020. Explaining knowledge distillation by quantifying the knowledge. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12925–12935.

Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4793–4801. IEEE.

Dongha Choi, HongSeok Choi, and Hyunju Lee. 2022. Domain knowledge transferring for pre-trained language model via calibrated activation boundary distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1669.

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

Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. 2021. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34:11920–11932.

John R Hershey and Peder A Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. *arXiv e-prints*, pages arXiv–2103.

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.

Congcong Jiang, Tieyun Qian, and Bing Liu. 2022. Knowledge distillation for discourse relation analysis. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 210–214, New York, NY, USA. Association for Computing Machinery.

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

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. *arXiv preprint arXiv:2106.03192*.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. 2020. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on Machine Learning*, pages 5958–5968.

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

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 Comput. Surv.*, 55(9).

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. On the importance of word and sentence representation learning in implicit discourse relation classification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

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*.

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

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

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

Ieva Staliunaite, Philip John Gorinski, and Ignacio Iacobacci. 2021. Improving commonsense causal reasoning by adversarial training and data augmentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13834–13842. AAAI Press.

Zhiyuan Tang, Dong Wang, and Zhiyong Zhang. 2016. Recurrent neural network training with dark knowledge transfer. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5900–5904.

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

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

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

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 *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11486–11494. AAAI Press.

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

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

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

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2019. Model compression with multi-task knowledge distillation for web-scale question answering system. *arXiv preprint arXiv:1904.09636*.

Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. 2022. Prompt-based connective prediction method for fine-grained implicit discourse relation recognition. *arXiv preprint arXiv:2210.07032*.

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

A The CoNLL 2016 Shared Task (CoNLL16)

The CoNLL 2016 shared task (Xue et al., 2016) provides more abundant annotation than PDTB for shallow discourse parsing. The PDTB section 23 and Wikinews texts following the PDTB annotation guidelines were organized as the test sets. CoNLL16 merges several labels of PDTB. For example, `Contingency.Pragmatic cause` is merged into `Contingency.Cause.Reason` to remove the former type with very few samples. Finally, there is a flat list of 14 implicit sense classes to be classified, detailed senses as shown in the first column of Table 10.

B Answer Sets on Three Datasets

In this section, we present the answer sets we select on PDTB 2.0/3.0 and CoNLL16 datasets, as illustrated in table 8, 9 and 10. In addition, we found that there are several data samples with two senses. In our data statistics and experiments process, we uniformly considered the first sense of these samples as their golden label for avoiding ambiguity.

Top-level	Second-level	Answer Set
Comparison	Concession	*although*, *nevertheless*
	Contrast	*but*, *however*
Contingency	Cause	*because*, *so*, *therefore*, *thus*
	Pragmatic cause	*since*
Expansion	Alternative	*instead*, *or*
	Conjunction	*and*, *furthermore*
	Instantiation	*instance*
	List	*first*
	Restatement	*specifically*
Temporal	Asynchronous	*previously*, *then*
	Synchrony	*simultaneously*

Table 8: Mapping between implicit discourse relation labels and connectives on PDTB 2.0 dataset, which has four top-level and 11 second-level senses. The answer set of top-level senses is a union set of second-level.

C Performance on Fine-grained IDRR

To better evaluate the performance of our method on fine-grained implicit discourse relation recognition, we compare it with three previous competitive models at each second-level sense of PDTB datasets [6]. As exhibited in Table 11 and 12, our

[6]Given the test set of PDTB 2.0 only covers 11 types of discourse relations, we restrict our results to a statistical analysis of these 11 discourse relations in this study. The same procedure was followed for the PDTB 3.0 dataset.

Top-level	Second-level	Answer Set
Comparison	Concession	*although*, *nevertheless*
	Contrast	*but*, *however*
	Similarity	*similarly*
Contingency	Cause	*because*, *so*
	Condition	*if*
	Purpose	*for*
Expansion	Substitution	*instead*
	Manner	*by*, *thereby*
	Level-of-detail	*specifically*
	Conjunction	*and*
	Instantiation	*instance*
	Equivalence	*namely*
Temporal	Asynchronous	*previously*, *then*
	Synchrony	*simultaneously*

Table 9: Mapping between implicit discourse relation labels and connectives on PDTB 3.0 dataset, which has four top-level and 14 second-level senses. The answer set of top-level senses is a union set of second-level.

