Augmented Test-time Adaptation for Implicit Discourse Relation Recognition

Anonymous ACL submission

Abstract

Implicit discourse relation recognition (IDRR), the task of identifying the overt coherence relation between two text spans, is a task that involves deep semantic processing and general knowledge. Previous efforts mostly focus on the data-rich news domain. Despite the success of large language model (LLM) prompting in a range of reasoning tasks, recent works found that IDRR cannot be solved under zero- or few-shot settings at all. This work proposes to exploit the generation power of LLMs and augment training data with synthetic DR instances. In particular, in order to generalize IDRR to other domains, we apply offline test-time adaptation to inform the model with contrastive DR features that are expected in the target domain during test-time. Experimental results show that this method is effective in cross-domain IDRR without any labelled data in the target domain. This simple strategy could be generally applied to other text classification tasks.

1 Introduction

IDRR is a task to identify the covert discourse relation (DRs) between two given text spans (the *Arg1* and *Arg2*) in the absence of a specific marker such as *because* or *moreover*. It is a challenging task both for humans (Hoek et al., 2021) and models (e.g., 55.30% F1 and 63.57% accuracy on PDTB 3.0 in Jiang et al., 2023). One hurdle that hinders the general application of DR parsing to downstream tasks is the domain gap (Atwell et al., 2022; Liu and Zeldes, 2023). Most IDRR models are trained on the largest discourse annotated resource, the Penn Discourse Treebank (PDTB 3.0, Prasad et al., 2019), which are news articles from the 90s, and only a few works have evaluated cross-domain performance of IDRR (Shi and Demberg, 2019; Kurfalı and Östling, 2021). The generalizability of IDRR models to other domains remains unknown.

Zero- or few-shot prompting of large pretrained language models (LLMs) has the potential to solve domain-general IDRR because they have reached or surpassed humans in various reasoning tasks (e.g., Mao et al., 2023; Bang et al., 2023; Gilardi et al., 2023; Törnberg, 2023). Nonetheless, recent works reported that IDRR by prompting was not successful at all: the best-performing model using *gpt4* only reaches 26.12% F1 and 36.84% accuracy on PDTB 3.0, even with in-context examples and strategic prompt designs (Chan et al., 2023a; Yung et al., 2024). In contrast , previous work has shown that even the less advanced *gpt2* can *generate* reasonable DR continuations (Ko and Li, 2020). This suggests that the LLMs do have discourse knowledge, but it may not be demonstrated through a *classification* task.

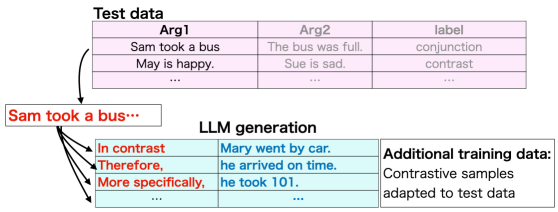

Figure 1: Summary of our proposed test-time adaptation framework for IDRR. The blue texts are the continuation (Arg2) generated by the LLM and the red texts are the prompts used in the generation. The gold labels of the test data are not accessible during training.

Motivated by this, we propose to use LLMs to generate training data for IDRR, with a particular focus on domain adaptation. To achieve this, we make use of the test-time augmentation and adaptation techniques that are common in computer vision (e.g., He et al., 2016; Wang et al., 2022; Gandelsman et al., 2022). As shown in Figure 1, the LLM is prompted to generate various DR continuations based on each *Arg1* of the *test set*. The generated DRs, which contain contrastive DR features adapted to the characteristics of the test-time instances, are appended to the training data of an IDRR model. Our experiment results show that this method improves a baseline model on cross-

domain IDRR. This framework can be applied to real-life scenarios to label test data of other domains in general text classification tasks.

