An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition

Kazumasa Omura[1]*, **Fei Cheng**[1], **Sadao Kurohashi**[1,2]
[1]Graduate School of Informatics, Kyoto University, Japan
[2]National Institute of Informatics, Japan
{omura, feicheng, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

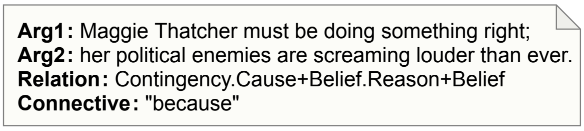
Implicit Discourse Relation Recognition (IDRR), which is the task of recognizing the semantic relation between given text spans that do not contain overt clues, is a long-standing and challenging problem. In particular, the paucity of training data for some error-prone discourse relations makes the problem even more challenging. To address this issue, we propose a method of generating synthetic data for IDRR using a large language model. The proposed method is summarized as two folds: extraction of confusing discourse relation pairs based on false negative rate and synthesis of data focused on the confusion. The key points of our proposed method are utilizing a confusion matrix and adopting two-stage prompting to obtain effective synthetic data. According to the proposed method, we generated synthetic data several times larger than training examples for some error-prone discourse relations and incorporated it into training. As a result of experiments, we achieved state-of-the-art macro-F1 performance thanks to the synthetic data without sacrificing micro-F1 performance and demonstrated its positive effects especially on recognizing some infrequent discourse relations.

Keywords: implicit discourse relation recognition, synthetic data generation, confusion matrix

1. Introduction

In order to comprehend the meaning of natural language text, it is essential to understand not only the meanings of individual sentences but also the semantic relations between them. Such semantic relations are called *discourse relations*. Automatic recognition of discourse relations has been actively studied due to its applicability to natural language understanding (Omura and Kurohashi, 2022; Bhargava and Ng, 2022) and various natural language processing (NLP) tasks (Saito et al., 2019; Tang et al., 2021).

Penn Discourse Treebank (PDTB) (Prasad et al., 2008) is one of the representative corpora regarding discourse relations. It has been built by annotating the Wall Street Journal with discourse relations between adjacent text spans named *arguments*. An example is shown in Figure 1; (hereafter, we express an argument pair as *Arg1* and *Arg2*.) The arguments in the example do not contain any discourse connectives.[1] Such examples are called *implicit* discourse relations.



Arg1: Maggie Thatcher must be doing something right;
Arg2: her political enemies are screaming louder than ever.
Relation: Contingency.Cause+Belief.Reason+Belief
Connective: "because"

Figure 1: Example from PDTB. PDTB defines at most three levels of hierarchical discourse relations. In the example, *Relation* is delimited by periods, and top-, second-, and third-level relations are "Contingency", "Cause+Belief", and "Reason+Belief", respectively. Note that the higher the level, the coarser the granularity. In addition, some discourse connectives are assigned to lexicalize the relations. Regarding implicit discourse relations, the annotated connectives are not present in arguments.

Implicit Discourse Relation Recognition (IDRR) is a long-standing and challenging problem. Even large language models (LLMs), which have achieved unprecedented performance on a variety of NLP tasks, still cannot solve the problem in a straightforward manner.[2] In addition to the complexity of IDRR itself, i.e., requiring deep reasoning due to the lack of overt clues, the paucity of training data for some error-prone discourse relations makes the problem even more challenging.

A naive solution to the aforementioned problem is to increase the number of annotated examples. However, it is not practical due to requiring cautious annotations by experts. Turning our attention to automatic generation of training data, synthetic data generation using language models has achieved

*Current affiliation is Nikkei Inc.

[1] A word or phrase that indicates certain discourse relation such as "and", "but", "for example", and so forth.

[2] We investigated the few-shot performance of GPT-3.5 and GPT-4 in IDRR and confirmed that it is far behind the fine-tuning performance of much smaller pre-trained language models, which is described in Section 3.2.

some success recently (Puri et al., 2020; Yang et al., 2020; Schick and Schütze, 2021; Liu et al., 2022a). There is room for exploration of their generative capabilities to generate argument pairs that have a given discourse relation, although the low few-shot performance of LLMs in IDRR is problematic.

In this study, we explore synthetic data generation for IDRR using an LLM. We first conduct preliminary experiments to confirm the paucity of training data for some error-prone discourse relations. Based on the preliminary results, we propose a method of generating synthetic data for these error-prone discourse relations using an LLM. Specifically, it is summarized as two folds: extraction of confusing discourse relation pairs based on false negative rate and synthesis of data focused on the confusion. We demonstrate the performance gain by incorporating the synthetic data into training.

The proposed method has two key points. First, we utilize a confusion matrix to obtain effective synthetic data. We address the data scarcity problem of some error-prone discourse relations by generating synthetic data based on a confusion matrix.

Second, we devise a method of generating synthetic data. It is probably ineffective to straightforwardly generate synthetic data for IDRR using an LLM due to the low few-shot performance. We presume that it is attributed to the number of discourse relations. In other words, it is challenging for an LLM to learn and distinguish numerous discourse relations from few-shot examples. On the other hand, it is relatively easy to learn a single discourse relation from few-shot examples. Thus, we decompose the process of generating synthetic data into two stages so that an LLM is required to learn only a single discourse relation in each stage. Further details are described in Section 4.1.

The contributions of this study are summarized as follows:

- We propose an error-driven method of generating synthetic data for IDRR using an LLM.

- According to the proposed method, we generated synthetic data several times larger than training examples for some error-prone discourse relations.

- Thanks to the synthetic data, we achieved state-of-the-art (SOTA) macro-F1 performance without sacrificing micro-F1 performance and demonstrated its positive effects especially on recognizing some infrequent discourse relations.[3]

2. Related Work

2.1. Improving IDRR

As seen in Figure 1, PDTB has two major characteristics: discourse relations are defined hierarchically and lexicalized by discourse connectives. Many previous studies on improving IDRR have exploited these characteristics.

