Synthetic Data Augmentation for Cross-domain Implicit Discourse Relation Recognition

Frances Yung, Varsha Suresh, Mansoor Ahmad, Zaynab Reza, Vera Demberg
Saarland University, Saarbrücken, Germany
{frances|vsuresh|mahmad|zreza|vera}@lst.uni-saarland.de

Abstract

Implicit discourse relation recognition (IDRR) – the task of identifying the implicit coherence relation between two text spans – requires deep semantic understanding. Recent studies have shown that zero-/few-shot approaches significantly lag behind supervised models However, LLMs may be useful for synthetic data augmentation, where LLMs generate a second argument following a specified coherence relation. We applied this approach in a cross-domain setting, generating discourse continuations using unlabelled target-domain data to adapt a base model which was trained on source-domain labelled data. Evaluations conducted on a large-scale test set revealed that different variations of the approach did not result in any significant improvements. We conclude that LLMs often fail to generate useful samples for IDRR, and emphasize the importance of considering both statistical significance and comparability when evaluating IDRR models.

1 Introduction

IDRR is the task of identifying the covert discourse relation (DRs) between two given text spans (the arguments: *Arg1* and *Arg2*) in the absence of a specific discourse connective (DC) such as *because* or *moreover*. Explicit labelling of discourse structure is beneficial to inform LLMs in summarization tasks (Li et al., 2016; Ishigaki et al., 2019; Xiao et al., 2020; Dong et al., 2021; Liu et al., 2023; Liu and Demberg, 2024). However, IDRR is challenging both for humans (Hoek et al., 2021) and models (SOTA, 56.50% F1 and 64.87% accuracy in Zeng et al., 2024), particularly in a cross-domain setting (Shi and Demberg, 2019; Atwell et al., 2022; Liu and Zeldes, 2023; Pyatkin et al., 2023).

Prompting of large pre-trained language models (LLMs), despite the human- or even superhuman-level performance in various reasoning tasks (e.g., Mao et al., 2023; Bang et al., 2023; Gilardi et al., 2023; Törnberg, 2023), was found to be not successful in IDRR. Few-shot prompting using *GPT-4* only reaches 30.90% F1 and 29.40% accuracy on the Penn Discourse Treebank (PDTB 3.0, Prasad et al., 2019), and 28.87% F1 and 32.67% accuracy on the multi-domain DiscoGeM corpus (Scholman et al., 2022), even with task-specific prompt engineering (Chan et al., 2024; Yung et al., 2024; Omura et al., 2024). On the other hand, a previous study demonstrated that LLMs can instead be used to generate synthetic data to augment the PDTB 3.0, improving the performance of classes that the baseline struggles to predict (Omura et al., 2024).

This work explores the application of synthetic data augmentation for cross-domain IDRR. Using raw texts from different target domains, we prompt LLMs to generate discourse continuations that express specific DRs. The generated data is then used to adapt the base model trained on human-annotated data of the source domain, which is the PDTB 3.0. We experimented with outputs of different LLMs, prompt templates, screening strategies and various advanced methods of domain adaptation, and evaluated on the entire DiscoGeM corpus, which is 4 times the size of the test set used in previous work (Omura et al., 2024). We only found marginal differences between models adapted to target-domain synthetic data compared with the base model. The benefit of generating synthetic data is unclear compared with direct application of the cross-domain base model, or using it to pseudo-label target-domain data. We have derived the following insights from the experimental results:

1. Synthetic data augmentation by LLMs does not improve IDRR under a cross-domain setting, where annotated target-domain data is unavailable for training nor validation. The synthetic data quality relies on the screening by a base model trained on large-scale source-

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 172–182

domain annotated data.

2. Previous positive results in in-domain IDRR may have been overly optimistic, as improvements were observed only under specific configurations, and exhibited high variability.

3. Fine-grained discourse inference remains a significant challenge for LLMs, compared with NLU tasks such as binary sentiment or topic classification (Ubani et al., 2023; Piedboeuf and Langlais, 2025).