2 Related Work

2.1 Implicit DR recognition

Various methodologies have been proposed for implicit DR recognition, including earlier works of feature engineering (Pitler et al., 2009; Park and Cardie, 2012) and later approaches based on deep neural networks (Shi and Demberg, 2019; Kishimoto et al., 2020; Guo et al., 2020; Jiang et al., 2021). Recent advances based on prompt-tuning on pre-trained LLMs successfully improved the SOTA (Xiang et al., 2022; Zhou et al., 2022; Zhao et al., 2023; Wu et al., 2023). However, attempts to classify implicit DR by zero/few-shot prompting were not fruitful - both a standard multiple-choice template (Chan et al., 2023a) and multi-step templates with verification questions (Yung et al., 2024) could only achieve half the performance of the supervised SOTA models.

One challenge of IDRR is the subtle distinction among different DR types, such as LEVEL-OF-DETAILS and INSTANTIATION. PDTB organizes the DR types in a three-level hierarchy, in which related DRs are grouped together. Based on the label hierarchy, several previous approaches successfully incorporated the dependence and contrast between the DR senses in IDRR models (Long and Webber, 2022; Wu et al., 2022; Jiang et al., 2023; Chan et al., 2023b). The current approach instead teaches the model the fine-grained difference of DR types using contrastive training samples that are adapted to the test-time context and domain.

2.2 Synthetic data generated by LLMs

Supervised learning algorithms rely on labeled data as the training objectives but manual annotation is time- and cost-intensive. Using controlled text generation (Hu et al., 2017) of LLMs through instructional prompts, large-scale training data were created for a range of NLP tasks, such as question-answering (Puri et al., 2020), textual similarity identification (Schick and Schütze, 2021), natural language understanding (Meng et al., 2022; Liu et al., 2022a) and commonsense reasoning (Yang et al., 2020).

Several previous works have explored whether LLMs can produce DRs correctly. Ko and Li (2020) reported that in 83% of the cases, the DR continuation generated by *gpt2*, prompted by a given *Arg1* and connective, were agreed by at least 3 out of 5 human annotators. However, the LLM has to be informed of the intended DR to be generated in order to lexicalize it correctly (Stevens-Guille et al., 2022), and in some cases, even the annotated labels in discourse resources are not fine-grained enough (Yung et al., 2021). To summarize, LLMs are aware of discourse coherence in their generation and are capable of generating DR arguments based on a given DR label or connective.

2.3 Test-time augmentation and adaptation

In image recognition, test-time adaptation (TTA) is increasing used for domain adaptation. The objective is to adapt a trained model to the distribution shift of the unlabelled test data before making a prediction (Liang et al., 2020; Sun et al., 2020; Wang et al., 2020; Iwasawa and Matsuo, 2021; Gandelsman et al., 2022). One way to perform TTA is by test-time augmentation, which is a common technique to improve model robustness (Krizhevsky et al., 2012; He et al., 2016; Matsunaga et al., 2017; Minh et al., 2018; Wang et al., 2022). For example, Zhang et al. (2022) adapt a model's parameters by minimizing the entropy of the model's output distribution across the augmentations of the test data. Chen et al. (2022) made use of a contrastive loss, using synthetically augmented test samples as extra positive examples.

TTA has been applied in question-answering (Banerjee et al., 2021; Ye et al., 2022, 2023), sentiment analysis and natural language inference (Antverg et al., 2022), structured data classification (Wei et al., 2023), machine translation evaluation (Zhan et al., 2023) and speech processing (Kim and Kim, 2021; Lin et al., 2022; Dumpala et al., 2023). Liu et al. (2022b) propose to select in-context examples that are semantically similar to the test samples when prompting LLMs. We instead use synthetic samples based on the test instances to train an LLM-based IDRR model.

3 Methodology

3.1 Contrastive test-time augmentation

We apply the TTA framework for cross-domain IDRR under the assumption that labelled training data and the whole unlabelled test data are both available.

As shown in Figure 1, each instance of the test data is augmented with one synthetic continuation

of the *Arg1* for each DR type to be augmented. The synthetic *Arg2*s, paired with the test-time *Arg1*s, are appended to the training data of the model. On the other hand, the dev set is not augmented and remains unchanged; the synthetic data are not used as tuning objectives. The gold labels of the test items are not accessible during training, as in an actual application scenario.