Cross-level Senses	Answer Set
Comp.Concession	*although*
Comp.Contrast	*but*, *however*
Cont.Cause.Reason	*because*, *as*
Cont.Cause.Result	*so*, *thus*, *consequently*
Cont.Condition	*if*
Exp.Alternative	*unless*, *or*
Exp.Alternative.Chosen alternative	*instead*, *rather*
Exp.Conjunction	*and*, *while*
Exp.Exception	*rather*
Exp.Instantiation	*instance*, *example*
Exp.Restatement	*specifically*
Temp.Asynchronous.Precedence	*then*
Temp.Asynchronous.Succession	*previously*
Temp.Synchrony	*meanwhile*

Table 10: Mapping between implicit discourse relation labels and connectives on CoNLL16 dataset which has 14 cross-level implicit senses.

Second-level Senses	BMGF-RoBERTa	LDSGM	PCP	CP-KD
Comp.Concession	0.0	0.0	0.00	**10.00**
Comp.Contrast	59.75	63.52	62.50	**67.44**
Cont.Cause	59.60	64.36	66.78	**67.66**
Cont.Pragmatic cause	0.0	0.0	0.0	0.0
Exp.Alternative	60.0	63.46	60.00	**66.67**
Exp.Conjunction	**60.17**	57.91	54.16	60.14
Exp.Instantiation	67.96	72.60	70.29	**77.06**
Exp.List	0.0	8.98	**27.03**	15.38
Exp.Restatement	53.83	58.06	59.91	**61.50**
Temp.Asynchronous	56.18	56.47	56.47	**66.67**
Temp.Synchrony	0.0	0.0	0.0	0.0
Macro F1	37.95	40.49	41.55	**44.77**

Table 11: Macro F1 scores on PDTB2-second senses.

Second-level Senses	BMGF-RoBERTa	LDSGM	PCP	CP-KD
Comp.Concession	57.47	**65.57**	55.96	56.65
Comp.Contrast	**55.10**	44.44	50.88	52.86
Comp.Similarity	0.0	0.0	40.00	**66.67**
Cont.Cause	67.88	68.38	68.75	**71.90**
Cont.Cause+Belief	0.0	0.0	**8.70**	0.0
Cont.Cause+SpeechAct	0.0	**64.36**	0.0	0.0
Cont.Condition	64.00	11.11	70.97	**85.71**
Cont.Purpose	95.03	91.94	91.11	**95.56**
Exp.Conjunction	59.28	**66.47**	62.69	65.91
Exp.Disjunction	0.0	0.0	**33.33**	0.0
Exp.Equivalence	16.36	4.00	**40.00**	10.53
Exp.Instantiation	69.64	70.65	71.37	**74.24**
Exp.Level-of-detail	52.37	43.05	53.41	**59.57**
Exp.Manner	28.57	30.00	42.11	**57.89**
Exp.Substitution	42.11	**70.00**	52.83	64.29
Temp.Asynchronous	65.02	62.67	**70.19**	68.37
Temp.Synchronous	29.03	25.93	**35.48**	21.82
Macro F1	41.28	37.44	49.87	**50.12**

Table 12: Macro F1 scores on PDTB3-second senses.

CP-KD method supersedes the prior state-of-the-art models in the majority of second-level senses, barring a few exceptions such as Exp.List and Exp.Conjunction.

Notably, our approach procures significant enhancements in several categories already demonstrating robust performance, such as Comp.Contrast and Exp.Instantiation. The improvements in these categories indicate that the novel approach of transforming the implicit discourse relation recognition task into a connective prediction task, followed by employing knowledge distillation to capture intrinsic connective associations, is highly effective.

Furthermore, the CP-KD method demonstrates an exceptional capacity to handle complex implicit relations, as evidenced by its superior performance in categories like Comp.Similarity and Cont.Condition in the PDTB 3.0 dataset, and Comp.Concession in the PDTB 2.0 dataset. This underlines the effectiveness of a combined approach of Prompt Learning and knowledge distillation in tackling intricate implicit discourse relations.

Additionally, our CP-KD method maintains a high degree of stability across various discourse relations, as shown by its consistently competitive performance across different relation types. This attribute reaffirms the CP-KD method's robust recognition capability across a diverse range of implicit discourse relations.