4. Manual analysis shows that the valid generated samples are often highly prototypical examples of the DR class. In contrast, DRs in real texts tend to be more ambiguous and rely on indirect inference.

2 Related Work

2.1 Synthetic data generated by LLMs

Supervised learning algorithms rely on labeled data as training objectives but manual annotation is time- and cost-intensive. Using controlled text generation (Hu et al., 2017) with LLMs through instructional prompts, large-scale training data was created for a range of NLP tasks, such as question-answering (Puri et al., 2020), textual similarity identification (Schick and Schütze, 2021), NLU (Meng et al., 2022; Liu et al., 2022), commonsense reasoning (Yang et al., 2020), and dialogue classification (Sharma and Feldman, 2023), etc.

Recent findings indicate that while some tasks benefit from synthetic augmentation, others do not. (Ubani et al., 2023; Møller et al., 2024; Piedboeuf and Langlais, 2025). For example, Ubani et al. (2023) shows that synthetic data augmentation has been shown to significantly improve performance in movie review classification (positive vs. negative) and question type identification (e.g., whether a question seeks a "reason" or a "number"). However, in ambiguous tasks like irony detection, models trained with synthetic data often underperform the baseline (Piedboeuf and Langlais, 2025).

2.2 Implicit DR recognition

The current SOTA of IDRR models are mostly based on fine-tuning or prompt-tuning of the RoBERTa model (Xiang et al., 2022; Zhou et al., 2022; Zhao et al., 2023; Jiang et al., 2023b; Zeng et al., 2024). Cross-domain IDRR remains challenging and understudied. We started our experiments with GOLF (Jiang et al., 2023b), one of the SOTA models trained on PDTB, but found no significant improvement over a standard RoBERTa model. Several previous studies have found that models do not generalize well to out-of-domain data (Shi and Demberg, 2019; Atwell et al., 2021; Liu et al., 2021; Scholman et al., 2021; Kurfalı and Östling, 2021; Liu and Zeldes, 2023; Braud et al., 2023; Li et al., 2024). Furthermore, attempts to classify implicit DR by zero/few-shot prompting were not fruitful – both a standard multiple-choice template (Chan et al., 2024) and multi-step templates with verification questions (Yung et al., 2024) result in performance significantly below that of supervised models.

Despite the low performance in IDRR, LLMs have been found to be capable of generating DR arguments based on a given DR label or DC (Ko and Li, 2020; Stevens-Guille et al., 2022). Ko and Li (2020) reported that in 83% of the cases, the DR continuation generated by *gpt2*, prompted by a given *Arg1* and a DC, 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).

Omura et al. (2024) introduced an approach to augment the PDTB with synthetic data. Specifically, given an *Arg1* from the original PDTB training set and a DR label, an LLM is prompted to generate an alternative *Arg2*. The generated samples undergo a secondary filtering stage via few-shot prompting to discard ambiguous cases. During model training, a weighted loss is applied to balance the original and synthetic samples. The augmentation strategy targets the most confusing DR classes, i.e. those with the lowest recall on the PDTB validation set. Performance gains were observed when augmenting the top-3 most confusing classes, but not top-1 nor top-5, depending on the model size, raising some concerns about the robustness of these findings.

Compared to a standard RoBERTa$_{base}$ model (Liu et al., 2019), the reported improvements were modest: accuracy increased from 64.2 to 64.8, and macro-F1 improved from 57.1 to 59.5. However, results varied across the 3 runs, with fluctuations ranging from ± 0.4 to ± 1.6 points.