Adding evenly distributed samples to the training set distorts the original label distribution. On the one hand, it is less relevant in a cross-domain scenario where the target distribution is unknown. On the other hand, some distributional features are universal in discourses (e.g. CONTRAST is unlikely to be more frequent than CONJUNCTION in any discourse). We take a balanced approach of augmenting only the more frequent relation types by TTA. More specifically, we create contrastive samples only for the most frequent 7 inter-sentential relations out of the 14 (see Table 5). Next, we describe how the synthetic *Arg2*s are generated.

3.2 Discourse continuation generation

To harness the LLM's strength of left-to-right generation, we prompt the LLM to produce the continuation of a discourse prefix. More specifically, given the *Arg1* of an implicit DR and a connective that unambiguously marks a relation sense, the LLM is prompted to generate the *Arg2*.

The prompt template is shown in Figure 2. The instruction is a simple generation task that does not ask for specific linguistic knowledge. Using a connective instead of a relation label to prompt the continuation effectively avoids the generation of explicit relations that are marked by connectives. Each prompt contains one domain-specific example of the relation sense to be generated. The purpose of the example is mainly to illustrate the desired format of the output.

Most of the *Arg2* is the next consecutive sentence in our experimental data except the PURPOSE, CONDITION and MANNER relations in PDTB 3.0, which are mostly annotated within the same sentence[1]. We prompt the continuation of the same sentence for these three types but the continuation of the next sentence for all other relation types. Corresponding examples and connectives are provided, e.g. *by means of* for MANNER and *in addition,* for CONJUNCTION.

[1] PURPOSE, CONDITION and MANNER have 99%, 98% and 73% intra-sentential relations respectively while the rest of the relations in PDTB 3.0 only have 5% on average.

```
###Instructions###
Complete the sentence, and don't generate more than one sentence.

###Example###
Q: The Artist has his routine. He spends his days sketching passers-by, or trying to. Later, ...
A: at night he returns to the condemned building he calls home.

###Your task###
Q: The brokerage firms learned a lesson the last time around. Therefore, ...
A:
```

Figure 2: Prompt used to generate DR continuations.

The connectives used to prompt the continuation for each sense are shown in Table 4. Each level-2 relation is prompted by two alternative connectives, from which one is sampled in each prompt of that relation. The two choices reflect the frequency of the corresponding level-3 relations[2].

To evaluate the quality of the generation, the generated *Arg2*s paired with the corresponding *Arg1*s are classified by an IDRR parser trained on the original unaugmented data (i.e. the baseline model). Based on the 20482 generations from the PDTB 3.0 test set (see Table 5), we found that the agreement between the predicted and intended labels is 60% when evaluated on level-1 and 31% on level-2.

Two generation examples are shown in Figure 3 and 4. We observed that the unmatched level-2 labels are often alternative interpretations of the relation (e.g. 2 and 7 in Figure 3 and 9 and 12 in Figure 4); or likely due to the parser's error (e.g. 5 in Figure 3). We decided not to filter any generations based on the parser's prediction in order to maintain the intended configuration of the augmentation and to teach the model DR signals that the baseline fails to detect.

4 Experiment

We tested our proposed method on multi-domain discourse-annotated resources: PDTB 3.0 (news domain; expert annotation) and DiscoGeM (Europarl, novel and Wikipedia domains; crowd-sourced). More details on the data are given in Section C.

We compare the performance of a baseline model trained on the original data with a model

[2] e.g. the two LEVEL-OF-DETAIL connectives are both for ARG2-AS-DETAIL because ARG1-AS-DETAIL is much rarer ($< 10\%$); while both REASON and RESULT connectives are used for CAUSE because they are evenly distributed.

trained additionally on the synthetic TTA augmentations under a cross-domain setup: the model is trained on the PDTB training set (plus DiscoGem test-set adapted augmentation) and tested on the DiscoGeM test set. For comparison, we also evaluate the method under an in-domain setting: the model is trained on the PDTB training set (plus PDTB test-set adapted augmentation) and tested on the PDTB test set.