D Results of Different Hyperparameters

In deep learning models, particularly those employing techniques like knowledge distillation, performance can be sensitive to the choice of hyperparameters and random seed (Cho and Hariharan, 2019). To scrutinize this effect, we conducted experiments with ten consecutive random seeds, varying the hyperparameters α in the range of 0.3 to 0.7 and τ from 1 to 5 on the PDTB 2.0 top-level senses. Table 13 and table 14 show the average and overall results, respectively, for different combinations of hyperparameters and random seeds.

Variations in the results can be attributed to the stochastic nature of deep learning model training and the specific dynamics induced by knowledge distillation. The balance between learning from soft targets (teacher's predictions) and hard targets (original ground truth labels) - governed by the hyperparameters - and the model weights initialization (controlled by the random seed) can significantly influence the optimization trajectory and final model performance.

While initial results were reported with a single random seed, we believed it necessary to demonstrate the effect of these variables on our CP-KD method. Despite the observed fluctuations, our model outperforms the state-of-the-art on average, attesting to the robustness and superiority of our approach. This analysis underscores the importance of thorough hyperparameter studies in future research for ensuring reproducibility and robustness of results in implicit discourse relation recognition. Other than the results in this section, experiments were performed with the first random seed.

	α	τ=1		τ=2		τ=3		τ=4		τ=5	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
average 10 seeds	0.3	74.57	68.18	74.23	67.66	74.42	67.90	74.46	67.85	74.31	67.88
	0.4	74.59	68.28	74.53	68.19	74.38	67.92	74.30	67.80	74.28	67.79
	0.5	74.42	67.84	74.37	67.71	74.12	67.55	74.08	67.59	74.04	67.59
	0.6	74.16	67.70	74.14	67.39	74.21	67.67	74.04	67.53	74.07	67.56
	0.7	73.85	67.08	73.93	67.28	74.01	67.22	73.93	67.29	73.78	67.05

Table 13: Average results of different hyperparameters over 10 seeds on PDTB 2.0 top-level senses.

