THE UNIVERSITY *of* EDINBURGH

Enhancing Implicit Discourse Relation Recognition by Exploiting Label Inter-relations

Wanqiu Long

Doctor of Philosophy

UKRI CDT in Natural Language Processing

School of Informatics

University of Edinburgh

2025

Abstract

Implicit Discourse Relation Recognition (IDRR) is a fundamental yet challenging task in discourse parsing, as it involves identifying rhetorical, semantic and/or pragmatic relationships between text spans in the absence of explicit connectives such as "because" or "however". While recent advances leveraging pre-trained language models and prompt-based learning have improved performance, most existing approaches treat discourse sense labels as flat and independent categories. This neglects the rich structural information embedded in annotation frameworks like the Penn Discourse Treebank (PDTB), where discourse relations are organized hierarchically and can co-occur in some ways.

The central claim of this thesis is that structured groupings of discourse senses — as encoded in the sense hierarchy — can serve as an effective structural prior to guide model training, particularly by shaping how label distances are represented and learned. This thesis proposes methods to enhance IDRR by focusing on two kinds of label inter-relations: the hierarchical relations and co-occurrence-based label inter-relations. First, we introduce a contrastive learning framework that utilizes the PDTB sense hierarchy to guide the selection of semantically meaningful negative examples during training, thereby encouraging the model to learn finer-grained distinctions between closely related senses. Second, we integrate hierarchical information into a prompt-based learning paradigm through a prototype-based verbalizer, which aligns label representations with the sense hierarchy. This approach is further extended to support zero-shot cross-lingual IDRR, demonstrating effectiveness across both monolingual and cross-lingual scenarios. Third, we explore multi-label classification frameworks to handle cases where multiple discourse relations simultaneously hold between a single pair of text spans — an under-addressed yet prevalent phenomenon in real-world discourse. Incorporating hierarchical sense information also improves the accuracy of multi-label predictions.

Extensive experiments demonstrate the effectiveness of our approaches that consider the label inter-relations. The results show that explicitly modeling label hierarchies improves model performance in both single-label classification and multi-label classification scenarios. This work advances our understanding of how structural relationships between discourse relations can be effectively utilized in computational models, while also highlighting the importance of handling multi-label cases in discourse relation recognition.

Finally, this thesis outlines several promising directions for future work. One avenue is to extend the proposed approaches to broader datasets, including discourse annotations from alternative frameworks such as RST and eRST, as well as texts from diverse domains and languages, to better assess the generalizability of the methods. Another direction involves integrating argument span detection with discourse relation recognition into a unified framework, thereby advancing toward more realistic and end-to-end discourse parsing systems. Additionally, future work may explore guiding Large Language Models (LLMs) to better represent discourse relations and understand the label relationships between senses by leveraging the hierarchical organization of discourse senses. Together, these directions point to the broader goal of capturing the complexity of coherence more effectively in natural language.

Lay Summary

When we read or write, we need to understand how ideas connect — whether one sentence explains another, adds more information, or presents a contrast. These relationships between parts of a text help us follow the flow of discourse and make sense of what we read. For example, even without an explicit connective like "so", we can easily infer a causal relationship between the two sentences: "It was raining heavily" and "I decided to stay indoors". Teaching models to recognize such implicit connections is a difficult but important task in natural language processing.

This thesis focuses on improving how models recognize these discourse relations, especially in situations where no clear connectives like "because" or "however" are used to signal them. A key insight is that while discourse relations can be annotated independently, many discourse annotation frameworks organize them into structured systems, such as hierarchies that group related relations together. For instance, "Reason" and "Result" are grouped under the broader "Cause" relation in the PDTB framework.

By helping models understand and use this structure rather than treating each label as unrelated, we can improve their ability to correctly classify discourse relations. This thesis develops different approaches that take the structure of senses into account during training. The research also looks at cases where more than one relation holds at the same time — such as when a sentence both explains something and contrasts it with a previous sentence. Most prior systems ignore this complexity, but human writing often contains these multiple relations between clauses or sentences.

The work presented here improves models' performance in identifying discourse relations, which can be helpful for many real-world applications such as summarization, translation systems, and reading comprehension. By enabling models to better understand how text spans are connected, this research can contribute to building more coherent and context-aware language technologies.

Acknowledgements

I feel incredibly fortunate for the unwavering support I received throughout this journey.

First and foremost, I would like to express my deepest gratitude to my primary supervisor, Prof. Bonnie Webber. Her willingness to take me on as a PhD student opened opportunities I could never have imagined. Throughout my doctoral journey, her profound knowledge, sharp insights, and thoughtful guidance have been invaluable. She has consistently pushed me to think critically, and her support has left a lasting influence on both my research and the way I approach academic work.

I am also deeply thankful to my second supervisor, Dr. Siddharth Narayanaswamy, for his invaluable guidance through both technical challenges and broader research questions. I am greatly appreciative of Prof. Mark Steedman for his crucial feedback and direction during my first year. My sincere thanks also go to Prof. Frank Keller, Prof. Adam Lopez, and Prof. Hannah Rohde for their constructive comments during my annual reviews. I am especially grateful to my thesis examiners, Prof. Adam Lopez and Prof. Leila Kosseim, for their thoughtful and thorough review of my thesis.

A PhD is not only an academic pursuit but also a personal journey shaped by community. I am grateful to my 2020 cohort for the sense of belonging they brought. Special thanks to Matthias Lindemann and Amr Keleg for their thoughtful feedback on my work. To Verna Danker, Agostina Calabrese, Siqi Sun, Danyang Liu, Shangmin Guo and Zheng Zhao—thank you for your companionship. Your practical advice and emotional support meant more than words can express. I also appreciate Sally Galloway and Bjorn Ross for their administrative help and for the friendly conversations that made the department feel like home.

Despite the geographical distance, my best friend Yun Xia in China has remained a constant source of encouragement, reminding me that true friendship knows no boundaries. I am equally grateful to my family for their love and unwavering belief in me throughout these years. To my partner—your love, patience, and steady faith have been my anchor. You have shown me the importance of staying grounded and taking action rather than overthinking.

A special thanks goes to my father and grandfather. Though they are no longer with

me, their selfless love and strength continue to guide me. Their memory has been a constant source of light in moments of doubt, reminding me to trust myself and keep moving forward.

To everyone who walked beside me on this journey—thank you.

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.

(*Wanqiu Long*)

Table of Contents

Chapter 1

Introduction

Understanding texts, whether spoken or written, goes beyond interpreting individual sentences or clauses in isolation. A crucial component of comprehension is the ability to recognize how different parts of a text are connected to form a coherent whole. This coherence, often referred to as textual coherence, is established through various linguistic mechanisms such as lexical cohesion (word relationships and repetition) (Morris and Hirst, 1991), referential coherence (pronouns and references) (Garnham et al., 1982), and discourse structure as captured by discourse parsing frameworks (Biran and McKeown, 2015; Li and Huang, 2023; Chi and Rudnicky, 2022). While coherence is a property of texts, the ability to recognize and make use of it plays a central role in enabling NLP systems to achieve deeper text understanding.

Discourse parsing is the computational task of automatically identifying discourse structures within a text. It typically involves two key components: discourse segmentation (Sediqin and Argamon, 2025; Saveleva et al., 2021), which divides a text into meaningful discourse units, and discourse relation recognition, which determines the rhetorical, semantic or pragmatic relations between these units (Varachkina and Pannach, 2021; Metheniti et al., 2024; Bourgonje and Demberg, 2024). Within discourse parsing, discourse relation recognition plays a critical role in understanding how different discourse units interact. Discourse relations help establish coherence by indicating how one piece of information (expressed in paragraphs, sentences, clauses, or smaller units of discourse) relates to another. These relations explain how different parts of a text are connected, whether through causality, contrast, elaboration, or other mechanisms. For example,

(1) [Sears faces an especially daunting challenge on the eve of the Christmas shopping season]$_1$, [everyday pricing in the current environment doesn't work]$_2$.

In this example, the situation described by the clause labeled "2" provides an explanation for the situation described by the clause labeled "1". Discourse relations are relations that hold between textual spans, which capture essential aspects of the coherence of a text. These relations are essential not only for understanding discourse coherence, but also serve as valuable resources for applications in computational linguistics and natural language processing.

By facilitating a deeper understanding of how ideas are interconnected within texts, discourse relations empower various NLP (Natural Language Processing) tools and applications to perform more effectively and produce results that better understand the context and meaning. Automatically identifying the senses that hold between sentences, clauses or smaller discourse units can be useful for downstream NLP tasks such as text summarization (Pu et al., 2023; Rennard et al., 2024; Huang and Kurohashi, 2021), dialogue systems (Chen et al., 2023), question answering (Du et al., 2023) and event relation extraction (Tang et al., 2021).

Given the broad applicability and effectiveness of discourse relations across diverse NLP tasks, a solid foundation for understanding and modeling these relations is essential. These tasks are supported by well-established discourse annotation frameworks. The key discourse annotation frameworks include the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008; Webber et al., 2019), Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), Segmented Discourse Representation Theory (SDRT) (Sporleder and Lascarides, 2008), the Cognitive approach to Coherent Relations (CCR) (Sanders et al., 1992), and Questions Under Discussion (QUD) (Kuppevelt, 1995; Grindrod and Borg, 2019; Ko et al., 2022), among others.

Frameworks such as RST and PDTB categorize discourse relations into finer-grained groupings to capture nuanced semantic and pragmatic distinctions. For instance, within the original RST (Mann and Thompson, 1988), the broad category of "Cause" is further refined into subtypes including "Volitional Cause", "Non-Volitional Cause", "Volitional Result", and "Non-Volitional Result", highlighting distinctions in the presence of intentional action and whether the outcome was deliberately brought about. Similarly, PDTB-3 (Webber et al., 2019) also adopts a structured hierarchy for the "Cause" relation, distinguishing sister relations such as "Cause+Belief" and "Cause+SpeechAct" based on pragmatic nuances, and daughter relations like "Reason", "Result", and "Nega tive-Result" that represent more precise causal semantics.

These hierarchical distinctions can inform the design of computational models for Discourse Relation Recognition. While early methods use the label hierarchy simply to

indicate what sense labels were available, ignoring the use of the hierarchical relations between the sense labels of discourse relations, recent studies have highlighted the benefits of explicitly utilizing discourse relation hierarchies or groupings to enhance model performance. For example, Wu et al. (2022) demonstrated improvements by exploiting dependencies between coarse- and fine-grained discourse labels.

Furthermore, computational approaches to Implicit Discourse Relation Recognition (IDRR) have advanced significantly in recent years. Recent methods commonly use pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), which are fine-tuned on discourse relation datasets, often with additional modeling strategies to enhance performance. More recently, prompt-based methods have emerged as the new state-of-the-art. These methods reformulate the classification task as a cloze-style task by incorporating prompts with masked tokens, allowing the model to better exploit its pre-trained knowledge. Approaches such as ConnPrompt (Xiang et al., 2022b), PCP (Zhou et al., 2022), and DiscoPrompt (Chan et al., 2023) have demonstrated notable improvements in IDRR performance.

The central claim of this thesis is that structured groupings of discourse senses — as encoded in the sense hierarchy — can serve as an effective structural prior to guide model training, particularly by shaping how label distances are represented and learned. Leveraging these hierarchical label relations enables models to better distinguish implicit discourse relations, thereby improving both accuracy and robustness. This claim is substantiated by incorporating the PDTB sense hierarchy into standard fine-tuning and prompt-based learning paradigms, which systematically integrate label structure into the training process across different modeling settings. In addition, the thesis addresses an often-overlooked aspect of discourse annotation: the presence of instances assigned multiple discourse sense labels. Prior work has largely treated these as separate single-label examples during training, and evaluation typically considers a prediction correct if any of the gold labels are matched. However, this practice overlooks the interplay between multiple valid discourse relations and fails to distinguish between instances where multiple senses co-occur versus those where only one applies. To address this gap, the thesis demonstrates that adopting multi-label classification frameworks provides a more faithful treatment of such instances. Furthermore, incorporating the PDTB sense hierarchy within these frameworks leads to measurable improvements in model performance.

1.1 Motivation

Discourse annotation frameworks such as PDTB-3 and RST rely on sense labels to characterize how coherence is realized in text. However, these sense labels are far from discrete or mutually independent categories. Instead, they exhibit definitional inter-dependencies, hierarchical organization, and patterns of co-occurrence, etc., all of which contribute to the complexity of discourse coherence. In addition, Stede et al. (2019) introduce lexicons of discourse connectives for various languages such as English (Das et al., 2018), German (Bourgonje and Stede, 2020), Chinese (Wan et al., 2024) and Dutch (Bourgonje et al., 2018), highlighting how connectives can share similar meanings yet signal different discourse relations. These patterns of semantic overlap and functional differentiation among connectives also reflect a key dimension of the broader network of inter-relations in discourse semantics.

For instance, in PDTB-3, the causal labels "Reason" and "Result" are definitionally interlinked: a reason is the antecedent condition leading to an event or state, whereas a result is the subsequent event or state triggered by that cause. Likewise, "Contrast" and "Similarity" form a conceptual opposition, with the former highlighting differences and the latter emphasizing commonalities. Moreover, certain relations in PDTB-3, such as "Asynchronous.Precedence" and "Asynchronous.Succession", are inherently directional: if one event precedes another, reversing that relationship is not logically valid. These multidimensional connections constitute the theoretical basis of discourse annotation frameworks, shaping annotation guidelines and guiding computational modeling.

In our work, we focus on those discourse relations expressed implicitly where explicit connectives (like "because" or "however") or other expressions that can indicate the relations (like "the reason is" or "leading to") are absent. Models must uncover the underlying relation from the interactions of the two discourse units, making Implicit Discourse Relation Recognition (IDRR) especially challenging. Although recent advances in neural networks and pretrained language models have improved IDRR performance, two kinds of label inter-relations remain insufficiently exploited: hierarchical label relations and co-occurrence label relations.

Hierarchical label relations are constructed by annotators as an artificial classification mechanism. The hierarchical structure within these frameworks further organizes sense labels from broad categories to increasingly finer-grained subtypes. Annotators may begin with a coarse-grained category (e.g., "Expansion"), then specify subtypes

like “Equivalence” or “Instantiation”. Such hierarchical classification not only aids annotation consistency by reducing ambiguity and guiding label choices, but also offers analysts the flexibility to annotate at varying levels of detail.