Utilizing Relation Hierarchy This kind of approach has been on the rise recently. For instance, Long and Webber (2022) introduced contrastive learning and utilized the relation hierarchy to select hard negatives, assuming it is difficult to classify discourse relations that have the same higher-level ones. However, we demonstrate an encoder-only language model such as RoBERTa (Liu et al., 2019) is apt to confuse infrequent discourse relations with frequent ones rather than misclassify discourse relations that have the same higher-level ones (cf. Section 3.3). Jiang et al. (2023) also developed the contrastive framework to learn the relation hierarchy and similarity between examples simultaneously, but the same can be pointed out. Wu et al. (2022) showed the effectiveness of learning to generate labels along the relation hierarchy. This method may suffer error propagation from mispredicted top-level discourse relations.

Utilizing Discourse Connectives Several studies have been devoted to learning implicit discourse relations through discourse connectives for some time. For instance, Nie et al. (2019) and Kishimoto et al. (2020) have reported the performance gain by performing an additional pre-training task to predict masked discourse connectives. Other studies such as Xiang et al. (2022) and Zhou et al. (2022) introduced prompt-based learning and utilized annotated discourse connectives as verbalizers. As implicit discourse relations are mentioned without discourse connectives, it is also worth considering approaches not relying on discourse connectives.

Other Approaches Xu et al. (2018) introduced active learning to obtain argument pairs that contain omittable discourse connectives (Rutherford and Xue, 2015) for data augmentation. Jiang et al. (2021) performed joint learning of classification and generation, aiming to deepen the model's understanding of discourse relations through generating arguments. To the best of our knowledge, no studies have been conducted on synthetic data generation for IDRR using an LLM.

[3] The synthetic data and code are available at https://github.com/ku-nlp/sdg4idrr.

2.2. Synthetic Data Generation for NLP tasks

After the advent of pre-trained language models, an increasing number of studies have attempted to utilize them for synthetic data generation. For instance, Schick and Schütze (2021) synthesized 121k sentence pairs for Semantic Textual Similarity using GPT-2 XL (Radford et al., 2019) and achieved superior performance with the synthetic data only. Liu et al. (2022a) incorporated human-in-the-loop into synthetic data generation for Natural Language Inference and built a dataset consisting of 108k examples using GPT-3 (Brown et al., 2020). In addition, synthetic data generation has been conducted for other NLP tasks including Question Answering (Puri et al., 2020), Commonsense Reasoning (Yang et al., 2020), and so forth. While recent studies lean toward improving few-shot performance with synthetic data (Meng et al., 2023; Dai et al., 2023), we aim to improve fine-tuning performance of encoder-only language models in IDRR considering the relatively low few-shot performance of LLMs.

3. Preliminaries

The proposed method is motivated by preliminary experimental results. We first report them.

3.1. Task Settings

As there are several variations of preprocessing and evaluation protocols regarding PDTB (Kim et al., 2020), we explicate task settings used in our experiments (Section 3.2, 3.3, and 5).

Version of PDTB PDTB has been updated several times over the years. While the previous version (PDTB-2) (Prasad et al., 2008) has been conventionally used so far, the latest version (PDTB-3) (Prasad et al., 2019) has improved in both quantity and quality of annotations. We adopt PDTB-3 taking into account that more annotated examples are available for generating synthetic data.

Label Set Label sets vary by the version of PDTB and the level of relations to classify. We address the fine-grained classification of second-level relations and follow Kim et al. (2020) to define a label set for the task. Specifically, we formulate IDRR as 14-way classification using only the labels with more than 100 examples.

Data Partitioning PDTB consists of 25 sections, and we need to partition them to build a dataset. For a fair comparison with previous studies, we adopt the conventional partition introduced by Ji and Eisenstein (2015), where we use sections 2-20, 0-1, and 21-22 as training, development, and test splits, respectively. For convenience, we call it *PDTB dataset*. The statistics of the PDTB dataset are organized in Table 1.

Relation	Train	Synthetic Data		Dev	Test
		Unfiltered	Filtered		
Sync.	435	2,501	1,286	33	43
Async.	1,007	-	-	105	108
Cause	4,475	-	-	449	406
Cause+B.	159	940	331	13	15
Purp.	1,092	-	-	96	89
Cond.	150	-	-	18	15
Conc.	1,164	-	-	105	97
Cont.	741	-	-	91	63
Conj.	3,586	-	-	299	237
Equiv.	254	1,167	771	25	30
Inst.	1,166	-	-	118	128
Level.	2,601	-	-	274	214
Manner	615	-	-	28	53
Sub.	343	-	-	32	32

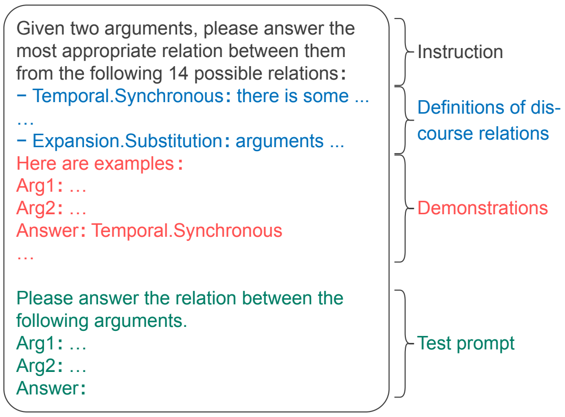
Table 1: Statistics of the PDTB dataset and synthetic data. Regarding multi-labeled examples, we counted the labels separately. For space limitation, we abbreviate the name of each discourse relation. As synthetic data may vary by model, we show the statistics of the synthetic data generated from the confusion matrix in Figure 3 as a representative. The statistics of synthetic data for RoBERTa$_{\text{LARGE}}$ are included in Appendix A.2.