Following the setup of previous works (Kim et al., 2020; Xu et al., 2023), we conduct 14-way classification of the most frequent level-2 types in PDTB. Only 9 of the 14 types actually occur in the DiscoGeM test set. Even so, in the cross-domain setup, we train a 14-way parser on PDTB to parse the DiscoGeM test set. This corresponds to a realistic scenario where the target domain's distribution is unknown. The data distributions of different experimental setups are shown in Table 5.

We use the base GOLF model (Jiang et al., 2023) as our baseline. This state-of-the-art model uses a global and a local contrastive loss that teach the model to distinguish between positive and negative examples, which suits our approach of contrastive augmentation. The models are run by the codes provided by the authors. The synthetic discourse continuations are generated by *gpt-4* using the API provided by OpenAI, queried in February 2024.

5 Results

First, we look at the results under the cross-domain setting, where the model is trained on PDTB and tested on DiscoGeM, data coming from different other domains. Table 1 compares the accuracy and macro F1 of the model with and without TTA synthetic data. The scores are an average of 3 runs. We see improvements on both the 4-way top level and the 14-way second level, confirming the effectiveness of our proposed method.

DiscoGeM (cross-domain)	**4-way** F1	ACC	**14-way** F1	ACC
BL (Jiang et al., 2023)	54.38	62.77	31.86	50.08
+ TTA (freq)	**54.33**	**63.35**	**34.00**	**51.11**

Table 1: Comparison between the baseline (BL) and training with TTA synthetic data of 7 senses (TTA freq) per test *Arg1* of DiscoGeM.

For comparison, we applied the TTA method also under an in-domain setting. Table 2 compares the baseline model with TTA on "all" 14 or only the "freq"uent 7 relation types. Two other SOTA models are included for comparison. We see that TTA does not improve the performance. Especially, there is a significant drop on the macro F1.

Table 3 shows the breakdown of the F1 per relation sense. It can be observed that TTA facilitates the model to infer the less frequent relations in each category (e.g. LEVEL-OF-DETAIL instead of CONJUNCTION, CONTRAST instead of CONCESSION). However, the rarer relations (e.g. SUBSTITUTION) are not predicted accurately. This suggests that the quality of the synthetic data for these senses might not be sufficient.

PDTB 3.0 (in-domain)	**4-way** F1	ACC	**14-way** F1	ACC
Zhao et al. (2023)	69.06	73.27	52.73	63.09
Long and Webber (2022)	70.05	**75.31**	**57.62**	**64.68**
BL (Jiang et al., 2023)	**70.88**	75.03	55.30	63.57
+ TTA (all)	67.50	73.59	51.24	62.61
+ TTA (freq)	67.00	73.55	50.54	62.86

Table 2: Comparison between the baseline (BL) and training with TTA synthetic data of 14 senses (TTA all) or the more frequent 7 senses (TTA freq) per test *Arg1* of PDTB 3.0. Results of the cited models are quoted from the source papers.

test on:	PDTB		DiscoGeM	
	BL	+TTA	BL	+TTA
CONJUNCTION	**64.09**	61.09	**54.21**	51.27
LEVEL OF DETAIL	52.6	**55.09**	35.3	**39.04**
INSTANTIATION	**72.53**	69.2	23.24	**30.5**
MANNER	**63.53**	34.48	-	-
SUBSTITUTION	**66.67**	48.57	11.07	**12.9**
EQUIVALENCE	**25.39**	13.82	-	-
CAUSE	**69.47**	66.76	56.94	**59.96**
PURPOSE	71.6	**85.75**	-	-
CAUSE+BELIEF	0	0.03	-	-
CONDITION	66.67	**68.82**	-	-
CONCESSION	**59.09**	54.71	30.12	**32.00**
CONTRAST	43.33	**54.4**	19.57	**20.28**
ASYNCHRONOUS	**68.79**	65.44	**56.28**	52.79
SYNCHRONOUS	**41**	29.4	0	**7.26**
Macro F-1	**54.63**	50.54	31.86	**34.00**

Table 3: Per-class F1 comparing BL and TTA (freq)

6 Conclusion

Cross-domain IDRR is challenging and understudied. We proposed a test time adaptation method using synthetic discourse continuations and experimentally demonstrated the effectiveness of the approach. The prompted synthetic discourses are often more straightforward than the real ones, which often ambiguously signal multiple senses. We plan to explore how to generate or select better synthetic samples for IDRR.

