Masked Language Model Helps Implicit Discourse Relation Recognition

Zizhe Wang

Master of Science
School of Informatics
University of Edinburgh
2024

Abstract

This study employs a pre-trained Masked Language Model (MLM) to explore Implicit Discourse Relation Recognition (IDRR), enhancing the automated identification of discourse relations without explicit connectives. Our approach adapts and fine-tunes the Amazon-EMAT model to predict multi-token discourse connectives, integrating them with a three-tier discourse relationship hierarchy, capturing the sense of discourse relations through mapping. By training the model on datasets containing explicit multi-token discourse connectives, we have generalized our findings to implicit discourse relations, significantly improving accuracy compared to traditional classification models. The results validate the effectiveness of using the MLM task for discourse analysis.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zizhe Wang*)

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Professor Bonnie Webber and Dr. Xixian Liao, for their meticulous guidance and support in my academic journey. Their mentoring has greatly enhanced my research capabilities. I consider myself very fortunate to have encountered such dedicated and excellent mentors during my postgraduate studies, and their guidance has been a crucial force in my academic growth.

I would also like to thank my friends for their encouragement and companionship. During times of anxiety and uncertainty, your support gave me the courage to keep moving forward. Your presence made this academic journey less lonely and filled with strength.

Lastly, I wish to extend my deepest gratitude to my parents. Thank you for your unconditional support, especially your financial assistance, which allowed me to pursue my studies at the University of Edinburgh, one of the world's leading institutions, without any financial burden. Your love and support have been my greatest motivation to forge ahead.

Table of Contents

Chapter 1

Introduction

Discourse typically refers to a series of clauses, sentences, or paragraphs within an article that convey its content. Discourse analysis involves examining and specifying the structure of these components. A subtask of discourse analysis, Discourse Relation Recognition (DRR), seeks to determine the discourse relations between two segments of text or arguments.

Predictable patterns of discourse relations are an important element of discourse coherence, so research on discourse relations is a crucial aspect of studying discourse coherence. Automatically identifying the meanings conveyed between sentences or clauses is highly valuable for downstream NLP tasks, such as machine reading comprehension [1], machine translation [2], sentiment analysis [3], text summarization [4], and event relation extraction [5]. The meanings of discourse relations are further categorized into predefined types which researchers have organized into a three-level hierarchy, ranging from broad top-level categories like Temporal, Contingency, Comparison, and Expansion to more specific meanings beneath them [6, 7]. Often, the relationship between two discourse segments is explicitly indicated by connectives. However, it is quite common for two text segments to indicate a discourse relationship without using an explicit connective. The process of identifying and categorizing these hidden relationships between segments that lack explicit connectives is referred to as Implicit Discourse Relation Recognition (IDRR). Although it is a particularly challenging task, substantial advancements have been achieved in IDRR research in recent years.

Since the advent of large-scale pre-trained language models, many researchers have attempted to apply these models to the IDRR task. The work described in this paper is inspired by recent research from Amazon [8], which proposed a method Extended-Matrix (EMAT) based on a BERT variant designed to overcome the limitation of

traditional BERT models, which can only predict a single token. Our approach adapts this model as part of our methodology, focusing on multi-token connectives, which are less ambiguous than single-token connectives, and eliminating the need for training from scratch.

By adding a limited number of multi-token connectives to the decoder's vocabulary and corresponding entries to the output prediction matrix, our method first predicts inter-sentential multi-token discourse connectives. Then, by mapping these connectives to the three-tiered discourse relation hierarchy, we achieve the goal of predicting discourse relations for the IDRR task. The results demonstrate that this MLM-based approach significantly improves the accuracy of predicting implicit discourse relations. Future research may focus on refining these techniques and exploring their applicability across more diverse datasets and languages, potentially broadening the scope and impact of this research in the field of discourse analysis.

Chapter 2

Background

2.1 Discourse Relation Recognition

As research at the levels of words, phrases, and sentences becomes increasingly deep and mature, more scholars are shifting their focus to discourse-level studies. Discourse generally refers to a cohesive and meaningful linguistic unit composed of a series of components that convey a complete and coherent message. Within a discourse, clauses are not haphazardly arranged but possess a definite structure and semantic relationships. Only by analyzing these structures and relationships can one achieve a thorough understanding and analysis of the discourse. Discourse structure analysis, also known as discourse parsing, is a core task in natural language processing [9, 10]. Since the framework represented by the Penn Discourse Treebank (PDTB) [6, 7] was released, providing a shallow representation of discourse structures that allows for the independent annotation of each discourse relation apart from others, it has attracted significant research attention [9, 10]. Discourse Relation Recognition (DRR), a subtask of discourse parsing, aims to understand the semantic connections between discourse units (also known as 'arguments', including phrases, sentences, and text segments) within a discourse. In addition to relational words and arguments, the sense of relations is typically classified into types. For example, in the PDTB-3, senses are artificially divided into three levels: class, type, and subtype. The top-level (class) is predefined as a comparison, contingency, expansion, and temporal as shown in Appendix A.

2.2 Implicit Discourse Relation Recognition

DRR can typically be divided into Explicit Discourse Relation Recognition (EDRR) and IDRR, depending on whether there are explicit connectives between the arguments. In the PDTB, the presence and interpretation of relationships between two arguments have been assessed and annotated by multiple experts.

For the EDRR task, explicit connectives can be directly extracted and classified into a certain relational meaning. For the IDRR task, since there are no explicit connectives, it is necessary to first detect the implicit relations which annotators can do by first manually inserting a connective and then labeling the sense they take it to convey. This implies that identifying implicit meanings often presents a challenge.

2.2.1 IDRR based on Machine Learning

Commonly, IDRR relies on classification methods to categorize relations. Initially, conventional machine learning techniques were used, treating adjacent spans of text, called 'arguments' (Arg1 and Arg2) as pairs of parameters. These pairs are input to predict the type of relational meaning existing between them. The basic process is illustrated in Fig 2.1.

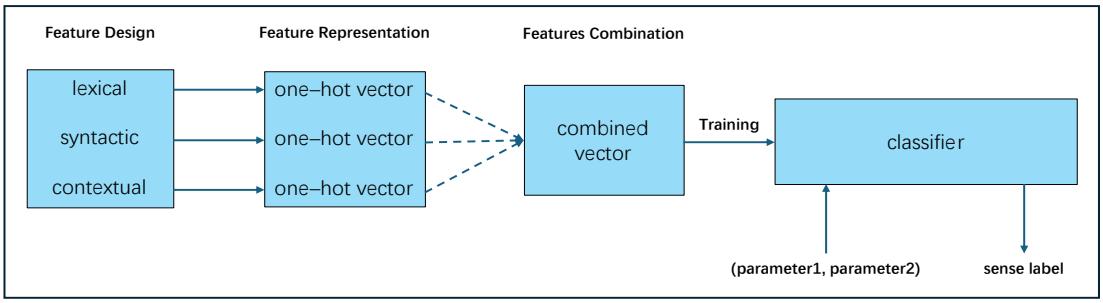

Figure 2.1: IDRR based on Machine Learning

Early machine learning approaches primarily focused on feature engineering, which involves the construction and selection of representative features for text classification. These features are typically categorized into several types: lexical features [11] and syntactic features [12], which primarily analyze the word, phrase, and structure within sentences; and contextual features [13], which pertain to the broader textual context beyond words or sentences. The integrated application of these features is crucial for a deep understanding and effective classification of textual content.

