Using Subtext to Enhance Generative IDRR

Zhipang Wang, Yu Hong*, **Weihao Sun, Guodong Zhou**
School of Computer Science and Technology, Soochow University, Suzhou, China
{zhipangwang, tianxianer, whsun16}@gmail.com; gdzhou@suda.edu.cn

Abstract

Implicit Discourse Relation Recognition (abbr., IDRR) is a NLP task of classifying argument pairs into different types of semantic relations. Arguments contain subtexts, some of which are beneficial to the perception of semantic relations. However, subtexts are connotative. The neural IDRR model fails to be aware of them without being given pertinent prompts. In this paper, we leverage LLaMA to generate subtexts for argument pairs, and verify the effectiveness of subtext-based IDRR. We construct an IDRR baseline using the decoder-only backbone LLaMA, and enhance it with subtext-aware relation reasoning. A confidence-diagnosed dual-channel network is used for collaboration between in-subtext and out-of-subtext IDRR. We experiment on PDTB-2.0 and PDTB-3.0 for both the main-level and secondary-level relation taxonomies. The test results show that our approach yields substantial improvements compared to the baseline, and achieves higher $F1$-scores on both benchmarks than the previous decoder-only IDRR models. We make the source codes and data publicly available.[1]

1 Introduction

IDRR determines the semantic relation between arguments when the in-between connective is absent (Prasad et al., 2008). For example, it outputs the relation "*Concession*" for the arguments $Arg1$ and $Arg2$ in 1), where the possible connective "*however*" is not given in the source text:

1) **Arg1**: *The new rate will be payable Feb. 15.*
Arg2: *A record date hasn't been set.*
Relation: Concession

Encoder-only language models such as RoBERTa (Long and Webber, 2022; Wu et al., 2023; Cai et al., 2024) and XLNet (Jiang et al., 2024) have been used for IDRR, where multi-class relation classification is conducted by linear layers with Softmax. Meanwhile, both T5 (Jiang et al., 2021; Chan et al., 2023) and the decoder-only Large Language Mdoels (LLMs) like GPT-3.5 (Chan et al., 2024) and GPT-4 (Yung et al., 2024) have also been verified for IDRR, where relations are properly generated conditioned on prompts and/or CoT. Significant improvements are reported in these arts.

Subtext hasn't been considered in the study of IDRR. Though, it is potentially useful for enhancing the IDRR models. A subtext is characterized by the metaphorical meaning hidden in the arguments. For example, the subtext of the two arguments in 1), most probably, is "*the rate should be recorded earlier though it hasn't been*". Such a subtext is more explicit or even straight in revealing the *Concessive* relation. Accordingly, we suggest that subtext can be used as a crucial evidence for enhancing the perception of implicit relation.

In this paper, we explore the method of applying subtexts, and systematically investigate the effectiveness upon LLM-based generative IDRR. The effort we made is to provide a preliminary study and stimulate innovative researches in subtext-based IDRR enhancement. Specifically, our contributions are summarized as follows:

- We first propose a Subtext-based Confidence-diagnosed Dual-channel Network (SCDN) for IDRR. In SCDN, subtext is generated by LLM. Confidence comparison is conducted to reconcile in-subtext and out-of-subtext IDRR.

- We verify the effectiveness of SCDN on the benchmarks PDTB-2.0 and 3.0 (Webber et al., 2019) . We report varied influences cased by the settings of prompting, confidence diagnosis, subtext generation and augmentation.

*Corresponding author
[1]https://github.com/ZpWang-AI/IDRR_Subtext.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 464–473
July 27 - August 1, 2025

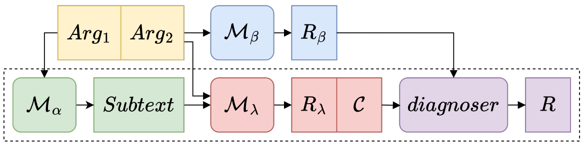


Figure 1: Architecture of SCDN.

2 Approach

Figure 1 shows the architecture of SCDN, which is constructed with three LLMs $\mathcal{M}_\alpha$, $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$. $\mathcal{M}_\alpha$ serves to generate the subtext for the given arguments. $\mathcal{M}_\beta$ takes the arguments as input and uses them as the only reliance for relation reasoning. $\mathcal{M}_\lambda$ combines the generated subtext and arguments, and infers the relation according to all of them. A probabilistic diagnosis model (*diagnoser*) is used to reconcile the decisions from $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$ based on confidence estimation.