7 Limitations

There are other configurations of data augmentation, using test-time adapted synthetic data or not, that were not explored in this work, such as augmentation in different proportions or synthetic sample selection. The performance of other IDRR models using the proposed method could be different from the baseline model we chose. Other ways to prompt the discourse continuation, as well as generation by other LLMs could have led to different outcomes and impacted the overall findings.

8 Ethical consideration

OpenAI's data collection complies with privacy laws[3]. We used the PDTB 3.0 corpus, which is licensed under the LDC User Agreement and the DiscoGeM corpus, which is publicly available on GitHub.

References

Omer Antverg, Eyal Ben-David, and Yonatan Belinkov. 2022. IDANI: Inference-time domain adaptation via neuron-level interventions. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 21–29, Hybrid. Association for Computational Linguistics.

Katherine Atwell, Anthony Sicilia, Seong Jae Hwang, and Malihe Alikhani. 2022. The change that matters in discourse parsing: Estimating the impact of domain shift on parser error. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 824–845, Dublin, Ireland. Association for Computational Linguistics.

Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. 2021. Self-supervised test-time learning for reading comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1200–1211, Online. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

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

[3] https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023b. 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.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305.

Sri Harsha Dumpala, Chandramouli Shama Sastry, and Sageev Oore. 2023. Test-time training for speech. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. 2022. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35:29374–29385.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Fengyu Guo, Ruifang He, Jianwu Dang, and Jian Wang. 2020. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7822–7829.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jet Hoek, Merel CJ Scholman, and Ted JM Sanders. 2021. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440.

Congcong Jiang, Tieyun Qian, Zhuang Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. 2021. Generating pseudo connectives with mlms for implicit discourse relation recognition. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*, pages 113–126. Springer.

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.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Sunwoo Kim and Minje Kim. 2021. Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 176–180. IEEE.

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, Marseille, France. European Language Resources Association.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from GPT-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59, Dublin, Ireland. Association for Computational Linguistics.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Murathan Kurfalı and Robert Östling. 2021. Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, pages 1–10, Online. Association for Computational Linguistics.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR.

Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. 2022. Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition. *arXiv preprint arXiv:2203.14222*.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022b. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. 2017. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *International skin imaging collaboration challenge at the international symposium on biomedical imaging (ISBI)*.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Tran Ngoc Minh, Mathieu Sinn, Hoang Thanh Lam, and Martin Wistuba. 2018. Automated image data preprocessing with deep reinforcement learning. *arXiv preprint arXiv:1806.05886*.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 108–112, Seoul, South Korea. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec,

Singapore. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Symon Stevens-Guille, Aleksandre Maskharashvili, Xintong Li, and Michael White. 2022. Generating discourse connectives with pre-trained language models: Conditioning on discourse relations helps reconstruct the PDTB. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 500–515, Edinburgh, UK. Association for Computational Linguistics.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. 2022. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation.

Kangda Wei, Sayan Ghosh, Rakesh Menon, and Shashank Srivastava. 2023. Leveraging multiple teachers for test-time adaptation of language-guided classifiers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7068–7088, Singapore. Association for Computational Linguistics.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 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, 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, 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 *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 55–66. Springer.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Hai Ye, Yuyang Ding, Juntao Li, and Hwee Tou Ng. 2022. Robust question answering against distribution shifts with test-time adaption: An empirical study. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6179–6192, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hai Ye, Qizhe Xie, and Hwee Tou Ng. 2023. Multi-source test-time adaptation as dueling bandits for extractive question answering. *arXiv preprint arXiv:2306.06779*.

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

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

Frances Yung, Merel Scholman, and Vera Demberg. 2021. A practical perspective on connective generation. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 72–83, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–820, Toronto, Canada. Association for Computational Linguistics.

Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642.

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.