Not all discourse annotation frameworks adopt a hierarchical labeling structure. While some utilize a flat, non-hierarchical approach, the hierarchical organization reveals the label relationship between the sense labels. Most fundamentally, there are strict subsumption relations where higher-level categories encompass their lower-level subcategories. Second, labels under the same top level category often share fundamental characteristics but differ in fine-grained functions. For example, “Condition” and “Otherwise” in RST are grouped under the same broader category. However, despite them being sisters, they are less likely to be compatible and they rarely co-occur between the same discourse units. This is because the two relations often serve mutually exclusive functions in a given context: “Condition” typically sets up a hypothetical scenario or prerequisite, while “Otherwise” presents an alternative outcome in the absence of that condition. Therefore, the hierarchical organization of senses can indicate the relational “distance” between sense labels.

Meanwhile, it is not uncommon for multiple discourse relations to exist between two discourse units. This is because, from different interpretive perspectives, multiple relations can be identified between the same pair of discourse units. For example, in PDTB-3, labels such as “Expansion.Manner” and “Contingency.Purpose” may frequently co-occur, as the way something is done (“Manner”) is often closely related to why it is done (“Purpose”). In contrast, “Expansion.Equivalence” and “Expansion.Instantiation” are much less likely to co-occur, as they serve different discourse functions: “Equivalence” emphasizes similarity or rephrasing, whereas “Instantiation” moves from general to specific.

While the annotation of multiple discourse relations is not universally supported across frameworks, the presence of multiple discourse relations reflects the inherent complexity of discourse relations. Capturing this complexity is beneficial for accurate discourse understanding and analysis, which in turn can support downstream tasks. It also facilitates the model to learn both discourse relations and the relationship between different discourse relations.

By exploiting the hierarchical relations between the senses and developing frameworks that support multiple concurrent relations, the computational models can better learn both individual discourse relations and their inter-relationships, thus promoting deeper discourse understanding and more robust downstream task performance. Nev-

ertheless, the majority of previous approaches overlook these hierarchical and multi-label concurrent relations, treating each label as an isolated category.

Prior to our work, only a few studies, such as Wu et al. (2022), have demonstrated performance gains by explicitly exploiting coarse- and fine-grained label dependencies, highlighting the untapped potential of modeling these relationships more systematically. Moreover, for those instances with multiple annotated labels, previous work on discourse relation recognition treat them as separate and different examples during training, and at test time, a prediction matching one of the gold types is taken as the correct answer. Treating multi-label discourse instances as separate examples risks oversimplifying the complexity of discourse relations. Real-world texts often involve multiple overlapping relations, and separating them may obscure how these relations interact. This approach can result in loss of contextual information, introduce training ambiguity, and hinder generalization.

This thesis presents an empirical investigation into how the inter-relations among sense labels—especially hierarchical structures and co-occurrence relations—can be considered to improve Implicit Discourse Relation Recognition (IDRR). We hypothesize that explicitly exploiting the label inter-relations can lead to better predictions across label granularity, and a better reflection of the complex semantics involved in discourse understanding. To test this, we propose a series of methods grounded in modern neural architectures, including contrastive learning, prompt-based learning, and multi-label classification, incorporating hierarchical label information.

We apply our approaches only to implicit discourse relation because implicit relations are more difficult to predict and thus stand to benefit most from the additional structure provided by the sense hierarchy. Explicit relations are typically easier to classify due to the presence of connectives, but the label hierarchy could still help disambiguate senses, though likely to a lesser extent, and it would be interesting future work to explore these benefits systematically.

Our experiments demonstrate that explicitly modeling the inter-relations among labels improves classification accuracy in both single-label and multi-label settings for IDRR. These improvements hold across monolingual and cross-lingual scenarios. We also outline future extensions to explicit relations and other frameworks such as RST and eRST.

1.2 Research Questions

Based on the observations discussed in Section of Motivation, we consider two core hypotheses.

- **Further exploiting the label inter-relations can enhance sense labeling for IDRR with the PDTB Sense Hierarchy in a single-label classification framework.**
  We demonstrate how the PDTB sense hierarchy can guide model learning in single-label settings by controlling the distance between senses during training, improving performance on implicit discourse relation recognition (IDRR).

- **Multi-label classification methods are feasible for IDRR, and incorporating label inter-relations further improves sense labeling.**
  We examine multiple multi-label classification approaches and enhance them with inter-sense hierarchical information, enabling the model to better distinguish fine-grained categories while capturing co-occurrence patterns.

1.3 Thesis Outline

The thesis is structured as follows:

In Chapter 2, we will provide an overview of discourse relations and the frameworks used for their annotation. The chapter first introduces the concept of discourse relations, discusses different theoretical perspectives, and compares various discourse annotation frameworks in terms of the types of discourse relations, their structures that organize senses and multi-sense annotation schemes. Then we discuss why we select PDTB as the sole focus for our experiments. The chapter also gives details for PDTB discourse framework and reviews existing approaches for modeling implicit discourse relations in PDTB. It summarizes the progress as well as the limitations of the methods.

Chapter 3 explores how incorporating the PDTB sense hierarchy into fine-tuning can improve implicit discourse relation recognition. It introduces a contrastive learning framework designed to enhance sense labeling by leveraging hierarchical label interrelations. The methodology involves sentence encoding, data augmentation, and structured negative pair selection. The experimental setup evaluates the proposed approach against baseline models and alternative methods for selecting negative examples. Experimental results demonstrate the effectiveness of contrastive learning which makes

use of the sense hierarchy in refining discourse relation classification. The chapter concludes with an ablation study to analyze the contribution of different components.

Chapter 4 first discusses the motivation behind using prompt-based learning to improve implicit discourse relation recognition (IDRR) by leveraging the PDTB sense hierarchy. Then, this chapter introduces prototype learning and a hierarchical prototype-based verbalizer, which align discourse relation representations with the PDTB hierarchy. Additionally, it presents a cross-lingual prototype transfer mechanism for zero-shot IDRR, enabling better generalization across languages without requiring extensive labeled data. The experimental setup evaluates performance on both monolingual and zero-shot transfer scenarios, followed by an in-depth analysis of learned prototypes and an ablation study to assess the contributions of different components.

Chapter 5 explores the use of multi-label classification for implicit discourse relation recognition (IDRR) and examines whether incorporating the PDTB sense hierarchy can improve sense labeling for multi-label classification method. It introduces different multi-label classification methods and explores strategies for leveraging hierarchical label relationships. The experimental results show that multi-label classification methods do not compromise the performance for single-label prediction while being able to predict multi-labels. Moreover, applying the PDTB sense hierarchy can help improve sense labeling for multi-label classification methods.

While Chapters 3-5 are based on previously published work, here we present further detail and more extensive analysis. Chapter 6 discusses the limitations of the proposed methods, focusing on data constraints and methodological challenges. For future work, it suggests extending these methods to a broader range of datasets, including those from different annotation frameworks, genres, and languages, to improve generalization. Furthermore, rather than only focusing on implicit discourse relations, applying the methods to other types of discourse relations such as explicit discourse relations could further validate their effectiveness. To further improve the model, it is essential to explore more advanced models and techniques, particularly to address unresolved challenges such as data imbalance, the difficulty of distinguishing multi-label examples from single-label ones, and accurately identifying multiple discourse labels within a given instance.

Chapter 2

Background

This chapter surveys leading theories and annotation frameworks for discourse relations and explains how they inform the modeling choices in this thesis. We first introduce major theories of discourse (RST, SDRT, CCR, D-LTAG, QUD, and eRST) and compare what each represent with respect to structure, signaling, and multiplicity of senses. We then justify our choice of the Penn Discourse Treebank (PDTB) as the annotation framework for our experiments on both single-label and multi-label implicit relation recognition. Throughout, we highlight how insights from other discourse annotation frameworks shape our modeling decisions (e.g., the types of discourse relations, the structure that organizes the senses, multi-sense annotation, and other pragmatic dimensions). Finally, we focus on discussing the Penn Discourse Treebank (PDTB) annotation scheme and review existing computational approaches for implicit discourse relation recognition in PDTB. This background provides the context for our work on enhancing discourse relation recognition by exploiting label inter-relations in both single-label and multi-label classification scenarios.

2.1 Theories of Discourse Relations

Understanding how textual segments connect to form coherent meaning is a central concern in discourse analysis. Prominent theories differ in what they treat as the core representation of coherence (hierarchical structure, dialogic intent, cognitive dimensions, or connective-anchored composition) and therefore make different commitments about what counts as a relation and how it should be modeled. Below we briefly introduce six influential approaches.

2.1.1 Rhetorical Structure Theory

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) posits that coherence in monologic text (i.e., text produced by a single author, such as essays or articles, as opposed to dialogic text involving conversational exchange) arises from rhetorical relations between minimal units of discourse. These relations reflect how one span (the satellite) supports, elaborates, or explains another (the nucleus), capturing both content and communicative intent.

To represent these rhetorical dependencies, RST adopts a hierarchical tree structure, where leaf nodes correspond to elementary discourse units (EDUs), and internal nodes represent higher-level spans connected by rhetorical relations. The notion of nuclearity encodes the relative importance of each span. While many relations are asymmetric, RST also accounts for multinuclear structures, in which all spans are of equal importance. This tree-based structure enables clear, interpretable annotations and has been widely applied to discourse corpora in English (Carlson et al., 2001), Spanish (da Cunha et al., 2011), and Chinese (Peng et al., 2022). A single primary tree, however, constrains how parallel or cross-cutting relations can be represented, since each discourse unit must belong to exactly one node in the tree, making it difficult to capture co-existing relations.

2.1.2 Segmented Discourse Representation Theory

Segmented Discourse Representation Theory (SDRT) (Sporleder and Lascarides, 2008) views discourse as a dynamic, dialogic process (i.e., an interaction involving multiple speakers, as opposed to monologic text authored by a single writer) in which coherence is constructed through relations between speech acts, which are communicative actions such as assertions, questions, or requests performed by a speaker. These relations, such as “Explanation” or “Contrast”, encode pragmatic functions and are sensitive to speaker intention and discourse context.

To model this complexity, SDRT uses directed acyclic graphs (DAGs), which allow for non-hierarchical, overlapping, and long-distance connections between discourse units. This structure supports the representation of relations that cannot be easily captured in tree form, such as shared support or parallel dependencies. By integrating dynamic semantics and formal pragmatics, SDRT offers a flexible and expressive framework for representing richly connected discourse structures.

2.1.3 Cognitive Approach to Coherence Relations

The Cognitive approach to Coherence Relations (CCR) (Sanders et al., 1992) seeks to explain how readers mentally process and classify discourse relations. Rather than starting from structural assumptions, CCR defines relations in terms of cognitive primitives, including polarity (positive/negative), basic operation (causal/additive), source of coherence (objective/subjective), and segment order (basic/non-basic). An optional fifth dimension, temporality, may also be considered. For example, in analyzing a causal relationship, CCR would examine whether it is positive (expected result) or negative (contrary-to-expectation), whether it represents a strong causal link or a weaker additive one, whether it is based on real-world causality (objective) or speaker reasoning (subjective), and whether the cause precedes or follows the effect in the text.

This feature-based classification system highlights the psychological plausibility of coherence judgments. Unlike RST or SDRT, CCR does not enforce a fixed representational format or hierarchy, focusing instead on the mental processes underlying discourse interpretation.

2.1.4 Discourse Lexicalized Tree-Adjoining Grammar

Discourse Lexicalized Tree-Adjoining Grammar (D-LTAG) (Forbes-Riley et al., 2003) extends the formalism of lexicalized Tree-Adjoining Grammar (Webber and Joshi, 1998) to the discourse level. It conceptualizes discourse connectives as predicates and clauses as their arguments, capturing predicate-argument structures beyond the sentence level. This view grounds explicit signaling (e.g., *because*, *however*) in a predicate–argument formalism and underlies the connective-centered perspective of the PDTB (Prasad et al., 2008; Webber et al., 2019).

In D-LTAG, discourse structures are built using elementary trees, anchored by lexical items such as such as “because” or “then”. These trees capture basic discourse relations (e.g., “Causality”) and can also be combined to extend or refine existing structures, allowing discourse to be represented in a flexible yet systematic way.

2.1.5 Question Under Discussion Theory

The Question Under Discussion (QUD) framework (Kuppevelt, 1995) models discourse as a sequence of implicit or explicit questions and their answers. Each segment of text contributes to resolving an open question, guiding the flow and relevance of

discourse content.

Rather than specifying structural constraints, QUD theories emphasize discourse intent and information structure. For instance, an "Elaboration" can be viewed as answering "How?" or "In what way?", while a "Contrast" answers "What are the differences?" This perspective has been influential in computational dialogue modeling and coherence assessment, offering a functional view of relation types grounded in inferable discourse goals.

2.1.6 Enhanced Rhetorical Structure Theory

Enhanced Rhetorical Structure Theory (eRST) (Zeldes et al., 2025) builds on the foundation of RST and retains hierarchical organization to represent discourse prominence but introduces secondary edges that allow multiple overlapping relations between spans, improving the ability to model complex discourse functions. Importantly, eRST also explicitly anchors linguistic signals that mark discourse relations, such as connectives, lexical cohesion, and syntactic patterns. This facilitates more systematic and signal-driven annotation and modeling.

Taken together, each of these theories has contributed valuable insights to the study of discourse, offering different perspectives on how text segments connect and relate to each other. RST (and eRST) employs a hierarchical tree structure to represent discourse relations; accordingly, this thesis adopts a similar notion of hierarchical organization to structure the PDTB sense space (Level-1 $\rightarrow$ Level-2 $\rightarrow$ Level-3). SDRT's allowance for multiple concurrent relations motivates the use of multi-label settings and co-occurrence modeling; CCR's cognitively motivated dimensions (polarity, causality, subjectivity, order) align with PDTB-3's belief/speech-act tags; QUD's goal-oriented view informs the treatment of concurrent answers to different informational needs; D-LTAG's predicate–argument treatment of connectives provides a basis for handling explicit versus implicit signaling; and eRST's secondary edges and signal anchoring motivate the exploration of multi-label settings and the focus on implicit relations, where relations must be inferred without explicit signals. Together, these insights situate the present work within a broader landscape of discourse theories, showing how multiple traditions converge to shape the modeling decisions in this thesis.