2.2.2 IDRR based on Pre-trained Learning

In recent years, deep learning technologies based on neural networks, exemplified by pre-trained models (PLM) such as BERT [14], RoBERTa [15], GPT [16], and T5 [17], have revolutionized feature engineering. After being trained on extensive corpora, these models are capable of capturing rich linguistic features, significantly enhancing adaptability and generalization to new tasks. Unlike the past practice of designing neural architectures from scratch for each specific task, it is now possible to quickly adapt these pre-trained models to a variety of downstream tasks through transfer learning. This is done by fine-tuning the models and adding output layers tailored for specific tasks. This approach reduces the need for manual feature design, thereby increasing processing efficiency and improving model performance.

For instance, in 2019, Shi and Demberg[18] leveraged a general Pre-trained Language Model (PLM) and conducted additional pre-training specifically tailored to domain-specific texts. They demonstrated that Next Sentence Prediction (NSP) aids in IDRR classification both within and across domains. Building on this, in 2020, Kishimoto et al. [19] extensively explored the performance of PLMs on IDRR downstream tasks, further affirming their effectiveness.

In the same year, Jiang and He [20] innovatively combined PLMs with recurrent neural networks (RNNs) to develop a new model architecture. This hybrid approach aimed to capitalize on the strengths of both PLMs and RNNs, particularly in handling sequences and contextual information more dynamically.

Additionally, some researchers have focused on refining the attention mechanisms within PLMs without altering their core architecture. One approach involved implementing penalty-based loss recalibration methods in the classifier component to enhance the learning process of attention mechanisms [21]. Concurrently, Jiang et al. [22] proposed a loss function inspired by contrastive learning, designed to deeply explore and comprehensively model the multilevel discourse relationships within texts. This method aims to improve the granularity and accuracy of relationship modeling by exploiting the comparative discrepancies between different discourse levels, thereby enhancing the overall interpretative power of the model. However, these methods still directly explore IDRR through classification results.

2.3 Masked language models[1]

The BERT model [14] has substantially improved the performance of numerous natural language processing tasks by pre-training on two fundamental tasks: Masked Language Modeling (MLM) and NSP. The NSP task helps the model implicitly learn discourse connectivity by predicting the logical relationships between sentences. Subsequent research [18] has validated this improvement in inter-sentence IDRR attributed to the NSP task.

However, additional research, including investigations into RoBERTa [15] and SpanBERT [23], suggests that excluding the NSP loss during training could lead to improved outcomes. This indicates that comprehending discourse relations might depend more heavily on the semantic insights obtained directly from the MLM task, rather than on a simple multi-classification task. In natural language text, it is common for multiple discourse relations to coexist within a single context. Moreover, in the PDTB, certain instances are annotated with dual discourse relations for a single connective. However, research on IDRR that employs this multi-classification approach typically predicts only one category at a time. Consequently, researchers are now exploring the transformation of the IDRR task into a generative one. A study [24] demonstrates that generating target sentences describing discourse relations allows for a deeper comprehension and articulation of the implicit connections between sentences. However, this T5-based approach requires significant training resources.

Further research [25] has introduced a technique for directly predicting discourse connectives. Through the generation and prediction of discourse connectives, this method is capable of more accurately capturing and expressing the subtle nuances in discourse relations. However, this approach continues to depend on classification outputs, as it is limited to predicting single-token connectives, which frequently possess multiple senses, complicating the accurate determination of a specific sense. Despite the possibility of ambiguity, their preliminary findings indicate that pseudo-connectives generated by MLMs may indeed aid in IDRR tasks. However, we know that multi-token connectives are generally less ambiguous than single-token connectives. For example, ‘as’ on its own can convey any of 15 senses, while multi-token connectives like ‘as soon as’ can convey 2 senses and ‘as a result’ can only convey 1 sense[7].

Building on these findings, it is a logical progression to utilize MLMs to directly predict connective phrases for identifying implicit discourse relations. Recent research

[1]The content of this section is largely derived from the author’s IPP with some modifications.

conducted by Amazon [8] introduced an approach employing MLMs to predict multi-token connectives, which would then allow more flexibility in text-based question-answering. This approach does not necessitate training from scratch but instead involves the addition of a limited number of multi-token entries to the decoder's vocabulary, along with corresponding entries in the output prediction matrix. This significantly reduces the number of parameters while outperforming current state-of-the-art models for multi-token completion.

2.4 Data Expansion

Regardless of the method used, implicit discourse relation recognition requires a significant amount of labeled data for training models. However, the availability of annotated corpora is restricted. Manually annotating data is a time-intensive process and, in most cases, requires both domain expertise and specialized knowledge. Therefore, considering the costs, we decided to automatically generate extra training data by removing discourse connectives from examples of explicit discourse relations.

Marcu and Echihabi [26] adopted this approach. However, indiscriminate use of these artificially generated data could extra training data by removing discourse connectives from examples of explicit discourse relations. Consequently, some researchers are endeavoring to identify the connections between explicit and implicit discourse relations. One approach is to employ various techniques to filter out samples and features that are truly valuable for implicit discourse recognition from existing corpora. For sample filtering, researchers use statistical analysis methods to assess the contextual differences in discourse relations. By calculating the omission rate of connectives and the contextual differences, datasets containing explicit relations are selected, providing valuable corpora for model training [27].

Additionally, clustering methods such as Single Centroid Clustering (SCC) can optimize the efficiency and effectiveness of model training by identifying the most representative samples from large datasets [28]. In terms of feature handling, the Teacher-Student Model uses a blend of explicit and implicit data to train a general teacher model. Through knowledge distillation techniques, features are transferred from the teacher model to a student model specifically for implicit discourse relation recognition [29]. Furthermore, Ji et al. [30] employed domain adaptation techniques by sharing features between the source and target domains (explicit and implicit) and adjusting the feature representation and distribution of explicit and implicit samples to

achieve feature alignment and distribution alignment, thereby reducing inter-domain differences. The comprehensive application of these methods helps enhance the accuracy and generalizability of implicit discourse relation recognition models.

Another approach leverages the inherent commonalities and differences between languages to generate implicit discourse relation datasets cross-linguistically. This method is based on the differences in expression forms when conveying the same meaning across different languages, where content expressed implicitly in one language might find its explicit counterpart in another [31]. For instance, researchers utilize the significantly higher frequency of connective omission in Chinese compared to English [32], and English compared to French [33], to generate suitable datasets. These inter-lingual differences enable the expression forms of the same information to be comparable across languages, providing a basis for constructing and expanding multilingual implicit discourse relation datasets.

Chapter 3

Methodology

3.1 Data Preprocessing

To predict implicit discourse relations within texts, scholars face a significant challenge due to the lack of explicit connectives linking sentences, which complicates the direct usage of such sentences in training neural networks. This limitation leads to a scarcity of annotated data, which becomes a major bottleneck for most neural network-based methods in this domain.

One commonly used method to address this challenge, discussed in previous Section 2.4, is known as dataset expansion. However, this technique demands substantial preparatory efforts and considerable computational resources, which can be impractical for many research settings.