random seed	α	τ=1		τ=2		τ=3		τ=4		τ=5	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
20221026	0.3	74.57	67.86	74.57	68.16	74.76	68.49	74.67	68.26	74.86	67.98
	0.4	**75.43**	**68.86**	74.76	67.73	74.67	67.56	74.57	67.77	74.86	68.15
	0.5	75.14	68.76	74.95	68.30	74.38	67.89	74.38	67.45	74.95	68.01
	0.6	74.57	68.28	74.47	67.61	74.86	68.09	74.76	68.30	74.76	68.50
	0.7	74.38	68.25	74.38	68.11	74.76	68.40	74.09	67.11	74.38	67.99
20221027	0.3	73.80	66.91	73.52	66.71	74.09	67.69	73.80	66.73	73.90	67.19
	0.4	73.42	68.05	**74.47**	**68.80**	74.09	66.45	74.19	66.63	74.00	66.54
	0.5	73.71	66.72	74.19	66.95	73.90	67.66	73.80	66.50	73.90	68.26
	0.6	73.71	66.32	73.71	66.06	73.90	66.14	73.42	65.89	73.80	66.84
	0.7	73.33	67.31	73.42	66.12	73.04	65.27	73.61	65.43	73.23	65.36
20221028	0.3	74.19	67.99	73.61	67.44	73.61	67.60	73.71	67.27	73.71	67.64
	0.4	74.57	67.87	74.00	68.07	74.19	68.14	74.00	67.93	73.80	67.30
	0.5	74.67	68.31	73.90	66.57	74.38	68.18	74.28	67.23	74.57	68.32
	0.6	74.09	67.54	74.67	67.96	**75.05**	**68.91**	74.19	67.51	74.67	67.81
	0.7	73.52	66.61	74.19	67.05	75.14	68.11	74.57	68.15	74.09	66.98
20221029	0.3	74.00	68.70	74.38	67.43	74.00	68.66	74.38	67.84	74.28	67.90
	0.4	73.80	67.61	**74.38**	**68.88**	74.28	68.63	73.80	68.41	73.90	68.23
	0.5	73.80	68.02	74.00	68.09	74.00	68.04	74.28	68.32	73.90	67.91
	0.6	74.00	68.47	74.19	68.60	74.19	68.89	74.09	68.66	74.28	68.86
	0.7	73.90	66.91	73.71	68.00	73.71	68.01	73.52	67.85	73.61	67.91
20221030	0.3	75.05	68.12	74.67	67.53	74.38	67.57	74.57	67.73	74.09	67.25
	0.4	**74.86**	**68.61**	74.57	67.79	73.61	67.53	74.00	66.88	73.71	66.53
	0.5	73.80	67.73	74.00	67.33	73.61	66.46	73.52	67.59	73.23	67.02
	0.6	73.33	67.14	73.23	66.80	73.14	66.55	73.33	66.90	73.42	67.00
	0.7	73.52	67.44	73.23	67.01	73.33	67.02	73.42	66.84	73.33	66.73
20221031	0.3	**75.05**	**69.56**	74.09	66.59	75.05	67.66	75.53	68.48	74.86	67.57
	0.4	75.05	69.31	74.19	66.73	74.28	66.91	73.90	66.68	74.09	66.52
	0.5	74.00	65.67	73.90	66.06	73.42	64.75	73.80	66.42	73.42	65.19
	0.6	74.09	66.20	73.71	65.66	73.71	66.69	73.61	66.60	73.42	66.40
	0.7	73.33	65.06	73.71	66.28	73.90	65.25	74.00	67.20	73.23	66.10
20221032	0.3	74.67	67.24	74.09	67.91	74.47	68.00	74.09	67.67	73.80	67.30
	0.4	74.28	66.73	74.19	67.71	74.09	67.47	74.28	67.81	74.38	68.11
	0.5	74.19	66.35	**74.76**	**67.28**	73.90	67.42	74.09	67.62	74.00	66.23
	0.6	74.09	67.23	74.28	67.21	74.67	67.24	74.09	66.96	74.67	67.75
	0.7	73.90	66.47	74.28	67.72	73.71	66.52	74.19	67.23	74.09	67.02
20221033	0.3	74.28	67.28	74.09	67.66	74.28	67.22	74.57	67.16	74.47	67.73
	0.4	**74.57**	**68.51**	74.67	67.69	74.57	68.02	74.28	67.97	74.09	67.96
	0.5	74.57	68.09	74.57	68.23	74.19	67.15	73.90	67.35	74.19	68.29
	0.6	74.00	67.71	73.80	66.62	73.80	67.53	74.00	67.72	73.70	66.95
	0.7	74.09	67.38	73.71	66.61	73.90	67.04	73.71	66.69	73.33	66.24
20221034	0.3	74.95	68.42	74.38	68.02	74.38	66.72	74.09	67.83	74.00	68.78
	0.4	75.33	69.25	75.14	69.06	74.86	68.95	74.67	68.06	74.38	68.33
	0.5	**75.81**	**69.76**	74.95	69.01	74.76	68.61	74.19	67.93	73.80	67.25
	0.6	75.05	68.96	74.57	68.21	74.28	67.68	74.19	67.56	73.90	67.18
	0.7	73.80	66.77	74.28	67.21	74.09	67.66	74.09	67.63	74.09	67.35
20221035	0.3	75.14	69.74	74.86	69.13	75.14	69.43	75.14	69.54	75.14	69.42
	0.4	74.57	68.03	74.95	69.42	75.14	69.57	75.33	69.89	**75.62**	**70.26**
	0.5	74.47	69.02	74.47	69.24	74.67	69.34	74.57	69.48	74.47	69.39
	0.6	74.67	69.11	74.76	69.17	74.47	68.93	74.67	69.17	74.38	69.20
	0.7	74.76	68.63	74.38	68.70	74.47	68.93	74.09	68.78	74.38	68.83

Table 14: Results of different hyperparameters and random seeds on PDTB 2.0 top-level senses.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In Section 6 Conclusion and Limitations.

☒ A2. Did you discuss any potential risks of your work?
Our paper is a foundational research. In our paper, we aims to utilize knowledge distillation to mine the internal correlations between connective and sense labels to address the implicit discourse relation recognition. We cannot think of any potential risks of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 Introduction.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

In Section 4 Experiment.

☑ B1. Did you cite the creators of artifacts you used?
In Section 4.3 Implementation Details.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section 4.2 Baselines.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Section 3 Methods. Apache License 2.0 gives permission on Commercial use, Modification, Distribution, Patent use and Private use.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The dataset we applied is a commonly used open-source benchmarks datasets in the field of shallow discourse parsing.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
In Section 3 Methods and Section 4.1 Datasets.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 4.1 Datasets.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

In Section 4.4 Experimental Results and Analysis.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In Section 4.3 Implementation Details.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 4.3 Implementation Details.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 4.4 Experimental Results and Analysis.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In Section 4.5 Ablation Study.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.