Aspect	RST	SDRT	PDTB	QUD	CCR	eRST
Relation Types	Not distinguished	Not distinguished	Explicit, Implicit, AltLex, EntRel, etc.	Not distinguished	Not distinguished	Explicit, Implicit, AltLex
Multi-Sense Support	No	Yes	Yes	Yes	Yes	Yes
Sense Hierarchy	Two-level hierarchy	Flat	Three-level hierarchy	No	Flat	Two-level hierarchy

Table 2.1: Comparison of Discourse Relation Frameworks

2.2 Comparative Analysis of Discourse Annotation Frameworks

Discourse annotation presents challenges that go beyond simply recording surface features of text, as it is guided by theories about how discourse elements relate to each other. Rather than uncovering relations independently, annotation encodes the types of connections that a given theory defines between parts of a text. It thus reflects how the applied framework interprets the construction of meaning, coherence, or structure, instead of relying on isolated observations. Building upon foundational theories of discourse relations (Section 2.1), researchers have developed various annotation frameworks to apply these theoretical constructs in practice and capture the complexity of discourse structure.

We compare frameworks along three key aspects that affect modeling and evaluation: the types of discourse relations, the structural organization of sense labels, and the support for multi-sense annotation. Table 2.1 provides a summary comparison, and further details are presented in the following subsections. Finally, we discuss why we select PDTB as the framework for our experiments.

2.2.1 Types of Discourse Relations

Regarding the distinction of discourse relation types, frameworks like RST, SDRT, QUD, and CCR do not explicitly categorize discourse relations based on their signaling mechanisms. In contrast, the Penn Discourse Treebank (both PDTB-2 and PDTB-3) and Enhanced Rhetorical Structure Theory (eRST) stand out for their emphasis on how discourse relations are expressed in text. Both PDTB-2 and PDTB-3 categorize relations into Explicit (signaled by connectives like “because”, “however”) and Im-

plicit (where connectives are absent but could be inferred). Additionally, PDTB-2 introduces categories like AltLex (Alternative Lexicalization), which captures relations signaled by alternative ways of expressing discourse relations (e.g., "The reason for this is...") (Prasad et al., 2010). These distinctions, further expanded in PDTB-3 with categories like AltLex-C (ways of expressing discourse relations with lexical-syntactic constructions) and Hypophora (Question-Answer pairs), provide a fine-grained analysis of discourse relations, particularly valuable for computational and language processing tasks. Table 2.2 shows the distribution of the relations in the PDTB-2 and the PDTB-3. We can see that implicit discourse relations account for a substantial portion of annotated discourse relations: approximately 39% in PDTB-2 and 40% in PDTB-3. Furthermore, compared with PDTB-2, PDTB-3 shows a substantial increase in the number of annotated implicit instances (from 16,047 to 21,732).

While RST does not explicitly distinguish between relations signaled by connectives and those conveyed implicitly, eRST offers a more comprehensive approach by supporting a wide variety of discourse relations and integrating insights from both RST and PDTB. Combining RST's hierarchical structure with PDTB-like attention to signals, eRST expands the concept of discourse signals beyond traditional connectives.

Drawing on the perspective of the RST Signaling Corpus (Das and Taboada, 2017), eRST aims for a more exhaustive inventory of discourse relation signals, directly anchoring them to relevant tokens in the text. This inventory includes but goes beyond signal types corresponding to PDTB's AltLex, which capture alternative lexicalizations of discourse relations. By integrating these diverse signal types, eRST can effectively distinguishes between explicit, implicit, and AltLex relations.

2.2.2 Structure that Organizes Senses

The structure that organizes sense labels plays a crucial role in both manual annotation and computational discourse analysis. RST and PDTB provide a structured framework that helps annotators choose appropriate relations consistently, ensuring better agreement and interpretability. Moreover, such sense hierarchy facilitates machine learning models in capturing relation similarities. Besides, fine-grained senses contribute to a more refined differentiation between various discourse relations, capturing fine-grained distinctions that may otherwise be overlooked in a simpler classification system.

Different discourse frameworks vary in how they organize discourse relation labels.

RelType	PDTB-2	PDTB-3
Explicit	18,452	24,235
Implicit	16,047	21,732
AltLex	785	1,497
AltLexC	–	135
EntRel	5,208	5,530
Hypophora	–	146
NoRel	252	283
Total	40,974	54,187

Table 2.2: Distribution of Relation Types in PDTB-2 and PDTB-3

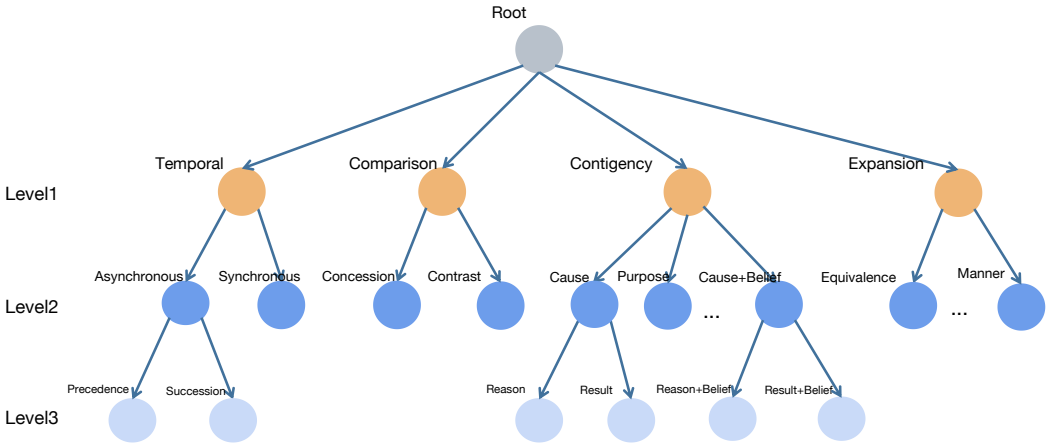


Figure 2.1: The PDTB-3 Sense Hierarchy

PDTB employs a three-level hierarchy, classifying relations into classes, types, and subtypes (see Figure 2.1). This structure was originally designed to allow annotators' flexibility in specifying relations at different levels of granularity. Over time, it became more of a guideline to aid annotators in selecting appropriate sense relations.

Beyond semantic relations which describe the logical or content-based connections between discourse segments (e.g., causal, temporal, or contrastive links), PDTB also accounts for pragmatic relations, which capture the speaker's communicative intentions or attitudes toward the content, through specialized pragmatic sense tags. While PDTB-2 introduced pragmatic variants of relations such as "Pragmatic Cause", "Pragmatic Condition", "Pragmatic Contrast", and "Pragmatic Concession" to capture discourse relations, PDTB-3 adopted a more fine-grained approach. It replaces these

RST Relations	
Circumstance	**Antithesis and Concession**
Solutionhood	Antithesis
Elaboration	Concession
Background	**Condition and Otherwise**
Enablement and Motivation	Condition
Enablement	Otherwise
Motivation	**Interpretation and Evaluation**
Evidence and Justify	Interpretation
Evidence	Evaluation
Justify	**Restatement and Summary**
Relations of Cause	Restatement
Volitional Cause	Summary
Non-Volitional Cause	**Other Relations**
Volitional Result	Sequence
Non-Volitional Result	Contrast
Purpose	

Table 2.3: Label Hierarchy from original RST (Mann and Thompson, 1988)

with explicit annotations that indicate whether a belief state—the speaker's mental attitude or degree of commitment toward a proposition—or a speech act—an utterance intended to perform a communicative function such as asserting, questioning, or requesting—is associated with either argument of a relation. For instance, relations such as "Cause+Belief" or "Concession+SpeechAct" signal that the interpretation involves either the speaker's belief or a communicative intention linked to one of the arguments. By integrating both semantic and pragmatic distinctions, the hierarchical label systems in PDTB offer a comprehensive framework for discourse analysis and ensure a more detailed and contextually aware representation of discourse relations.

The RST label sense hierarchy generally comprises two levels: a higher-level category of rhetorical relations and finer-grained subcategories. In the original RST taxonomy (Table 2.3), some labels have finer-grained subcategories. For instance, "Cause" is further divided into "Volitional Cause", "Non-Volitional Cause", "Volitional Result", "Non-Volitional Result". This fine-grained level can capture more distinctions between different types of relations.

The Enhanced Rhetorical Structure Theory (eRST) adopts a hierarchical label tax-

onomy that is directly derived from the Georgetown University Multilayer (GUM) corpus (Zeldes, 2017). This inventory includes 32 labels, which are used to define and categorize discourse relations. The GUM corpus's labeling system, as utilized by eRST, features both fine-grained and coarse-grained sense types.

With regard to SDRT and CCR, the relations are typically represented as a flat set rather than being organized into a hierarchical structure like in RST or PDTB. SDRT primarily annotates semantic and pragmatic discourse relations. In SDRT, relations such as "Narration", "Contrast", and "Explanation" are not further categorized into sub-types. Similarly, CCR adopts a flat structure by modeling discourse relations through binary cognitive features, such as causal/additive or positive/negative.

As for QUD, it does not employ traditional discourse relation labels. Instead, discourse relations in QUD are expressed as free-form questions rather than predefined relation label. QUD structures are flexible and do not rely on a fixed taxonomy of relations, focusing instead on how each utterance addresses or raises questions in the discourse context. However, there are some standard questions that one always finds in a QUD approach, including Explanation questions (e.g., "Why?"), which seek reasons or motivations; Clarification or Specification questions (e.g., "Which one?" or "Who exactly?"), which request more precise information; and Contrast questions (e.g., "What's the difference?"), which aims to highlight the distinctions between two entities or events. These recurring question types help shape the flow of information and coherence in discourse.

2.2.3 Multi-Sense Annotation

Discourse annotation frameworks vary in their treatment of multiple relations between the same discourse units. While some frameworks impose a strict single-relation annotation, others allow multiple relations to coexist, capturing different dimensions of meaning within the same discourse span. Furthermore, frameworks that support multi-label discourse relations adopt different annotation strategies. Understanding these distinctions is essential for improving discourse modeling and enhancing the applicability of discourse frameworks in linguistic and computational tasks.

2.2.3.1 Multi-Sense Annotation in PDTB

With regard to PDTB annotation, annotators can assign multiple relations to an example when they believe they all hold simultaneously. In PDTB-3, such multiple-label

annotations occur in approximately 4.8% of the relations. The following example for multiple label annotation for PDTB is from PDTB-3.

(2) [In the past decade, Japanese manufacturers concentrated on domestic production for export]. [In the 1990s, spurred by rising labor costs and the strong yen, these companies will increasingly turn themselves into multinationals with plants around the world].

This example demonstrates the simultaneous presence of two distinct implicit discourse relations. The annotator wanted to capture the sense that the first sentence described an earlier time period and the second, a subsequent time period. But the annotator also wanted to capture the sense that despite Japan's concentration on domestic production for export in the earlier time period, the future will be different – i.e. conceding what was true in the 1980s, and asserting that the future will be different. Therefore, the pair of arguments in the example is linked by two discourse relations that capture different dimensions of the relationship between them. Understanding these complex relations is crucial for the interpretation of discourse.

Apart from identifying standalone multiple implicit/explicit discourse relations in the PDTB dataset, annotators were also allowed to indicate implicit relations that occur alongside explicit or AltLex relations (Rohde et al., 2018). Illustrative instances of such cases are presented in Example (3) and Example (4).

(3) [The company posted record profits last year], [but it plans to lay off 1,000 employees next quarter].

Example (3) expresses an explicit "Concession" relation via the connective "but", highlighting the unexpected shift from strong financial performance to future layoffs. At the same time, the event sequence implies a "Temporal.Asynchronous.Precedence" relation which indicates that the event in the first argument happened before the event in the second argument, where the record profits precede the planned layoffs in the example. In the PDTB-3, such examples are annotated with both "Comparison.Concession .Arg2-as-denier" and "Temporal.Asynchronous.Precedence", capturing both the logical and temporal dimensions of the relation. The dot-separated naming convention used here reflects the hierarchical organization of the sense inventory: the first component denotes a top-level class (e.g., "Temporal" or "Comparison"), the second narrows the subtype (e.g., "Asynchronous" vs. "Synchronous"), and the third specifies the most fine-grained sense (e.g., "Precedence" vs. "Succession").

(4) [She didn't respond to any of our follow-up messages], [this led us to believe that she was no longer interested in the position].

Example (4) demonstrates the co-occurrence of an AltLex and an implicit discourse relation. The phrase "this led us to" functions as an AltLex—an alternative lexicalization of an explicit causal connective such as "therefore" or "as a result". In examples like "She didn't respond to any of our follow-up messages, this led us to believe that "she was no longer interested", annotators may infer a "Contingency.Cause+Belief" relation, where the second span conveys the speaker's belief or interpretation grounded in the preceding situation. At the same time, the event sequence also licenses a "Temporal.Asynchronous.Precedence" relation, with the initial event (lack of response) preceding the inferred belief or conclusion.

The two examples thus also show that co-occurring relations may be (i) multiple semantic relations describing distinct semantic dimensions in Example (3), or (ii) a mix of semantic and pragmatic relations when one link reflects an explicit, logical connection and the other is inferred from speaker intent as in Example (4), "Temporal.Asynchronous.Precedence" (semantic) alongside annotators may infer a "Contingency.Cause+Belief"" (pragmatic).

2.2.3.2 Multi-Sense Representation in SDRT

As for SDRT, it allows for multiple relations between discourse units, as illustrated in Example (5) from (Sporleder and Lascarides, 2008). This example is similar to Examples (2), (3), and (4), all of which involve multi-sense interpretations between discourse units.

(5) π_1: John bought an apartment.
π_2: But he rented it (out to others).

In this example, unit π_2 is posited to have both a "Narration" relation and a "Contrast" relation with π_1. SDRT distinguishes between coordinating (e.g., "CONTRAST") and subordinating (e.g., "ELABORATION") discourse relations, but these relations can coexist within the same discourse unit.

