A Multi-Task and Multi-Label Classification Model for Implicit Discourse Relation Recognition

Nelson Filipe Costa and **Leila Kosseim**
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
nelsonfilipe.costa@mail.concordia.ca
leila.kosseim@concordia.ca

Abstract

We propose a novel multi-label classification approach to implicit discourse relation recognition (IDRR). Our approach features a multi-task model that jointly learns multi-label representations of implicit discourse relations across all three sense levels in the PDTB 3.0 framework. The model can also be adapted to the traditional single-label IDRR setting by selecting the sense with the highest probability in the multi-label representation. We conduct extensive experiments to identify optimal model configurations and loss functions in both settings. Our approach establishes the first benchmark for multi-label IDRR and achieves SOTA results on single-label IDRR using DiscoGeM. Finally, we evaluate our model on the PDTB 3.0 corpus in the single-label setting, presenting the first analysis of transfer learning between the DiscoGeM and PDTB 3.0 corpora for IDRR.

1 Introduction

Implicit discourse relation recognition (IDRR) is one of the most challenging tasks in computational discourse analysis. Its goal is to identify the sense of discourse relations connecting text arguments in the absence of explicit discourse connectives, such as *but* and *because*. The Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008a) organizes discourse senses across three hierarchical levels with increasing degrees of detail (Webber et al., 2019). However, understanding the sense of a discourse relation can be a complex and subjective task without the guidance of discourse connectives. Consider the two arguments of the implicit discourse relation below:

Arg1: The lights flickered.

Arg2: The power went out.

In this example, it is not clear whether the two arguments of the relation are related by a temporal or causality relation. One could interpret that the lights flicked and *then* the power went out, or that the lights flickered *because* the power went out. This ambiguity illustrates the biggest challenge of IDRR. Despite its difficulty, IDRR plays a crucial role in downstream tasks that rely on text coherence. In summarization, for example, it helps preserve the logical flow and rhetorical intent of the source text. In dialogue systems, it supports coherence and intent modeling by inferring how utterances relate to each other, improving dialogue discourse parsing (Li et al., 2024; Thompson et al., 2024).

To date, most research on IDRR has relied on the different iterations of the PDTB corpus (Prasad et al., 2008b, 2019). However, despite significant efforts, SOTA performance on IDRR has plateaued in recent years at F1-Scores of 71.59 at level-1 and 57.62 at level-2 (Long and Webber, 2022; Liu and Strube, 2023; Chan et al., 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024). One possible reason may be the inherently subjective nature of implicit discourse interpretation, which can be difficult to capture using predominantly single-label annotated corpora (Stede, 2008; Rohde et al., 2016; Scholman and Demberg, 2017; Hoek et al., 2021). The idea of possible multiple relations between discourse arguments had already been considered in the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) and the ambiguity in the annotation of implicit relations has been further highlighted by the challenges in mapping them across discourse frameworks (Demberg et al., 2019; Costa et al., 2023). In light of these challenges, recent studies have advocated for the multi-label annotation of implicit relations to better capture their nuanced and complex nature in discourse corpora (Yung et al., 2019; Pyatkin et al., 2020; Scholman et al., 2022a,b; Pyatkin et al., 2023; Yung and Demberg, 2025). This shift in perspective has led to the creation of DiscoGeM (Scholman et al., 2022a) - the first multi-label annotated corpus of implicit discourse relations.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 231–245
Aug 25–27, 2025. ©2025 SIGDIAL

Following recent work in NLP (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Basile et al., 2021; Fornaciari et al., 2021; Uma et al., 2021; Plank, 2022; Jiang and de Marneffe, 2022), we see disagreement between humans annotators not as noise, but as an opportunity to enhance the relevance of IDRR models by capturing the complexity of human interpretations. This paper addresses the subjectivity in IDRR by proposing a multi-task classification model capable of learning both multi-label and single-label representations of implicit discourse relations. Our model is trained on the DiscoGeM corpus and uses a multi-task architecture to leverage the interdependence between senses and jointly learn sense representations across all three sense levels in the PDTB 3.0 hierarchy. The main contributions of this work are:

- We present the first multi-label approach in IDRR which produces probability distributions over all possible sense labels of a discourse relation using the DiscoGeM corpus. Our multi-label model can then be adapted to the traditional single-label task by selecting the label with the highest probability.
- We conduct extensive experiments, comparing different pre-trained language models as encoders of our model and evaluating how different loss functions impact performance in multi-label and single-label IDRR.
- We establish the first benchmark on multi-label IDRR and achieve SOTA results on single-label IDRR using DiscoGeM.
- We present an in-depth analysis on the potential of transfer learning between the DiscoGeM and the PDTB 3.0 corpora on single-label IDRR.

2 Previous Work

Despite the recent advances in natural language understanding, IDRR remains one of the most challenging tasks in discourse analysis. Most previous work addressed IDRR as a single-label classification task by fine-tuning (Long and Webber, 2022; Liu and Strube, 2023) or prompt-tuning (Chan et al., 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024) pre-trained language models (PLMs). More precisely, Long and Webber (2022) apply contrastive learning and augment training data with explicit connectives from PDTB 3.0 metadata to fine-tune their model, while Liu and Strube (2023) propose a two-step pipeline that first generates an explicit connective for each relation to fine-tune a classifier on the modified input. Chan et al. (2023) inject hierarchical structure and connective-based explanations into prompts, enabling joint predictions across all three PDTB sense levels. Zhao et al. (2023) address data scarcity in IDRR by using a parameter-efficient prompt-tuning framework that incorporates hierarchical label guidance into the verbalizer. Building on this, Zeng et al. (2024) propose a prompt-tuning approach that integrates both global and local hierarchical label information into the verbalizer to improve output alignment with pre-trained objectives. Finally, Long and Webber (2024) introduce a prototype-based verbalizer informed by the PDTB 3.0 sense hierarchy, combining contrastive and prototype learning to eliminate the need for manually designed verbalizers. While most of these approaches (Long and Webber, 2022; Liu and Strube, 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024) use RoBERTa$_{\text{base}}$ (Liu et al., 2019) as their PLM, Chan et al. (2023) uses T5$_{\text{base}}$ (Raffel et al., 2020). Another line of work tried to solve single-label IDRR by directly prompting large language models (LLMs) through prompt-engineering (Chan et al., 2024; Yung et al., 2024). However, both works show that the results obtained through directly prompting LLMs in zero-shot and few-shot settings are still far behind from those obtained through fine-tuning and prompt-tuning PLMs.

The shift toward multi-label annotation has stirred an initial wave of research in multi-label IDRR. For instance, Long et al. (2024) used the 4.9% of implicit relations in the PDTB 3.0 corpus that are annotated with two senses to build a model capable of predicting up to two senses per instance. However, given that the 95.1% of PDTB 3.0 annotations remain single-label, their model predominantly produces single-label predictions. In contrast, Yung et al. (2022) and Costa and Kosseim (2024) also use the multi-label DiscoGeM corpus but convert its annotations into a single-label format during training. Thus, to date, no previous work has truly captured the full multi-label representation of implicit discourse relations - with the exception of our work on multi-lingual IDRR (Costa and Kosseim, 2025), where we consider multi-label IDRR in a multi-lingual setting incorporating hierarchical learning in the training of our classification model and comparing it against LLMs via direct prompting with few-shot learning. Similar to other works (Chan et al., 2024; Yung et al., 2024), we found that prompting LLMs leads to worst results compared to the fine-tuning of PLMs such as RoBERTa in the context of IDRR. Therefore, we do not consider prompting LLMs in this paper.