An alternative strategy is to use data augmentation techniques. This method has been notably applied in the study of the Penn Discourse Treebank (PDTB-2), where traditional machine learning models were trained on sections 2-21 and tested on section 23. This latter section contains 761 implicit discourse relations, which, although useful, represents a small sample size [13, 34, 35]. The limited size of the test set can obscure whether observed improvements in model performance are due to genuine methodological enhancements or merely coincidental fits to specific attributes of the test data.

To address this, researchers used cross-validation techniques within this constrained dataset to better generalize findings without the need to significantly expand the dataset. This method allows for more efficient utilization of available data and helps in assessing the robustness and general applicability of new features or models, providing a more reliable basis for evaluating improvements in the field of discourse analysis.

Nevertheless, we aim not to restrict ourselves solely to PDTB data nor to complicate data acquisition excessively. Research by Kishimoto et al. [19] demonstrates that tailored text for discourse classification, supplemented with additional pre-training and using samples with explicit connectives for training, significantly benefits the recognition of implicit discourse relations. Inspired by this, for our experimental design, we chose a dataset constructed from Wikipedia dumps by HuggingFace, extracting 687,469 entries with explicit connectives, to fine-tune the model.

BERT's training process consists of two stages: pre-training and fine-tuning. In the pre-training stage, BERT acquires contextual information and understands the relationships between two adjacent sentences using a large unlabelled corpus. In the fine-tuning stage, BERT is trained on a task-specific dataset and modifies the pre-trained representations for downstream tasks. Although additional pre-training steps may be beneficial for identifying implicit discourse relations if a substantial corpus is available, given BERT was specifically trained on Wikipedia ($\tilde{2}$.5B words) and Google's BooksCorpus ($\tilde{8}$00M words), and considering computational costs, we focus on fine-tuning to adapt to the downstream task of implicit discourse relation recognition.

To determine a suitable training set for our study, we initially compiled a list of multi-token connectives. This list was sourced from the Appendix A of PDTB-3 (shown in Appendix A) and from the Connective-Lex [36]. Our focus was specifically on predicting multi-token discourse connectives that occur between sentences (inter-sentential). We manually removed any connectives that functioned as within sentences. The final list comprised 47 such connectives, and we have included this specific list in the Appendix B of the current paper.

Our goal is to learn about discourse relations from explicit connectives, which mark conceptual relations between two sentences.

Fig 3.1 showcases sentences that illustrate both explicit and implicit discourse relations. In the raw text, explicit connectives are clearly marked, while implicit connectives are inferred and inserted by annotators.

For instance, in the first example, The relation is marked with a complex sense including 'Temporal.Asynchronous.Succession' and 'Contingency.Cause.Reason'. This indicates a time-based sequence that also provides a causal explanation. This demonstrates that explicit connectives can sometimes have multiple senses.

For model training, we need to process the sentences into the format of Example 3.1:

[CLS] [Arg1] [SEP] [MASK] [Arg2] (3.1)

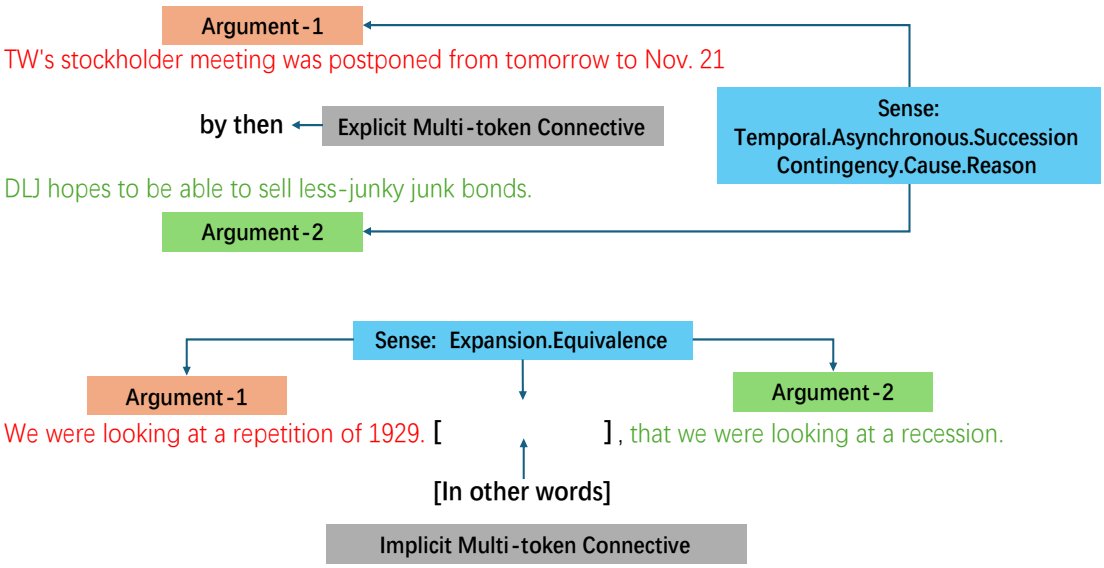

Figure 3.1: Examples of discourse relation annotation with explicit and implicit multi-token connectives in the PDTB corpus

Here, we employ a specific format to process sentences so that the model can learn logical connections within the text. This format includes special markers: [CLS] at the beginning of each sentence, representing the context of the entire sentence; [Arg1] and [Arg2] representing the first and second arguments in a discourse relation, respectively; [SEP] used to separate these arguments; [MASK] serving as a placeholder for the connective that the model needs to predict. Through this structured input, the model can more effectively learn and predict the explicit or implicit connections between the arguments in the text.

As previously discussed, we hypothesize that training with explicit argument pairs will effectively generalize the identification of implicit argument pairs. Therefore, for the test set, we utilize sentences annotated with implicit discourse relations from the most recent Penn Discourse Treebank (PDTB-3) to evaluate the accuracy of our predictions, using the same preprocessing method.

3.2 Mask-Filling Task

Our research methodology is based on the decoder matrix of an extended BERT model's Masked Language Model (MLM) to handle multi-token explicit discourse relation connectives shown in Fig 3.2.