2.2.3.3 Expanding RST: Multi-Sense Annotation in eRST

RST limits the relationship between two text spans (or units) to a single relation. In cases where multiple relations seem applicable, the annotator is expected to select

the most likely or appropriate one (Taboada and Mann, 2006). When annotators encounter text spans that could potentially be linked by multiple relations, they must engage in a systematic selection process to identify the most plausible relation based on several key criteria: the writer's primary communicative goal, contextual support from surrounding text, structural coherence within the broader discourse, and explicit linguistic markers (Mann et al., 1989). However, some researchers have pointed out that this assumption of a single dominant relation may oversimplify naturally occurring discourse. In particular, Liu et al. (2023) suggest that certain RST parser outputs, previously considered errors, may in fact reflect concurrent relations, the instances where multiple discourse relations are simultaneously involved between the same pair of spans. These concurrent relations are not captured in traditional gold-standard trees, which enforce a single-label constraint, highlighting a potential limitation of the standard RST annotation framework.

Unlike traditional RST, eRST enables multiple relations to hold between the same nodes in the tree. Figure 2.2 illustrates how eRST captures discourse relations beyond the primary rhetorical structure. In the primary tree, [23–27] is linked to [22] via an "Evaluation.Comment" relation, following the RST hierarchy where [22] is the nucleus. However, the discourse connective "but" at the start of [23] signals a concession relation that is not captured in the primary tree. To address this, eRST adds a secondary edge (in blue), linking [22] and [23–24] to encode the concession. This example highlights how secondary edges allow eRST to represent additional semantic or pragmatic relations such as "Concession" that the primary tree misses. The primary tree in eRST follows the structure of Rhetorical Structure Theory (RST) and maintains a rhetorical relation framework. This means it emphasizes hierarchical nuclearity, where discourse units are categorized into nuclei and satellites. Secondary edges are permitted to connect any two nodes in the tree and introduce relations that break the hierarchical tree constraints. These relations are often motivated by semantic or pragmatic considerations.

2.2.3.4 Multi-Sense Support in CCR and QUD

CCR's attribute-based system inherently supports multi-label discourse relations by decomposing them into multiple co-occurring cognitive attributes. For instance, in CCR, a "Concession" relation (e.g., "It was raining, but they still went for a walk.") is characterized as a negative causal relation (denial of expectation), while also exhibiting an objective source of coherence (as it describes facts rather than subjective reasoning).

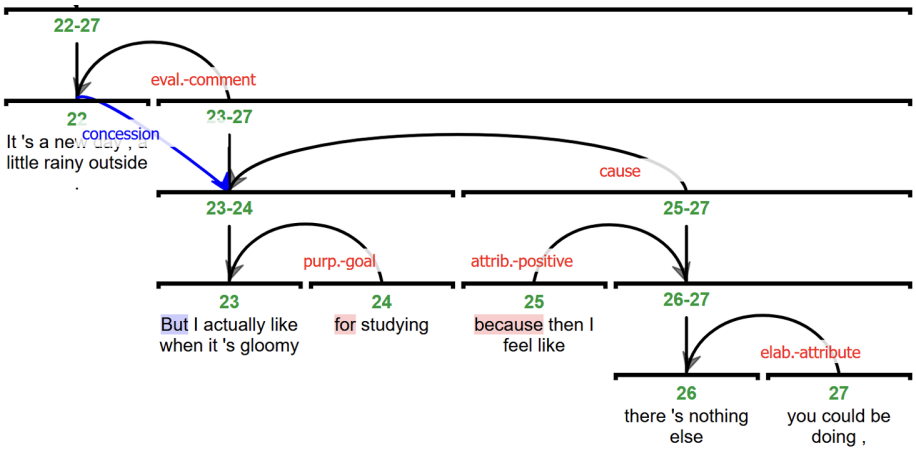


Figure 2.2: Illustration of primary tree and secondary edge in eRST (adapted from eRST (Zeldes et al., 2025))

Additionally, such relations may carry a temporal aspect if they indicate chronological progression. This approach aligns with multi-sense annotation frameworks like eRST and PDTB, where discourse units may simultaneously participate in different relation types.

As for QUD, Ko et al. (2023) find that sometimes a question could be interpreted in terms of different RST relations. Take the following example from (Ko et al., 2023),

> (6) **[anchor]** According to a preliminary National Weather Service summary, Monday's tornado was a top-end EF5, with top winds of 200 to 210 miles per hour (mph), and was 1.3 miles wide.
>
> **[QUD]** How long did the tornado last?
>
> **[answer]** It was tracked on the ground for 50 minutes - an eternity for a tornado - and its damage zone is more than 17 miles wide.
>
> **RST labels that could work:** "Evidence", "Proportion", "Elaboration.Additional", "Manner".

While the first clause in the sentences of answer provides a direct response to the generated question, the second clause responds to a different question. Ko et al. (2023) point out that this single question-answer pair can be analyzed through multiple RST relations as defined in the fine-grained relation set of the RST Discourse Treebank (Carlson et al., 2001, 2002): "Evidence" (supporting the tornado's intensity), "Proportion" (contextualizing the duration), "Elaboration.Additional" (providing new information), and "Manner" (describing how the event unfolded).

The above comparison highlights what we can consider in selecting an IDRR framework: (i) explicit treatment of signaling (explicit, implicit, AltLex)(ii) a hierarchical and fine-grained sense inventory and (iii) support for multi-sense annotation with established benchmarks. We can use these to guide our experimental choice.

2.2.4 Why We Adopt PDTB

We adopt PDTB as our annotation framework for pragmatic reasons. First, PDTB offers a stable, fine-grained sense inventory augmented in PDTB-3 with pragmatic dimensions (belief and speech-act), which supports consistent supervision and analysis. Second, because our experiments focus on the task Implicit Discourse Relation Recognition (IDRR), we benefit from PDTB's explicit/implicit partition, which enables connector-agnostic inference while maintaining a clear explicit vs. implicit distinction. Third, PDTB corpora are widely used community benchmarks for IDRR, with established splits, enabling controlled comparisons with prior work. In addition, using both versions, the PDTB-2 and the PDTB-3, allows us to assess robustness across sense inventories and annotation revisions. Finally, PDTB supports multi-sense (multi-label) annotation, which can support co-occurrence modeling.

We stress that alternative frameworks remain valuable and inform our modeling. RST and SDRT bring tree structure-based and graph-based strengths, respectively, but they do not partition relations by signaling mechanism in the way PDTB does. QUD and CCR are likewise insightful, yet they do not provide a fixed label taxonomy (QUD is question/answer–driven; CCR is cognitive feature-based), which complicates supervised evaluation and cross-study comparability. eRST is promising particularly in their support for multi-sense annotation and separation of implicit relation and explicit relation types, although we do not use it as it was introduced after the core stages of this work and thus fall outside the scope of our current experiments. For availability, the senses inventory, comparability, and multi-label support, we therefore conduct our experiments under the PDTB-2 and the PDTB-3, while viewing eRST-based evaluation as future work.

2.3 Penn Discourse Treebank (PDTB) Annotation

As the Penn Discourse Treebank (PDTB) serves as the annotation framework for this thesis, we will give a detailed introduction for it. In this section, we first describe

how PDTB defines and annotates discourse units, which form the basic elements of discourse relations. We then give more details on its hierarchical sense annotation system, while we have mentioned PDTB captures the semantic or pragmatic relationships between different types of discourse relations through a three-level hierarchy in Section 2.2.2.

2.3.1 Discourse Units

In the PDTB, discourse units are defined as text spans that participate in forming discourse relations, with each relation involving two arguments (Arg1 and Arg2). The framework adopts a flexible approach to unit definition, allowing relations to hold between various textual spans. For explicit relations, the only constraint is that the arguments must be within the same document, meaning they can occur between adjacent or non-adjacent spans. In contrast, implicit relations are more constrained, as they primarily hold between adjacent sentences where a discourse connective could be inferred. This constraint is imposed by the PDTB annotation guideline rather than being a theoretical limitation of implicit relations themselves.

Additionally, discourse relations can occur in two main configurations:

1. **Inter-sentential:** Between adjacent sentences
2. **Intra-sentential:** Within sentences, including:
    - Between VPs or clauses conjoined by punctuation:

      [Stocks closed higher in Hong Kong, Manila, Singapore, Sydney and Wellington]$_1$, [but were lower in Seoul]$_2$. [wsj_0231]
    - Between a free adjunct or free to-infinitive and its matrix clause:

      [Banks need a competitive edge]$_1$ [to sell their products]$_2$. [wsj_0238]
    - Between a marked syntactic construction and its matrix clause :

      [Bell Atlantic posted a strong earnings gain for the third quarter]$_1$, [as did Southern New England Telecommunications]$_2$. [wsj_1728]

PDTB-3 expanded the annotation of intra-sentential relations found in the PDTB-2, adding approximately 5.66k new implicit instances and 5.67k new explicit instances (Webber et al., 2019). This expansion significantly improves the coverage of discourse relations within sentences, allowing for a more comprehensive understanding of how discourse is structured at a finer granularity and supporting more comprehensive modeling of discourse coherence.

While argument identification—the task of determining the precise text spans (Arg1 and Arg2) that participate in a discourse relation—is an essential step in discourse parsing, it remains a challenging problem in its own right (Kong et al., 2014; Lin et al., 2014). In the PDTB framework, arguments can be non-adjacent, span clauses or sentences, and vary in syntactic structure, which makes reliable automatic identification difficult, especially for implicit relations.

In this work, we assume that the discourse arguments have already been identified and focus exclusively on the task of discourse relation sense classification, predicting the type of discourse relation that holds between two given spans. This assumption is consistent with prior work that isolates sense classification from argument segmentation in order to study the specific challenges of label prediction (Ji and Eisenstein, 2015; Shi and Demberg, 2017; Kim et al., 2020; Xiang et al., 2022b). While we acknowledge that this simplification abstracts away a key part of the full discourse parsing pipeline, it allows us to focus on evaluating the benefits of hierarchical label structure and representation learning for sense recognition. We discuss potential directions for integrating argument identification into our framework in Section 6.2.

2.3.2 Hierarchical Sense Annotation

While Section 2.2.2 has indicated that PDTB employs a three-level hierarchical taxonomy for annotating discourse relations, as illustrated in Figure 2.1, we will give more details in this subsection. The PDTB sense hierarchy is not merely a set of labels but is supported by definitions and usage guidelines, including illustrative examples for each sense as described in the PDTB Annotation Manual Webber et al. (2019). This ensures consistent and reliable annotation across annotators. Importantly, annotators can be told (in the Annotation Manual) to give precedence to one label in cases where more than one label could apply. For example, they are told in the Manual that examples that satisfy the conditions for both “Contrast” and “Concession”, should be labeled as “Concession”.

The three-level hierarchical label structure in the PDTB captures both coarse-grained and fine-grained semantic relationships between text spans. This hierarchical organization not only attempts to facilitate annotation consistency but also reveals important semantic and pragmatic relationships between different types of discourse relations. The sense hierarchy in PDTB consists of three levels, starting with the most general Level-1 classes: “Temporal”, “Contingency”, “Comparison”, and “Expansion”. Each

of these classes is then refined into more specific Level-2 types that capture distinct semantic relationships. For example, in PDTB-3, Temporal relations are subdivided into "Synchronous" and "Asynchronous" types, while "Contingency" includes relations such as "Cause" and "Condition". At Level-3, the finest granularity level, relations are further specified to capture detailed semantic distinctions, such as the division of "Cause" into "Reason" and "Result" to reflect the directionality of causal relations. While it might appear that PDTB's Level-3 senses simply reflect directionality rather than finer-grained semantic distinctions, they nonetheless play an essential role in capturing argument-specific roles within asymmetric relations. These tags such as "Reason" or "Result" are not merely structural markers but enrich the interpretive granularity beyond what Level-2 provides. Unlike RST's two-tier sense hierarchy, PDTB's third level introduces a systematic mechanism to distinguish how the same relation type manifests depending on argument order and discourse function. This added specificity contributes to more accurate discourse interpretation and can inform computational models for argument role prediction, pragmatic inference, and relation generation. Thus, Level-3 senses do not always define new relation types, they provide a functionally meaningful extension to the sense hierarchy.

Figure 2.3 and Figure 2.4 illustrate the sense hierarchy in the PDTB-2 and the PDTB-3 respectively. The sense hierarchy used in the PDTB-3 differs somewhat from that used in the PDTB-2, with additions motivated by the needs of annotating intra-sentential relations and changes motivated by difficulties that annotators had in consistently using some of the senses in the PDTB-2 hierarchy. Because of the differences in these two hierarchies, we use the PDTB-2 hierarchy for PDTB-2 data and the PDTB-3 hierarchy for PDTB-3 data respectively.

This hierarchical structure embodies several important relationships between labels. Most fundamentally, there are strict subsumption relations where higher-level categories encompass their lower-level subcategories. For instance, in the PDTB-3, a relation which is only labeled as "Reason" at Level-3 must belong to "Cause" at Level-2 and "Contingency" at Level-1. Wu et al. (2020, 2022) have utilized such dependency between Level-1 and Level-2 senses by modeling the prediction process hierarchically, predicting the Level-1 sense between arguments first, and then selecting from Level-2 senses that are valid children of the predicted Level-1 label.