3 Our Approach

In this work, we introduce a novel multi-task classification model for IDRR that simultaneously predicts sense distributions across all three levels in the PDTB 3.0 framework. As illustrated in Figure 1, the model processes a concatenated pair of discourse arguments (ARG1+ARG2) using RoBERTa (Liu et al., 2019) as the PLM encoder[1]. The resulting contextualized embedding is passed through a linear transformation and dropout layer and subsequently fed to three distinct classification heads - each corresponding to a sense level in the hierarchy. In the multi-label setting, each head outputs a probability distribution over the available senses at its respective level. In the single-label setting, we apply an additional pooling layer that selects the sense with the highest probability from each distribution. We use the Adam optimizer (Kingma and Ba, 2015) to minimize the loss function, which we calculate as the sum of the individual losses of each classification head.

Multi-Label Classification. For each multi-label classification head, we compute the loss using the mean absolute error (MAE) loss function (see Equation 2 in Appendix B), which we found to yield better performance (as detailed in Section 5.1). Following previous work in multi-label classification for NLP (Pyatkin et al., 2023; van der Meer et al., 2024), we evaluate model performance in this setting using the Jensen-Shannon (JS) distance (Lin, 1991) to measure the similarity between the predicted and target probability distributions.

Single-Label Classification. In the single-label setting, we compute the loss for each classification head using the cross-entropy (CE) loss function (see Equation 1 in Appendix B), which emphasizes the correct classification of the highest probability sense label (as discussed in Section 5.1). We evaluate model performance in this setting using the

[1]We also experimented with other PLMs, but RoBERTa achieved the best performance (see Section 5.1).

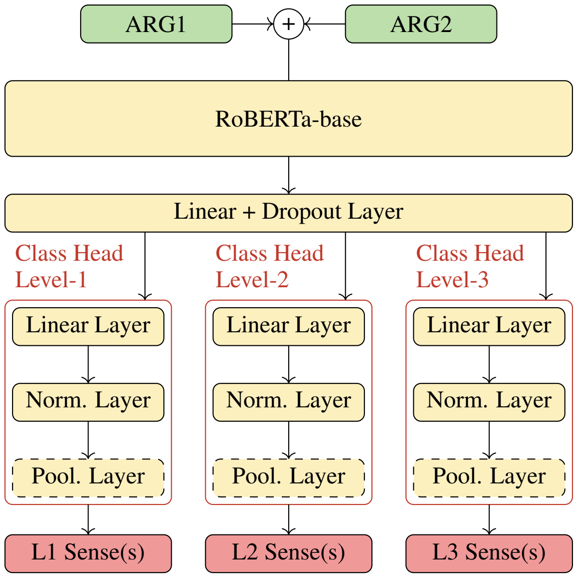

Figure 1: Architecture of our multi-task classification model for IDRR. Given a pair of discourse arguments as input, the model generates one output per sense level in the PDTB 3.0 framework. In the multi-label setting, each output is a probability distribution over the senses at the corresponding level. In the single-label setting, the sense with the highest probability is selected via the pooling layers (indicated by dashed lines).

weighted F1-score based on the majority label in the predicted and target sense distributions.

4 Data Preparation

We used two different discourse annotated corpora in this work: DiscoGeM (Scholman et al., 2022a) and PDTB 3.0 (Prasad et al., 2019). Our model was trained exclusively on the DiscoGeM corpus and evaluated on both DiscoGeM (for multi-label and single-label classification) and on PDTB 3.0 (for single-label classification). Although the two corpora differ in annotation methodologies, both follow the same discourse framework (Webber et al., 2019). This shared foundation allowed us to directly evaluate the transfer learning potential of training our model on DiscoGeM and evaluating it on PDTB 3.0 in single-label IDDR.

4.1 DiscoGeM

The DiscoGeM corpus contains 6,807 implicit discourse relations drawn from four textual genres: 2,800 relations from political texts, 3,060 from literary texts, 645 from encyclopedic texts and 302 relations extracted from the PDTB 3.0 corpus. Each relation was independently annotated by at least 10 crowdworkers and the resulting annotations were aggregated to form a probability distribution over 29 discourse senses in the PDTB 3.0 framework - the BELIEF and the SPEECH-ACT level-2 senses were not annotated in DiscoGeM.

Since the DiscoGeM corpus did not include the CAUSE+BELIEF sense from the standard set of 14 level-2 senses proposed by Kim et al. (2020) for IDRR classification in the PDTB 3.0 framework, we replaced it with the SIMILARITY sense (the most frequent among the remaining level-2 senses) to preserve a 14-label set. In addition to classifying senses at level-1 and level-2, we also incorporated level-3 sense labels when available. When no level-3 sense was available, we defaulted to their corresponding level-2 sense as replacement. Table 8 in Appendix A shows the distribution of senses across all levels in the original DiscoGeM annotations and in our adapted label set.

For single-label classification, we replaced the multi-label sense distribution of each discourse relation by the sense with the highest score - referred to as the majority label. To reduce computational complexity and maintain consistency across model evaluations, we used the same training, validation and testing splits in both the multi-label and the single-label settings. We split 70% of the corpus for training, 10% for validation and 20% for testing. To ensure a balanced distribution, we kept the same distribution of majority labels across all data splits (see Figure 2 in Appendix A). We opted for a fixed split of the data, instead of multiple folds for cross-validation, to streamline the experimentation process as we conducted an extensive number of experiments.

4.2 PDTB 3.0

The PDTB 3.0 corpus was annotated by expert annotators and consists of 53,631 discourse relations extracted from Wall Street Journal news articles - from which 21,827 are implicit. Since there are currently no other benchmarks in multi-label IDRR using DiscoGeM and the majority of research in IDRR relies on the single-label annotated PDTB 3.0 corpus, we also evaluated our model on the traditional single-label classification task using different test splits of the PDTB 3.0 corpus through transfer learning.

To allow transfer learning between the two corpora, we used the same set of 14 level-2 senses described in Section 4.1. Following the common approach to IDRR, we only kept the level-1 and level-2 senses in the PDTB 3.0. To compare our work against SOTA models in single-label IDRR, we replicated the two commonly used Lin (Lin et al., 2009) and Ji (Ji and Eisenstein, 2015) test splits - the former uses section 23 in the PDTB 3.0 corpus as the test set, while the latter uses sections 21-22. However, since recent works highlighted the limitations of using such small test sets to draw meaningful conclusions (Shi and Demberg, 2017; Kim et al., 2020), we also generated test splits following the cross-validation scheme proposed by Kim et al. (2020). Table 9 in Appendix A shows the total number of level-2 sense instances on the different single-label test sets.