Handling of Multi-labeled Examples Regarding multi-labeled examples, we follow a common practice (Ji and Eisenstein, 2015; Qin et al., 2017). Specifically, during the training phase, we convert them into separate examples. During the evaluation phase, a prediction is regarded as correct if it matches one of the labels.[4]

3.2. Few-shot Performance of LLMs

Few studies attempted to employ LLMs for IDRR except Chan et al. (2023), which investigated the zero-shot performance of GPT-3.5 on PDTB-2. We also investigated the few-shot performance of GPT-3.5 and GPT-4 (OpenAI, 2023) on PDTB-3.

[4] We found there are two implementations of this. Let us consider the case where a model predicts "A" to an example with the labels "A" and "B". One implementation overwrites the prediction with "A" and "B", while the other discards the label "B" of the example. This may cause discrepancies in the total number of true positives and labels among studies. In this study, we confirmed the implementation in a compared method and adopted the former implementation.



Given two arguments, please answer the most appropriate relation between them from the following 14 possible relations:
- Temporal.Synchronous: there is some ...
...
- Expansion.Substitution: arguments ...
Here are examples:
Arg1: ...
Arg2: ...
Answer: Temporal.Synchronous
...

Please answer the relation between the following arguments.
Arg1: ...
Arg2: ...
Answer:

Figure 2: Prompt template for few-shot learning in second-level relation classification on PDTB-3.

Setting	Micro-F1	Macro-F1
GPT-3.5 few-shot	23.2	19.0
GPT-4 few-shot	29.4	30.9
RoBERTa$_{\text{BASE}}$ (Vanilla)	64.2	57.1

Table 2: Experimental results of few-shot learning in second-level relation classification on PDTB-3. The vanilla fine-tuning performance of RoBERTa$_{\text{BASE}}$ is taken from Table 4.

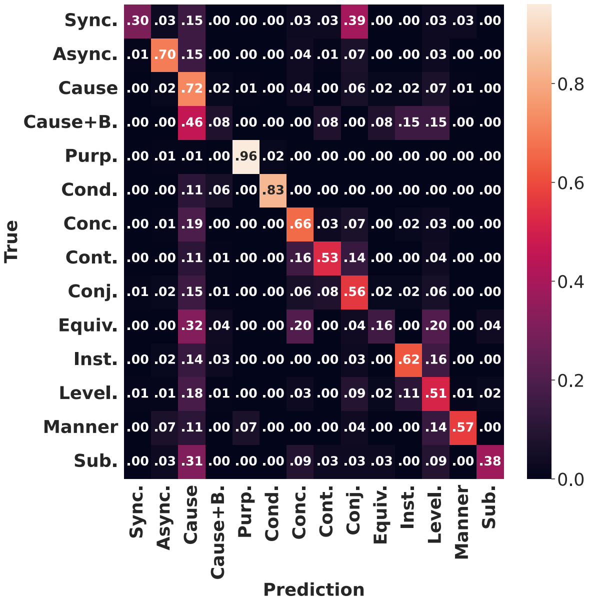
3.2.1. Experimental Settings

As mentioned in Section 3.1, we address the 14-way classification of second-level relations. Figure 2 illustrates the prompt template for few-shot learning on the task. We instructed LLMs to generate one of the labels given definitions of discourse relations and demonstrations.

For the LLMs, we employed the snapshots of GPT-3.5 and GPT-4 from June 13th, 2023 (a.k.a "gpt-3.5-turbo-16k-0613" and "gpt-4-0613"). We retrieved K nearest neighbors of a test example from training examples for each discourse relation and used the $K \times 14$ examples as demonstrations referring to Liu et al. (2022b). We made use of the RoBERTa$_{\text{LARGE}}$-based supervised SimCSE[5] (Gao et al., 2021) for retrieving nearest neighbors and set K to 8 considering the token limit of the LLMs. We used the test split of the PDTB dataset for evaluation and evaluated the model by micro-F1 and macro-F1.

3.2.2. Results

Table 2 shows the few-shot performance of GPT-3.5 and GPT-4. Despite providing more than 100 examples as demonstrations, the few-shot performance is far behind the fine-tuning performance of the RoBERTa$_{\text{BASE}}$ model.



	Sync.	Async.	Cause	Cause+B.	Purp.	Cond.	Conc.	Cont.	Conj.	Equiv.	Inst.	Level.	Manner	Sub.
Sync.	.30	.03	.15	.00	.00	.00	.03	.03	.39	.00	.00	.03	.03	.00
Async.	.01	.70	.15	.00	.00	.00	.04	.01	.07	.00	.00	.03	.00	.00
Cause	.00	.02	.72	.02	.01	.00	.04	.00	.06	.02	.02	.07	.01	.00
Cause+B.	.00	.00	.46	.08	.00	.00	.00	.08	.00	.08	.15	.15	.00	.00
Purp.	.00	.01	.01	.00	.96	.02	.00	.00	.00	.00	.00	.00	.00	.00
Cond.	.00	.00	.11	.06	.00	.83	.00	.00	.00	.00	.00	.00	.00	.00
Conc.	.00	.01	.19	.00	.00	.00	.66	.03	.07	.00	.02	.03	.00	.00
Cont.	.00	.00	.11	.01	.00	.00	.16	.53	.14	.00	.00	.04	.00	.00
Conj.	.01	.02	.15	.01	.00	.00	.06	.08	.56	.02	.02	.06	.00	.00
Equiv.	.00	.00	.32	.04	.00	.00	.20	.00	.04	.16	.00	.20	.00	.04
Inst.	.00	.02	.14	.03	.00	.00	.00	.00	.03	.00	.62	.16	.00	.00
Level.	.01	.01	.18	.01	.00	.00	.03	.00	.09	.02	.11	.51	.01	.02
Manner	.00	.07	.11	.00	.07	.00	.00	.00	.04	.00	.00	.14	.57	.00
Sub.	.00	.03	.31	.00	.00	.00	.09	.03	.03	.03	.00	.09	.00	.38

Figure 3: Normalized confusion matrix of the RoBERTa$_{\text{BASE}}$ model. We applied row normalization to the confusion matrix so that each element represents sensitivity or false negative rate.

3.3. Confusion Matrix of Encoder Model

In order to identify the propensity for error in a commonly used model, we analyzed the confusion matrix.