2.1 Subtext and Relation Generators

Our baseline is the generator $\mathcal{M}_\beta$ which performs out-of-subtext IDRR. We prompt it by Question Answering (QA). Given the arguments $\dot{A}$ and $\ddot{A}$, we combine them with a question Q_β of "*what is the relation between arguments*": I_β=[{$\dot{A}$,$\ddot{A}$}; Q_β]. We feed I_β into $\mathcal{M}_\beta$ to generate a relation label.

To fulfill the in-subtext IDRR, we construct a subtext generator $\mathcal{M}_\alpha$. Its input is formed by the arguments and a prompting question Q_α of "*what is the implicit meaning*": I_α=[{$\dot{A}$,$\ddot{A}$}; Q_α]. There isn't any constraint applied to subtext generation (i.e., $\mathcal{M}_\alpha(I_\alpha)$) such as the length of subtext. Further, we build the generator $\mathcal{M}_\lambda$ to perform in-subtext IDRR. It uses both subtext and arguments as input, and combines them with a multi-choice question Q_λ: I_α=[{$\dot{A}$,$\ddot{A}$}; {S}; Q_λ]. The question Q_λ is designed as "*what is the relation between arguments given subtext*", which allows $\mathcal{M}_\lambda$ to generate a relation label in the manner of multi-choice QA (Yung et al., 2024) as follows.

2) Q_λ: What is the relation of $\dot{A}$ and $\ddot{A}$ given S?
A. *Contingency* B. *Expansion*
C. *Temporality* D. *Comparison*

In our experiments, we uniformly use LLaMA3-8B-Instruct (Dubey et al., 2024) to construct the generators $\mathcal{M}_\alpha$, $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$. Due to the zero-resource situation that there isn't any ground-truth subtext provided in PDTB-2.0 and 3.0, we train the subtext generator $\mathcal{M}_\alpha$ by teacher-student knowledge distillation (Hu et al., 2023). GPT-3.5-turbo (Brown et al., 2020) is used as the teacher.

2.2 Confidence Diagnoser

It is unavoidable that the subtext-based generator encounters two problems, including 1) arguments inherently don't contain a subtext, and 2) the generated subtext is unqualified. To relieve the problems, we use a diagnoser to reconcile $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$, where $\mathcal{M}_\beta$ conducts out-of-subtext IDRR, while $\mathcal{M}_\lambda$ additionally uses subtext for in-subtext IDRR.

Assume $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$ output the relations R_β and R_λ respectively. The diagnoser first verifies the reliability of R_λ. Confidence score $\mathcal{C}$ is measured for verification. $\mathcal{C}$ is an average logistic probability over all the tokens output by $\mathcal{M}_\beta$. Each $c_i \in \mathcal{C}$ is the non-normalized probability estimated by the logistic function in the final layer of LLaMA3:

$$\mathcal{C} = \overline{\sum_{t_i \in R_\beta} log_{\mathcal{M}_\beta}(t_i)} \tag{1}$$

On this basis, the diagnoser verifies whether $\mathcal{C}$ is larger than the type-specific threshold θ. If larger, the diagnoser determines that R_λ is reliable for output, otherwise the prediction R_β of $\mathcal{M}_\beta$ is adopted. In our experiments, we provide an exclusive threshold for each relation type in the taxonomy of PDTB. They are obtained by empirical observation upon the IDRR performance obtained on the training set. More details of threshold settings and performance curves can be found in Appendix A.

3 Experiments

3.1 Dataset and Evaluation Metrics

We experiment on two versions of discourse relation analysis datasets, including PDTB-2.0 (Prasad et al., 2008) and 3.0 (Webber et al., 2019). We follow the previous work to use sections 0-22 for IDRR, where sections 2-20 are used for training, and sections 0-1 are used as the development set, while 21-22 for testing. Appendix B shows the data statistics in all the datasets.

Multi-class Macro-F1 (F_1) and accuracy rate (Acc) are used as the evaluation metrics.