Beyond these strict hierarchical constraints, relations within the same parent category often share semantic properties while exhibiting varying degrees of compatibility. Notably, the incompatibility between relations tends to increase at finer granularity

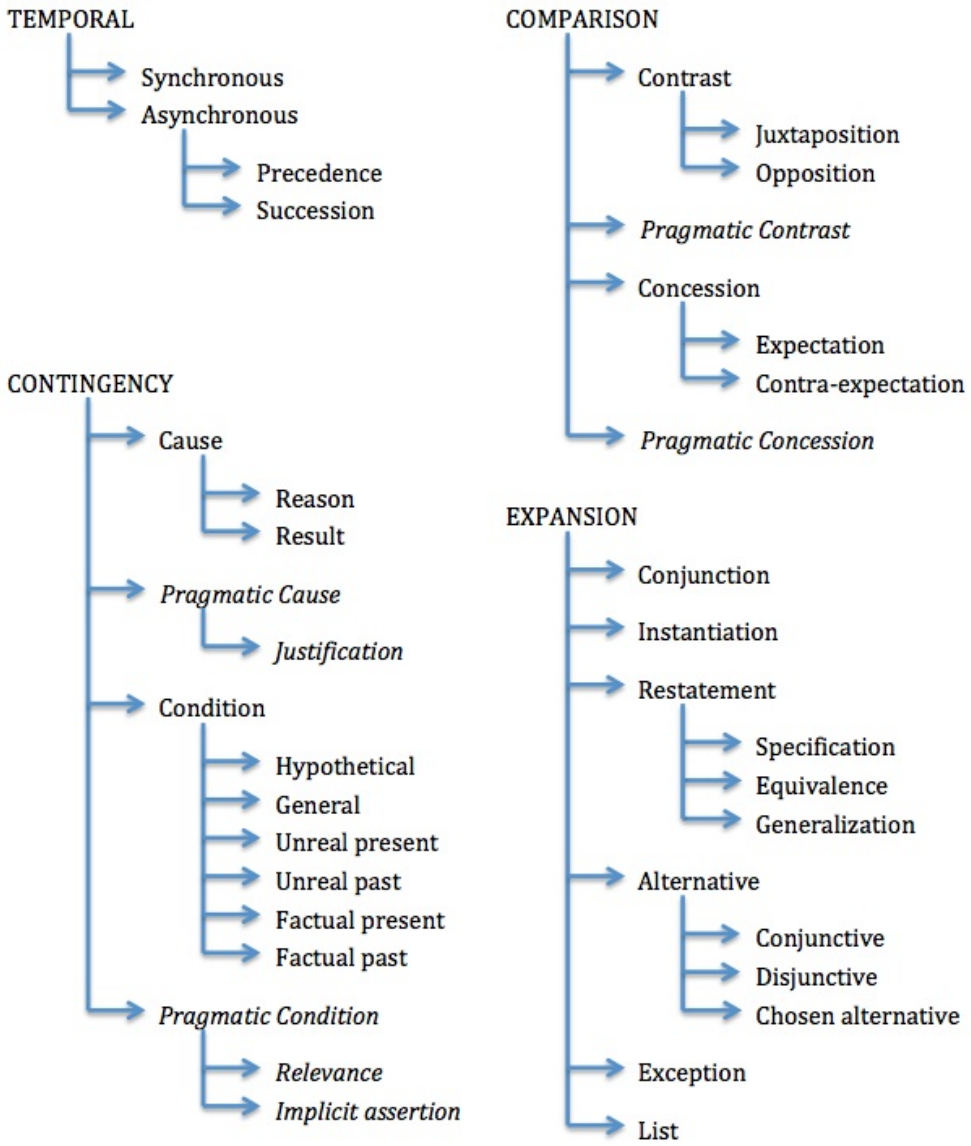


Figure 2.3: The PDTB-2 Sense Hierarchy(Prasad et al., 2008). Level-1 senses represent broad semantic classes (e.g., *Contingency*, *Comparison*), while Level-2 senses provide more specific subtypes (e.g., *Cause*, *Contrast*). Some Level-2 senses are further refined into Level-3 distinctions (e.g., *Reason*, *Result*).

levels. For example, in the PDTB-3, at Level-3, while "Result" and "Reason" under "Cause" share the common property of expressing causality, they are mutually exclusive as they represent opposing directions of the causal relationship. Similar incompatibility patterns can be observed at Level-2 senses, such as between "Conjunction" and "Disjunction" under "Expansion" in the PDTB-3, where the fundamental semantic properties of these relations make them mutually exclusive. Specifically, "Conjunction" implies additive continuity, while "Disjunction" signals contrastive alternation.

Level-1	Level-2	Level-3
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
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	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

Figure 2.4: The PDTB-3 Sense Hierarchy (Webber et al., 2019). Level-1 senses indicate high-level categories. Level-2 senses refine these into more specific discourse relations, and Level-3 senses are used for asymmetric relations to specify the role of the arguments (e.g., ARG1-as-Denier, ARG2-as-Goal).

These incompatibility relationships provide additional structural constraints that can be valuable for relation recognition.

In addition, similarity between relations is evident in how they align conceptually within their parent categories. For instance, in the PDTB-3, within “Expansion” at Level-2, “Expansion.Level of detail” refers to cases where the second argument pro-

vides a finer-grained description or elaboration of the first one, often adding specific details to a more general statement, while "Expansion.Instantiation" refers to cases where the second argument presents a concrete example or instance of the general concept introduced in the first segment. These two relations are similar, as they share the property of providing additional details or specifications.

The hierarchical nature of PDTB's annotation scheme presents both opportunities and challenges for discourse relation recognition. The structured representation of relations reflects their semantic organization, providing constraints and similarity patterns that could potentially inform computational approaches. In Chapter 3, 4, and 5 of this thesis, we demonstrate how these hierarchical relations can be leveraged to improve the recognition of discourse relations.

2.4 Modeling Implicit Discourse Relations in PDTB

Discourse relations have been exploited across diverse NLP tasks. In text summarization, they support content selection and ordering for coherent outputs; for example, Huang and Kurohashi (2021) incorporate discourse relation-aware graphs to capture inter-sentential links for extractive summaries. In dialogue systems, discourse relations help preserve logical flow and enable natural topic shifts; Chen et al. (2023) show that recognizing the correct discourse relation improves consistency and coherence in multi-turn interactions. In question answering, discourse relation types such as causation, elaboration, and contrast connect evidence to questions; Du et al. (2023) design sentence-level discourse graphs to mediate logical interaction between questions and conditions. Finally, in event relation extraction, discourse relations can help reveal event–event links; Tang et al. (2021) jointly train discourse relation recognition with event relation extraction to reduce reliance on explicit connectives and labeled data, successfully identifying implicit event relations.

While discourse relations broadly benefit many downstream applications, this thesis focuses on *implicit* discourse relation recognition (IDRR) in the PDTB, where relations are not overtly signaled by connectives. Accurately identifying such relations remains challenging because systems must infer nuanced semantic and pragmatic links beyond surface markers. The PDTB provides a well-established testbed for studying IDRR, and a range of modeling approaches has emerged over the past decade.

Early neural models learned argument representations directly from data (Chen et al., 2016a; Liu and Li, 2016; Dai and Huang, 2018), with attention mechanisms

and context-aware encoders further enriching representations. However, most of these approaches treated labels as flat and mutually exclusive, overlooking two central properties of PDTB annotations: (i) a hierarchical sense hierarchy and (ii) the possibility that multiple relations co-occur between the same pair of arguments.

Recent work has leveraged pretrained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), leading to significant performance improvements. The primary approaches can be categorized into:

- **Direct Fine-tuning:** Studies like (Shi and Demberg, 2017; Liu et al., 2020; Wu et al., 2020, 2022) demonstrated the effectiveness of fine-tuning pretrained models for discourse relation recognition.

- **Prompt-based Methods:** Recent work (Xiang et al., 2022b; Zhou et al., 2022) explored prompt-based learning approaches, framing discourse relation recognition as a connective prediction task.

Despite strong results, most PLM-based systems still do not explicitly leverage the PDTB sense hierarchy, and they typically assume a single gold label per instance. Only Wu et al. (2020, 2022) have tried to leverage the dependence between the Level-1 and Level-2 labels. They attempted to assign a Level-1 sense that holds between arguments, and then only considered as possible Level-2 senses, those that are daughters of the Level-1 senses. However, their method treats the hierarchy as a hard constraint during inference and does not incorporate it into the learning process itself.

In our prior work (Long and Webber, 2022), we proposed the first approach to incorporate hierarchical structure into training by guiding contrastive learning with hierarchy-aware negative sampling. This line of work was later extended by some subsequent studies such as Jiang et al. (2023), who model the full label hierarchy as a graph, Lian et al. (2024), who introduced learnable class centers to encode label similarity by using the sense hierarchy, and Wu et al. (2024), who further develop this direction by applying contrastive learning across multiple model layers with hierarchical constraints at different levels of abstraction. Although these works explore novel architectural improvements, they all continue to treat implicit discourse relation recognition as a single-label classification problem—limiting their ability to capture the co-occurrence of discourse senses.

To address this limitation, our more recent work Long et al. (2024) was the first to treat Implicit Discourse Relation Recognition (IDRR) as a multi-label classification

task using the PDTB-3. We compare various modeling strategies and demonstrated that multi-label models not only match the performance of single-label counterparts but also can capture sense co-occurrence patterns. Building on our work, Costa and Kosseim (2024) propose a multi-task, multi-label architecture trained on DiscoGeM (Scholman et al., 2022), a crowdsourced dataset with soft label distributions. Their model learns from these distributions and transfers effectively to PDTB-3, underscoring the increasing relevance of multi-label formulations and supporting the broader direction initiated by our work.

However, neither of these studies investigates whether incorporating the sense hierarchy can further improve sense labeling performance in the multi-label classification setting. This limitation is addressed in our thesis work (see Chapter 5). This thesis builds on both strands, specifically, the multi-label learning and the use of the hierarchical label structure, investigating how incorporating the PDTB sense hierarchy into a multi-label classification framework can improve implicit discourse relation prediction. To our knowledge, this is the first work to systematically combine both perspectives.

2.5 Conclusion

This chapter has provided a comprehensive overview of discourse relations and their various theoretical frameworks. We began by examining the fundamental introduction of discourse relations and their crucial role in understanding textual coherence. The discussion covered several major theoretical approaches, including RST, SDRT, CCR, D-LTAG, PDTB, QUD, and eRST, highlighting how each framework contributes unique perspectives to our understanding of discourse structure and coherence.

A comparative examination of these frameworks revealed significant differences in how they define relation types, the structures of senses, and multi-sense annotation. While frameworks such as RST, SDRT, QUD, and CCR do not explicitly classify discourse relations based on their signaling mechanisms, PDTB distinguishes them into categories such as Explicit, Implicit, AltLex, and Entity-based relations. The structures of organizing the senses also differ, with PDTB employing a hierarchical organization of classes, types, and subtypes, while the relations in other frameworks are represented either in a flat set or in fine-grained and coarse-grained levels. Furthermore, some frameworks support multi-sense annotation, allowing multiple relations between the same discourse units (e.g., SDRT and eRST), whereas others impose a

single-relation constraint, requiring annotators to select the most salient relation (e.g., traditional RST). These variations in relation categorization, label structuring, and multi-senses flexibility have important implications for discourse analysis and computational modeling. Additionally, we discuss the reasons why we use PDTB as the discourse framework for our experiments.

The examination of PDTB annotation highlights how its hierarchical organization of discourse relations captures relationships between different types and subtypes, guiding human annotation and revealing label dependencies that could enhance computational models. However, early methods for implicit discourse relation recognition, despite improving performance, fail to fully utilize this structure. Most approaches treat the task as a flat multi-class classification or handle hierarchy levels independently, ignoring inter-level dependencies. Additionally, many models focus solely on single-label classification, neglecting the annotated multi-label instances. This gap between PDTB's rich annotation scheme and existing models motivates our exploration of methods that effectively leverage label hierarchies for both single-label and multi-label discourse relation classification.

Chapter 3

Enhancing Sense Labeling in Fine-tuning with the PDTB Sense Hierarchy

Building on the foundational understanding of discourse relations and their annotation introduced in Chapter 2, this chapter reflects on how hierarchical label structures can be meaningfully leveraged to improve sense labeling in implicit discourse relation recognition. While the PDTB sense hierarchy provides a rich, multi-level organization of discourse senses, prior work has largely overlooked its potential in guiding model learning. Our exploration reveals that explicitly incorporating this structure into the fine-tuning process can lead to improved performance.

One of the key insights from this chapter is that hierarchical relationships among labels can inform the selection of semantically meaningful negative examples in contrastive learning. By aligning the contrastive objective with the semantic distances encoded in the label hierarchy, we observed that the model learns more discriminative representations. This approach improves performance without the need for additional supervision or more complex architectures, highlighting the underutilized value of label hierarchies in training objectives. Our findings contribute to a broader understanding of how structured label information can support more effective representation learning. This chapter draws on work originally presented in Long and Webber (2022).

3.1 Motivation

While pre-trained language models (Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019) have dramatically improved performance on discourse relation recognition (Shi and Demberg, 2019; Liu et al., 2020; Kishimoto et al., 2020), they still struggle with implicit relations where no explicit connectives guide the interpretation. This challenge is particularly evident in cases which appear similar on the surface in some way while belonging to distinctly different categories in the sense hierarchy. Consider these three examples from the PDTB-3 whose second argument (Arg2) appears similar to the others:

(7) [“Valley National ”isn’t out of the woods yet]$_1$. [The key will be whether Arizona real estate turns around or at least stabilizes.]$_2$. (From wsj_1865)

(8) [The House appears reluctant to join the senators]$_1$. [A key is whether House Republicans are willing to acquiesce to their Senate colleagues’ decision to drop many pet provisions.]$_2$. (From wsj_2372)

(9) [Japanese culture vs. American culture is irrelevant]$_1$. [The key is how a manager from one culture can motivate employees from another]$_2$. (From wsj_1092)

Despite their surface similarities, these examples illustrate an important distinction in discourse relations. In Examples (7) and (8), the annotators took the second argument (Arg2) as providing more details about the the claim in the first argument (Arg1) — the sense called “Expansion.Level-of-detail.Arg2-as-detail”, while in Example (9), they took the “how a manager from one culture motivates employees from another ” in the second argument (Arg2) as expressing an alternative to “Japanese culture vs. American culture” in the first argument (Arg1) as an explanation for the situation under discussion. This sense is called “Expansion.Substitution.Arg2-as-substitution”.

This challenge of distinguishing between semantically distinct discourse relations that may share similar surface-level lexical or syntactic features such as having the same or similar predicates in Arg2 prompted us to explore new approaches. Recent advances in contrastive learning have shown promising results across various NLP tasks by learning to differentiate between similar and dissimilar examples (Kim et al., 2021; Zhang et al., 2021; Yan et al., 2021). The core principle of contrastive learning — minimizing the distance between similar instances (positive examples) while maximizing the distance to dissimilar instances (negative examples) — has proven effective in constructing meaningful representations.