3.3.1. Experimental Settings

We investigated the confusion matrix of the RoBERTa$_{\text{BASE}}$ model, which has been employed in most recent studies. We fine-tuned the RoBERTa$_{\text{BASE}}$ pre-trained model[6] on the PDTB dataset and computed a confusion matrix on the development split. Training details and hyper-parameters are described later in Section 5.1.2 and 5.1.4, respectively.

3.3.2. Results

We define the degree of confusion by false negative rate considering the class imbalance as seen in Table 1. Figure 3 illustrates the normalized confusion matrix of the RoBERTa$_{\text{BASE}}$ model. Several non-diagonal elements indicate high degree of confusion, i.e., much room for improvement. Furthermore, it is observable from Table 3 that RoBERTa$_{\text{BASE}}$ is apt to confuse infrequent discourse relations such as "Cause+Belief" and

[5] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

[6] https://huggingface.co/roberta-base

Ground Truth	Prediction
Contingency.Cause+B	Contingency.Cause
Temporal.Sync.	Expansion.Conj.
Expansion.Equiv.	Contingency.Cause
Expansion.Sub.	Contingency.Cause
Expansion.Equiv.	Comparison.Conc.

Table 3: Top-5 confusing discourse relation pairs in the RoBERTa$_{\text{BASE}}$ model. For space limitation, we abbreviate the name of each second-level relation.

"Equivalence" with frequent ones rather than misclassify discourse relations that have the same higher-level ones.

4. Synthetic Data Generation for IDRR

Based on the preliminary results, we propose an error-driven method of generating synthetic data for improving fine-tuning performance of an encoder-only language model in IDRR.

4.1. Proposed Method

The proposed method of generating synthetic data consists of the following three steps (cf. Figure 4)

1. Extract top-*k* confusing discourse relation pairs based on false negative rate.
2. For each confusing discourse relation pair (R_{true}, R_{pred}), retrieve training examples that have the relation R_{true} as the source of synthetic data.
3. Synthesize data based on the retrieved examples using an LLM.

The following paragraphs explicate each step.

STEP1: Extraction of Confusing Discourse Relation Pairs The first step is to extract confusing discourse relation pairs based on a confusion matrix. As described in Section 3.3, we fine-tune a model, calculate a confusion matrix on the development split, and extract top-*k* confusing discourse relation pairs based on false negative rate. We utilize false negative rate as the degree of confusion to treat infrequent and frequent discourse relations equally.

STEP2: Retrieval of Training Examples Next, we prepare the source of synthetic data. We utilize training examples as the source judging it is difficult to generate argument pairs that have a given discourse relation from scratch. Specifically, for each confusing discourse relation pair (R_{true}, R_{pred}), we retrieve all the training examples that have the relation R_{true} in preparation for the following synthesis process.

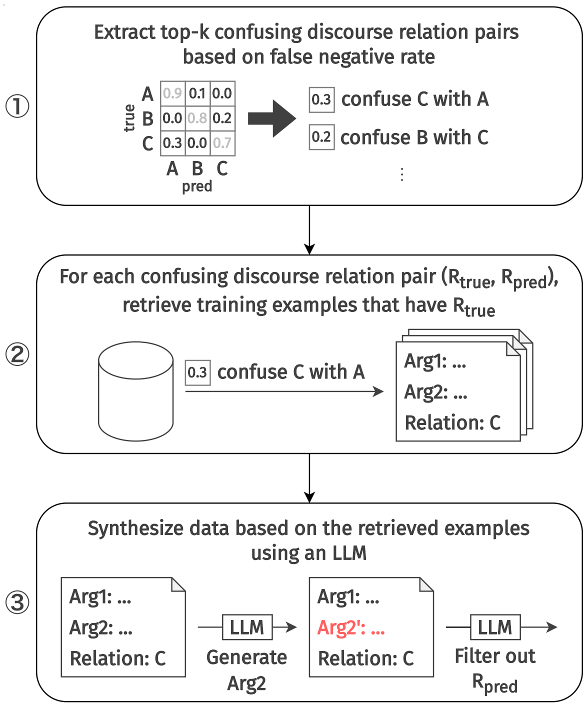
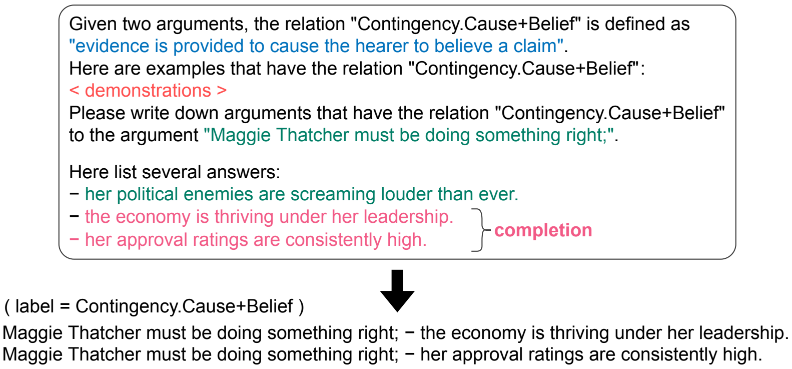


Figure 4: Overview of the proposed method.

STEP3: Synthesis of Data Then, we synthesize data focused on resolving the confusion. As mentioned in Section 1, we adopt two-stage prompting to synthesize data (Figure 5). Specifically, we first instruct an LLM to generate a candidate list of Arg2 given Arg1, original Arg2, and the definition sentence of the relation R_{true}. We synthesize Arg2 considering the unidirectionality of language models. Figure 6 demonstrates the aforementioned process using the example in Figure 1. Synthetic data can be obtained by splitting completion by the item mark "- " and combining each split with Arg1 and the label of R_{true}. In the second stage, we ask an LLM whether each pair of Arg1 and synthetic Arg2 has the relation R_{pred} or not. Regarding the demonstrations for learning the relation R_{true}/R_{pred}, we use *K* nearest neighbors of a source example referring to Liu et al. (2022b), which are retrieved from the training examples that have R_{true}/R_{pred}.