3.2 Implementation Details

We use AdamW optimizer (Loshchilov and Hutter, 2019) to optimize LLaMA3. For subtext generation $\mathcal{M}_\alpha$, the learning rate is set to 1e-4. A 5-epoch training process is conducted. For the relation generators $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$, the learning rates are uniformly set to 5e-5, and the best checkpoint is reached based on F_1 within 10 epochs. Both

Method	Backbone Model	Parameters	PDTB2		PDTB3	
			F_1	*Acc*	F_1	*Acc*
ChatGPT (Chan et al., 2024)	GPT-3.5-turbo	-	36.11	44.18	-	-
PIDRA (Yung et al., 2024)	GPT-4	-	-	-	47.53	52.84
FCL (Long and Webber, 2022)	RoBERTa-base	125M	69.60	72.18	70.05	75.31
CP-KD (Wu et al., 2023)	RoBERTa-base	125M	68.86	75.43	72.07	77.00
CP-KD (Wu et al., 2023)	RoBERTa-large	355M	**71.88**	**76.77**	**75.52**	**78.56**
SCIDER (Cai et al., 2024)	RoBERTa-base	125M	67.00	72.11	-	-
OTMT (Jiang et al., 2024)	XLNet-large	355M	64.46	72.34	-	-
CG-T5 (Jiang et al., 2021)	T5-base	223M	57.18	65.54	-	-
DiscoPrompt (Chan et al., 2023)	T5-base	223M	65.79	71.70	-	-
DiscoPrompt (Chan et al., 2023)	T5-large	738M	70.84	75.65	-	-
IICOT (Lu et al., 2023)	Flan-T5-base	248M	65.26	71.13	69.79	73.98
IICOT (Lu et al., 2023)	Flan-T5-large	783M	69.23	76.04	73.06	**77.46**
Baseline	Llama3-8B-Instruct	8.03B	66.72	73.90	70.71	75.31
SCDN (ours)	Llama3-8B-Instruct	8.03B	**71.14**	**78.20**	**73.33**	76.93

Table 1: Performance on PDTB 2.0/3.0. Encoder-only PLMs, decoder-only LLMs and T5-based encoder-decoder models are considered. The best results are separately marked in bold for encoder-only and decoder-only models.

Model	PDTB2		PDTB3	
	F_1	*Acc*	F_1	*Acc*
Out-of-subtext	66.72	73.90	70.71	75.31
In-subtext	70.56	77.82	72.79	76.32
SCDN	**71.14**	**78.20**	**73.33**	**76.93**

Table 2: Test results in ablation study.

employ a weight decay of 1e-2, a batch size of 1, and gradient accumulation over 8 steps. We don't extensively tune the hyperparameters.

All experiments are performed on a NVIDIA A100 GPU. Our model implementations are based on PyTorch [2] and the Transformers library[3].

3.3 Main Results

We compare SCDN to the recently-proposed advanced models, including 1) decoder-only ChatGPT (Chan et al., 2024) and PIDRA (Yung et al., 2024), 2) encoder-only FCL (Long and Webber, 2022), CP-KD (Wu et al., 2023), SCIDER (Cai et al., 2024), and OTMT (Jiang et al., 2024), as well as 3) T5-based CG-T5 (Jiang et al., 2021), DiscoPrompt (Chan et al., 2023), and IICOT (Lu et al., 2023). The decoder-only models take full advantage of prompt engineering for IDRR. The encoder-only models are effective in representation learning for relation understanding. T5-based models combine the advantages. More contributions of these arts are summarized in Appendix C.

Table 1 shows the comparison results on the test set, where the 4-way relation classification performance on the main relation taxonomy is reported. It can be observed that our SCDN achieves higher $F1$-score than both the decoder-only and T5-based IDRR models. However, it still fails to outperform the encoder-only models. This is partially attributed to hallucination of LLaMA3 and off-topic results it generated.

Besides, we conduct experiments for the 2nd-level taxonomy, where each main relation type is divided into fine-grained relation senses. For example, the relation type "*Contingency*" contains the senses of "*Conditionality*" and "*Causality*". In this experiment, SCDN shows promising performance, which is reported in Appendix D due to page limit.

3.4 Ablation Study

To provide a direct insight into the influence of subtexts, we conduct an ablation study. Three IDRR models are considered in it, including 1) **out-of-subtext** generator $\mathcal{M}_\beta$ which is separately fine-tuned without using subtexts, 2) **in-subtext** generator $\mathcal{M}_\lambda$ which additioanlly uses the generated subtexts during fine-tuning, and 3) **SCDN** which uses both $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$, and reconciles them with the confidence-based diagnoser.