Previous work has indicated that contrastive learning can help select good negative examples (Joshua et al., 2021; Wang et al., 2021; Suresh and Ong, 2021). As described in Section 2.3.2, discourse relations in the PDTB are organized in a three-level hierarchical structure. This structure can indicate the similarities between labels. Despite this rich hierarchical organization, most previous approaches have treated these discourse senses as a flat list of labels, failing to utilize the valuable structural information inherent in the hierarchy. While Wu et al. (2022) made initial progress by leveraging the dependencies between Level-1 and Level-2 labels, improving the model prediction consistency across Level-1 and Level-2, we further utilize the hierarchical structure for selecting negative examples in contrastive learning during training.

Our objectives extend beyond merely improving performance metrics. We aim to demonstrate that a deeper understanding and utilization of discourse sense hierarchies can lead to more accurate models for implicit relation recognition.

3.2 Methodology

In this work, we use a multi-task learning framework, which consists of classification tasks and a contrastive learning task. As Figure 3.1 illustrates, we first use a simple multi-task model based on RoBERTa-base (Liu et al., 2019), and then we develop a contrastive learning algorithm where the sense hierarchy is used to select positive and negative examples.

As part of our broader goal to structure the representation space according to the discourse sense hierarchy, we explore data augmentation strategies that can help the model learn more robust and semantically informed representations. In particular, we investigate whether generating more examples by using the PDTB Meta-data can improve its ability to distinguish between closely related senses. Practically, augmentation also increases the supply of informative positive and negative pairs—critical for contrastive objectives (Chen et al., 2020; Khosla et al., 2020).

Detailed descriptions of these components follow. We first provide general background on contrastive learning, then describe how we integrate it with the sense hierarchy, and finally present our augmentation method and its role in the training pipeline.

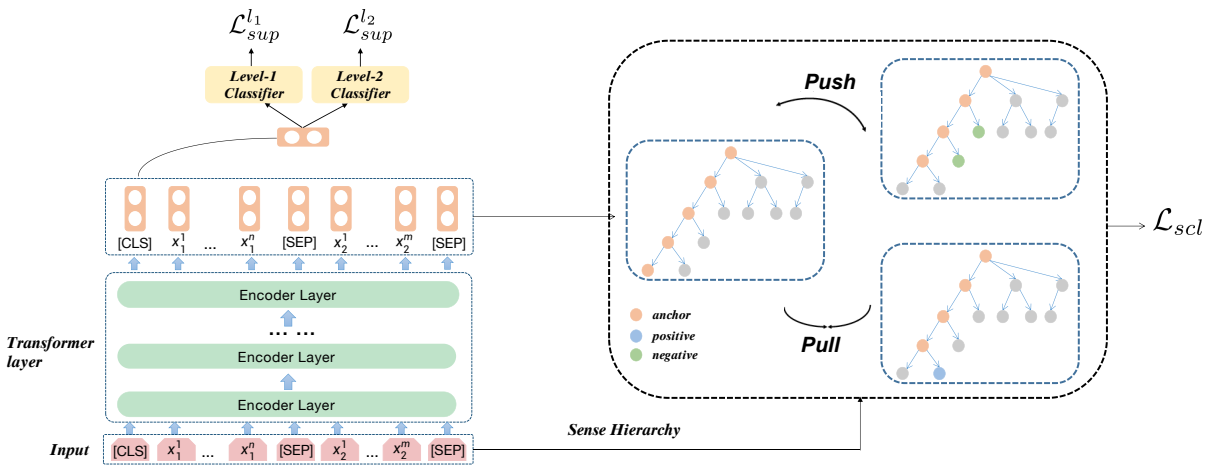


Figure 3.1: **The overall architecture of our model.** When given an example, we search the positive and negative examples in a training batch based on the sense hierarchy of the PDTB. We narrow the distances among examples from the same types at Level-2 or Level-3 and increase the distances among examples from different types at Level-2 and Level-3.

3.2.1 Contrastive Learning

In contrastive learning, each training instance is treated as an anchor and compared to a set of positive examples (semantically similar instances) and negative examples (semantically dissimilar instances). The objective is to learn an embedding space in which anchors are pulled closer to their positives and pushed away from their negatives.

Let a mini-batch contain N pairs $\{(x_i, x_i^{pos})\}_{i=1}^{N}$, where x_i is the i-th anchor which is the reference example we start from when computing similarity and x_i^{pos} is its corresponding positive example. We denote by $p(i)$ the index of x_i^{pos} in the batch, $h_i \in \mathbb{R}^d$ the embedding of x_i, and $h_{p(i)}$ the embedding of x_i^{pos}. Let $\text{sim}(\cdot,\cdot)$ denote cosine similarity, and let $\tau > 0$ be a temperature parameter controlling the concentration of the similarity distribution. Smaller τ makes the similarity distribution sharper (more concentrated on the highest-similarity items), while larger τ makes it flatter (more uniform). This controls how strongly the model focuses on the most similar items.

The standard InfoNCE loss (van den Oord et al., 2018) is defined as:

$$\mathcal{L}_{\text{con}} = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp(\text{sim}(h_i, h_{p(i)})/\tau)}{\sum_{a=1, a\neq i}^{N} \exp(\text{sim}(h_i, h_a)/\tau)}. \tag{3.1}$$

Here, the numerator measures the similarity between the anchor and its positive, while the denominator aggregates similarities with all other embeddings in the batch

(including the positive), encouraging the model to assign relatively higher similarity to positives than to negatives.

3.2.2 Supervised Contrastive Learning

The supervised contrastive loss (Gunel et al., 2021) extends Eq. 3.1 to make use of label information. Let y_i denote the class label of anchor i. Let $P(i) = \{a \neq i \mid y_a = y_i\}_{i=1}^{N}$ denote the set of positive indices for anchor i, with $|P(i)| = N_{y_i} - 1$ where N_{y_i} is the number of samples in the batch with label y_i.

The supervised contrastive loss is:

$$\mathcal{L}_{\text{scl}} = \frac{1}{N} \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(h_i, h_p)/\tau)}{\sum_{a=1, a \neq i}^{N} \exp(\text{sim}(h_i, h_a)/\tau)}. \tag{3.2}$$

This formulation averages over all positives of the same label for each anchor, ensuring that the loss is invariant to the number of positives per class. Intuitively, it encourages embeddings of all same-class samples to cluster together, while pushing them away from different-class samples.

3.2.3 Sentence Encoder

Each annotated discourse relation in the PDTB consists of two text spans—referred to as *arguments*—and one or more relational senses that hold between them. Following common practice, we represent each example by concatenating its two arguments into a single sequence and encoding it using RoBERTa (Liu et al., 2019).

Specifically, given the first argument Arg1 and the second argument Arg2, we construct the input sequence:

$$\text{[CLS]},\ \text{Arg1},\ \text{[SEP]},\ \text{Arg2},\ \text{[SEP]}$$

where [CLS] and [SEP] are RoBERTa's special tokens marking the start and separation of segments. This sequence is tokenized and passed through RoBERTa, and we take the contextual representation of the [CLS] token from the final hidden layer as the embedding $h_i \in \mathbb{R}^d$ for instance i. This embedding serves as the input to both supervised classification layers and the contrastive learning objective.

This setup allows the encoder to jointly model both arguments in a single context, enabling it to capture semantic and discourse-level dependencies between them.

3.2.4 Data Augmentation Using PDTB Meta-data

We leverage the explicit connective meta-data provided in the PDTB for each implicit discourse relation (see Webber et al., 2019). This meta-data records one or two explicit connectives that annotators judged could have signaled the annotated sense. For example, an implicit "Cause" relation might have the connective "because" recorded as a possible signal.

While Patterson and Kehler (2013) used these connectives as targets for a multi-class connective prediction task, we use them differently: for each annotated implicit relation, we create an additional augmented training example by inserting the recorded connective into the second argument, immediately after the [SEP] token. Figure 3.2 illustrates this for the connective "In contrast".

Formally, if the original instance has arguments $(\text{Arg1}, \text{Arg2})$ and an associated connective c, the augmented instance becomes:

$$[\text{CLS}],\ \text{Arg1},\ [\text{SEP}],\ c\ \text{Arg2},\ [\text{SEP}]$$

This augmented example inherits the same discourse sense label as the original.

Because each implicit relation in PDTB meta-data has at least one recorded connective and at most two[1], this augmentation approximately doubles the number of training examples. Concretely, if a batch originally contains N instances, augmentation increases the batch size to at least $2N$, providing more positive and negative pairs for contrastive learning.

Intuitively, this augmentation makes the model more robust by exposing it to both purely implicit examples and its variants with inserted connectives, encouraging it to learn sense-relevant semantic patterns regardless of whether a connective is present. More importantly, this might help the model better encode distinctions between closely related senses — a key requirement for controlling label distances in the embedding space according to the hierarchical structure.

e_{pos}: The government includes money spent on residential renovation. Dodge doesn't.
e^*_{pos}: The government includes money spent on residential renovation. In contrast, dodge doesn't.

Figure 3.2: An example with inserted connective: the connective word is "In contrast".

[1]PDTB allows a maximum of two senses per relation, each with its own connective meta-data.

3.2.5 Positive and Negative Pair Selection under the PDTB Sense Hierarchy

In our setting, each PDTB-annotated discourse relation has a hierarchical label: Level-1 (l_1), Level-2 (l_2), and possibly Level-3 (l_3) for asymmetric senses. Only terminal-level labels are used for annotation: Level-2 for symmetric senses and Level-3 for asymmetric senses. For example, "Temporal.Asynchronous" (Level-2) has two Level-3 children: "Precedence" and "Succession".

We define positives and negatives based on this hierarchy:
Positive set $\mathcal{P}(i)$: For an anchor e_i with labels (l_1^i, l_2^i, l_3^i), the positives are all other instances in the batch with the same Level-2 label if l_3 does not exist, or the same Level-3 label if it does:

$$\mathcal{P}(i) = \{\, e \in \mathcal{B} \mid l_2^e = l_2^i \text{ or } l_3^e = l_3^i \,\}. \tag{3.3}$$

For example, if the label is "Temporal.Asynchronous.Precedence", positives include all other "Temporal.Asynchronous.Precedence" instances.
Negative set $\mathcal{N}(i)$: Negatives are chosen as *sister senses*: instances with the same Level-1 label but different labels in lower levels:

$$\mathcal{N}(i) = \{\, e \in \mathcal{B} \mid l_1^e = l_1^i \ \wedge\ (l_2^e \neq l_2^i \ \vee\ (l_3^i \text{ exists } \wedge\ l_3^e \neq l_3^i)) \,\}. \tag{3.4}$$

For instance, for "Temporal.Asynchronous.Precedence", negatives include "Temporal.Asynchronous.Succession" and "Temporal.Synchronous".

This selection strategy explicitly uses the PDTB hierarchy to construct more informative positive and negative pairs, promoting fine-grained label discrimination. For each anchor, the positives in the batch are treated as similar and the negatives are the rest of the batch. If negatives aren't in the same batch as the positives, the model has nothing to contrast against and can't learn to push them apart.

3.2.6 Overall Training Objective

We jointly optimize supervised classification at Level-1 and Level-2 with hierarchy-aware supervised contrastive learning. Let $\mathcal{L}_{\text{sup}}^{(1)}$ and $\mathcal{L}_{\text{sup}}^{(2)}$ be the cross-entropy losses at Level-1 and Level-2 respectively, and $\mathcal{L}_{\text{ccl}}$ be the supervised contrastive loss in Eq. 3.2. The final loss is:

$$\mathcal{L} = \mathcal{L}_{\text{sup}}^{(1)} + \mathcal{L}_{\text{sup}}^{(2)} + \beta\, \mathcal{L}_{\text{ccl}}, \tag{3.5}$$

where $\beta \geq 0$ balances the contribution of the contrastive objective. Setting $\beta = 0$ yields a purely supervised model, while $\beta > 0$ incorporates hierarchy-aware contrastive learning as an auxiliary signal.

3.3 Experimental Setting

3.3.1 Datasets and Baselines

Besides providing label sense hierarchy, the Penn Discourse TreeBank (PDTB) also frequently serves as a dataset for evaluating the recognition of discourse relations. The earlier corpus, PDTB-2 (Prasad et al., 2008) included 40,600 annotated relations, while the later version, PDTB-3 (Webber et al., 2019) includes an additional 13K annotations, primarily intra-sentential, as well as correcting some inconsistencies in the PDTB-2. Table 3.1 and Table 3.2 present the statistics of Level-1 relations and Level-2 relations. It is important to note that we only use the Level-2 senses having sufficient data as the target Level-2 labels and their data as our data source by following previous work (Ji and Eisenstein, 2015; Xiang et al., 2022b; Dai and Huang, 2018; Shi and Demberg, 2017). Senses such as "Condition+SpeechAct", "Concession+SpeechAct", and "Disjunction" in PDTB-3, as well as "Pragmatic Condition" and "Exception" in PDTB-2, are excluded due to insufficient training data, with each category containing fewer than 50 instances.

Because of the differences in these two hierarchies, we use the PDTB-2 hierarchy for PDTB-2 data and the PDTB-3 hierarchy for PDTB-3 data respectively. The datasets are kept separate throughout all experiments, and train/validation/test splits are performed independently for PDTB-2 and PDTB-3. **To ensure comparability with existing work, we follow earlier work (Ji and Eisenstein, 2015; Bai and Zhao, 2018; Liu et al., 2020; Xiang et al., 2022a) using Sections 2-20 of the corpus for Training, Sections 0-1 for Validation, and Sections 21-22 for testing**. With regard to those instances with multiple annotated labels, we also follow previous work (Qin et al., 2016). Such instances are treated as separate examples during training. At test time, a prediction matching one of the gold types is taken as the correct answer. Implicit relation recognition is usually treated as a classification task. While 4-way (Level-1) classification was carried out on both PDTB-2 and PDTB-3, more detailed 11-way (Level 2) classification was done only on the PDTB-2 and 14-way (Level 2) classification, only on the PDTB-3.

Label	*n*
Comparison	2,518
Contingency	7,583
Expansion	10,833
Temporal	1,828
Comparison.Concession	1,494
Comparison.Contrast	983
Contingency.Cause	5,785
Contingency.Cause+Belief	202
Contingency.Condition	199
Contingency.Purpose	1,373
Expansion.Conjunction	4,386
Expansion.Equivalence	336
Expansion.Instantiation	1,533
Expansion.Level-of-detail	3,361
Expansion.Manner	739
Expansion.Substitution	450
Temporal.Asynchronous	1,289
Temporal.Synchronous	539

Table 3.1: Label counts for level-1 senses and level-2 senses that have more than 100 annotated instances in PDTB-3.