4.2. Generation of Synthetic Data

According to the proposed method, we generated three synthetic data from top-1, 3, and 5 confusing discourse relation pairs to examine the effect of *k* in later experiments. We fixed the value of *K*, the number of nearest neighbors for learning the relation R_{true}/R_{pred}, to 8 referring to Min et al. (2022). For the LLM, we employed GPT-4 (a.k.a "gpt-4-0613"). Table 1 includes the statistics of the synthetic data generated from top-3 confusing discourse relation pairs for RoBERTa$_{\text{BASE}}$ as a representative. The



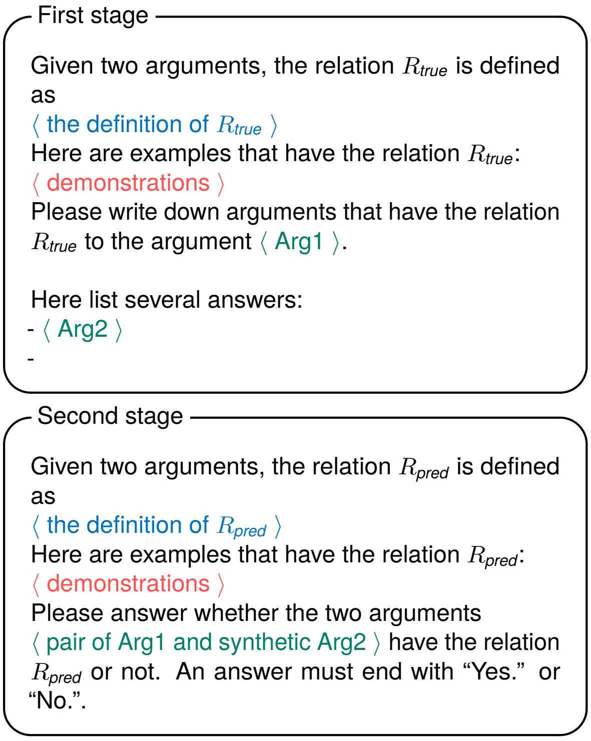
Given two arguments, the relation "Contingency.Cause+Belief" is defined as "evidence is provided to cause the hearer to believe a claim".
Here are examples that have the relation "Contingency.Cause+Belief":
< demonstrations >
Please write down arguments that have the relation "Contingency.Cause+Belief" to the argument "Maggie Thatcher must be doing something right;".

Here list several answers:
– her political enemies are screaming louder than ever.
– the economy is thriving under her leadership.
– her approval ratings are consistently high.

completion

(label = Contingency.Cause+Belief)
Maggie Thatcher must be doing something right; – the economy is thriving under her leadership.
Maggie Thatcher must be doing something right; – her approval ratings are consistently high.

Figure 6: Illustration of the first stage of synthetic data generation using the example in Figure 1. Definitions of discourse relations are taken from the PDTB-3 annotation manual.[7]



First stage

Given two arguments, the relation R_{true} is defined as
⟨ the definition of R_{true} ⟩
Here are examples that have the relation R_{true}:
⟨ demonstrations ⟩
Please write down arguments that have the relation R_{true} to the argument ⟨ Arg1 ⟩.

Here list several answers:
- ⟨ Arg2 ⟩
-

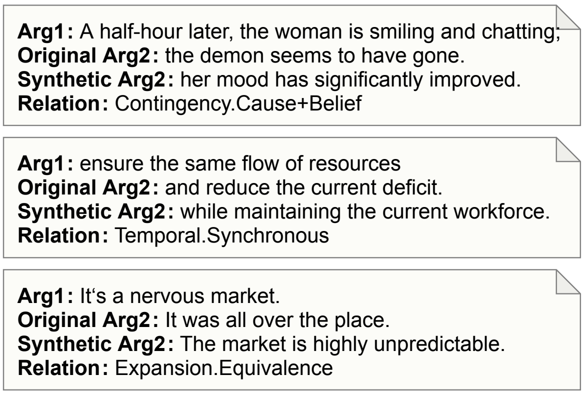
Second stage

Given two arguments, the relation R_{pred} is defined as
⟨ the definition of R_{pred} ⟩
Here are examples that have the relation R_{pred}:
⟨ demonstrations ⟩
Please answer whether the two arguments ⟨ pair of Arg1 and synthetic Arg2 ⟩ have the relation R_{pred} or not. An answer must end with "Yes." or "No.".

Figure 5: Prompt templates for an LLM. We adopt two-stage prompting to generate synthetic data. R_{true} and R_{pred} represent ground-truth and mispredicted discourse relations, respectively.

total cost of generating and filtering synthetic data was about $390.

Analysis of Synthetic Data In order to analyze the quality of synthetic data quantitatively, we sampled 30 examples each for the "Cause+Belief", "Synchronous", and "Equivalence" relations and manually verified them. We selected these three relations because they were top confusing discourse relations in all the experimental settings we tested. As a result of manual verification, 20, 20, and 23 examples of the "Cause+Belief", "Synchronous", and "Equivalence" relations were judged as valid, which appears to be acceptable as synthetic data.



Arg1: A half-hour later, the woman is smiling and chatting;
Original Arg2: the demon seems to have gone.
Synthetic Arg2: her mood has significantly improved.
Relation: Contingency.Cause+Belief

Arg1: ensure the same flow of resources
Original Arg2: and reduce the current deficit.
Synthetic Arg2: while maintaining the current workforce.
Relation: Temporal.Synchronous

Arg1: It's a nervous market.
Original Arg2: It was all over the place.
Synthetic Arg2: The market is highly unpredictable.
Relation: Expansion.Equivalence

Figure 7: Examples of synthetic data.

We also analyzed the examples for each relation qualitatively. "Cause+Belief" is required that one argument expresses some belief, and the other provides its justification. As is the example in Table 7, synthetic Arg2s are sometimes factual and inconsistent with original Arg2 when it expresses some belief. One of the possible remedies is to utilize third-level relation to select examples whose Arg1 expresses some belief.