Table 2 shows the test results for the main relation taxonomy. It proves that the utilization of

[2]https://github.com/pytorch/pytorch
[3]https://github.com/huggingface/transformers

Subtext Generation Model	F_1	Acc
GPT-3.5-turbo	71.55	75.98
LLaMA3 w/o Distill	71.07	75.37
LLaMA3 w/ Distill (Partial)	71.66	76.19
LLaMA3 w/ Distill (Whole)	**72.79**	**76.32**

Table 3: Contributions from different subtext generators.

Prompt	LLM	F_1	Acc
P_1	GPT-3.5-turbo	35.88	41.79
P_2	GPT-3.5-turbo	38.30	42.76
P_3	GPT-3.5-turbo	**39.24**	**43.72**
P_1	GPT-4-turbo	44.84	50.07
P_2	GPT-4-turbo	**46.29**	**52.21**

Table 4: Reliability of Prompts (on PDTB 3.0).

subtexts yields various levels of improvements. Appendix E provides a case study to show the effect.

3.5 Comparison among Subtext Generators

The qualified subtexts are crucial for SCDN. We investigate the subtexts generated by different LLMs, and verify their effects by our in-subtext model. There are three types of LLMs considered, including 1) GPT-3.5-turbo, 2) LLaMA3-8B-Instruct, 3) LLaMA3 which is strengthened by teacher-student knowledge distillation. During distillation, the subtexts generated by GPT-3.5-turbo for all training data are specified as **Guidance Data** from teacher. We use two different-sized guidance data: **Partial** and **Whole**. In the "Partial" case, we only adopt the guidance data which enables the subtext-based relation generator $\mathcal{M}_\lambda$ to output correct results. In the "Whole" case, all the guidance data is used.

Table 3 shows the performance of in-subtext IDRR models on the test set of PDTB 3.0, where different subtext generators are used. It can be observed that, compared to LLaMA3 (w/o distillation), GPT-3.5 enables the in-subtext model to perform better. Furthermore, no matter whether "Partial" or "Whole" guidance data is used, knowledge distillation causes improvements, and the latter case improves the in-subtext model more substantially. It is surprising that distillation allows the weaker LLaMA3 to be more contributive than its teacher GPT-3.5. The possible reason is because that LLaMA3 takes the advantage of itself when absorbing beneficial experience from GPT-3.5.

3.6 Prompts for Subtext Generation

The reliability of subtexts relies heavily on the design methods of prompts. For example, if we didn't remind LLMs of the ultimate purpose (i.e., being applied for IDRR), they fail to provide reliable subtexts. In our experiments, we evaluate different prompts as follows:

- P_1: It contains Q_1 and Q_2 (Section 2.1) that ask for subtext generation and IDRR in turn.
- P_2: It replaces the key words in P_1 with their different synonyms, e.g., "*subtext*" is replaced with "*implicit meaning*".
- P_3: It expands P_2 by adding a prefix to Q_1, where the prefix is an additional question about "*whether there truly is a subtext in the considered argument*". This prompt helps to avoid a forcible subtext generation.

Table 4 shows the IDRR performance on the development set when the above prompts are separately used, where GPT-3.5 and 4.0 are considered during the validation process. It can be observed that both synonym replacement and non-forcible subtext generation yield improvements. Accordingly, P_2 has been introduced into our SCDN for optimization. However, P_3 fails to be used in SCDN as it actually causes performance degradation. For example, by P_3, the in-subtext "Single" generator $\mathcal{M}_\beta$ obtains a $F1$-score of 71.7% on the test set of PDTB 3.0, causing a performance reduction of about 1.1% (compared to the "Single" case in Table 2). This implies that LLaMA3 in $\mathcal{M}_\beta$ isn't able to effectively perform reasoning with a relatively-complex Chain-of-Thought (CoT). Besides, P_3 also causes severe performance reduction when it is used in SCDN. This is because that a limited number of subtexts are generated by GPT-3.5 due to the constraint from P_3, and thus GPT-to-LLaMA distillation falls into the low-resource scenario.

4 Conclusion

In this paper, we verify that the utilization of subtexts helps to strengthen the LLMs-based generative IDRR. Experiments demonstrate that reconciliation of in-subtext and out-of-subtext IDRR is effective. We also exhibit that distilling a lightweight LLM-based subtext generator is contributive, when the prompt doesn't raise a complex CoT.