Based on the architectural diagram, we can describe the process as the following

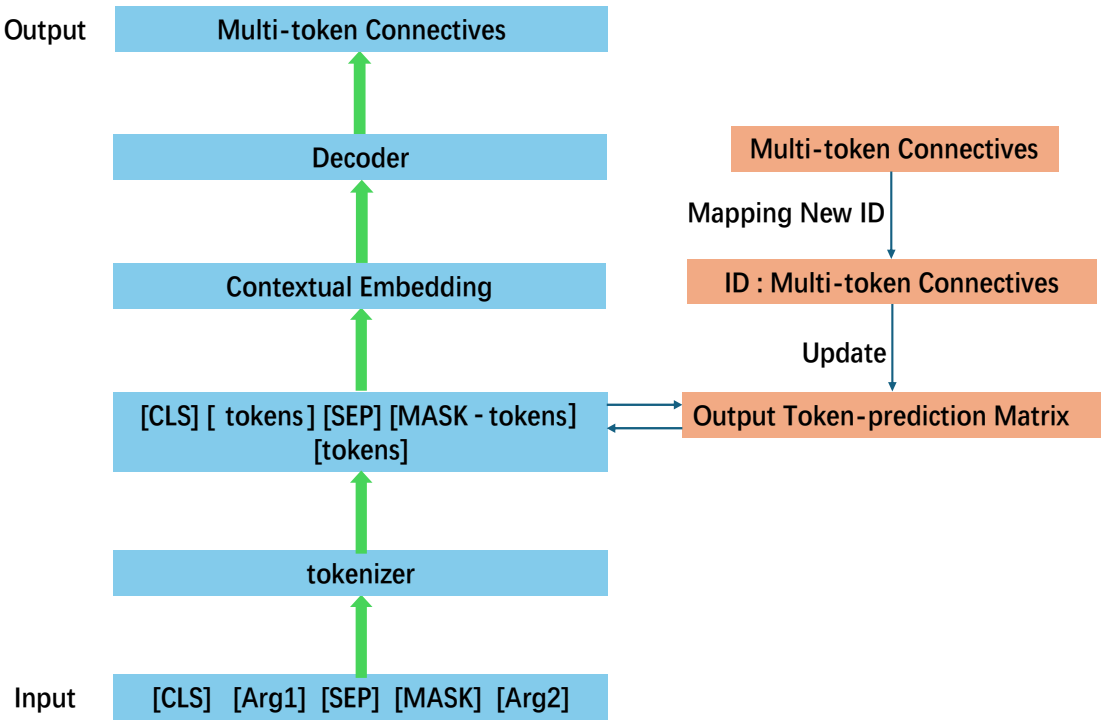

Figure 3.2: Structure of Our Model

steps, with the parameter settings explained at the end.

Loading Datasets and Model: Initially, the dataset described in Section 3.1 is loaded alongside a pre-trained BERT model. This stage is crucial to ensure that our model is primed to process new inputs and adapt effectively during the retraining phase.

Reading Vocabulary Files: Subsequently, we engage in processing a vocabulary file specifically curated to include multi-token discourse connectives, which are not typically present in the model's original lexicon. These discourse connectives are essential for recognizing discourse relation

Mapping and Extending the Lexicon: For each new vocabulary item in the file, if it is not already in the model's lexicon, we assign it a new embedding vector. These vectors are generated using a dimensionality aligned with the pre-existing model architecture to ensure compatibility and optimal integration. The new embeddings are then added to the model's output prediction matrix, thus enabling the prediction of these multi-token connectives.

Integration of New Embedding Vectors: The newly created embedding vectors are integrated directly into the model's output token prediction matrix. This direct insertion bypasses traditional tokenization and vocabulary-matching processes typically used in NLP, allowing for a more nuanced understanding and generation of text.

Optimization and Performance Tuning: With a configuration of 4 GPUs and 16 parallel processes. Fine-tuning is conducted with a batch size of 128 and a learning rate of 0.0001, employing the Adam optimizer to ensure efficient convergence. This phase spans 2 epochs and focuses on achieving an optimal balance between training speed and model accuracy, ensuring robust performance.

3.3 Mapping Task

The ultimate goal of predicting multi-token connectives is to identify the corresponding senses of discourse relations. Thus, our task extends beyond merely forecasting the connectives themselves; we also aim to map the relevant senses associated with these connectives from Appendix A of PDTB-3 as shown in Appendix C.

Utilizing 5103 manually annotated instances of multi-token connectives and their corresponding senses from PDTB-3, we establish these as our gold standard for calculating accuracy metrics.

We acquire the top-k (where k = 1, 3, 5, and 10) connectives. However, for the senses, we only consider the top-1 and top-2 corresponding senses. Ultimately, this method yields results formatted similarly to Table 3.1, encompassing about 5,000 entries, which facilitates the final evaluation (listing only top-1 as an example, the format for top-k connectives and their corresponding senses remains consistent).

3.4 Evaluation Metric

To comprehensively evaluate the predictive efficacy of our model, we employ accuracy as the primary metric of assessment.

Initially, we evaluate the accuracy of multi-token discourse connectives predicted by the model. The accuracy is defined as follows: for each data point i in the dataset, if any of the top k predictions pred_i^1 to pred_i^k (where $k = 1$ or 5) matches the manually annotated gold-standard connective gold_i, the accuracy for that item is scored as 1; otherwise, it is 0. The index j traverses these top k predictions for each sample.

Let n be the total number of data points in the dataset. For each data point i, where $i = 1, 2, \ldots, n$, let gold_i denote the gold-standard connective, and pred_i^j represent the j-th prediction out of the top k predictions for that data point, where $j = 1, 2, \ldots, k$.

Define the accuracy for each data point i as:

Attribute	**Details**
Sentence	Businesses were borrowing at interest rates higher than their own earnings. [CONNECTIVE] What we're seeing now is the wrenching readjustment of asset values to a future when speculative-grade debt will be hard to obtain rather than easy.
Top-1 Connective	After all
Golden Connective	As a result
Top-1 Sense Level 3	Contingency.Cause+Belief.Reason+Belief, Expansion.Conjunction, Expansion.Level-of-detail.Arg2-as-detail
Golden Sense Level 3	Contingency.Cause.Result
Top-1 Sense Level 2	Expansion.Level-of-detail, Contingency.Cause+Belief, Expansion.Conjunction
Golden Sense Level 2	Contingency.Cause
Top-1 Sense Level 1	Expansion, Contingency
Golden Sense Level 1	Contingency

Table 3.1: Analysis of Connective and Sense Usage in Context

$$\text{acc}_i = \begin{cases} 1 & \text{if } \exists j \in \{1,\ldots,k\} : \text{pred}_i^j = \text{gold}_i \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Subsequently, we map the first prediction of our model to the corresponding sense. If the gold-standard senses are a subset of the predicted senses, then the method of measuring accuracy is similar to that used for calculating the accuracy of connectives. Given the three-tiered semantic hierarchy in PDTB-3, this method of calculation is uniformly applied to all three levels. Based on the table you provided, here's a more precise way to phrase your statement:

Subsequently, we map the first prediction of our model to the corresponding sense. If the gold-standard senses are a subset of the predicted senses, then the method of measuring accuracy is similar to that used for calculating the accuracy of connectives. Given the three-tiered semantic hierarchy in PDTB-3, this method of calculation is uniformly applied to all three levels. However, level 3 accuracy is defined as the highest possible precision achievable for each sense, which may include some senses classified at Level 2. This definition applies because certain discourse connectives do not have a corresponding level 3 subtype, and only senses that are not symmetric possess a level 3. For example, in the 'Temporal' class, the 'Synchronous' type does not have a level 3 subtype. The formula for the accuracy at each level is defined as:

$$\text{acc}_i^{(l)} = \begin{cases} 1 & \text{if } \exists j \in \{1,\ldots,k\} : \text{pred}_{i,l}^j = \text{gold}_{i,l} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

The accuracy measure $\text{acc}_i^{(l)}$ represents the accuracy of the i-th instance at level l. It equals 1 if any prediction in the set $\{\text{pred}_{i,l}^1, \ldots, \text{pred}_{i,l}^k\}$ matches the gold standard $\text{gold}_{i,l}$; otherwise, it equals 0.