Regarding "Synchronous", we observed GPT-4 was apt to include discourse connectives such as "while" to establish the relation. Although such examples are valid, this may cause shortcut learning (Geirhos et al., 2020), which raises the need for refining instructions.

Synthetic data of "Equivalence" was often judged as valid. One of the possible reasons is that the relation is regarded as a kind of paraphrasing and is relatively easy to understand for the LLM.

[7] https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf

5. Experiments

We conducted experiments to examine the effectiveness of incorporating synthetic data into training.

5.1. Experimental Settings

5.1.1. Data and Model

We used the PDTB dataset and synthetic data generated by the proposed method (cf. Section 3.1, 4.2). These statistics are organized in Table 1.

We evaluated the performance of the RoBERTa (Liu et al., 2019) model to compare with previous studies. We employed the base-[6] and large-size[8] pre-trained models hosted on Hugging Face Hub.

5.1.2. Training Details

During the training phase, we minimize the standard softmax cross-entropy loss. When incorporating synthetic data into training, we minimize the weighted sum of the losses of training examples and synthetic data, which is expressed by the following equations.

$$H = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(x)}}{\sum_{y' \in [Y]} e^{f_{y'}(x)}}$$

$$L = H_{\text{training}} + \lambda \times H_{\text{synthetic}}$$

where N is a batch size, Y is a set of classes, $f_y(x)$ is the logit for the class y, and λ is the weight for synthetic data.

During the evaluation phase, we evaluate the model by micro-F1 and macro-F1. We measure the performance on the development split per epoch and adopt the model parameters with the best dev performance (macro-F1) for evaluation on the test split.

5.1.3. Compared Methods

We adopted the following methods for comparison.

Vanilla In this setting, we merely fine-tune models without synthetic data.

Logit Adjustment (Menon et al., 2021) Based on the results of Table 3, we speculate the synthetic data generated by the proposed method is effective in learning long-tail discourse relations. Thus, we compare logit adjustment with our proposed method as a baseline of learning long-tail classes. This method adjusts logits when computing the standard softmax cross-entropy loss so that the rarer the class, the greater the loss. The above is expressed by the following equation.

$$H = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(x)+\tau \log \pi_y}}{\sum_{y' \in [Y]} e^{f_{y'}(x)+\tau \log \pi_{y'}}}$$

where π_y is an estimate of the class prior and τ is the temperature. We used the class frequencies on the training examples as π_y and set τ to 1.0 referring to the authors' report.

Long and Webber (2022) This is the state-of-the-art (SOTA) method in macro-F1 performance on the same task settings as ours. As mentioned in Section 2.1, they achieved superior performance by introducing contrastive learning. They also used additional training examples generated by inserting annotated discourse connectives between arguments.

Liu and Strube (2023) This is the state-of-the-art (SOTA) method in micro-F1 performance on the same task settings as ours. They achieved superior performance by introducing joint learning of generating discourse connectives between given argument pairs and predicting discourse relations based on them.

Jiang et al. (2023) We compare this as one of the methods using RoBERTa$_{\text{LARGE}}$. As mentioned in Section 2.1, they introduced the contrastive framework.

5.1.4. Hyper-Parameters

Regarding baselines, we performed grid search of learning rate from {5e-6, 1e-5, 2e-5} and selected the one that achieved the best macro-F1 on the development split. When incorporating synthetic data into training, we used the same hyper-parameters but performed grid search of λ, the weight for synthetic data, from {0.5, 0.25}. As we generated three synthetic data from top-1, 3, and 5 confusing discourse relation pairs, we adopted the one that achieved the best macro-F1 on the development split. Specifically, we used the synthetic data generated from top-3 and top-5 confusing discourse relation pairs for RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$, respectively. The specific values of hyper-parameters are included in A.3.

5.2. Results

Table 4 shows the experimental results of (second-level) IDRR on the PDTB dataset. As the synthetic data focuses on learning infrequent discourse relations, it might cause the forgetting of frequent discourse relations and deteriorate micro-F1. Despite

[8] https://huggingface.co/roberta-large

Model	Setting		Micro-F1	Macro-F1
GPT-3.5	few-shot		23.2	19.0
GPT-4	few-shot		29.4	30.9
RoBERTa$_{\text{BASE}}$	Vanilla		$64.2_{\pm 1.2}$	$57.1_{\pm 0.4}$
	Logit Adjustment (Menon et al., 2021)		$62.1_{\pm 1.5}$	$59.0_{\pm 0.5}$
	Long and Webber (2022)		64.7	57.6
	Liu and Strube (2023)		$\mathbf{65.5}^{\dagger}_{\pm \mathbf{0.4}}$	$54.9^{\dagger}_{\pm 0.8}$
	Ours	+synthetic data (unfiltered)	$64.5_{\pm 0.8}$	$58.4_{\pm 1.2}$
		+synthetic data (filtered)	$64.8_{\pm 1.0}$ $(64.0_{\pm 0.8})$	$\mathbf{59.1}_{\pm \mathbf{1.5}}$ $(\mathbf{57.5}_{\pm \mathbf{1.6}})$
RoBERTa$_{\text{LARGE}}$	Vanilla		$67.7_{\pm 0.5}$	$60.9_{\pm 1.6}$
	Logit Adjustment (Menon et al., 2021)		$64.8_{\pm 0.7}$	$61.4_{\pm 0.6}$
	Jiang et al. (2023)		$66.4^{\dagger}$	$60.1^{\dagger}$
	Ours	+synthetic data (unfiltered)	$67.4_{\pm 0.6}$	$62.2_{\pm 1.2}$
		+synthetic data (filtered)	$\mathbf{68.8}_{\pm \mathbf{0.4}}$ $(\mathbf{68.1}_{\pm \mathbf{0.3}})$	$\mathbf{62.4}_{\pm \mathbf{1.5}}$ $(\mathbf{61.6}_{\pm \mathbf{0.8}})$

Table 4: Experimental results of (second-level) IDRR on PDTB-3. The scores are the mean and standard deviation over three runs with different random seeds. The difference between “+synthetic data (unfiltered)” and “+synthetic data (filtered)” is whether or not to apply the filtering by an LLM in the second stage. † denotes that the handling of multi-labeled examples might be different from ours (cf. Section 3.1). For these studies, we also evaluated our model in their manner and reported the results in parentheses.