Top-1	Top-3	Top-5	Top-10
92 (30.5%)	164 (54.3%)	197 (65.2%)	215 (71.2%)

Table 1: Number of instances where the reference label in the PDTB 3.0 was found within the top-k labels in DiscoGeM for the set of 302 co-annotated relations.

4.3 Corpora Agreement

Since a total of 302 implicit discourse relations taken from the PDTB 3.0 were also independently annotated in DiscoGeM, we calculated the annotation agreement between the two corpora over this overlapping set of relations. Table 1 reports the number of relations where the PDTB 3.0 reference label matches the majority label (top-1) in DiscoGeM, or appears within its top-3, top-5, or top-10 majority labels. The fact that for 28.8% of the jointly annotated relations, the reference label from the PDTB 3.0 corpus was not selected by at least one of the annotators of DiscoGeM illustrates the variability in human interpretation of implicit discourse relations - even when annotations are based on the same underlying framework.

5 Experiments

To optimize our model (illustrated in Figure 1), we explored multiple model configurations and loss functions using the validation set of DiscoGeM. All of the results reported in the following sections are the average scores of three different runs with different random starts for 10 epochs and with a batch size of 16. We ran our experiments on a 32-core compute node with 512GB of RAM. All of the code used can be found on GitHub[2].

[2] https://github.com/nelsonfilipecosta/Implicit-Discourse-Relation-Recognition

PLM	Loss	JS Distance ↘ (Multi-Label)			F1-Score ↗ (Single-Label)		
		Level-1	Level-2	Level-3	Level-1	Level-2	Level-3
BERT	MAE	0.314 ± 0.004	0.461 ± 0.007	0.535 ± 0.004	63.34 ± 1.41	46.44 ± 0.82	39.20 ± 1.82
	CE	0.328 ± 0.004	0.563 ± 0.002	0.634 ± 0.004	63.49 ± 1.38	51.42 ± 0.58	47.65 ± 0.92
DistilBERT	MAE	0.330 ± 0.006	0.472 ± 0.003	0.543 ± 0.003	61.29 ± 0.96	39.29 ± 0.66	31.75 ± 0.55
	CE	0.341 ± 0.005	0.571 ± 0.006	0.639 ± 0.003	60.28 ± 0.58	45.55 ± 1.52	41.19 ± 1.50
RoBERTa	MAE	$\mathbf{0.304 \pm 0.010}$	$\mathbf{0.448 \pm 0.006}$	$\mathbf{0.530 \pm 0.004}$	65.12 ± 1.20	51.81 ± 1.54	42.99 ± 1.09
	CE	0.317 ± 0.003	0.561 ± 0.003	0.630 ± 0.005	$\mathbf{65.48 \pm 0.82}$	$\mathbf{53.51 \pm 0.39}$	$\mathbf{49.55 \pm 0.11}$
DistilRoBERTa	MAE	0.318 ± 0.001	0.463 ± 0.003	0.539 ± 0.004	61.30 ± 0.38	42.75 ± 1.63	34.96 ± 1.29
	CE	0.329 ± 0.002	0.564 ± 0.005	0.636 ± 0.003	65.02 ± 0.69	52.83 ± 0.71	47.94 ± 0.88

Table 2: Results of experimenting with different pre-trained language models (PLMs) and different loss functions in multi-label classification (JS distance) and in single-label classification (weighted F1-score). The results were averaged across three different runs with random starts. Values in bold show the best score for each metric.

5.1 Model Selection

Motivated by prior work suggesting that PLMs trained with a next sentence prediction (NSP) objective perform better on IDRR (Shi and Demberg, 2019), we compared BERT (Devlin et al., 2019), which includes the NSP objective, against RoBERTa (Liu et al., 2019), which omits it. RoBERTa has also been widely adopted in SOTA single-label IDRR models (Long and Webber, 2022; Liu and Strube, 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024), making it a strong baseline. To further explore efficiency-performance trade-offs, we also evaluated distilled variants of both models (Sanh et al., 2019). Each PLM configuration was evaluated using both the mean absolute error (MAE) and cross-entropy (CE) loss functions. The results of these experiments are summarized in Table 2.

As shown in Table 2, RoBERTa consistently outperforms BERT across both loss functions, yielding lower Jensen-Shannon (JS) distances in the multi-label setting and higher F1-scores in the single-label setting across all sense levels. These results suggest that the performance gains of using larger models outweigh the previously reported advantages of using models pre-trained on the NSP objective in IDRR (Shi and Demberg, 2019). In addition, both BERT and RoBERTa outperform their respective distilled variants, reaffirming the benefit of using full-sized models for this task. The results in Table 2 also show that the choice of loss function plays an important role. Models trained with MAE consistently achieve lower JS distances in the multi-label setting, while model trained with CE achieve higher F1-scores in the single-label setting. We present a more detailed analysis on the performance of each loss function in Appendix B, as well as the results of training the different models with the mean squared error (MSE) and the Huber loss - which led to slightly worse results as shown in Table 10 of the appendix. Based on these findings, we selected RoBERTa as the backbone encoder of our model, using MAE loss for multi-label classification and CE loss for single-label classification.

5.2 Fine-Tuning

We fine-tuned our model separately with different learning rates for multi-label and single-label IDRR. Table 3 shows the results of fine-tuning our model with the MAE loss in the multi-label setting and with the CE loss in the single-label setting. Under both settings, we observed performance improvements with learning rates of 1×10^{-5} and 5×10^{-6}. To further enhance model performance with these two learning rates, we also experimented incorporating two decay functions: linear decay and cosine annealing with warm restarts (Loshchilov and Hutter, 2017). For linear decay, the learning rate was gradually reduced over the first 5 epochs to half its initial value. In the cosine annealing schedule, the learning rate oscillated between its original value and half that value over two complete cycles across 10 epochs. As shown in Table 3, the cosine annealing strategy with a learning rate of 1×10^{-5} achieved the best results for multi-label classification, while the linear decay schedule with a learning rate of 5×10^{-6} proved most effective in the single-label setting.