At last, we then compute the average accuracy across all data points in the dataset, as shown in Formula 3.3 :

$$\text{accuracy} = \frac{1}{n}\sum_{i=1}^{n} \text{acc}_i^{(l)} \tag{3.3}$$

Chapter 4

Results and Analysis

4.1 Connectives

Amazon-EMAT [8] utilizes data from two sources: Wikipedia and the BookCorpus [37] for its Multi Token Completion prediction of top-k accuracies (k = 1,3,5,10) as demonstrated in Table 4.1. Additionally, this table also presents the model's accuracy on the PDTB-3 IDRR corpus.

The Amazon-EMAT objective is for a general MTC, which views phrases as NP chunks or entities, retaining phrases that appear 500 times or more in the corpus, leaving us with approximately 93K phrases, which is significantly larger than our vocabulary and is not limited to any specific domain. Our results can attest to the model's enhanced performance in recognizing discourse relations.

Top-k	**Top-1 Accuracy (%)**	**Top-3 Accuracy (%)**	**Top-5 Accuracy (%)**	**Top-10 Accuracy (%)**
Amazon–EMAT	12.64%	20.48%	24.63%	30.65%
Our	24.50%	38.60%	48.16%	61.38%

Table 4.1: Performance by Top-k Accuracy Levels

4.2 Senses

4.2.1 Overall Multi-Level Sense Analysis

Table 4.2 displays the model's accuracies at Level-1, Level-2, and Level-3 on the PDTB-3 corpus. Previous studies have either focused on the older PDTB (PDTB-2) version or only recorded Level-1 and Level-2 of sensed types, whereas we have meticulously

documented the accuracy corresponding to each level. As shown in Table 4.2, when we utilize only the top-1 predicted multi-token connective for mapping, our performance in the four-way classification at the first level surpasses previous studies.

However, the performance at the second level falls short compared to the approach proposed by Long and Webber, which employs a semantic hierarchy to select contrastive learning samples for the task of recognizing implicit discourse relations. Considering that during the expert annotation phase of PDTB-3, 1-2 connectives are also labeled, we additionally recorded the discourse relation senses represented by the top-2 most likely predicted multi-token connectives. When mapping with the predicted top-2 multi-token connectives, the results demonstrate better performance than earlier systems.

Model	**Top-level Accuracy (%)**	**Second-level Accuracy (%)**	**Third-level Accuracy (%)**
Liu and Li [38]	57.67	N/A	N/A
Chen et al. [39]	57.33	N/A	N/A
Lan et al. [40]	57.06	N/A	N/A
Ruan et al. [41]	58.01	N/A	N/A
BiLSTM [42]	60.45	N/A	N/A
BERT [42]	64.04	N/A	N/A
Long and Webber [43]	75.31	64.68	N/A
Our (Top-1)	79.23	56.09	51.96
Our (Top-2)	**91.75**	**76.84**	**73.80**

Note: N/A indicates that data for the respective accuracy level is not available.

Table 4.2: Model Accuracy Comparison Across Different Levels of Complexity

We observe that when obtaining the senses corresponding to the top-2 connectives, the accuracies at the top level of discourse relations approach those of expert annotations. One possible reason, as mentioned in Section 2.1, is that whether based on PDTB-2 or PDTB-3, most previous research utilized implicit discourse relation cases as the data source, only varying in how the training, validation, and test sets were divided. In contrast, we utilized all cases of implicit discourse relations from PDTB-3 exclusively for our test set. Our training and development data comprised additional explicit discourse relation instances, which greatly exceeded the volume of data available for other studies on implicit discourse relations. For brevity, the training and development dataset will be referred to as the Train-Dev dataset hereafter.

4.2.2 Specific Multi-Level Sense Analysis

We also investigated the accuracy corresponding to each sense type at different levels. Table 4.3 displays the accuracy of the four general sense types (Level 1) on the PDTB-3 corpus.

Sense Type	Top-1 Accuracy(%)	Top-2 Accuracy(%)
Expansion	87.55	97.13
Contingency	69.59	86.24
Comparison	32.23	60.90
Temporal	39.58	62.50

Table 4.3: Accuracy for Level 1: Top-1 and Top-2 Analysis

It has been observed that the highest prediction accuracy is achieved when the discourse meaning is categorized as 'Expansion', while the lowest accuracy occurs when categorized as 'Temporal'. Subsequently, we examined the distribution of discourse meanings within both the Train-Dev dataset and the test dataset. The apparent cause of this discrepancy in accuracy seems to be the uneven distribution of data.

Thus, we attempted to explore the distribution of discourse senses, considering that we extracted contexts from the Wikipedia corpus using multi-token connectives. In order to classify each context according to its corresponding sense, when the connective might be mapped to multiple senses, we employed the following method: initially calculating the distribution of annotated proportions of explicit discourse connectives in PDTB-3. We presume that both the Wall Street Journal corpus (the data source of PDTB) and the Wikipedia corpus represent real-world text distributions, and we assume these distributions are similar.

Let C be the multi-token connective, and let D_1 and D_2 represent two different corpora, which are PDTB-3 and our explicit discourse connectives Train-Dev dataset (source from Wikipedia).

For each sense S of the connective C in D_1, calculate the proportion $p(S|C,D_1)$ as follows:

$$p(S|C,D_1) = \frac{\text{Number of instances of } S \text{ with } C \text{ in } D_1}{\text{Total instances of } C \text{ in } D_1} \tag{4.1}$$

Assume the distribution of senses in D_2 mirrors D_1. For the total instances of C in D_2 denoted as $n(C,D_2)$, the expected number of instances for each sense S in D_2,

denoted $n(S,C,D_2)$, is calculated as:

$$n(S,C,D_2) = p(S|C,D_1) \times n(C,D_2) \tag{4.2}$$

The specific distribution results are shown in Table 4.4:

Table 4.4: Sense Distribution of Multi-token Connectives in Train-Dev Dataset

Multi-token Connectives	Exp. (%)	Cont. (%)	Temp. (%)	Comp. (%)
after all	50.0	50.0		
after that				100.0
along with	100.0			
and then	100.0			
as a result		100.0		
as an alternative	100.0			
as well	64.71			35.29
at that point			100.0	
at that time			100.0	
at the same time			100.0	
at the time			100.0	
but then				100.0
but then again				100.0
by comparison				100.0
by contrast				100.0
by the way	50.0			50.0
by then		12.5	87.5	
even before			50.0	50.0
even before then			50.0	50.0
even then			50.0	50.0
for example	100.0			
for instance	100.0			
for one	100.0			
for one thing	50.0	50.0		

Continued on next page

Table 4.4 continued from previous page

Multi-token Connectives	Exp. (%)	Cont. (%)	Temp. (%)	Comp. (%)
in addition	100.0			
in any case				100.0
in any event	100.0			
in contrast				100.0
in essence	100.0			
in fact	87.06			12.94
in other words	100.0			
in particular	100.0			
in short	100.0			
in sum	100.0			
in the end	36.36	18.18	27.27	18.18
in the meantime			93.33	6.67
in the meanwhile			100.0	
in this way		100.0		
in turn	33.33	33.33	33.33	
just in case		100.0		
later on			100.0	
more accurately	100.0			
no matter				100.0
on the contrary				100.0
on the other hand				100.0
on the other				100.0
quite the contrary	100.0			
that is	100.0			

For instance, in the dataset of PDTB-3 explicit discourse connectives with the connective 'as well,' let $C =$ 'as well', and senses $S_1 =$ 'Expansion' and $S_2 =$ 'Comparison' in D_1 (PDTB-3) with proportions for 'Expansion' at 35% and 'Comparison' at 65%, the formula would be applied to the Wikipedia corpus as follows:

For $n(\text{'as well'}, \text{Wikipedia}) = 3428$:

$$n(\text{'Expansion'}, \text{'as well'}, \text{Wikipedia}) = 0.35 \times 3428 \approx 1200$$

$$n(\text{‘Comparison’}, \text{‘as well’}, \text{Wikipedia}) = 0.65 \times 3428 \approx 2228$$

Finally, we obtained the results depicted in Figure 4.1 .