In the future, we will investigate the default subtext which isn't implied in the given argument. On the contrary, it derives from common-sense knowledge. Accordingly, we will convert binary arguments analysis to triplet, where the default subtext is regarded as the third non-negligible argument.

5 Limitations

This study proves the effectiveness of utilizing subtexts for enhancing IDRR. Nevertheless, a deeper analysis upon subtexts is still required. Our findings reveal that, in some cases, the shareable subtext is implied in one argument but irrelevant to the other, where the irrelevant argument appears as the noise during detecting the subtext. In some other cases, the default subtext occurs, which is not implied in any argument but derives from the common-sense knowledge. However, the subtext generation method in this paper cannot deal with these two problems. In the future, we will firstly study the common sense based default subtext generation. On this basis, we will convert the conventional binary arguments analysis to triplet, where the default subtext is used as a supplementary argument. This work will encounter the issues of 1) how to determine whether a pair of arguments are relevant to some default subtexts, and what they are, 2) how to detect and generate default subtexts conditioned on common-sense knowledge, and 3) how to reconcile the utilization of a triple of arguments and assign proper attention to them during relation discrimination.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under No.62376182.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mingyang Cai, Zhen Yang, and Ping Jian. 2024. Improving implicit discourse relation recognition with semantics confrontation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8828–8839, Torino, Italia. ELRA and ICCL.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

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

Zujun Dou, Yu Hong, Yu Sun, and Guodong Zhou. 2021. CVAE-based re-anchoring for implicit discourse relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1275–1283, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan, et al. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chengming Hu, Xuan Li, Dan Liu, Haolun Wu, Xi Chen, Ju Wang, and Xue Liu. 2023. Teacher-student architecture for knowledge distillation: A survey. *Preprint*, arXiv:2308.04268.

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. 2024. One general teacher for multi-data multi-task: A new knowledge distillation framework for discourse relation analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:239–249.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? *Preprint*, arXiv:2404.00999.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Yuxiang Lu, Yu Hong, Zhipang Wang, and Guodong Zhou. 2023. Enhancing reasoning capabilities by instruction learning and chain-of-thoughts for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5634–5640, Singapore. Association for Computational Linguistics.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

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)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11486–11494.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jing Xu, Ruifang He, Haodong Zhao, Huijie Wang, and Lei Zeng. 2023. Dual hierarchical contrastive learning for multi-level implicit discourse relation recognition. In *Natural Language Processing and Chinese Computing*, volume 14303, pages 55–66. Springer Nature Switzerland, Cham.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. *Preprint*, arXiv:2402.04918.

Lei Zeng, Ruifang He, Haowen Sun, Jing Xu, Chang Liu, and Bo Wang. 2024. Global and local hierarchical prompt tuning framework for multi-level implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7760–7773, Torino, Italia. ELRA and ICCL.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Relation	PDTB2-Top	PDTB3-Top
Comparison	31.06	29.97
Contingency	35.21	33.45
Expansion	31.22	29.83
Temporal	31.11	28.53

Table 5: Optimal thresholds for the main relation types.

A Threshold for Confidence Diagnosis

We set up a threshold for each relation type in the taxonomies of IDRR. For example, there are 4 thresholds provided for the four main IDRR relation types (i.e., *Expansion*, *Temporality*, *Contingency*, and *Comparison*). The adoption of different thresholds is because that the varying token lengths of relation labels cause unbalanced ranges of average confidence scores.

Let us consider the relation type T as an example. To seek for the optimal threshold θ_T for T, we empirically observe the T-oriented IDRR performance curve obtained when different optional values are used as thresholds. Within this empirical observation, the IDRR generator $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$ is used to predict relations and confidence scores for all the instances that hold a relation of T in the training set. And the accuracy Acc_T is used as the performance metric, which is calculated as follows:

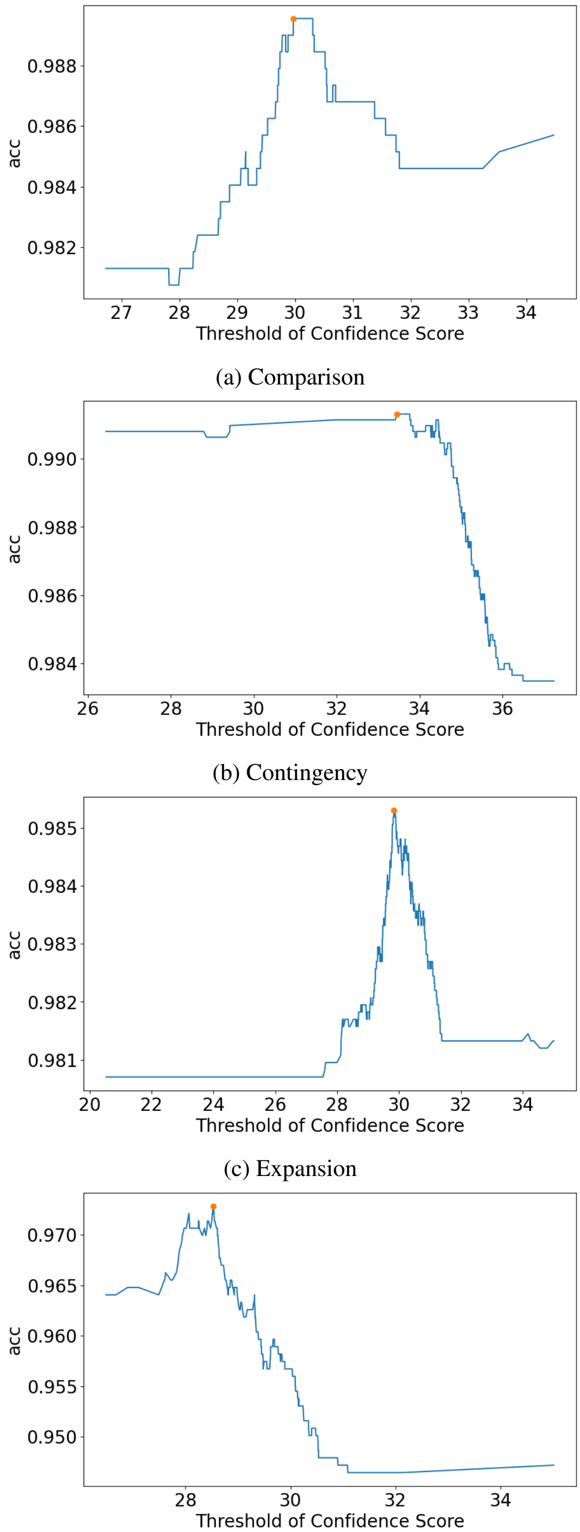
$$Acc_T(\check{\theta}) = \frac{n}{|D_T|} \tag{2}$$

where, $\check{\theta}$ is an optional threshold which is sequentially sampled from the range of confidence scores in the training set. n is the number of argument pairs which are given a positive relation prediction by $\mathcal{M}_\beta$ and $\mathcal{M}_\lambda$, and $|D_T|$ is the number of all argument pairs which hold the relation T.

We adopt the optimal threshold θ by maximum likelihood estimation on all optional thresholds:

$$\theta_T = \arg\max_{\check{\theta}_T \in \check{\theta}_{all}} (Acc_T(\check{\theta}_T)) \tag{3}$$

Figure 2 shows the curves of Acc_T changing with different thresholds in PDTB 3.0. We also present the specific values of the finally adopted thresholds θ in Table 5, 6 and 7, where Table 5 provides the thresholds for the main relation types (labeled as "Top"), while Table 6 and 7 give the thresholds for the relation senses in the secondary-level taxonomies (labeled as "Second").



(a) Comparison

(b) Contingency

(c) Expansion

(d) Temporal

Figure 2: Acc_T on the training dataset with varying thresholds. The final selected thresholds are marked with red dots, which corresponds to the highest Acc_T.

Relation	PDTB2-Sec
Comparison.Concession	28.10
Comparison.Contrast	29.10
Contingency.Cause	30.79
Contingency.Pragmatic cause	29.05
Expansion.Alternative	31.68
Expansion.Conjunction	28.02
Expansion.Instantiation	30.80
Expansion.List	28.48
Expansion.Restatement	27.56
Temporal.Asynchronous	28.12
Temporal.Synchrony	28.25

Table 6: Thresholds for all relation senses of PDTB 2.0.

Relation	PDTB3-Sec
Comparison.Concession	28.01
Comparison.Contrast	30.75
Contingency.Cause	32.01
Contingency.Cause+Belief	29.41
Contingency.Condition	30.89
Contingency.Purpose	30.31
Expansion.Conjunction	27.16
Expansion.Equivalence	28.31
Expansion.Instantiation	30.05
Expansion.Level-of-detail	28.44
Expansion.Manner	29.08
Expansion.Substitution	27.96
Temporal.Asynchronous	28.03
Temporal.Synchronous	29.65

Table 7: Thresholds for all relation senses of PDTB 3.0.