A Connectives for DR prompting

Relation	Connective
CONJUNCTION	In addition, | Furthermore,
LEVEL-OF-DETAIL	More specifically, | In particular,
INSTANTIATION	For example, | For instance,
MANNER	by | by means of
SUBSTITUTION	Instead, | Rather than that,
EQUIVALENCE	In other words, | That is to say,
CAUSE	It is/was because | Therefore,
PURPOSE	in order | so as
CAUSE+BELIEF	As an evidence, | This justifies that
CONDITION	if | if it is/was
CONCESSION	Nonetheless, | Nevertheless,
CONTRAST	On the other hand, | In contrast,
ASYNCHRONOUS	Later, |Subsequently,
SYNCHRONOUS	Simultaneously, | Meanwhile,

Table 4: The discourse connectives used to prompt different DR types.

B Generation examples

Arg1: In this building lights were beginning to show.
Original *Arg2*: It was rapidly growing dusk. (CAUSE)
gpt4 generation:

1. CONJUNCTION: the sound of chatter and laughter could be heard echoing through the hallways. (IDRR model: CONJUNCTION)
2. LEVEL-OF-DETAIL: The warm glow from the third floor window indicated someone was still awake. (IDRR model: INSTANTIATION)
3. INSTANTIATION: The office on the third floor had its windows brightly lit, indicating that someone was working late. (IDRR model: INSTANTIATION)
4. CAUSE: People were starting to arrive for the evening event. (IDRR model: CONJUNCTION)
5. CONCESSION: The eerie silence made it feel as though it was completely deserted. (IDRR model: CAUSE)
6. CONTRAST: The adjacent structure was still shrouded in darkness. (IDRR model: CONTRAST)
7. ASYNCHRONOUS: People started to arrive for the evening event. (IDRR model: SYNCHRONOUS)

Figure 3: An example of the generated continuations prompted by a given *Arg1* in the DiscoGeM test set and a connective, and the predictions of the BL cross-domain model, given the *Arg1* and the generated *Arg2*.

Arg1: The Federal Reserve signaled its willingness to provide liquidity.
Original *Arg2*: The interest rate on its loans to major banks inched downward early in the day. (LEVEL-OF-DETAIL)

gpt4 generation:

1. CONJUNCTION: it also indicated a readiness to cut interest rates if necessary. (IDRR model: CONJUNCTION)
2. LEVEL-OF-DETAIL: it indicated that it would inject more money into the financial system if necessary to prevent a market crash. (IDRR model: LEVEL-OF-DETAIL)
3. INSTANTIATION: it announced plans to buy back government bonds to inject more money into the economy. (IDRR model: INSTANTIATION)
4. MANNER: lowering interest rates. (IDRR model: MANNER)
5. SUBSTITUTION: it may consider implementing quantitative easing measures to stimulate the economy. (IDRR model: LEVEL-OF-DETAIL)
6. EQUIVALENCE: the Federal Reserve indicated its readiness to supply more money into the economy. (IDRR model: LEVEL-OF-DETAIL)
7. CAUSE: investors can expect a more stable financial market in the near future. (IDRR model: CAUSE)
8. PURPOSE: to stabilize the financial markets during periods of economic uncertainty. (IDRR model: PURPOSE)
9. CAUSE+BELIEF: they announced a series of emergency lending programs to support the economy during the financial crisis. (IDRR model: INSTANTIATION)
10. CONDITION: needed to stabilize the economy. (IDRR model: CAUSE)
11. CONCESSION: the markets remained volatile. (IDRR model: CONCESSION)
12. CONTRAST: some economists are concerned that this could lead to inflation. (IDRR model: CONCESSION)
13. ASYNCHRONOUS: it announced a series of emergency lending programs to support the economy. (IDRR model: LEVEL-OF-DETAIL)
14. SYNCHRONOUS: the stock market responded positively, with major indices showing significant gains. (IDRR model: CAUSE)

Figure 4: An example of the generated continuations prompted by a given *Arg1* in the PDTB test set and a connective, and the predictions of the BL model, given the *Arg1* and the generated *Arg2*.

C Data

We use two discourse-annotated resources in our experiments: PDTB 3.0 and DiscoGeM.