6 Results

In this section, we present the results of our experiments in two parts. Section 6.1 reports the performance of our models on the DiscoGeM test set for both multi-label and single-label IDRR and

LR	Decay	JS Distance ↘ (Multi-Label with MAE)			F1-Score ↗ (Single-Label with CE)		
		Level-1	Level-2	Level-3	Level-1	Level-2	Level-3
$1e^{-4}$	–	0.386 ± 0.004	0.509 ± 0.003	0.569 ± 0.005	36.07 ± 1.67	21.61 ± 0.58	10.05 ± 0.99
$5e^{-5}$	–	0.339 ± 0.041	0.478 ± 0.039	0.554 ± 0.016	64.71 ± 2.39	49.76 ± 1.65	45.52 ± 1.59
$1e^{-5}$	–	0.304 ± 0.010	0.448 ± 0.006	0.530 ± 0.004	$\mathbf{65.48 \pm 0.82}$	53.51 ± 0.39	49.55 ± 0.11
	Linear	0.305 ± 0.009	0.459 ± 0.007	0.547 ± 0.004	64.70 ± 0.35	54.32 ± 2.15	48.96 ± 0.81
	CosAn	$\mathbf{0.299 \pm 0.004}$	$\mathbf{0.447 \pm 0.008}$	$\mathbf{0.529 \pm 0.008}$	65.15 ± 0.61	53.99 ± 1.17	49.56 ± 1.25
$5e^{-6}$	–	0.303 ± 0.006	0.461 ± 0.005	0.548 ± 0.005	65.45 ± 0.36	53.67 ± 0.45	49.77 ± 1.15
	Linear	0.315 ± 0.006	0.480 ± 0.007	0.566 ± 0.007	64.84 ± 1.14	$\mathbf{55.86 \pm 1.10}$	$\mathbf{50.34 \pm 1.55}$
	CosAn	0.306 ± 0.003	0.469 ± 0.002	0.558 ± 0.004	65.40 ± 1.04	54.97 ± 1.85	49.62 ± 1.26
$1e^{-6}$	–	0.348 ± 0.003	0.522 ± 0.005	0.613 ± 0.003	60.66 ± 1.38	47.66 ± 0.64	40.96 ± 0.48

Table 3: Results of experimenting with different learning rates and decay functions in multi-label classification (JS distance) using the MAE loss and in single-label classification (weighted F1-score) using the CE loss. The results were averaged across three different runs with random starts. Values in bold show the best score for each metric.

Section 6.2 presents the transfer learning results of our single-label model evaluated on various test splits of the PDTB 3.0 corpus.

6.1 Results on DiscoGeM

Table 4 presents the performance of our models (trained with MAE and CE loss functions) on the DiscoGeM test set for both multi-label and single-label IDRR. Results in the multi-label setting are reported using Jensen-Shannon (JS) distance, while results in the single-label setting are reported using weighted F1-score. As there is currently no prior work on full multi-label IDRR, we include a baseline for comparison. This baseline emulates the DiscoGeM annotation protocol by generating ten single-label predictions per relation, sampled from the probability distribution of each sense at each level, and then averaging them into a probability distribution. In the single-label setting, we compare our results against prior work evaluated on DiscoGeM (Yung et al., 2022; Costa and Kosseim, 2024). As shown in Table 4, our model consistently outperforms the random baseline across all levels and achieves a substantial improvement over the current SOTA in level-2 classification (Yung et al., 2022). At level-3, our model performs at par with the best reported results from Costa and Kosseim (2024). It is worth noting that their model was specifically trained to predict level-3 senses, whereas our model is designed to generalize across all three sense levels simultaneously.

6.2 Transfer Learning Results on PDTB 3.0

The results of transfer learning from DiscoGeM to PDTB 3.0 on single-label IDRR are presented in Table 5. We compare our model against SOTA approaches in this task (Long and Webber, 2022; Liu and Strube, 2023; Zhao et al., 2023; Zeng et al., 2024; Long and Webber, 2024) and Yung et al. (2022). As shown in Table 5, our model has a lower performance than those trained on PDTB 3.0. However, it is important to emphasize that our model was trained exclusively on DiscoGeM and did not see any PDTB 3.0 data in its training. The lower performance in this zero-shot transfer learning setting, where no fine-tuning is done on the target corpus, can also be explained by thelimited agreement between the two corpora - as shown in Table 1, the majority sense in DiscoGeM only matched the PDTB 3.0 sense on 92 (30.5%) of the co-annotated discourse relations. This also explains the contrast in performances shown in Tables 4 and 5 between our model and the one from Yung et al. (2022), which was trained on the two corpora. Our model performs significantly better when evaluated on DiscoGeM, but worst when evaluated on the PDTB 3.0. These findings suggest that, to improve cross-corpus generalization, future work should explore intermediate fine-tuning on PDTB 3.0 after pretraining on DiscoGeM.

7 Analysis

To further evaluate the performance of our model, we conduct two additional analyses. In Section 7.1, we perform a per-sense evaluation to assess the ability of the model to correctly classify individual senses at both level-1 and level-2. In Section 7.2, we evaluate the cross-level coherence of the predictions by examining how well the predicted labels at level-1 and level-2 align with the sense hierarchy defined in the PDTB 3.0 framework. For ease of in-

Models	JS Distance ↘ (Multi-Label)			F1-Score ↗ (Single-Label)		
	Level-1	Level-2	Level-3	Level-1	Level-2	Level-3
Baseline	0.519 ± 0.002	0.636 ± 0.001	0.714 ± 0.002	34.72 ± 1.63	22.42 ± 1.34	18.81 ± 0.80
(Yung et al., 2022)	–	–	–	–	23.66 ± 1.19	–
(Costa and Kosseim, 2024)	–	–	–	–	–	$\mathbf{51.38 \pm n/a}$
Ours w/ MAE	$\mathbf{0.299 \pm 0.002}$	$\mathbf{0.446 \pm 0.003}$	$\mathbf{0.523 \pm 0.002}$	65.39 ± 0.54	50.13 ± 1.54	41.60 ± 0.59
Ours w/ CE	0.323 ± 0.003	0.564 ± 0.003	0.634 ± 0.002	$\mathbf{65.89 \pm 1.35}$	$\mathbf{55.99 \pm 1.73}$	$\mathbf{50.82 \pm 1.26}$

Table 4: Final results on the test set of DiscoGeM in multi-label classification (JS distance) and in single-label classification (weighted F1-score). The results were averaged across three different runs with random starts. Values in bold show the best score for each metric.

Models	F1-Score (PDTB 3.0)					
	Lin		Ji		Cross	
	Level-1	Level-2	Level-1	Level-2	Level-1	Level-2
Yung et al. (2022)	–	–	–	38.07 ± 2.25	–	–
Long and Webber (2022)	–	–	$70.05 \pm n/a$	$\mathbf{57.62 \pm n/a}$	–	–
Liu and Strube (2023)	–	–	71.15 ± 0.47	54.92 ± 0.81	$\mathbf{70.06 \pm 1.72}$	$\mathbf{55.26 \pm 1.32}$
Zhao et al. (2023)	–	–	$69.06 \pm n/a$	$52.73 \pm n/a$	–	–
Zeng et al. (2024)	–	–	$\mathbf{71.59 \pm n/a}$	$56.50 \pm n/a$	–	–
Long and Webber (2024)	–	–	$71.19 \pm n/a$	$52.91 \pm n/a$	–	–
Ours w/ MAE	47.17 ± 1.16	27.89 ± 1.21	43.88 ± 1.06	26.29 ± 1.71	45.53 ± 1.23	29.30 ± 1.49
Ours w/ CE	$\mathbf{50.44 \pm 0.81}$	$\mathbf{34.38 \pm 0.86}$	49.43 ± 1.47	33.11 ± 1.57	49.72 ± 1.45	33.21 ± 1.67

Table 5: Transfer learning results on different test splits of the PDTB 3.0 in single-label classification (weighted F1-score). The Lin and Ji results were averaged across three different runs with random starts, while the Cross results were averaged across all 12 folds. Values in bold show the best score for each level.

terpretation, we consider only single-label results.