3 Experiment

3.1 Data

We set out to use synthetic discourse samples to adapt an IDRR model trained on source-domain data, which is the PDTB 3.0 (Prasad et al., 2019), to predict implicit DRs in the target domains, which are the sub-corpora in DiscoGeM 1.5 (Scholman et al., 2022; Yung and Demberg, 2025). The target domains include *Europarl* (EP), *Wikipedia* (WK) and *novel* (NV). We experiment based on a real-life scenario, where labelled target domain data is *unavailable*, i.e. the DiscoGeM data is used only for testing. However, it is assumed that the target domain is known during testing. The label distribution of the data we used and more details about the data can be found in Section A in the Appendix.

For the generation of the synthetic data, we collected raw texts from similar sources as the target domains: EP texts from the Europarl Direct Corpus (Koehn, 2005; Cartoni and Meyer, 2012); WK texts from the Wikipedia[1] and NV texts from the Opus Book Corpus (Tiedemann, 2012), omitting EP proceedings, WK articles and novels that are included in DG. While DG's data contains various translation directions, we only used original English texts as in the PDTB. 4000 sentences from each domain are randomly sampled for synthetic data generation, which is described in the next subsection.

3.2 Methodology

Table 1 summarizes the different methodological variants we explored. To harness the LLM's strength of left-to-right generation, we prompt the LLM to produce the continuation of a discourse prefix as in Omura et al. (2024). Given a sentence (the *Arg1*) from the raw text of the target domain and a DR label, an LLM is prompted to generate the following sentence (the *Arg2*). One in-context example selected from DiscoGeM is provided. We experimented with two different prompt templates.

The synthetic data was generated using three open-source LLMs, including: *Mistral-7B-Instruct-v 0.2* (Jiang et al., 2023a), *Llama3.1 8B-Instruct* (Dubey et al., 2024), and *gemma2 9B* (Team Gemma et al., 2024). We also considered *Deepseek-V3 7B* (DeepSeek-AI, 2024) but initial inspection of the synthetic data revealed that the generations were particularly noisy (e.g. the generated *Arg2*s were often exact repetition paraphrase of the given *Arg1*s). We thus did not include this model in the current experiment.

The generated DRs then undergo a selection process to remove noisy instances. In particular, the DC-prompt actually prompts the generation of *explicit* relations. The relation between the two arguments could shift when the explicit DC is removed (Sporleder and Lascarides, 2008; Liu et al., 2024). Since the zero-shot performance of LLMs significantly lags behind supervised models (Yung et al., 2024), we trained a RoBERTa$_{base}$ model (Liu, 2019) on the PDTB 3.0 to predict the DR of the synthetic samples and compare the prediction with the intended label. We compared three screening strategies, balancing sample quality and diversity.

The screened synthetic DR instances (statistics in Table 6 in the Appendix) are then used as domain-specific data to adapt the source-domain model. Similarly, we evaluated several methods and configurations, such as prefix tuning (Li and Liang, 2021) vs. simple data concatenation. We compare the proposed models with the baseline model trained on PDTB (i.e. the model for filtering), the SOTA GOLF model (Jiang et al., 2023b) trained on PDTB, and pseudo-labeling (Yarowsky, 1995). The pseudo-labelled data (DG$_{pseudo}$) are produced by using the baseline model to label adjacent sentence pairs in target-domain raw data. Sentence pairs in which an explicit DC was found at the beginning of the second sentence are excluded. This follows the preprocessing steps outlined in DiscoGeM: DCs within the first 5 tokens of the generated sentence are identified by string matching against a closed list (Scholman et al., 2022) and then excluded.

For each of the 14 Level-2 (see Section A) DR labels in PDTB 3.0, we generated one synthetic continuation using each LLM, based on the same set of $4,000$ randomly sampled raw sentences in each target domain. We also generate synthetic samples to the class SIMILARITY, which is exclusive to the target domain.[2] A total of $12,000$ pseudo-labelled instances per domain were used for comparison, roughly matching the size of the screened synthetic data (see Table 6). The combined data in the domain-mixed configuration was also down-sampled to approximately $10,000$ instances, while preserving the per-domain and per-class distribu-

[1] *"featured articles"* on the Wikipedia website as of 01.03.2020

[2] Since the baseline PDTB model does not classify the SIMILARITY, these synthetic samples never get through the *strict screen* but can possibly pass the *confusion* and *combination* screens.