PDTB 3.0 is the largest discourse resource in English annotated by trained annotators. The raw texts come from the news articles of the Wall Street Journal. Implicit relations are annotated between consecutive sentences as well as within a single sentences, if identified. 32 relation sense labels are defined which are arranged in a hierarchy, with 4 level-1 labels (EXPANSION, CONTINGENCY, COMPARISON and TEMPORAL), 22 level-2 labels, and 28 level-3 labels.

DiscoGeM is a crowdsourced corpus of implicit discourse relations in English containing texts from multiple genres: European Parliament preceedings (EP), Wikipedia articles (WK), and novels (NV). Each implicit DR between two consecutive sentences was annotated by 10 crowdworkers by the two-step DC insertion method (Yung et al., 2019) and using the labels from the PDTB 3.0 sense hierarchy.

Data split: In our IDRR experiments, sections $2-20$, $1-2$ and $21-22$ of PDTB 3.0 are used as the training, development and test sets (i.e. the Ji-data-split), while the provided data split of DiscoGem is used. Following previous works (Kim et al., 2020; Xu et al., 2023), we conduct 14-way classification of the most frequent level-2 labels. Only 9 of these 14 labels occur in the test set of DiscoGeM.

Instances that are not annotated with one of the 14 frequent labels (matched with the single sense label in the training/dev set and with multiple labels in the test set) and sub-labels not belonging to the 14 are excluded.

Multiple labels: In training and development, we use single sense label per instance, which are the *Conn1 SenseClass1* label of PDTB, the *Majority Sampled Label* of DiscoGeM, and the sense label used to prompt the synthetic discourse continuation.

During evaluation, the predicted label is compared against multiple labels per instance in the test set, which correspond to the *Conn1 SenseClass1* and *Conn2 SenseClass1* label of PDTB and the *Majority Distribution 40* label (all labels with 40% or more votes) of DiscoGeM. A match with any of the annotated labels is counted as a correct prediction. This setup follows previous works such as Long and Webber (2022) and Jiang et al. (2023).

To calculate the per-class F1 of the multi-label items, the TP counts are normalized by the number of multiple labels. For example, if the predicted label is A while the gold labels are A and B, the TP of A is increased by 0.5. The gold counts of A and B are both increased by 0.5.

	Original data							Synthetic data		
	train	dev	test					TTA on PDTB		TTA on DG
	PDTB	PDTB	PDTB	DG_{ep}	DG_{nv}	DG_{wk}	DG_{sum}	all	freq.	freq.
EXP.CONJUNCTION	3584	298	236	144	178	62	384	1463	1463	1264
EXP.LEVEL-OF-DETAIL	2597	261	208	76	112	19	207	1463	1463	1264
EXP.INSTANTIATION	1166	116	124	35	15	8	58	1463	1463	1264
EXP.MANNER	615	14	17	0	0	0	0	1463	0	0
EXP.SUBSTITUTION	343	27	16	1	2	0	3	1463	0	0
EXP.EQUIVALENCE	254	25	30	0	0	0	0	1463	0	0
CONT.CAUSE	4475	448	404	230	151	21	402	1463	1463	1264
CONT.PURPOSE	1092	96	89	0	0	0	0	1463	0	0
CONT.CAUSE+BELIEF	159	13	15	0	0	0	0	1463	0	0
CONT.CONDITION	150	18	15	0	0	0	0	1463	0	0
COMP.CONCESSION	1163	103	97	29	43	5	77	1463	1463	1264
COMP.CONTRAST	741	82	54	7	15	4	26	1463	1463	1264
TEMP.ASYNCHRONOUS	1007	101	105	7	82	11	100	1463	1463	1264
TEMP.SYNCHRONOUS	435	33	45	0	7	0	7	1463	0	0
Total size	17781	1635	1463	529	605	130	1264	20482	10241	8848

Table 5: Distribution of the data for the 14-way PDTB parser. Distributions based on the single label per instance (*Conn1 SenseClass1* in PDTB and *majority_label_sampled* in DiscoGeM (DG))