B Statistics of PDTB datasets

We use the benchmark datasets PDTB-2.0 and 3.0 in our experiments, and follow the common practice (Ji and Eisenstein, 2015) to divide each of them into training (Train), validation (Dev) and test sets. The statistics in the datasets are shown in Table 8 and Table 9.

C Related Work

Recent research has demonstrated that PLMs outperform traditional machine learning methods for the IDRR task. Consequently, many studies have explored incorporating novel modules into the encoder-only transformer architecture to obtain better representations and extract more comprehensive features from the input. The new added modules includes Conditional Variational AutoEncoder (Dou et al., 2021), Graph Convolutional Network (Wu et al., 2022), Gated Recurrent Unit (Wu et al., 2022), and attention mechanism (Wu et al., 2022; Xiang et al., 2022a; Jiang et al., 2023).

On the other hand, some studies applied new training strategies like contractive learning (Long and Webber, 2022; Jiang et al., 2023; Xu et al., 2023; Zeng et al., 2024), knowledge distillation (Wu et al., 2023; Jiang et al., 2024), and extra pre-training (Wang et al., 2023).

Notably, Zhou et al. (2022) proposed a prompt-based approach that involves connective prediction and answer mapping. Their work paved the way for better leveraging of connectives, like enhancing the input with predicted connectives (Liu and Strube, 2023; Liu et al., 2024), or mapping the answers by connectives directly (Xiang et al., 2022b; Zhou et al., 2022; Wu et al., 2023; Wang et al., 2023; Zeng et al., 2024). Additionally, several studies investigated the potential of multi-level hierarchical information for IDRR. These works explored modeling the relationships between labels and fusing the global and local information within the multi-level hierarchical structure(Jiang et al., 2023; Xu et al., 2023; Zhao et al., 2023).

However, the potential of generative models and LLMs has been relatively underexplored for IDRR(Chan et al., 2024; Omura et al., 2024; Yung et al., 2024). Therefore, our work aims to address this gap by investigating how to effectively utilize LLMs' reasoning capabilities and incorporate additional relevant information into the input.

Relation	Train	Dev	Test
Comparison	1,894	191	146
Contingency	3,281	287	276
Expansion	6,792	651	556
Temporal	665	54	68
Total	12,632	1,183	1,046

Table 8: Data statistics of PDTB 2.0.

Relation	Train	Dev	Test
Comparison	1,830	190	154
Contingency	5,896	579	529
Expansion	7,941	748	643
Temporal	1,418	136	148
Total	17,085	1,653	1,474

Table 9: Data statistics of PDTB 3.0.

Method	Backbone Model	PDTB2-Sec		PDTB3-Sec	
		F_1	Acc	F_1	Acc
ChatGPT (Chan et al., 2024)	GPT-3.5-turbo	9.27	15.59	-	-
PIDRA (Yung et al., 2024)	GPT4	-	-	25.77	36.98
FCL (Long and Webber, 2022)	RoBERTa-base	49.66	61.69	57.62	64.68
CP-KD (Wu et al., 2023)	RoBERTa-base	44.77	64.00	50.12	66.21
CP-KD (Wu et al., 2023)	RoBERTa-large	47.78	66.41	52.16	67.84
SCIDER (Cai et al., 2024)	RoBERTa-base	-	59.62	-	-
OTMT (Jiang et al., 2024)	XLNet-large	-	61.06	-	-
CG-T5 (Jiang et al., 2021)	T5-base	37.76	-	-	-
DiscoPrompt (Chan et al., 2023)	T5-base	43.68	61.02	-	-
DiscoPrompt (Chan et al., 2023)	T5-large	49.03	64.58	-	-
SCDN (ours)	LLaMA3-8B-Instruct	46.38	62.46	55.04	64.35

Table 10: Performance of the secondary level classification on PDTB 2.0/3.0.

D Performance on Relation Senses

Besides of the relation types in the main-level taxonomy of PDTB, we additionally evaluate our models upon the secondary-level taxonomy. The latter taxonomy consists of the fine-grained relation senses. Table 10 shows the Macro $F1$-scores and accuracies of our models, as well as the previous work that reported the performance on the secondary-level taxonomy.