LLMs	*Mistral-7B-Instruct, Llama3.1 8B-Instruct, gemma2-9B*
prompt template (more details in Appendix B)	
DC-prompt	lexicalize the DR to be generated by a connective (DC), e.g. *because* for the CAUSAL relation
DR-prompt	directly prompt by the DR label, providing the definition on the annotation manual
screening method (more details in Appendix C)	
strict screen	only include samples where the intended DR matches the prediction by the base model
confusion screen	exclude samples where the predicted label is a frequent misprediction of the intended label
combi screen	combination: apply the confusion screen if the intended DR is rare (implicit DR types with $<= 5\%$ distribution in PDTB 3.0), otherwise the strict screen
adaptation model	
PDTB + DG_{syn}	a RoBERTa$_{base}$ model trained on a direct combination of the PDTB and synthetic data
PDTB $\rightarrow$ DG_{syn}	the PDTB-trained RoBERTa$_{base}$ model adapted to the synthetic data by prefix-tuning
PDTB $\rightarrow_{IV}$ DG_{syn}	include an invariance loss (Zhou et al., 2020; Tzeng et al., 2014) alongside the standard cross-entropy loss to encourage the model to learn features that are indistinguishable between real and synthetic data final loss: $L_{\text{CE}} - \lambda L_{\text{IV}}$, where λ is set to 0.1 based on a search from $\{0.1, 0.3, 0.5\}$
target-domain data configuration	
domain-specific	one model specifically adapted to the synthetic data of each domain
domain-mixed	a single model trained on the combined synthetic data of all domains, with a domain token prepended to the beginning of each sample (Yung et al., 2022)

Table 1: Variants of the synthetic DR augmentation approach explored in the experiment.

tions to ensure comparability.

All domain adaptation models are trained for 3 epochs with a learning rate of 1e-4, primarily chosen to prevent overfitting (Yang et al., 2024), as our validation set consists of silver data rather than the target data. We set the embedding dimension of the prefix-tuning parameters to 512 leading to $\approx$ 7M trainable parameters compared to $\approx$ 130M parameters required for full fine-tuning.

3.3 Results

Table 2 presents the major comparison of the model variants. We focused on models using generations from *Mistral*, as it had the highest screening pass rate, indicating superior generation quality. All results are averaged values based on 3 random seeds. For evaluation, we computed accuracy and class-wise F1 in line with previous works (e.g. Xue et al., 2015), where macro-F1 scores are averaged across all classes occurring in the test set. For items with multiple gold labels, predictions matching any of the gold labels are considered correct, and the unmatched alternative labels are excluded from the classwise F1 calculation. We assessed the statistical significance of the difference between each model and the baseline using t-tests conducted over the results from the 3 experimental runs.

It can be seen in Table 2 that none of the model variants consistently outperform the baseline across domains and evaluation metrics. Considering the variation across runs, most results do not show statistically significant differences from the baseline. This suggests that numerical differences, up to 2.7% points, are primarily due to network randomness rather than genuine improvements. The SOTA GOLF model also does not outperform the baseline on DiscoGeM, highlighting the challenge of cross-domain IDRR. No clear advantage is observed over the more straightforward pseudo-labeling method. The only consistent and significant observation is the under-performance of the *confusion* and *combi* screens. This proves that the synthetic samples are not helpful without strict guidance by a supervised model, trained on a large number of human-annotated data,

4 Discussion and conclusion

Contrary to the improvements reported in previous studies, our results did not confirm the benefits of synthetic data augmentation for IDRR in a cross-domain setting. This is in line with recent reports that synthetic samples generated by LLMs do not improve abstract and ambiguous tasks, such as irony detection (Piedboeuf and Langlais, 2025).