To evaluate the effectiveness of our proposed method, we compare it against a set of strong baselines. These baselines were selected based on their impact at the time, their use of different modeling strategies, and their relevance to either PDTB-2 or PDTB-3. Since most prior work focused on one dataset, we use different sets of baselines for each.

Baselines for PDTB-2: These models represent the state-of-the-art approaches on PDTB-2 before the release of the PDTB-3. We selected them to cover a range of modeling paradigms:

- Dai and Huang (2019): a neural model leveraging external event knowledge and coreference relations.
- Shi and Demberg (2019): a neural model that uses the inserted connectives to learn better argument representations.
- Nguyen et al. (2019): a neural model which predicts the labels and connectives. simultaneously.
- Guo et al. (2020): a knowledge-enhanced Neural Network framework.

Label	**n**
Comparison	2,291/2,503
Contingency	3,911/4,255
Expansion	8,249/8,561
Temporal	909/950
Comparison.Concession	223
Comparison.Contrast	1,210
Contingency.Cause	4,172
Contingency.Pragmatic cause	83
Expansion.Conjunction	3,534
Expansion.Instantiation	1,445
Expansion.Alternative	1,185
Expansion.List	400
Expansion.Restatement	3,206
Temporal.Asynchronous	697
Temporal.Synchrony	251

Table 3.2: Label counts for PDTB-2 Level-1 and 11 senses of Level-2 (label set commonly used in the literature for Level-2 classification). Level-1 classification is evaluated on Ji split (Ji and Eisenstein, 2015), so we list both the label counts in Ji split and the total label counts in the whole dataset.

- Kishimoto et al. (2020): a model applying three additional training tasks.
- Liu et al. (2020): a RoBERTa-based model which consists of a contextualized representation module, a bilateral multi-perspective matching module, and a global information fusion module.
- Jiang et al. (2021): a method that recognizes the relation label and generates the target sentence containing the meaning of relations simultaneously.
- Dou et al. (2021): a method using conditional variational autoencoder (VAE) to estimate the risk of erroneous sampling.
- Wu et al. (2022): a label dependence-aware sequence generation model.

These baselines cover knowledge-based, connective-prediction, multi-task, generation-style, and label-structured approaches. The work most closely related to ours is Wu et al. (2022), which exploits the dependency between Level-1 and Level-2 senses *at*

inference (restricting Level-2 candidates given a predicted Level-1 class). In contrast, we inject the PDTB sense hierarchy *into training* via hierarchy-aware objectives.

Baselines for PDTB-3: PDTB-3 was released in 2019, and few early models were directly developed for it. We rely on baselines from Xiang et al. (2022a), who reimplemented several strong PDTB-2 models on PDTB-3:

- Liu and Li (2016): a model that combines the two arguments' representation for stacked interactive attention.
- Chen et al. (2016a): a mixed generative-discriminative framework.
- Lan et al. (2017): a multi-task attention neural network.
- Ruan et al. (2020): a propagative attention learning model.
- Xiang et al. (2022a): a model that uses a Dual Attention Network (DAN).

These baselines allow us to position our method relative to both earlier neural models and more recent methods using pretrained language models.

3.3.2 Implementation Details

In our experiments, we use the pre-trained RoBERTa-base (Liu et al., 2019) as our Encoder. We adopt Adam (Kingma and Ba, 2015) with the learning rate of $3e{-}5$ and the batch size of 256 to update the model. The maximum training epoch is set to 25 and the wait patience for early stopping is set to 10 for all models. We clip the gradient L2-norm with a threshold 2.0. For contrast learning, the weight of positive examples is set to 1.6 and the weight of negative examples is set to 1. This setting also follows common practice in supervised contrastive learning, where giving slightly more emphasis to positive pairs helps the model focus on relevant semantic alignment. In addition, we experimented with different weighting values, and found that this setting yielded the best performance in our scenario. All experiments are performed with 1× 80GB NVIDIA A100 GPU.

3.3.3 Effects of the Coefficient β

As shown in Equation 3.5, the coefficient β is an important hyperparameter that controls the relative importance of supervised loss and contrastive loss. Thus, we vary β from 0 to 2.4 with an increment of 0.2 each step, and inspect the performance of our model using different β on the validation set.

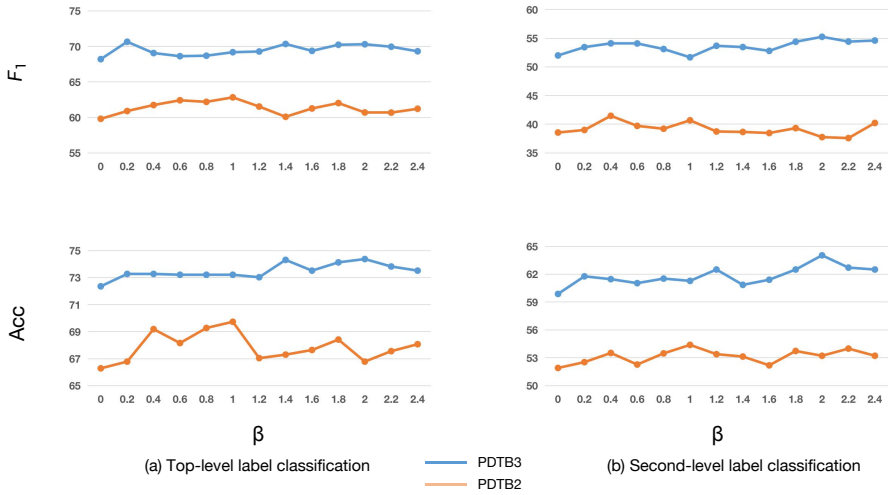


Figure 3.3: Effects of β on the validation set.

From the Figure 3.3, we can find that, compared with the model without contrastive learning (β = 0), the performance of our model at any level is always improved via contrastive learning. For the PDTB-2, when β exceeds 1.0, the performance of our model tends to be stable and declines finally. Thus, we directly set β = 1.0 for all PDTB-2 related experiments thereafter. For PDTB-3, the Accuracy and F1 score of the validation set reach the highest point at β = 2.0. Therefore we choose β = 2.0 for all related experiments.

We have considered three ways of investigating why there is such a difference in the optimal weighting coefficient. First, compared with the PDTB-2, the PDTB-3 contains about 6,000 more implicit discourse relations. Secondly, although the sense hierarchies of both the PDTB-2 and the PDTB-3 have three levels and have the same senses at Level- 1, several changes at Level-2 and Level-3 due to difficulties found in annotating certain senses. Moreover, the implicit relations that occur within sentences that rarely occur across sentences weren't annotated in the PDTB-2. In the PDTB-3, many more discourse relations are annotated within sentences. Liang et al. (2020) report quite striking difference in the distribution of sense relations inter-sententially vs. intra-sententially between PDTB-2 and PDTB-3. Therefore, these major differences in the PDTB-3 and the PDTB-2 might cause the fluctuation of the coefficient value.

Model	PDTB-2				PDTB-3			
	Top Level		Second Level		Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Dai and Huang (2019)	59.66	52.89	48.23	33.41	-	-	-	-
Shi and Demberg (2019)	61.42	46.40	47.83	-	-	-	-	-
Nguyen et al. (2019)	-	53.00	49.95	-	-	-	-	-
Guo et al. (2020)	57.25	47.90	-	-	-	-	-	-
Kishimoto et al. (2020)	65.26	58.48	52.34	-	-	-	-	-
Liu et al. (2020)	69.06	63.39	58.13	-	-	-	-	-
Jiang et al. (2021)	-	57.18	-	37.76	-	-	-	-
Dou et al. (2021)	70.17	65.06	-	-	-	-	-	-
Wu et al. (2022)	71.18	63.73	60.33	40.49	-	-	-	-
Liu and Li (2016)	-	-	-	-	57.67	46.13	-	-
Chen et al. (2016)	-	-	-	-	57.33	45.11	-	-
Lan et al. (2017)	-	-	-	-	57.06	47.29	-	-
Ruan et al. (2020)	-	-	-	-	58.01	49.45	-	-
Xiang et al. (2022) (BiLSTM)	-	-	-	-	60.45	53.14	-	-
Xiang et al. (2022) (BERT)	-	-	-	-	64.04	56.63	-	-
Ours	**72.18**	**69.60**	**61.69**	**49.66**	**75.31**	**70.05**	**64.68**	**57.62**

Table 3.3: Experimental results on the PDTB-2 and the PDTB-3. For the PDTB-2, we compare with strong baselines from previous work. We report Accuracy (Acc) and Macro-F1 for both the top-level (Level-1) and second-level (Level-2) classification. Bold values indicate the best performance for each column.

Model	PDTB-2				PDTB-3			
	Comp.	Cont	Exp.	Temp.	Comp.	Cont	Exp.	Temp.
Nguyen et al. (2019)	48.44	56.84	73.66	38.60	-	-	-	-
Guo et al. (2020)	43.92	57.67	73.45	36.33	-	-	-	-
Liu et al. (2020)	<u>59.44</u>	60.98	77.66	<u>50.26</u>	-	-	-	-
Jiang et al. (2021)	55.40	57.04	74.76	41.54	-	-	-	-
Dou et al. (2021)	55.72	<u>63.39</u>	**80.34**	44.01	-	-	-	-
Liu and Li (2016)	-	-	-	-	29.15	63.33	65.10	41.03
Lan et al. (2017)	-	-	-	-	30.10	60.91	64.03	33.71
Ruan et al. (2020)	-	-	-	-	30.37	61.95	64.28	34.74
Chen et al. (2016b)	-	-	-	-	27.34	62.56	64.71	38.91
Xiang et al. (2022a) (BiLSTM)	-	-	-	-	34.16	65.48	67.82	40.22
Xiang et al. (2022a) (BERT)	-	-	-	-	35.83	66.77	70.00	42.13
Ours	**65.84**	**63.55**	<u>79.17</u>	**69.86**	**63.30**	**78.60**	**79.91**	**58.39**

Table 3.4: The results for relation types at Level-1 on the PDTB-2 and the PDTB-3 in terms of Macro-F1 (%). Bold values indicate the best performance for each column, and underlined values indicate the second best.

3.4 Results

The experimental results (Tables 3.3–3.6) demonstrate that our method consistently outperforms strong baselines across different evaluation settings on both PDTB-2 and

Second-level Label	Liu et al. (2020)	Wu et al. (2022)	Ours
Temp.Asynchronous	56.18	56.47	**59.79**
Temp.Synchrony	0.00	0.00	**78.26**
Cont.Cause	59.60	64.36	**65.58**
Cont.Pragmatic cause	0.0	0.0	0.00
Comp.Contrast	59.75	**63.52**	62.63
Comp.Concession	0.0	0.0	0.00
Exp.Conjunction	**60.17**	57.91	58.35
Exp.Instantiation	67.96	72.60	**73.04**
Exp.Restatement	53.83	58.06	**60.00**
Exp.Alternative	60.00	**63.46**	53.85
Exp.List	0.0	8.98	**34.78**

Table 3.5: The results for relation types at Level-2 on the PDTB-2 in terms of Macro-F1 (%) (second-level multi-class classification).

PDTB-3. Table 3.3 shows the overall classification results of our method and the baselines at Level-1 and Level-2 on the both PDTB-2 and PDTB-3, with best performances highlighted in bold. For Level-1 classification, Table 3.4 presents the Macro-F1 scores across the four general sense types ("Comparison", "Contingency", "Expansion", and "Temporal").

At Level-1, our method achieves the highest accuracy and Macro-F1 on both datasets, with especially strong gains on PDTB-3. On the PDTB-2, our model achieves 72.18% accuracy and 69.60% Macro-F1 at the top level, significantly surpassing other methods, while on the PDTB-3, our model further improves, reaching 75.31% accuracy and 70.05% Macro-F1 at the top level Across the four general sense types ("Comparison", "Contingency", "Expansion", and "Temporal"), the Macro-F1 scores in Table 3.4 indicate that our approach handles diverse discourse relations robustly. For example, in PDTB-2, "Comparison" and "Temporal" see the largest improvements, suggesting that the representation effectively captures fine-grained semantic distinctions in these categories.

At the second level, our model outperform existing approaches in terms of the overall accuracy and Macro-F1 for both the PDTB-2 and the PDTB-3. Tables 3.5 and 3.6 present the second-level classification results on the PDTB-2 and the PDTB-3 in terms of Macro-F1 scores. Level-2 results further highlight the strength of our approach. On the PDTB-2 (Table 3.5), our model surpasses previous methods in most sense types, with substantial improvements in challenging categories such as "Temporal.Asynchronous" and "Temporal.Synchrony". Even in cases where performance is slightly lower (e.g., "Expansion.Conjunction"), our model remains competitive. On

Second-level Label	Ours
Temp.Asynchronous	66.35
Temp.Synchrony	41.38
Cont.Cause	71.38
Cont.Cause+Belief	0.0
Cont.Condition	74.07
Cont.Purpose	96.05
Comp.Contrast	56.91
Comp.Concession	60.11
Exp.Conjunction	61.70
Exp.Equivalence	11.43
Exp.Instantiation	69.83
Exp.Level-of-detail	55.34
Exp.Manner	78.43
Exp.Substitution	63.77

Table 3.6: The results of different relations on PDTB-3 in terms of Macro F1 (%) (second-level multi-class classification).

the PDTB-3, we provide the first comprehensive Level-2 results (Table 3.6), we do not compare them with previous work.

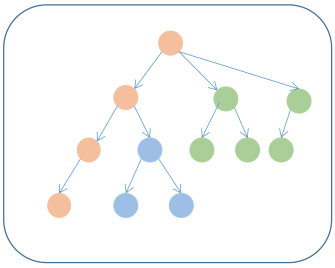
Overall, these findings validate the effectiveness of our proposed method. In the following subsections, we analyze two key factors driving these improvements: (1) the effect of different negative example selection strategies based on the sense hierarchy, and (2) the relative contribution of contrastive learning and our proposed data augmentation techniques.