Relation	RoBERTa$_{\text{BASE}}$			RoBERTa$_{\text{LARGE}}$	
	VNL	Ours	L&W	VNL	Ours
Sync.	34.4	32.6♠	41.4	35.5	38.1♠
Async.	66.8	68.0	66.4	72.9	76.0
Cause	69.3	69.8	71.4	74.1	74.8
Cause+B.	1.7	11.8♠	0.0	5.0	5.1♠
Purp.	94.8	93.6	96.1	95.8	95.2
Cond.	70.2	73.8	74.1	75.5	78.0
Conc.	60.1	61.7	60.1	63.3	63.9
Cont.	49.0	49.1	56.9	56.7	56.4
Conj.	60.6	60.0	61.7	62.9	65.0
Equiv.	21.6	34.1♠	11.4	25.3	31.4♠
Inst.	69.8	72.7	69.8	73.1	73.2
Level.	57.0	57.0	55.3	58.9	59.4
Manner	80.3	80.5	78.4	80.9	83.1♠
Sub.	63.7	62.3	63.8	72.7	73.9

Table 5: Experimental results for each discourse relation. VNL and L&W represent Vanilla and Long and Webber (2022), respectively. “Ours” corresponds to the “+synthetic data (filtered)” setting. ♠ denotes that the model was trained with synthetic data of the relation.

the concern, we achieved the SOTA macro-F1 performance without sacrificing micro-F1 performance in both RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$ thanks to the synthetic data.

Detailed results are organized in Table 5. Regarding RoBERTa$_{\text{BASE}}$, the synthetic data is actually effective in learning infrequent discourse relations such as the “Cause+Belief” and “Equivalence” relations. On the other hand, it does not work on the “Synchronous” relation. One of the possible reasons is that the synthetic data may contain some phrases that induce shortcut learning as discussed in Section 4.2. The above problem can be alleviated by refining the instruction so as not to include discourse connectives. Regarding RoBERTa$_{\text{LARGE}}$, the synthetic data is generally effective in learning the target discourse relations except “Cause+Belief”. As the size of synthetic data for “Cause+Belief” is relatively small, the model may not have been adequately trained on the relation.

Comparing “+synthetic data (filtered)” with “+synthetic data (unfiltered)”, we can see the solid performance gain thanks to the filtering. We presume some removed examples are harmful to learning discourse relations even though they are not always noisy.

5.3. Discussion

Effects of Top-*k* Table 6 shows the change in performance of RoBERTa$_{\text{BASE}}$ when varying how many confusing discourse relation pairs to extract. While synthetic data is generally effective in improving macro-F1, learning to resolve more confusion does not necessarily lead to overall performance improvement, which suggests the importance of selecting which confusion to focus on.

Prompting from Discourse Connectives Inspired by previous studies, we attempted to gener-

k	Micro-F1	Macro-F1
0	$64.2_{\pm 1.2}$	$57.1_{\pm 0.4}$
1	$63.6_{\pm 0.8}$	$57.5_{\pm 0.9}$
3	$\mathbf{64.8}_{\pm \mathbf{1.0}}$	$\mathbf{59.1}_{\pm \mathbf{1.5}}$
5	$63.9_{\pm 1.0}$	$58.4_{\pm 1.6}$

Table 6: Correspondence between the number of confusing discourse relation pairs to extract and the performance of RoBERTa$_{\text{BASE}}$.

ate synthetic data utilizing discourse connectives to examine their effectiveness. Let us explain based on Figure 6. We added an annotated discourse connective to the beginning of original Arg2 (i.e. "her political enemies ... " → "**because** her political enemies ... ") and made an LLM generate text that starts with the connective (i.e. "– {completion}" → "– **because** {completion}"). We incorporated the synthetic data generated by the above method and evaluated the model performance.

As a result, the performance of RoBERTa$_{\text{BASE}}$ was micro-F1 of 64.2 and macro-F1 of 58.7. The effects of discourse connectives in this setting are limited.

Single-Stage Augmentation Strategy We generated synthetic data by simply instructing an LLM to paraphrase argument pairs and investigated the performance gain by the synthetic data to compare with our strategy. In Figure 5, we modified the instruction to "Please write down paraphrases of <pair of Arg1 and Arg2> keeping the relation R_{true}" and obtained paraphrases of argument pairs. We incorporated the synthetic data and evaluated the model performance.

As a result, the performance of RoBERTa$_{\text{BASE}}$ was micro-F1 of 64.4 and macro-F1 of 57.3, which implies the importance of generating diverse Arg2.

6. Conclusions

We proposed a method of generating synthetic data for IDRR using an LLM, which consists of two main steps: extraction of confusing discourse relation pairs based on false negative rate and synthesis of data focused on the confusion. According to the proposed method, we generated synthetic data effective in IDRR while addressing the complexity of IDRR by two-stage prompting. Thanks to the synthetic data, we achieved the SOTA macro-F1 performance without sacrificing micro-F1 performance and demonstrated its effectiveness especially in recognizing some infrequent discourse relations.

We will explore another prompting strategy to improve the quality of synthetic data. In addition, we would like to consider a method to generate synthetic data from scratch.

7. Limitations

In this study, we only used GPT-4 for synthetic data generation and did not focus on whether our proposed method also works using smaller or larger language models. We should not suggest the possibility that other language models could be used. In addition, there is no denying that GPT-4 might be trained on data regarding IDRR, leading to the effectiveness of our proposed method. However, we speculate this issue is unlikely because the few-shot performance of GPT-4 on IDRR is low.