We also found high variance in the model performance. In particular, the F1 scores of the rare classes are unlikely to be significant due to the skewed data distribution. The evaluation methods for instances with multiple gold labels also vary across studies, leading to inconsistencies. For example, while some works, including the current study, discard unmatched alternative gold labels; other works, such as Omura et al. (2024), count them as true positives in the F1 calculation.

Since PDTB was annotated by experts and DG via crowdsourcing, discrepancies in annotation methods may have an impact on the results. Pyatkin et al. (2023) report a moderate 56.9% agree-

Model	LLM	tpl.	screen	tgt. domain data		EP		WK		NV	
				config.	size[3]	F1	Acc	F1	Acc	F1	Acc
Baseline PDTB	-	-	-	-	0	21.03	42.00	22.81	45.58	21.94	43.98
GOLF PDTB	-	-	-	-	0	21.30	42.05	23.98	46.29	21.20	42.93
PDTB $\rightarrow$ DG$_{\text{syn}}$	llama3	DC	strict	specific	8680	21.69	41.74	22.33	47.32	22.89	44.19
	gemma2	DC	strict	specific	8315	21.94	41.88	23.99	46.67	23.12	44.98
	mistral	DC	strict	specific	10546	21.47	40.54	24.42	47.05	22.68	44.88
	mistral	DC	confuse	specific	43214	11.90*	21.73*	16.94*	35.07*	15.47*	31.54*
	mistral	DC	combi	specific	18286	16.72*	32.47*	19.41*	41.90*	18.73*	39.05*
	mistral	DR	strict	specific	12376	21.62	42.16	**24.87**	47.59*	22.86	**46.67**
	mistral	DR	strict	mixed	10441	21.58	41.77	24.03	47.21	**23.19**	45.54
	mistral	DC	strict	mixed	10441	**22.40***	42.05	23.73	46.78	22.50	44.19
PDTB + DG$_{\text{syn}}$	mistral	DC	strict	specific	12356	21.32	39.72*	23.82	46.94	22.80	44.86
PDTB $\rightarrow_{IV}$ DG$_{\text{syn}}$	mistral	DC	strict	specific	12376	21.57	40.32	24.12	47.53	22.41	44.60
	mistral	DC	strict	mixed	10441	22.03*	41.46	22.38	47.05	22.50	43.74
PDTB + DG$_{\text{pseudo}}$	-	-	-	specific	12000	20.81	41.28	23.63	46.29	23.07*	44.12
PDTB $\rightarrow$ DG$_{\text{pseudo}}$	-	-	-	specific	12000	20.71	**42.37**	23.42	**47.64***	22.30	44.29
	-	-	-	mixed	10000	21.78*	42.01	23.90	46.83	21.80	43.00
PDTB $\rightarrow_{IV}$ DG$_{\text{pse}}$	-	-	-	specific	12000	21.01	41.88	24.37	47.21	21.66	43.39

Table 2: Performance of model variants evaluated on the sub-corpora of DiscoGeM. The best scores are bolded. Significant differences from the baseline model, based on variations across runs, are marked with *. The synthetic data sizes of the domain-specific models are averaged across the models of the three domains.

ment between the original and DG-style crowdsourced labels of 300 PDTB items. A similar rate (57.7%) was found between expert annotations of implicit DRs (Zikánová et al., 2019), suggesting that crowdworkers align with expert judgments to a comparable degree.

In fact, interpretation of implicit DRs involves deep cognitive pro- cessing that is universally difficult for humans (Oza et al., 2009; Zhou and Xue, 2012; Poláková et al., 2013; Hoek et al., 2021). The major reason is that multiple interpretations are often possible based on the perspectives of the readers (Rohde et al., 2016; Scholman et al., 2022).