7.1 Per-Sense Results

To identify disparities in performance between individual labels and to assess how well the model is able to generalize across senses, we evaluate the single-label predictions of our model at level-1 and level-2. Table 6 shows the weighted F1-score of each predicted sense in the test set of DiscoGeM and in the different test sets of the PDTB 3.0.

As shown in Table 6, the per-sense performance of the model on the DiscoGeM test set at both level-1 and level-2 generally reflects the distribution of senses in the corpus (see Table 8 in Appendix A). Note that some certain senses — CONDITION, PURPOSE, EQUIVALENCE and MANNER — never appear as the majority label of relation in DiscoGeM and are therefore absent from the test set (see Figure 2 and Table 9 in Appendix A). Unsurprisingly, the model struggles to predict under-represented senses such as SYNCHRONOUS, CONTRAST, SIMILARITY and SUBSTITUTION, which appear infrequently in the training data. Addressing this imbalance, potentially through targeted data augmentation, could improve generalization on these under-represented senses. The confusion matrices in Figures 3 and 4 in Appendix C provide further details into the per-sense performance on the test set of DiscoGeM. The performance of the model on the PDTB 3.0 test sets, also shown in Table 6, is consistently lower than on DiscoGeM, mirroring the trends reported in Table 4 and Table 5. This performance gap is particularly pronounced for certain senses, such as TEMPORAL at level-1 and ASYNCHRONOUS and INSTANTIATION at level-2. Conversely, the model performs better at the level-2 CONTRAST sense. Since the model was not trained on any PDTB 3.0 data, these discrepancies can be explained by annotation inconsistencies across both corpora as highlighted in Section 4.3.

7.2 Sense Coherence across Levels

Table 7 shows the percentage of times a sense at level-1 was predicted with a coherent sense at level-2 (and vice-versa) in the test set of DiscoGeM and in the different test sets of the PDTB 3.0. This enables us to examine the extent to which the predictions of the model are consistent with the hierarchical sense structure defined in the PDTB 3.0.

Level-1	DiscoGeM	PDTB 3.0			Level-2	DiscoGeM	PDTB 3.0		
	Test	Lin	Ji	Cross		Test	Lin	Ji	Cross
Temporal	61.27 ± 1.34	4.32 ± 1.58	15.75 ± 4.50	16.23 ± 4.86	Synchronous	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
					Asynchronous	57.87 ± 0.84	6.01 ± 2.05	22.02 ± 4.42	23.16 ± 6.65
Contingency	58.67 ± 1.83	36.12 ± 7.17	34.27 ± 8.24	41.08 ± 2.87	Cause	65.36 ± 0.12	47.21 ± 1.93	46.40 ± 2.47	48.93 ± 2.07
					Condition	−	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
					Purpose	−	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Comparison	39.31 ± 3.15	32.65 ± 1.69	37.67 ± 5.20	33.72 ± 2.34	Concession	36.24 ± 4.55	29.70 ± 5.09	27.92 ± 1.75	19.46 ± 3.70
					Contrast	4.13 ± 2.94	11.10 ± 7.92	12.70 ± 9.35	10.43 ± 6.61
					Similarity	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Expansion	75.39 ± 1.15	64.44 ± 0.66	62.69 ± 0.38	63.48 ± 1.16	Conjunction	52.56 ± 0.66	42.30 ± 1.26	39.91 ± 1.97	44.45 ± 3.04
					Equivalence	−	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
					Instantiation	44.84 ± 0.18	12.15 ± 9.12	17.90 ± 12.72	15.62 ± 10.49
					Level-of-Detail	46.85 ± 3.28	33.61 ± 2.90	29.32 ± 2.70	29.27 ± 3.06
					Manner	−	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
					Substitution	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 6: Individual per-sense results for each sense in level-1 and level-2 on the test set of DiscoGeM and on different test splits of the PDTB 3.0 in single-label classification (weighted F1-score). Senses marked with "−" were not present in the DiscoGeM test set. The DiscoGeM and the PDTB 3.0 Lin and Ji results were averaged across three different runs with random starts, while the PDTB 3.0 Cross results were averaged across all 12 folds.

Level-1	DiscoGeM	PDTB 3.0			Level-2	DiscoGeM	PDTB 3.0		
	Test (%)	Lin (%)	Ji (%)	Cross (%)		Test (%)	Lin (%)	Ji (%)	Cross (%)
Temporal	92.94 ± 2.93	93.27 ± 6.31	90.93 ± 4.51	92.48 ± 4.93	Synchronous	n/a	n/a	n/a	n/a
					Asynchronous	90.00 ± 1.44	82.74 ± 6.76	82.44 ± 11.34	88.30 ± 8.08
Contingency	98.91 ± 0.20	100.00 ± 0.00	99.74 ± 0.45	100.00 ± 0.00	Cause	62.37 ± 4.92	38.19 ± 6.45	38.42 ± 8.27	29.93 ± 1.32
					Condition	−	n/a	n/a	n/a
					Purpose	−	n/a	n/a	n/a
Comparison	86.31 ± 1.22	73.81 ± 13.21	68.09 ± 10.47	58.45 ± 4.12	Concession	81.19 ± 2.83	89.15 ± 8.10	89.36 ± 8.98	97.93 ± 1.52
					Contrast	91.67 ± 14.43	90.77 ± 10.09	90.77 ± 11.11	92.68 ± 8.19
					Similarity	n/a	n/a	n/a	n/a
Expansion	70.70 ± 1.26	37.91 ± 10.82	65.02 ± 12.30	53.01 ± 2.21	Conjunction	97.61 ± 0.71	97.14 ± 0.03	3.20 ± 0.88	96.69 ± 0.96
					Equivalence	−	n/a	n/a	n/a
					Instantiation	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
					Level-of-Detail	99.82 ± 0.31	99.80 ± 0.35	100.00 ± 0.00	100.00 ± 0.00
					Manner	−	n/a	n/a	n/a
					Substitution	n/a	n/a	n/a	n/a

Table 7: Percentage of the instances that were classified with coherent senses at level-1 and level-2 on the test set of DiscoGeM and on different test splits of the PDTB 3.0 in single-label classification. Senses marked with "−" were not present in the DiscoGeM test set and senses marked with "n/a" were never predicted by the model. The DiscoGeM and the PDTB 3.0 Lin and Ji results were averaged across three different runs with random starts, while the PDTB 3.0 Cross results were averaged across all 12 folds.

As Table 7 shows, the Contingency sense at level-1 is very often predicted with a coherent level-2 sense (i.e., Cause, Condition or Purpose). Similarly, the level-2 senses Instantiation, Level-of-Detail and Conjunction are often predicted with a coherent level-1 sense. However, other senses, such as Comparison and Expansion at level-1 and Cause at level-2, are more often predicted with a contradicting sense. A possible explanation might be that these senses are inherently hard to distinguish and the model is biased towards the most represented sense. For instance, the level-2 senses Cause and Asynchronous are among the most often co-annotated pair of senses in the DiscoGeM and Cause appears approximately three times more often than Asynchronous (see Table 8 in Appendix A), the model is likely more biased towards Cause when distinguishing between these two senses. One possible method to generate more coherent predictions across sense levels, would be to share the predictions of lower-level senses to help inform the prediction of higher-level senses within the model.