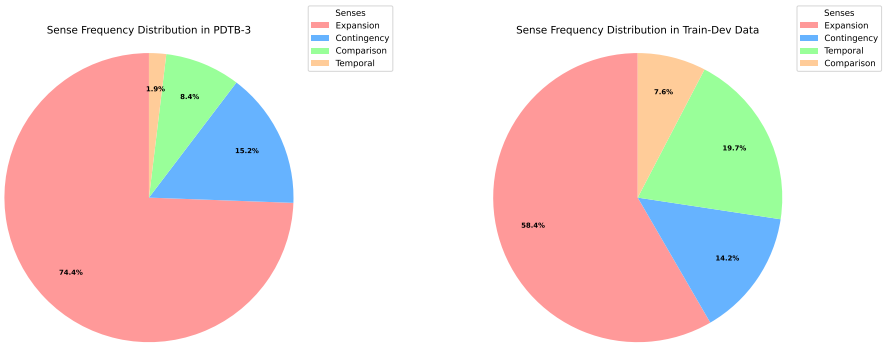

Figure 4.1: Sense Frequency Distribution in Train-Dev and PDTB-3 Dataset

Finally, Table 4.4 and Table 4.5 respectively illustrate the accuracy rates corresponding to levels 2 and 3 on the PDTB-3.

From this data, it is evident that there are significant differences in accuracy rates among different levels of discourse meaning classification. For instance, despite an overall improvement in accuracy at levels 2 and 3, certain specific types such as ‘Comparison.Contrast’ and ‘Temporal.Asynchronous’ exhibit lower accuracy rates. This may reflect their lower distribution in real-world data and the complexity involved in data annotation or interpretation.

We believe that enhancing the diversity and balance of Train-Dev samples, along with improvements to the structure of the classification model, could lead to higher predictive accuracy in future research.

Sense Type	**Top-1 Accuracy(%)**	**Top-2 Accuracy(%)**
Comparison.Contrast	30.33	57.35
Contingency.Cause	68.41	85.85
Expansion.Conjunction	52.43	84.98
Expansion.Equivalence	42.24	60.87
Expansion.Instantiation	56.78	72.70
Expansion.Level-of-detail	79.05	93.63
Temporal.Asynchronous	27.03	48.65
Temporal.Synchronous	32.20	54.24

Table 4.5: Accuracy for Level 2: Top-1 and Top-2 Analysis

Sense Type	**Top-1 Accuracy(%)**	**Top-2 Accuracy(%)**
Contingency.Cause.Result	56.75	76.67
Expansion.Instantiation.Arg2-as-instance	56.78	72.70
Expansion.Level-of-detail.Arg1-as-detail	37.09	63.64
Expansion.Level-of-detail.Arg2-as-detail	76.47	94.12
Temporal.Asynchronous.Precedence	27.03	48.65

Table 4.6: Accuracy for Level 3: Top-1 and Top-2 Analysis

Chapter 5

5.1 Conclusions

In this paper, we explore the main components of the DRR task, particularly emphasizing the significance of IDRR, which presents greater challenges. Enhancing the accuracy of IDRR is crucial not only for its own sake but also holds significant practical application in machine reading comprehension, text summarization, and other downstream tasks.

We have selected a robust dataset from Wikipedia, utilizing explicit discourse relation datasets to train our model. This strategic choice is aimed at overcoming the common challenge of insufficient annotated data available for training neural networks.

Our research demonstrates the transferability from explicit to implicit relations by training on explicit data for implicit task recognition. This approach not only simplifies the training process but also reduces reliance on extensive manual annotations, which are typically expensive and time-consuming. In this process, the dataset containing explicit discourse relations serves as a raw data source for subsequent research.

Our principal contribution involves adapting and enhancing Amazon's EMAT model, evaluated using the latest PDTB-3 dataset provided for implicit discourse relation recognition. The results show a significant improvement in the accuracy of recognizing implicit discourse relations. Considering the limited number of multi-word discourse connectives, thus our expanded vocabulary is quite limited, having a negligible impact on the model's parameter count while maintaining efficiency in enhancing performance.

As a result, This unique method of first predicting multi-word connectives and then mapping them to a three-tier discourse relation hierarchy has proven more effective than previous direct classification approaches, validating the efficacy of leveraging the Masked Language Model (MLM) task of pre-trained models for discourse relation recognition.

5.2 Limitations and Future Work

While the current study successfully addresses several aspects as section 5 of implicit discourse relation recognition, it also highlights its limitations. Future research can build upon these insights to further refine and advance the field.

Data Expansion: In the selection of the training set for this paper, sentences containing explicit discourse connectives were directly retrieved, utilizing the 'Explicit Relation Senses and Explicit Connectives' in Appendix A of PDTB-3 as the gold annotations for Train-Dev dataset. For testing, 'Implicit Relation Senses and Implicit Connectives' from Appendix C of PDTB-3 were used as the gold annotations. Previous studies and our own research have confirmed that training tasks using explicit connectives indeed facilitate the recognition of implicit discourse relations. However, as mentioned in Section 2.4, indiscriminate use of these artificial data can degrade the performance of implicit discourse relation identification, preventing optimal results. Our approach to obtaining the Train-Dev dataset in the most accessible manner was to reduce training costs and indeed resulted in performance improvements. However, future studies could explore using data augmentation techniques mentioned in Section 2.4 to further enhance the accuracy of implicit discourse relation identification.

Evaluation Metrics and Model Comparison: Since our approach involves directly predicting the connectives themselves and mapping connectives to relation senses based on Appendices A and C of PDTB-3, the mapped senses often extend beyond a single category, thereby rendering the task as non-binary and unsuitable for F1-score calculation. This introduces limitations in comparison with previous studies, which typically use both accuracy and F1-score as evaluation metrics. Consequently, our study can only compare improvements in predictive accuracy with previous research.

Linking Explicit and Implicit Discourse Relations: One limitation of our current approach lies in its handling of cases where explicit and implicit discourse relations coexist within the same sentence or passage.

For instance, in the example from the PDTB-3 annotation manual—We've got to get out of the Detroit mentality **and** be part of the world mentality—the explicit discourse relation is expressed through the connective **and** (*Expansion.Conjunction*), while an implicit discourse relation is conveyed through the word **instead** (*Expansion.Substitution*), which suggests a substitutional relationship not directly linked to the explicit conjunction. In this scenario, the explicit and implicit relations are independent but co-occur, creating a complex interplay that is challenging to capture accurately with traditional

methods.