We manually annotated a random subset of 100 synthetic samples and found a 65% agreement with the intended DRs in the prompts. 28 of the 35 disagreement were valid alternative interpretations. This shows that the generated DRs are valid but less ambiguous than natural ones, which often lack clear cues and allow multiple plausible readings. Figure 1 provides several examples: while alternative discourse relations can be inferred from the genuine DR samples, the synthetic examples tend to be clearer and more straightforward.

The limited effectiveness of synthetic data for IDRR may therefore be explained by the perspectivist nature of DR inference. Since prompting LLMs to generate specific relations tends to bias outputs of prototypical cases, a potential improvement could involve prompting examples that reflect multiple plausible senses, combined with co-occurrence-aware, multi-label screening. Future work could also explore additional factors, such as the effect of synthetic sample size and the impact of contrastive generation, such as pairing different *Arg1*s in each sample v.s. using a single *Arg1* with multiple continuations representing different DR senses.

1a) Genuine REASON/CONJUNCTION

Arg1: It is an honour and a pleasure to have the opportunity to present this report to Parliament today.
Arg2: It is on the very important subject of product liability on which the European Community first introduced legislation as long ago as 1985 in the form of a directive...

1b) Synthetic REASON

Arg1: Personally he had nothing to fear, for the convicts could not reach him in Granite House.
Arg2: He was securely locked within Granite House.

2a) Genuine ARG2-AS-INSTANT/ARG2-AS-DETAIL

Arg1: The history of agriculture began thousands of years ago.
Arg2: After gathering wild grains beginning at least 105,000 years ago, nascent farmers began to plant them around 11,500 years ago.

2b) Synthetic ARG2-AS-INSTANT

Arg1: Holmes mourned that the pony pennings of his day were only "a shadow of their former glory".
Arg2: Breeds such as Shire horses or Friesians, once prominent in England and the Netherlands respectively, could serve as examples.

Figure 1: Genuine DR examples from DiscoGeM 1.5 and screened synthetic samples generated by *Mistral*.

5 Limitation

A primary limitation of the current work is the lack of experimentation with the latest, more powerful and larger LLMs, such as *GPT-4o* and *Claude-3 Opus*, due to constraints in budget, time, and computational resources. These more advanced models may be capable of generating higher-quality discourse samples for augmentation, potentially leading to improved performance.

In addition, we did not extensively refine the prompt templates in order to generate discourses that are closer to natural examples. As discussed in Section 4, the generated samples were less ambiguous than naturally occurring discourses. A possible improvement could involve explicitly instructing LLMs to generate ambiguous examples or instances with multiple DR interpretations. However, intensive prompt engineering is necessary to generate high-quality ambiguous DR samples. Pseudo-labeling remains a more promising approach for capturing DR ambiguity, as it leverages real texts rather than synthetic ones.

Acknowledgements

This project is supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074).

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325, Singapore and Online. 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.

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*.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Bruno Cartoni and Thomas Meyer. 2012. Extracting directional and comparable corpora from a multilingual corpus for translation studies. In *Proceedings 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

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.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yue Dong, Andrei Mircea, and Jackie Chi Kit Cheung. 2021. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, Online. Association for Computational Linguistics.

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

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

Jet Hoek, Merel C.J. Scholman, and Ted J.M. 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, Tübingen, Germany. Association for Computational Linguistics.

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.

Tatsuya Ishigaki, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2019. Discourse-aware hierarchical attention network for extractive single-document summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 497–506, Varna, Bulgaria. INCOMA Ltd.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of* the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2219–2224, Lisbon, Portugal. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2023b. 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.

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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand.

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.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147, Los Angeles. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. 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.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, and Vera Demberg. 2023. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, 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? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. 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.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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. *CoRR*, abs/1907.11692.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. 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*.

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.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. LLM-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192, St. Julian's, Malta. 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.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Suman Meena, Dipti Misra Sharma, and Aravind Joshi. 2009. Experiments with annotating discourse relations in the hindi discourse relation bank. In *Proceedings of the 7th International Conference on Natural Language Processing (ICON-2009), Hyderabad, India*.