8 Conclusion

In this work, we proposed a novel multi-label framework for IDRR, addressing one of the most complex challenges in discourse analysis. We trained a multi-task classification model on the DiscoGeM corpus to simultaneously learn multi-label representations of discourse relations across all three sense levels in the PDTB 3.0 framework.

Our model can also be adapted to the traditional single-label IDRR setting by selecting the sense with the highest probability in the multi-label representation. We conducted extensive experiments to identify optimal model configurations and loss functions in both settings. Our approach establishes a first benchmark on multi-label IDRR and achieves SOTA results in single-label IDRR using DiscoGeM. Finally, we evaluated our model on the PDTB 3.0 corpus in the single-label setting, presenting the first analysis of transfer learning between the DiscoGeM and PDTB 3.0 corpora for IDRR. Our results show that zero-shot direct transfer learning between both corpora still needs further research.

Future work should explore the application of data augmentation techniques, such as paraphrasing, to determine whether augmenting underrepresented senses in the DiscoGeM could enhance the performance of the model on individual senses. Additionally, future research should investigate whether cascading information from higher-level to lower-level classification heads within the model could improve the coherence of its predictions across sense levels. Lastly, it would be worthwhile to examine whether a model trained to learn multi-label representations of discourse relations in DiscoGeM could be further fine-tuned on the PDTB 3.0 to achieve superior performance in single-label IDRR using the PDTB 3.0 corpus.

9 Limitations

Despite the promising results, there are a few limitations to our work. The first limitation of our work comes from the fact that the CAUSE+BELIEF sense was not annotated in the DiscoGeM corpus. Therefore, we could not use the standard 14-label set of second-level senses proposed for the PDTB 3.0 by Kim et al. (2020) and widely used in literature. Instead, we replaced it by the SIMILARITY sense - the next most available sense in the DiscoGeM corpus. Ideally, we would have replaced the CAUSE+BELIEF sense with another sense under the CONTINGENCY level-1 sense, but they were all already considered.

Our choice of pre-trained language model to generate the embeddings for each pair of discourse arguments also imposes a limitation to our work. Due to computational and time limitations, we could not explore the fine-tuning of larger models, such as LLaMA 3 (Grattafiori et al., 2024), nor explore prompting LLMs. However, in our other work (Costa and Kosseim, 2025), we show that the use of LLMs via direct prompting with few-shot learning for IDRR does not lead to better results. Finally, we would like to acknowledge that the extensive experimental analysis conducted in this work would not have been possible without access to a high-performance computing facility - which entails a non-negligible carbon footprint which we did not monitor during any of our experiments.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics (ACL).

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 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL'24)*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics (ACL).

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023. DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 35–57, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

Nelson Filipe Costa and Leila Kosseim. 2024. Exploring Soft-Label Training for Implicit Discourse Relation Recognition. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI'24)*, pages 120–126, St. Julians, Malta. Association for Computational Linguistics (ACL).

Nelson Filipe Costa and Leila Kosseim. 2025. Multi-Lingual Implicit Discourse Relation Recognition with Multi-Label Hierarchical Learning. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'25)*, Avignon, France. Association for Computational Linguistics (ACL).

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP'23)*, pages 344–352, Varna, Bulgaria.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics (ACL).

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*, pages 2591–2597, Online. Association for Computational Linguistics (ACL).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

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 1st Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Online. Association for Computational Linguistics (ACL).

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 (TACL)*, 3:329–344.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1357–1374.

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 (ACL'20)*, pages 5404–5414, Online. Association for Computational Linguistics (ACL).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, pages 1–15, San Diego, California, USA.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024. Dialogue Discourse Parsing as Generation: A Sequence-to-Sequence LLM-based Approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'24)*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics (ACL).

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, pages 343–351, Singapore. Association for Computational Linguistics (ACL).

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 (ACL'23)*, pages 15696–15712, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Wanqiu Long, N Siddharth, and Bonnie Webber. 2024. Multi-Label Classification for Implicit Discourse Relation Recognition. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL'24)*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics (ACL).

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 (EMNLP'22)*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

Wanqiu Long and Bonnie Webber. 2024. Leveraging Hierarchical Prototypes as the Verbalizer for Implicit Discourse Relation Recognition. *arXiv preprint arXiv:2411.14880*.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In

Proceedings of the 5th International Conference on Learning Representations (ICLR'17), pages 1–16, Toulon, France.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7:677–694.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008a. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie Webber. 2008b. Penn Discourse Treebank Version 2.0. LDC2008T05. Web Download. Philadelphia: Linguistic Data Consortium.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05. Web Download. Philadelphia: Linguistic Data Consortium.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 2804–2819, 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 (TACL)*, 11:1014–1032.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21(140):1–67.

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 (LAW'16)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Merel Scholman and Vera Demberg. 2017. Examples and Specifications that Prove a Point: Identifying Elaborative and Argumentative Discourse Relations. *Dialogue & Discourse*, 8(2):56–83.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022a. DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 3281–3290, Marseille, France. European Language Resources Association (ELRA).

Merel Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty, and Vera Demberg. 2022b. Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC'22)*, pages 2148–2156, Marseille, France. European Language Resources Association (ELRA).

Wei Shi and Vera Demberg. 2017. On the Need of Cross Validation for Discourse Relation Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*, pages 150–156, Valencia, Spain. Association for Computational Linguistics (ACL).

Wei Shi and Vera Demberg. 2019. Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*, pages 5790–5796, Hong Kong, China. Association for Computational Linguistics (ACL).

Manfred Stede. 2008. Disambiguating Rhetorical Structure. *Research on Language and Computation*, 6(3):311–332.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An Incremental

Discourse Parser. In *Findings of 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics (ACL).

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72(1):1385–1470.

Michiel van der Meer, Neele Falk, Pradeep K Murukannaiah, and Enrico Liscio. 2024. Annotator-Centric Active Learning for Subjective NLP Tasks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 18537–18555, Miami, Florida, USA. Association for Computational Linguistics (ACL).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. Technical report, University of Pennsylvania.

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. Julian's, Malta. Association for Computational Linguistics (ACL).

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 (CODI'22)*, pages 48–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics (ICCL).

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 for Computational Linguistics (ICCL).

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 (LAW'19)*, pages 16–25, Florence, Italy. Association for Computational Linguistics (ACL).