We are also aware that the prompting strategy is underexplored. We do not claim our proposed method is optimal but position it as one successful example.

Acknowledgements

We thank anonymous reviewers for their valuable comments. The first author was supported by Grant-in-Aid for JSPS Fellows #22J15958.

8. Bibliographical References

Prajjwal Bhargava and Vincent Ng. 2022. DiscoSense: Commonsense reasoning with discourse connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10295–10310, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. AugGPT: Leveraging ChatGPT for Text Data Augmentation.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

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

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

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.

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

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

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

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

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.

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

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023. Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab

Emirates. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Kazumasa Omura and Sadao Kurohashi. 2022. Improving commonsense contingent reasoning by pseudo-data and its application to the related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 812–823, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

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

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.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Jun Saito, Yugo Murawaki, and Sadao Kurohashi. 2019. Minimally supervised learning of affective events using discourse relations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5758–5765, Hong Kong, China. 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.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. 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 732–742, Online. Association for Computational Linguistics.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A Label Dependence-Aware Sequence Generation Model for Multi-Level Implicit Discourse Relation Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11486–11494.

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.

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 725–731, Brussels, Belgium. Association for Computational Linguistics.

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.

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.

9. Language Resource References

Rashmi Prasad and Alan Lee and Nikhil Dinesh and Eleni Miltsakaki and Geraud Campion and Aravind Joshi and Bonnie Webber. 2008. *Penn Discourse Treebank Version 2.0*. distributed via Linguistic Data Consortium, ISLRN 488-589-036-315-2.

Rashmi Prasad and Bonnie Webber and Alan Lee and Aravind Joshi. 2019. *Penn Discourse Treebank Version 3.0*. distributed via Linguistic Data Consortium, ISLRN 977-491-842-427-0.

A. Appendix

A.1. Definitions of Discourse Relations

Table 7 shows the definitions of discourse relations used in the experiments.

A.2. Statistics of Synthetic Data for RoBERTa$_{\text{LARGE}}$

Table 8 organizes the statistics of synthetic data for RoBERTa$_{\text{LARGE}}$.

A.3. Hyperparameters

Hyper-parameters used in the experiments are organized in Table 9.

A.4. Top-Level Relation Classification Performance

As we can also calculate the performance in top-level relation classification from the results of second-level relation classification, we report it in Table 11 for reference.

Relation	Train	Synthetic Data		Dev	Test
		Unfiltered	Filtered		
Sync.	435	2,501	1,286	33	43
Async.	1,007	-	-	105	108
Cause	4,475	-	-	449	406
Cause+B.	159	940	331	13	15
Purp.	1,092	-	-	96	89
Cond.	150	-	-	18	15
Conc.	1,164	-	-	105	97
Cont.	741	-	-	91	63
Conj.	3,586	-	-	299	237
Equiv.	254	1,167	203	25	30
Inst.	1,166	-	-	118	128
Level.	2,601	-	-	274	214
Manner	615	3,948	3,666	28	53
Sub.	343	-	-	32	32

Table 8: Statistics of the synthetic data for RoBERTa$_{\text{LARGE}}$.

Name	Value	
	RoBERTa$_{\text{BASE}}$	RoBERTa$_{\text{LARGE}}$
Epoch	20	
Batch size	32	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	2e-5	1e-5
Weight decay	0.01	
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	
λ	0.25	

Table 9: Hyper-parameters for IDRR on PDTB-3.

Relation	Definition
Temporal.Synchronous	there is some degree of temporal overlap between the events described by the arguments
Temporal.Asynchronous	one event is described as preceding the other
Contingency.Cause	the situations described in the arguments are causally influenced but are not in a conditional relation
Contingency.Cause+Belief	evidence is provided to cause the hearer to believe a claim
Contingency.Purpose	one argument presents an action that an agent undertakes with the purpose of the goal conveyed by the other argument being achieved
Contingency.Condition	one argument presents a situation as unrealized (the antecedent), which (when realized) would lead to the situation described by the other argument
Comparison.Concession	an expected causal relation is cancelled or denied by the situation described in one of the arguments
Comparison.Contrast	at least two differences between the arguments are highlighted
Expansion.Conjunction	both arguments, which don't directly relate to each other, bear the same relation to some other situation evoked in the discourse
Expansion.Equivalence	both arguments are taken to describe the same situation, but from different perspectives
Expansion.Instantiation	one argument describes a situation as holding in a set of circumstances, while the other argument describes one or more of those circumstances
Expansion.Level-of-detail	both arguments describe the same situation, but in less or more detail
Expansion.Manner	the situation described by one argument presents the manner in which the situation described by other argument has happened or been done
Expansion.Substituion	arguments are presented as exclusive alternatives, with one being ruled out

Table 7: Definitions of discourse relations. They are basically taken from the PDTB-3 anntotation manual, but we slightly modify that of Expansion.Conjunction.

Model	Setting		Micro-F1	Macro-F1
GPT-3.5	few-shot		41.5	36.5
GPT-4	few-shot		42.3	42.2
RoBERTa$_{\text{BASE}}$	Vanilla		$74.8_{\pm 0.6}$	$69.9_{\pm 0.8}$
	Ours	+synthetic data (unfiltered)	$75.1_{\pm 0.3}$	$70.0_{\pm 0.5}$
		+synthetic data (filtered)	$74.7_{\pm 0.2}$	$70.0_{\pm 0.2}$
RoBERTa$_{\text{LARGE}}$	Vanilla		$78.0_{\pm 1.1}$	$74.0_{\pm 1.2}$
	Ours	+synthetic data (unfiltered)	$77.5_{\pm 0.6}$	$73.6_{\pm 0.4}$
		+synthetic data (filtered)	$78.6_{\pm 0.4}$	$74.5_{\pm 0.5}$

Table 11: The performance in top-level relation classification calculated from the results of second-level relation classification. The scores are the mean and standard deviation over three runs with different random seeds. Note that the performance is optimized for second-level relation classification, not top-level, and there is probably room for improvement.