Frédéric Piedboeuf and Philippe Langlais. 2025. On evaluation protocols for data augmentation in a limited data scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3428–3443, Abu Dhabi, UAE. Association for Computational Linguistics.

Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. 2013. Introducing the Prague discourse treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing.

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.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany. 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.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Merel C. J. Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022. Discogem: A crowdsourced corpus of genre-mixed implicit discourse relations. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*, Marseille, France. European Language Resources Association (ELRA).

Ashwyn Sharma and David I Feldman. 2023. Team cadence at mediqa-sum 2023: Using chatgpt as a data augmentation tool for classifying clinical dialogue. In *CLEF (Working Notes)*, pages 1680–1687.

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.

Caroline Sporleder and Alex Lascarides. 2008. Using automatically labelled examples to classify rhetorical relations: An assessment. *Natural Language Engineering*, 14(3):369–416.

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.

Team Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Solomon Ubani, Suleyman Olcay Polat, and Rodney Nielsen. 2023. Zeroshotdataaug: Generating and augmenting training data with chatgpt. *arXiv preprint arXiv:2304.14334*.

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.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2020. Do we really need that many parameters in transformer for extractive summarization? discourse can help ! In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 124–134, Online. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Jianfei Yang, Hanjie Qian, Yuecong Xu, Kai Wang, and Lihua Xie. 2024. Can we evaluate domain adaptation models without target-domain labels? In *The Twelfth International Conference on Learning Representations*.

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.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Frances Yung and Vera Demberg. 2025. On crowdsourcing task design for discourse relation annotation. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 12–19, Abu Dhabi, UAE. International Committee on 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.

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.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13025–13032.

Yuping Zhou and Nianwen Xue. 2012. Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77.

Šárka Zikánová, Jiří Mírovskỳ, and Pavlína Synková. 2019. Explicit and implicit discourse relations in the prague discourse treebank. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, pages 236–248. Springer.

A Data

	PDTB 3.0		DG 1.5 (test)			
	train	dev	EP	WK	NV	ttl.
Expansion						
CONJUNC.	3584	298	314	177	323	814
LEV.-OF-DET.	2493	262	532	161	460	1153
INSTANT.	1117	116	212	37	94	343
MANNER	191	14	20	0	6	26
SUBSTITUT.	278	27	41	4	47	92
EQUIVAL.	252	25	50	2	38	90
DISJUNCT.	0	0	0	0	5	5
EXCEPTION	0	0	0	1	6	10
Contingency						
CAUSE	4469	450	885	86	857	1828
PURPOSE	1102	97	139	10	49	198
CAUSE+BEL.	157	13	0	0	0	0
CONDITION	152	18	85	3	19	107
Contrast						
CONCESSION	1164	103	229	23	186	438
CONTRAST	639	82	38	19	58	115
SIMILARITY	0	0	65	6	42	113
Temporal						
ASYNCHR.	985	102	18	70	536	624
SYNCHR.	433	34	73	16	158	247
Total	17016	1641	2704	615	2884	6203

Table 3: Distribution of the Level-2 classes, grouped under 4 Level-1 categories, in each data subset. DiscoGeM's distribution is based on the single majority label per sample

We use PDTB 3.0 (Prasad et al., 2019) as the source-domain data for training and tuning and DiscoGeM 1.5 (Scholman et al., 2022; Yung and Demberg, 2025) as the target-domain data for evaluation. Table 3 shows the distribution of all labelled data used in this study. PDTB 3.0 is the largest discourse resource in English annotated by trained annotators. The texts come from the news articles of the Wall Street Journal in the 90s. Implicit relations are annotated between consecutive sentences as well as within individual sentences, if identified. The relation labels are arranged in a 3-level hierarchy. We train our source-domain model to predict the 14 Level-2 labels with more than 10 instances in the test set, as in previous works (e.g. Kim et al., 2020). Sections $2-20$ and $1-2$ are used as for training and tuning respectively (Ji et al., 2015).