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'24)*, pages 7760–7773, Torino, Italia. European Language Resources Association (ELRA) and International Committee for Computational Linguistics (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 (ACL'23)*, pages 6477–6492, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

A Data Statistics

To standardize and ensure a fair comparison of results on IDRR, Kim et al. (2020) proposed a set of 14 level-2 senses in the PDTB 3.0 framework. However, the CAUSE+BELIEF sense in this set was not annotated in DiscoGeM. Therefore, we adapted the standard set to accommodate the existing level-2 senses in the corpus by replacing CAUSE+BELIEF with SIMILARITY. We then removed all senses not included in the standard set from the corpus and normalized the distribution values over the remaining senses following the L1 norm. This ensures the sum of each distribution adds to 1 for each instance while preserving the relative distance between senses. Table 8 shows the distribution of senses across all levels in the original DiscoGeM corpus and in our adapted set of 14 level-2 senses. Each value represents the sum of the corresponding sense across all instances. We then split 70% of the corpus for training, 10% for validation and 20% for testing. To ensure a balanced distribution, we kept the same distribution of majority labels across all data splits as shown in Figure 2 in Section 4.1. In the single-label setting, we replaced the full multi-label sense distribution of each discourse relation by its majority sense. Table 9 shows the level-2 majority senses in the test set of DiscoGeM and reference senses in the different test splits of the PDTB 3.0 (see Section 4.2).

B Loss Functions

We evaluated the performance of different loss functions on multi- and single-label IDRR. Considering $\hat{y}^h$ as the predicted distribution of the classification head h (with $h \in \{1, 2, 3\}$), y^h as the corresponding target distribution, C^h the number of senses at the sense level-h and N the number of instances in each batch, we evaluated the following loss functions:

- Cross-Entropy (CE)

$$loss^h(\hat{y}^h, y^h) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C^h} log\left(\frac{e^{\hat{y}^h_{n,c}}}{\sum_{i=1}^{C^h} e^{\hat{y}^h_{n,i}}}\right) y^h_{n,c} \quad (1)$$

Level-1	Sum		Level-2	Sum		Level-3	Sum	
	Original	Adapted		Original	Adapted		Original	Adapted
Temporal	584.3	619.5	Synchronous	95.8	102.7	-	-	-
			Asynchronous	488.5	516.8	Precedence	448.1	474.2
						Succession	40.5	42.7
Contingency	1,745.1	1,822.9	Cause	1,740.2	1,819.0	Reason	382.5	400.1
						Result	1,357.7	1,418.9
						NegResult	0.0	0.0
			Condition	1.2	1.2	ARG1-as-Cond	0.1	0.1
						ARG2-as-Cond	1.1	1.1
			Neg-Condition	1.1	0.0	ARG1-as-NegCond	1.0	0.0
						ARG2-as-NegCond	0.1	0.0
			Purpose	2.6	2.7	ARG1-as-Goal	1.7	1.7
						ARG2-as-Goal	0.9	0.9
Comparison	831.5	878.0	Concession	517.1	548.7	ARG1-as-Denier	169.3	179.9
						ARG2-as-Denier	347.7	368.8
			Contrast	202.4	213.4	-	-	-
			Similarity	112.0	116.0	-	-	-
Expansion	3,034.6	3,184.6	Conjunction	1,441.1	1,518.0	-	-	-
			Disjunction	2.9	0.0	-	-	-
			Equivalence	19.0	19.8	-	-	-
			Exception	1.4	0.0	ARG1-as-Exception	0.3	0.0
						ARG2-as-Exception	1.3	0.0
			Instantiation	388.8	403.8	ARG1-as-Instance	16.9	17.8
						ARG2-as-Instance	371.9	386.0
			Level-of-Detail	1,137.0	1,196.1	ARG1-as-Detail	160.8	170.2
						ARG2-as-Detail	976.2	1,025.9
			Manner	4.6	4.8	ARG1-as-Manner	1.3	1.3
						ARG2-as-Manner	3.3	3.4
			Substitution	39.7	42.1	ARG1-as-Substitution	0.0	0.0
						ARG2-as-Substitution	39.7	42.1

Table 8: Distribution of senses across all levels in the original DiscoGeM and in our adapted set of 14 level-2 senses (see Section 4.1). Each value represents the sum of the corresponding sense across all instances.

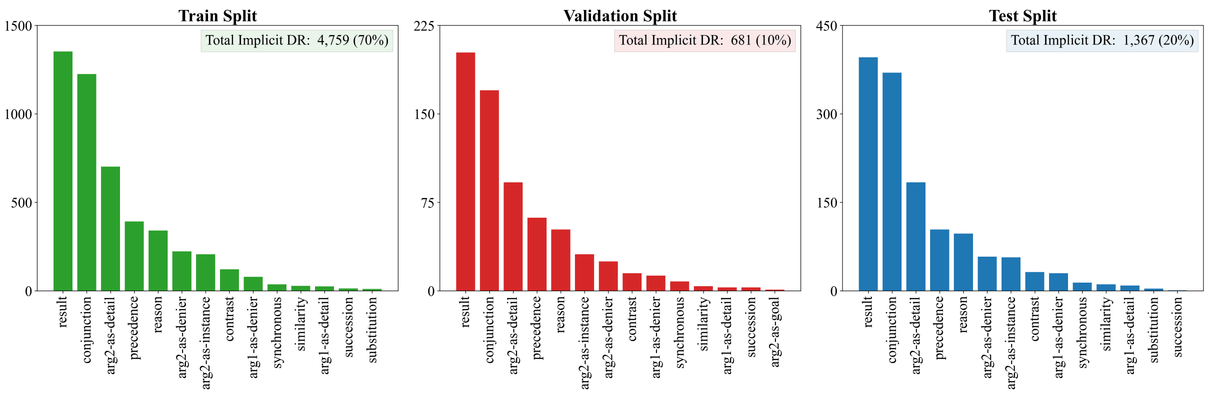

Figure 2: Distribution of majority label senses across all splits of the DiscoGeM corpus (see Section 4.1).

Level-1	Level-2	DiscoGeM	PDTB 3.0		
		Test	Lin	Ji	Cross
TEMPORAL	SYNCHRONOUS	13	19	43	42.85 ± 11.94
	ASYNCHRONOUS	103	33	105	99.54 ± 24.95
CONTINGENCY	CAUSE	515	271	406	465.38 ± 77.28
	CONDITION	0	10	15	16.31 ± 3.61
	PURPOSE	0	46	89	110.08 ± 22.60
COMPARISON	CONCESSION	104	83	98	121.23 ± 22.87
	CONTRAST	27	52	54	69.92 ± 18.99
	SIMILARITY	10	1	2	2.23 ± 1.48
EXPANSION	CONJUNCTION	336	161	236	349.38 ± 86.25
	EQUIVALENCE	0	24	30	27.54 ± 10.27
	INSTANTIATION	53	70	124	118.62 ± 24.97
	LEVEL-OF-DETAIL	202	194	208	263.69 ± 26.94
	MANNER	0	4	17	17.92 ± 6.49
	SUBSTITUTION	4	27	26	30.23 ± 8.28
Total		$1,367$	995	$1,453$	$1,734.92 \pm 271.89$

Table 9: Distribution of level-2 majority label senses across the test set of DiscoGeM (see Section 4.1) and the different test splits of the PDTB 3.0 (see Section 4.2). The values in PDTB 3.0 Cross correspond to the averaged distribution of majority label senses across all 12 test folds.