It can be found that our SCDN achieves promising performance, outperforming the T5-base based generative models. Nevertheless, SCDN has an obvious performance gap compared to the T5-large based DiscoPrompt (Chan et al., 2023). DiscoPrompt is a strong IDRR model which learns from the reliance between relations and implicit connectives during training. The implicit connectives are informative when being used as guidance during training, which allows T5-large to infer relations from additional perspectives. By contrast, we didn't use implicit connectives as guidance when fine-tuning SCDN. Besides, as shown in Table 11 and 12, some relation senses in the secondary-level taxonomy are given much less available training data than others in PDTB 2.0 and 3.0. More seriously, some arguments of such relation senses fail to be given a subtext by our subtext generator. This results in the insufficient training towards these relation senses when we fine-tune our in-subtext IDRR model and SCDN. And this causes severe performance degradation.

Relation	Number
Comparison.Concession	180
Comparison.Contrast	1,566
Contingency.Cause	3,227
Contingency.Pragmatic cause	51
Expansion.Alternative	146
Expansion.Conjunction	2,805
Expansion.Instantiation	1,061
Expansion.List	330
Expansion.Restatement	2,376
Temporal.Asynchronous	517
Temporal.Synchrony	147
Total	12,406

Table 11: Counts of secondary level relation types on the training set of PDTB 2.0.

E Case Study

To qualitatively assess the impact of subtexts on the reasoning ability of LLaMA3 for IDRR, we manually examine the test data as well as predictions of the **Out-of-subtext** and **In-subtext** models. The out-of-subtext IDRR model $\mathcal{M}_\beta$ predicts relations conditioned only on arguments, while the in-subtext IDRR model $\mathcal{M}_\lambda$ combines subtext and arguments, and uses them as clues for relation reasoning. The following analysis showcases the benefits associated with the use of subtexts:

Example 1

Argument 1: *I personally don't enjoy seeing players who I remember vividly from their playing days running about and being gallant about their deficiencies.*

Relation	Number
Comparison.Concession	1,170
Comparison.Contrast	639
Contingency.Cause	4,483
Contingency.Cause+Belief	157
Contingency.Condition	154
Contingency.Purpose	1,102
Expansion.Conjunction	3,584
Expansion.Equivalence	252
Expansion.Instantiation	1,117
Expansion.Level-of-detail	2,493
Expansion.Manner	191
Expansion.Substitution	278
Temporal.Asynchronous	985
Temporal.Synchronous	433
Total	17,038

Table 12: Counts of secondary level relation types on the training set of PDTB 3.0.

Argument 2: *People should be allowed to remember players as they were.*
Subtext: *The implicit meaning between the arguments is that people should be allowed to remember and cherish the memories of players from their playing days without seeing them participate in activities that may diminish their legacy or reputation. It suggests a desire to preserve the image and impact of these players in the eyes of the public.*
Ground-truth: Expansion
Ouf-of-subtext model: Comparison
In-subtext model: Expansion

Example 2

Argument 1: *For a family with taxable income of $50,000 to pay the same 28% incremental tax rate as a family with taxable income of $250,000.*
Argument 2: *The 33% bracket should apply to all income over the applicable level, not just the 5% rate adjustment amount.*
Subtext: *The implicit meaning between these arguments is that both are advocating for a more progressive tax system where higher-income individuals pay a higher percentage of their income in taxes. Argument 1 suggests that the current tax system is unfair because a family with a much lower income is paying the same percentage of taxes as a family with a much higher income. Argument 2 suggests that the tax brackets should be adjusted to ensure that higher-income individuals pay a higher tax rate on all of their income, not just the amount above a certain threshold. Both arguments imply that the current tax system is not equitable and needs to be restructured to address income inequality.*
Ground-truth: Comparison
Ouf-of-subtext model: Expansion
In-subtext model: Comparison

Example 3

Argument 1: *He believes in what he plays, and he plays superbly.*
Argument 2: *His recent appearance at the Metropolitan Museum, dubbed "A Musical Odyssey," was a case in point.*
Subtext: *The implicit meaning between these arguments is that the musician's belief in his music is evident in his exceptional performance at the Metropolitan Museum, showcasing his talent and skill as a musician.*
Ground-truth: Expansion.Instantiation
Ouf-of-subtext model: Expansion.Restatement
In-subtext model: Expansion.Instantiation