DiscoGeM 1.5 is a crowdsourced corpus of implicit discourse relations in English containing texts from multiple genres: European Parliament proceedings (EP), Wikipedia articles (WK), and literature (NV). Each relation is annotated by 10 crowdworkers using a connective insertion task. The label set is also based on the PDTB 3.0 label hierarchy, but only inter-sentential relations are annotated. We use the complete DG corpus for evaluation, except the instances that are labelled NO RELATION, which is not considered as a type of implicit DRs in PDTB 3.0.

In the training and tuning of all the models, we use a single label per instance (*Conn1SenseClass1* label of PDTB). The predicted labels are evaluated against multiple labels, which are defined as annotations with 40% or more votes.

B Prompt template

The exact DC-prompt and the DR-prompt templates are shown in Figure 2. Table 4 lists the connectives used in the DC-prompt for each DR label.

```
###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:
```

```
###Instructions###
Given two arguments, the relation "Conjunction" is defined as "both arguments, which don't directly relate to each other, bear the same relation to some other situation evoked in the discourse".
Here are examples that have the relation "Conjunction":
She, out of gratitude, had her arms wrapped around his neck as they slept. CONJUNCTION
Various articles of their clothing lay intermingled around the bed.

###Your task###
Please write down the second arguments that have the relation CONJUNCTION to the first argument: "And over the desert plain one heard only the moan of squalls through the broken trellises of the enclosures." Here list several second arguments:
```

Figure 2: Top: DC-prompt; bottom: DR-Prompt

C Screening methods

The *confusion screen* filters out samples where the base model's predicted label does not match the intended label but instead corresponds to a frequent

intended DR L'	DC in DC-prompt
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,
SIMILARITY	Similarly,
ASYNCHRONOUS	Later, |Subsequently,
SYNCHRONOUS	Simultaneously, | Meanwhile,

Table 4: The discourse connectives used in the DC-prompt for different DR types.

misclassification of the ground truth. For example, if the baseline model frequently misclassifies CAUSE+BELIEF as CAUSE, synthetic samples labeled as CAUSE+BELIEF but predicted as CAUSE are excluded by the confusion screen. Table 5 provides a complete mapping of the most common mispredictions, derived from the confusion matrix of the *RoBERTa-base* model evaluated on the PDTB 3.0 dev set. This screening method follows the strategy proposed in previous work (Omura et al., 2024), but replaces zero-shot prompting-based predictions with those obtained through supervised classification. Table 6 summarizes the

intended label L'	confuse(L')
CONJUNTION,LEVEL-OF-DETAIL SUBSTITUTION, EQUIVALENCE CAUSE+BELIEF, CONDITION CONCESSION, ASYNCHRONOUS	CAUSE
INSTANTIATION, MANNER, CAUSE	LEVEL-OF-DETAIL
SYNCHRONOUS, SIMILARITY	CONJUNCTION
PURPOSE	CONDITION
CONTRAST	CONCESSION

Table 5: Intended label L' vs confuse(L') used in the **confuse screen**. The generation is selected if $L_{pred} \neq$ confuse(L').

screened label distributions per different settings. The screened data size of the *mistral* generation is $10\% - 20\%$ larger, indicating higher agreement with the supervised model. On the other hand, the high selection rate of the confusion screen used in the previous work suggests a significantly more lenient selection process.

LLM	llama3	gemma2	mistral
prompt	DC		
screen	strict		
EP	9361	8464	10724
WK	9101	9068	10689
NV	7579	7415	10224
LLM	mistral		
prompt	DC	DC	DR
screen	confuse	smooth	strict
EP	39846	17337	11473
WK	44833	19434	12457
NV	44962	18087	13198

Table 6: Size of the synthetic data generated by different LLMs, prompts, and screens. There were 60000 instances (4000 $Arg1_{\text{raw}}$s $\times$ 15 DR types) generated in each case before screening.