Our current model does not fully address such intricate cases where multiple, potentially independent discourse relations are at play. This limitation suggests that future research should focus on refining our approach to better identify and label these layered discourse relations. Specifically, developing techniques to simultaneously recognize explicit connectives and derive implicit meanings from the broader context could enhance the model's capability in such scenarios.

Handling Multi-Sense Annotations: Another limitation is the model's ability to handle cases where implicit discourse relations are annotated with multiple senses. Annotators may perceive more than one valid sense within a single relation, especially in complex sentences. This multiplicity poses a challenge for models trained to identify a single, dominant sense. Future work could explore methods to incorporate multi-sense annotations into the training process, allowing the model to better capture the full spectrum of possible interpretations. This could involve augmenting the training data with examples specifically annotated for multiple senses or expanding our model into a multi-sense prediction framework that can predict multiple discourse relations simultaneously.

Data Volume and Pre-training: Despite opting for a sufficiently fine-tuned dataset considering cost issues, the volume of data used is still significantly less than that employed by the pre-trained models themselves. Adding additional pre-training steps could potentially enhance model performance. This approach would not only leverage more extensive data handling but also incorporate a broader context of training, which might improve the model's ability to generalize across different discourse relations and contexts.

Dataset Balance: The use of unbalanced Train-Dev dataset and test datasets impacts the accuracy of identifying different levels of discourse relation meanings. This imbalance can lead to models performing well on some discourse relation types while underperforming on others, especially those less represented in the Train-Dev dataset. Strategies such as stratified sampling and synthetic data generation might be employed in future research to address these disparities, thereby providing a more robust framework for discourse analysis across varied and complex datasets.

In summary, although this study makes significant strides in the field of implicit discourse relation recognition, it also identifies several areas for improvement that warrant further exploration in future research. The discussed limitations—spanning data expansion, evaluation metrics, the handling of complex discourse relations, multi-

sense annotations, and data balance—highlight the complexities inherent in this field and underscore the need for continued innovation. By addressing these challenges, future research can refine current models, improve the accuracy of discourse relation identification, and contribute to a more nuanced understanding of implicit discourse. Ultimately, these advancements will not only boost model performance but also deepen our comprehension of the intricate ways in which discourse relations are conveyed in natural language.

Bibliography

[1] Boris Galitsky, Dmitry Ilvovsky, and Elizaveta Goncharova. Relying on discourse analysis to answer complex questions by neural machine reading comprehension. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 444–453, Held Online, September 2021. INCOMA Ltd.

[2] Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. Assessing the discourse factors that influence the quality of machine translation. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 283–288, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[3] Liat Ein-Dor, Ilya Shnayderman, Artem Spector, Lena Dankin, Ranit Aharonov, and Noam Slonim. Fortunately, discourse markers can enhance language models for sentiment analysis. *CoRR*, abs/2201.02026, 2022.

[4] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[5] Ruixue Tang, Yanping Chen, Ruizhang Huang, and Yongbin Qin. Enhancing interaction representation for joint entity and relation extraction. *Cognitive Systems Research*, 82:101153, 2023.

[6] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In Nicoletta

Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[7] Rashmi Prasad, Bonnie Webber, and Aravind Joshi. Reflections on the Penn Discourse TreeBank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950, December 2014.

[8] Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. Simple and effective multi-token completion from masked language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

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

[10] Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. CoNLL 2016 shared task on multilingual shallow discourse parsing. In Nianwen Xue, editor, *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany, August 2016. Association for Computational Linguistics.

[11] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[12] Haoran Li, Jiajun Zhang, and Chengqing Zong. Predicting implicit discourse relations with purely distributed representations. In Maosong Sun, Zhiyuan Liu, Min Zhang, and Yang Liu, editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 293–305, Cham, 2015. Springer International Publishing.

[13] Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In Philipp Koehn and Rada Mihalcea, editors, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore, August 2009. Association for Computational Linguistics.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[16] 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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[18] Wei Shi and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 5790–5796, Hong Kong, China, November 2019. Association for Computational Linguistics.

[19] Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France, May 2020. European Language Resources Association.

[20] Dan Jiang and Jin He. Tree framework with bert word embedding for the recognition of chinese implicit discourse relations. *IEEE Access*, 8:162004–162011, 2020.

[21] Haoran Li, Jiajun Zhang, and Chengqing Zong. Implicit discourse relation recognition for english and chinese with multiview modeling and effective representation learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 16(3), mar 2017.

[22] Feng Jiang, Peifeng Li, and Qiaoming Zhu. Recognizing chinese discourse relations based on multi-perspective and hierarchical modeling. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[23] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans, 2020.

[24] Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[25] Congcong Jiang, Tieyun Qian, Zhuang Chen, Kejian Tang, Shaohui Zhan, and Tao Zhan. Generating pseudo connectives with mlms for implicit discourse relation recognition. In Duc Nghia Pham, Thanaruk Theeramunkong, Guido Governatori, and Fenrong Liu, editors, *PRICAI 2021: Trends in Artificial Intelligence*, pages 113–126, Cham, 2021. Springer International Publishing.

[26] Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 368–375, USA, 2002. Association for Computational Linguistics.

[27] Attapol Rutherford and Nianwen Xue. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

[28] Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. Implicit discourse relation recognition by selecting typical training examples. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 2757–2772, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[29] Congcong Jiang, Tieyun Qian, and Bing Liu. Knowledge distillation for discourse relation analysis. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 210–214, New York, NY, USA, 2022. Association for Computing Machinery.

[30] Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2224, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[31] Changxing Wu, Xiaodong Shi, Yidong Chen, Yanzhou Huang, and Jinsong Su. Bilingually-constrained synthetic data for implicit discourse relation recognition. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312, Austin, Texas, November 2016. Association for Computational Linguistics.

[32] Yuping Zhou and Nianwen Xue. PDTB-style discourse annotation of Chinese text. In Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbae Lee, and Jong C. Park, editors, *Proceedings of the 50th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 69–77, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[33] Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. Using explicit discourse connectives in translation for implicit discourse relation classification. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

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

[35] Attapol Rutherford, Vera Demberg, and Nianwen Xue. A systematic study of neural discourse models for implicit discourse relation. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291, Valencia, Spain, April 2017. Association for Computational Linguistics.

[36] Manfred Stede, Tatjana Scheffler, and Amália Mendes. Connective-lex: A web-based multilingual lexical resource for connectives. *Discours*, 24, 10 2019.

[37] Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, abs/1506.06724, 2015.

[38] Yang Liu and Sujian Li. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233, Austin, Texas, November 2016. Association for Computational Linguistics.

[39] Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Implicit discourse relation detection via a deep architecture with gated relevance network. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 1726–1735, Berlin, Germany, August 2016. Association for Computational Linguistics.

[40] Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[41] Huibin Ruan, Yu Hong, Yang Xu, Zhen Huang, Guodong Zhou, and Min Zhang. Interactively-propagative attention learning for implicit discourse relation recognition. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3168–3178, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[42] Wei Xiang, Bang Wang, Lu Dai, and Yijun Mo. Encoding and fusing semantic connection and linguistic evidence for implicit discourse relation recognition. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[43] Wanqiu Long and Bonnie Webber. Facilitating contrastive learning of discourse relational senses by exploiting the hierarchy of sense relations, 2023.