- Mean Absolute Error (MAE)

$$loss^h(\hat{y}^h, y^h) = \frac{1}{NC^h} \sum_{n=1}^{N} \sum_{i=1}^{C^h} |\hat{y}^h_{n,i} - y^h_{n,i}| \quad (2)$$

- Mean Squared Error (MSE)

$$loss^h(\hat{y}^h, y^h) = \frac{1}{NC^h} \sum_{n=1}^{N} \sum_{i=1}^{C^h} (\hat{y}^h_{n,i} - y^h_{n,i})^2 \quad (3)$$

- Huber Loss (Huber)

$$loss^k(\hat{y}^h, y^h) = \sum_{n=1}^{N} \sum_{i=1}^{C^h} \begin{cases} \frac{{\delta^h_{n,i}}^2}{2NC^h} & \text{, if } |\delta^h_{n,i}| < 1 \\ \frac{2|\delta^h_{n,i}|-1}{2NC^h} & \text{, otherwise} \end{cases} \quad (4)$$

with

$$\delta^h_{n,i} = \hat{y}^h_{n,i} - y^h_{n,i} \quad (5)$$

The MAE and MSE losses in Equations 2 and 3, respectively, aim to minimize the overall difference between the predicted and target distributions by penalizing errors across all possible labels. MAE minimizes the average absolute differences, leading to predictions that are closer to the target distribution in an averaged sense, while MSE places a larger penalty on larger errors due to its quadratic nature, which can result in a stronger emphasis on outliers. The Huber loss in Equation 4 combines the properties of MAE and MSE in function of the delta parameter defined in Equation 5. In other words, the Huber loss behaves like MAE for smaller errors and like MSE for larger errors. The results of experimenting with the MAE, MSE and Huber losses were therefore relatively similar. However, the MAE was slightly better and thus better suited for multi-label IDRR as shown in Table 2 in Section 5.1. Table 10 shows the results of experimenting with the MSE and Huber loss functions in multi-label IDRR (JS distance) and in single-label IDRR (weighted F1-score). Conversely, the CE loss in Equation 1 focuses on maximizing the probability of the label with the highest score in the target distribution which makes it better suited for single-label IDRR as shown in Table 2 in Section 5.1.

Model	Loss	JS Distance ↘ (Multi-Label)			F1-Score ↗ (Single-Label)		
		Level-1	Level-2	Level-3	Level-1	Level-2	Level-3
BERT	MSE	0.319 ± 0.005	0.462 ± 0.002	0.544 ± 0.005	63.82 ± 0.83	48.68 ± 0.16	41.36 ± 1.55
	Huber	0.318 ± 0.007	0.463 ± 0.001	0.542 ± 0.002	64.36 ± 0.48	49.45 ± 0.96	41.58 ± 1.08
DistilBERT	MSE	0.336 ± 0.007	0.481 ± 0.003	0.558 ± 0.003	60.10 ± 0.96	43.77 ± 0.86	37.66 ± 1.14
	Huber	0.337 ± 0.002	0.479 ± 0.003	0.557 ± 0.006	60.41 ± 0.86	44.06 ± 0.64	37.29 ± 0.28
RoBERTa	MSE	0.308 ± 0.012	0.455 ± 0.001	$\mathbf{0.534 \pm 0.005}$	$\mathbf{65.28 \pm 0.36}$	$\mathbf{53.94 \pm 1.80}$	$\mathbf{44.91 \pm 1.31}$
	Huber	$\mathbf{0.305 \pm 0.007}$	$\mathbf{0.450 \pm 0.004}$	0.539 ± 0.003	65.13 ± 0.60	53.39 ± 0.19	44.33 ± 0.64
DistilRoBERTa	MSE	0.313 ± 0.004	0.470 ± 0.005	0.555 ± 0.003	64.32 ± 1.99	50.48 ± 0.29	42.43 ± 0.22
	Huber	0.318 ± 0.004	0.472 ± 0.001	0.561 ± 0.004	64.11 ± 0.72	50.46 ± 0.73	42.41 ± 0.30

Table 10: Results of experimenting with different pre-trained language models and different loss functions in multi-label classification (JS distance) and in single-label classification (weighted F1-score). The results were averaged across three different runs with random starts. Values in bold show the best score for each metric.

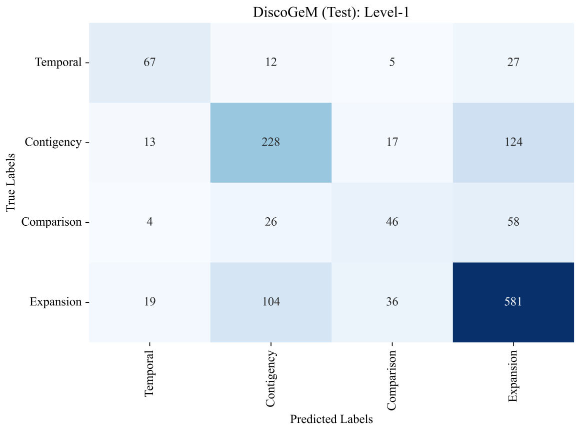

Figure 3: Confusion matrix for the individual per-class results of each sense in level-1 on the test set of DiscoGeM.

C Confusion Matrices

To provide further details into the per-sense performance on the test set of DiscoGeM in Section 7.1, we generated a confusion matrix at level-1 and level-2 for the results of a single run on the test set of DiscoGeM as shown in Figures 3 and 4, respectively. The confusion matrix in Figure 3 shows that, with the exception with COMPARISON, the most predicted label always aligns with the correct label. Nevertheless, the confusion matrix also shows that the model is more biased towards the most represented senses level-1 senses EXPANSION and CONTINGENCY (see Table 8 in Appendix A). At the level-2, the confusion matrix in Table 4 shows that the model is not able to predict less represented senses in DiscoGeM (see Table 8 in Appendix A), such as SYNCHRONOUS, SIMILARITY and SUBSTITUTION. However, for the other senses, the most predicted label often aligns with the correct label.

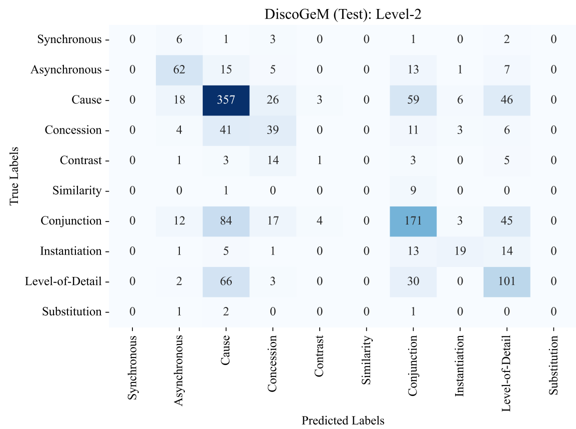

Figure 4: Confusion matrix for the individual per-class results of each sense in level-2 on the test set of DiscoGeM.