Appendix A

Level-1 (class)	**Level-2 (type)**	**Level-3 (subtype)**
TEMPORAL	SYNCHRONOUS	-
	ASYNCHRONOUS	PRECEDENCE,
		SUCCESSION
CONTINGENCY	CAUSE	REASON,
		RESULT,
		NEGResult
	CAUSE+Belief	REASON+Belief,
		RESULT+Belief
	CAUSE+SpeechAct	REASON+SpeechAct,
		RESULT+SpeechAct
	CONDITION	ARG1-AS-COND,
		ARG2-AS-COND
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND,
		ARG2-AS-NEGCOND
	PURPOSE	ARG1-AS-GOAL,
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER,
		ARG2-AS-DENIER
	CONTRAST	-
	SIMILARITY	-
EXPANSION	CONJUNCTION	-
	DISJUNCTION	-
	EQUIVALENCE	-
	EXCEPTION	ARG1-AS-EXCEPT,
		ARG2-AS-EXCEPT
	INSTANTIATION	ARG1-AS-INSTANCE,
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL,
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER,
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST,
		ARG2-AS-SUBST

Table A.1: PDTB-3 Sense Hierarchy

Appendix B

A	B-F	G-I	J-Z
after all	but then	in addition	later on
along with	but then again	in any case	more accurately
and then	by comparison	in any event	no matter
as a result	by contrast	in contrast,	on the contrary
as an alternative	by the way	in essence	on the other hand
as well	by then	in fact	on the other
at that point	even before	in other words	quite the contrary
at that time	even before then	in particular	that is
at the same time	even then	in short	
at the time	for example	in sum	
	for instance	in the end	
	for one	in the meantime	
	for one thing	in the meanwhile	
		in this way	
		in turn	

Table B.1: Multi-token connectives organized by starting letters

Appendix C

Table C.1: Explicit Connectives and their Senses

Explicit Connectives	**Senses**
after all	Contingency.Cause+Belief.Reason+Belief, Expansion.Conjunction, Expansion.Level-of-detail.Arg2-as-detail
after that	Temporal.Asynchronous.Succession
along with	Expansion.Conjunction
and then	Expansion.Disjunction
as a result	Contingency.Cause.Result, Contingency.Cause+Belief.Result+Belief, Expansion.Level-of-detail.Arg2-as-detail
as an alternative	Expansion.Disjunction
as well	Comparison.Similarity, Expansion.Conjunction
at that point	Temporal.Synchronous
at that time	Temporal.Synchronous
at the same time	Temporal.Synchronous, Expansion.Conjunction
at the time	Temporal.Synchronous
but then again	Comparison.Concession.Arg2-as-denier
but then	Comparison.Concession.Arg2-as-denier

Table C.1 – continued from previous page

Explicit Connectives	**Senses**
by comparison	Comparison.Contrast, Comparison.Concession.Arg2-as-denier, Expansion.Conjunction
by contrast	Comparison.Contrast, Comparison.Concession.Arg2-as-denier
by the way	Comparison.Contrast, Expansion.Conjunction
by then	Temporal.Asynchronous.Succession|Contingency. Cause.Reason, Temporal.Asynchronous.Succession
even before then	Temporal.Asynchronous.Succession|Comparison. Concession.Arg2-as-denier
even before	Temporal.Asynchronous.Precedence|Comparison. Concession.Arg1-as-denier
even then	Temporal.Asynchronous.Precedence|Comparison. Concession.Arg2-as-denier
for example	Expansion.Instantiation.Arg2-as-instance, Contingency.Cause.Reason, Expansion.Level-of-detail.Arg2-as-detail
for instance	Expansion.Instantiation.Arg2-as-instance, Expansion.Conjunction, Expansion.Level-of-detail.Arg2-as-detail
for one thing	Expansion.Instantiation, Contingency.Cause.Reason, Expansion.Conjunction, Expansion.Instantiation.Arg2-as-instance, Expansion.Level-of-detail.Arg2-as-detail
for one	Expansion.Instantiation, Expansion.Instantiation.Arg2-as-instance
in addition	Expansion.Conjunction, Expansion.Level-of-detail.Arg2-as-detail
in any case	Comparison.Concession.Arg2-as-denier

Table C.1 – continued from previous page

Explicit Connectives	**Senses**
in any event	Expansion.Conjunction, Expansion.Level-of-detail.Arg1-as-detail
in contrast	Comparison.Contrast
in essence	Expansion.Conjunction
in fact	Comparison.Concession.Arg2-as-denier, Comparison.Contrast, Expansion.Conjunction, Expansion.Instantiation.Arg2-as-instance, Expansion.Level-of-detail.Arg1-as-detail, Expansion.Level-of-detail.Arg2-as-detail, Contingency.Cause+Belief.Reason+Belief, Contingency.Cause+Belief.Result+Belief, Contingency.Cause.Reason, Contingency.Cause.Result, Expansion.Equivalence
in other words	Expansion.Equivalence, Comparison.Similarity, Contingency.Cause.Reason, Contingency.Cause.Result, Expansion.Conjunction, Expansion.Level-of-detail.Arg1-as-detail, Expansion.Level-of-detail.Arg2-as-detail
in particular	Expansion.Instantiation.Arg2-as-instance, Expansion.Level-of-detail.Arg2-as-detail, Expansion.Conjunction
in short	Expansion.Level-of-detail.Arg1-as-detail, Contingency.Cause+SpeechAct.Result+SpeechAct, Contingency.Cause.Reason, Contingency.Cause.Result, Expansion.Conjunction, Expansion.Equivalence, Expansion.Level-of-detail.Arg2-as-detail

Table C.1 – continued from previous page

Explicit Connectives	**Senses**
in sum	Expansion.Level-of-detail.Arg1-as-detail, Expansion.Conjunction, Expansion.Equivalence, Expansion.Level-of-detail.Arg2-as-detail
in the end	Comparison.Concession.Arg2-as-denier, Comparison.Contrast, Contingency.Cause.Result, Expansion.Conjunction, Expansion.Level-of-detail.Arg1-as-detail, Expansion.Level-of-detail.Arg2-as-detail, Temporal.Asynchronous.Precedence, Expansion.Equivalence
in the meantime	Temporal.Asynchronous.Succession, Temporal.Synchronous—Comparison.Contrast, Temporal.Synchronous, Temporal.Synchronous
in the meanwhile	Temporal.Synchronous
in this way	Contingency.Cause.Result
in turn	Temporal.Asynchronous.Precedence, Contingency.Cause.Result, Expansion.Conjunction, Expansion.Level-of-detail, Temporal.Asynchronous
later on	Temporal.Asynchronous.Precedence
more accurately	Expansion.Substitution.Arg2-as-subst
no matter	Comparison.Concession.Arg1-as-denier
on the contrary	Comparison.Contrast, Expansion.Level-of-detail.Arg2-as-detail
on the other hand	Comparison.Concession.Arg2-as-denier, Comparison.Contrast
on the other	Comparison.Concession.Arg2-as-denier, Comparison.Contrast
quite the contrary	Expansion.Substitution

Table C.1 – continued from previous page

Explicit Connectives	**Senses**
that is	Expansion.Equivalence, Expansion.Level-of-detail.Arg2-as-detail, Contingency.Cause.Reason, Contingency.Cause.Result, Expansion.Conjunction, Expansion.Level-of-detail.Arg1-as-detail