3.5 Analysis

3.5.1 Comparisons with Other Negatives Selecting Methods

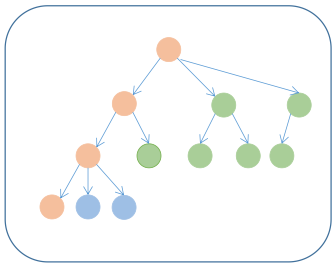
Before settling on our final hierarchy-aware negative-sampling method described in the methodology section, we systematically evaluated four plausible alternatives. As summarized in Figure 3.4, these variants differ in (i) the level at which negatives are drawn (Level 1 vs. Level 2), (ii) whether Level 3 distinctions are enforced, and (iii) whether positives are similarity-weighted. Across PDTB-2/3, our chosen strategies consistently delivered the strongest IDRR performance, so we adopt it for all subsequent experiments. Below we detail the alternatives and report their comparative results.

Concretely, Method 2 below uses examples with different labels at Level-2, while



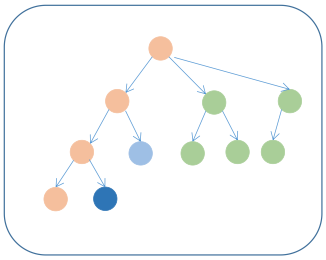
Positive: examples with same label at level-1.
Negative: examples with different labels at level-1.

(a) method 1



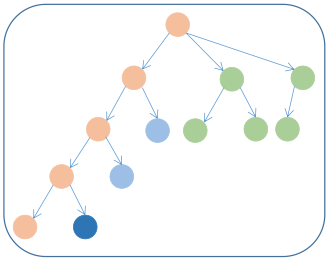
Positive: examples with same label at level-2.
Negative: examples with different labels at level-2.

(b) method 2



Positive: Examples with same label at level-1.
More weight are given to the examples
with same label at level-2.
Negative: Examples with different labels at level-1.

(c) method 3



Positive: Examples with same label at level-1,
more weight are given to the examples
with same label at level-2 or level-3.
Negative: Examples with different labels at level-1.

(d) method 4

Figure 3.4: Another four negative examples selected methods. **orange** ball represent anchor, **green** ball represent negative examples, and **blue** ball represent positive examples. **Darker blue** ball means more weight is given to more similar (potentially) positive examples.

Model	PDTB-2				PDTB-3			
	Top Level		Second Level		Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Method 1	68.91	65.04	58.61	46.27	73.25	68.00	61.17	55.58
Method 2	69.39	63.95	58.33	44.80	73.53	68.36	61.93	54.85
Method 3	69.39	66.53	58.61	39.20	72.49	67.49	60.77	54.33
Method 4	69.10	65.30	57.07	47.46	71.26	66.47	59.53	47.24
Ours	**72.18**	**69.60**	**61.69**	**49.66**	**75.31**	**70.05**	**64.48**	**57.62**

Table 3.7: Comparisons with other negatives defining methods.

Model	Comp.	Cont	Exp.	Temp.
Method 1	63.26	60.42	76.78	59.74
Method 2	60.78	60.82	77.89	56.30
Method 3	59.85	**65.18**	76.43	64.67
Method 4	57.25	61.73	77.30	64.90
Ours	**65.84**	63.55	**79.17**	**69.86**

Table 3.8: The results of relation types at level-1 on PDTB-2 in terms of F1 (%) (top-level multi-class classification).

Methods 1, 3 and 4 use examples with different labels at Level-1. With regard to the use of weight for Method 3 and Method 4, we aim to give more weight to more similar (potentially) positive examples based on the hierarchy. Specifically, we give more weight to the examples from the same Level-2/Level-3 type than to their sister types at Level-2/Level-3 when all of the examples from the same level-1 are positive examples. In addition, Method 4 exploits Level-3 labels, while Methods 1 to 3 only consider Level-1 and Level-2 labels.

In our experiments for alternative negative sampling methods, we use the same hyperparameters as in our main approach for consistency and fair comparison within a shared training setup. While we acknowledge that these methods might benefit from separate hyperparameter tuning, our goal here is to isolate the effect of different negative sampling strategies under controlled conditions. Additionally, it would be costly, as finding the best hyperparameters for each method would require substantial computational resources and time.

For Method 3 and Method 4, the weights of positive examples are set to 1.6 and

1.3 respectively, while the weight of negative examples remains 1. These weights were selected based on preliminary experiments aimed at roughly balancing the influence of positive and negative samples during training. Although not exhaustively tuned, they were sufficient to ensure stable training and reasonable performance across methods. Future work could further explore hyperparameter optimization tailored to each method to better assess their full potential.

In addition, in our setting, the number of negative examples per anchor is mainly influenced by the batch size (256 in our experiments) and the number of categories. While different negative sampling strategies may change the exact number of negatives for the same anchor, we keep the batch size fixed to ensure comparability across methods. A more systematic analysis of the interplay between negative construction and negative quantity—such as how varying batch sizes or class distributions impact learning—remains an important direction for future work.

As shown in Tables 3.7, our negative example selection strategy achieves the highest accuracy and Macro-F1 on the both PDTB-2 and the PDTB-3 for Level-1 and Level-2 classification. On the PDTB-2, our method outperforms the closest competitor (Method 4) by over 3 points in both metrics. At Level-2, the improvement is also clear, indicating the method's advantage in distinguishing fine-grained relations. On the PDTB-3, the gains are also pronounced, exceeding Method 2 by around 1.8 points in accuracy and 1.7 points in Macro-F1. At Level-2, our method outperform Method 2 by roughly 2.5–3 points. Category-level results at Level-1 (Table 3.8) show that our method achieves the highest F1 in most sense types on PDTB-2, particularly in "Comparison" , "Expansion" , and "Temporal" .

These results highlight the robustness of our approach in both top-level and fine-grained discourse relation classification, demonstrating its superior ability. Compared with Method 2, we utilize level-3 labels, which indicates the level-3 label information is helpful for the approach. The greatest difference between our method and other three methods is that our negative examples are only those sister types at level-2 or level-3, not including the examples from different level-1. On the contrary, the negative examples in those three methods are examples from other level-1 types. We suppose that this might make a too strong assumption that examples from different level-1 are very dissimilar. In PDTB datasets, some examples have been annotated with multiple labels. We found that in 99.26% of the 986 co-occurring labels, the labels belong to different Level-1 categories, indicating that cross-level-1 label co-existence is extremely common.

Although only a small portion of instances in the PDTB are explicitly annotated with multiple labels, many instances are likely to have multiple valid senses that are not fully annotated in the dataset.

For example, some examples annotated as "Temporal.Asynchronous" might have the sense of "Contingency.Cause" as well but lack this annotation. This means that in choosing negative examples, relations labeled "TEMPORAL.ASYNCHRONOUS" may closely resemble those labeled "CONTINGENCY.CAUSE" and therefore not be effective as negative examples. Specifically, for the following example from (Moens and Steedman, 1988) :

(10) **when** [they built the 39th Street bridge]$_1$, [they solved most of their traffic problems]$_2$.

If the connective "when" is replaced with "because", the sentence still sounds not strange. Therefore, regarding all examples from different level-1 as negative examples might have some negative impacts on learning the representations.

This incomplete annotation leads to a risk that examples labeled with different Level-1 categories might still share similarities. Consequently, treating instances from different Level-1 categories as negative pairs may introduce noisy or even contradictory training signals, potentially hurting the model's ability to distinguish fine-grained discourse relations.

3.5.2 Ablation Study

We have undertaken ablation studies for further investigating the effectiveness of our method. We compare the performance of RoBERTa, RoBERTa with multi-task learning but without incorporating contrastive learning, and our method using both multi-task learning and contrastive learning to investigate the effectiveness of our contrastive learning method. In addition, we compare model performance with and without our proposed data augmentation method to evaluate its effectiveness.

Effects of contrastive learning algorithm Figures 3.5 and 3.6 show that incorporating multi-task learning (MTL)—jointly predicting Level-1 and Level-2 labels from the same [CLS] representation—yields consistent improvements over separate predictions, confirming the dependency between the two levels. On the PDTB-2, RoBERTa-MTL boosts accuracy from 68.14% to 69.87% and Macro-F1 from 64.87% to 65.39%, while

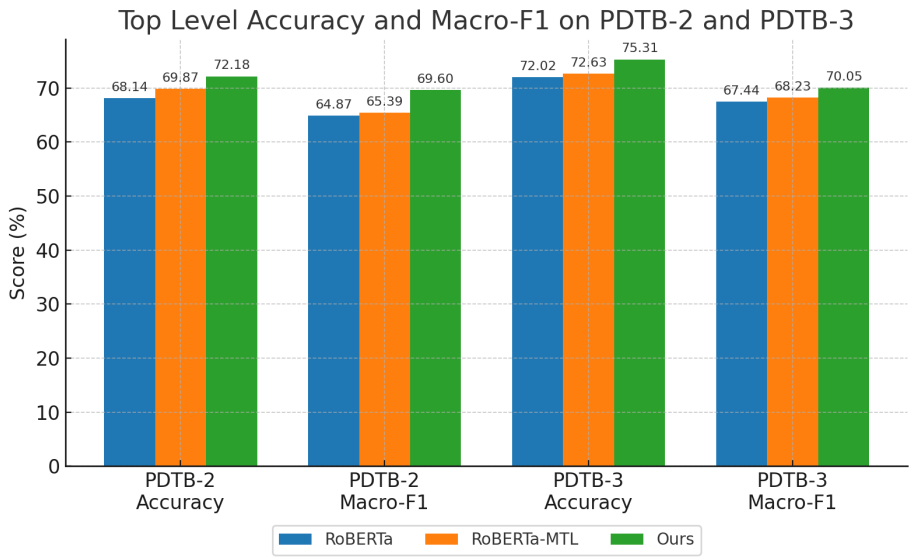


Figure 3.5: Top-level Accuracy and Macro-F1 comparison of RoBERTa, RoBERTa-MTL, and our method on PDTB-2 and PDTB-3.

Datasets	Model	Top Level		Second Level	
		Acc	Macro-F1	Acc	Macro-F1
PDTB-2	Ours	**72.18**	**69.60**	**61.69**	**49.66**
	-augmentation	71.70	67.85	59.19	45.54
PDTB-3	Ours	**75.31**	**70.05**	**64.68**	**57.62**
	-augmentation	73.32	69.02	63.24	51.80

Table 3.9: Effects of data augmentation.

on the PDTB-3, accuracy rises from 72.02% to 72.63% and Macro-F1 from 67.44% to 68.23%. These gains indicate that sharing representations across levels allows the model to capture complementary information from the hierarchy.

Building on MTL, our method further integrates contrastive learning, leading to a substantial performance jump. At Level-1, accuracy reaches 72.18% (+2.31 over RoBERTa-MTL) and Macro-F1 climbs to 69.60% (+4.21) on PDTB-2. On the PDTB-3, the improvements are similar, with 75.31% accuracy and 70.05% Macro-F1. At Level-2, our model also surpasses RoBERTa-MTL across both datasets, demonstrating that contrastive learning not only preserves the benefits of MTL but also strengthens class separation by leveraging the hierarchical structure.

Effects of data augmentation Table 3.9 compares the results with and without data augmentation for both PDTB-2 and PDTB-3, showing that data augmentation is beneficial in generating useful examples. With augmentation, our model achieves 72.18%

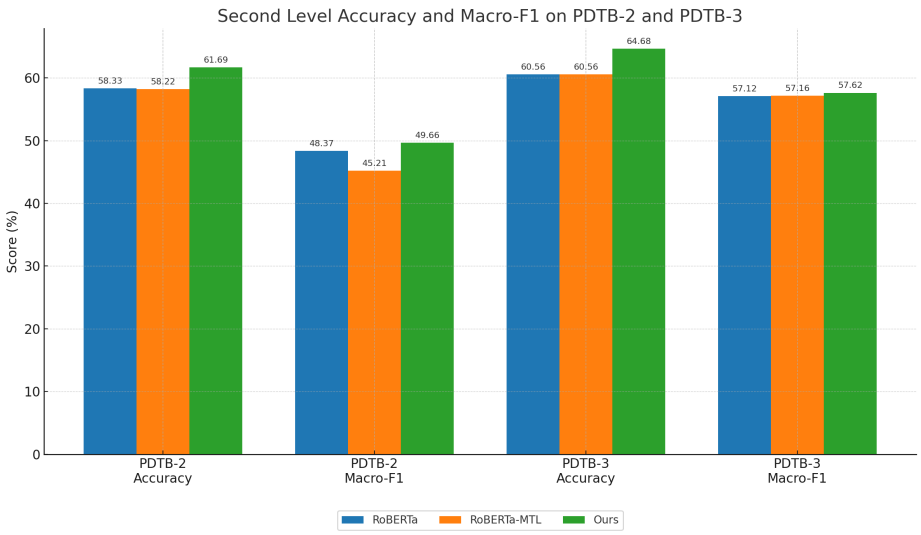


Figure 3.6: Second-level Accuracy and Macro-F1 comparison of RoBERTa, RoBERTa-MTL, and our method on PDTB-2 and PDTB-3.

accuracy and 69.60% Macro-F1 at the top level on PDTB-2, compared to 71.70% and 67.85% without augmentation. At the second level, accuracy drops from 61.69% to 59.19% and Macro-F1 from 49.66% to 45.54% without augmentation. A similar trend is observed on PDTB-3, where augmentation improves top-level accuracy from 73.32% to 75.31% and Macro-F1 from 69.02% to 70.05%.

Khosla et al. (2020) demonstrated that having a large number of hard positives and negatives in a batch improves performance. Given the large number of second-level classes — 11 for PDTB-2 and 14 for PDTB-3 — ensuring sufficient positive examples within a batch size of 256 is challenging. Without data augmentation, the model struggles to fully utilize contrastive learning, leading to a performance drop, particularly in second-level classification.

3.6 Conclusion

In this chapter, we present one of our proposed methods to use the hierarchical label relationships effectively in discourse relation recognition for single label classification framework, incorporating this kind of label inter-relation into the selection of negative examples for contrastive learning. Our experimental results demonstrate that this method achieves superior overall performance compared to previous approaches. Through a systematic comparison of different strategies for selecting negative examples based on the hierarchical structures, we uncovered important insights.

Despite the improvements introduced by the methods in this work, it is important to note that the overall performance remains relatively limited, indicating the ongoing challenge of discourse relation classification. This work lays the foundation for our subsequent research, presented in the next chapter, which builds upon and extends these initial findings. As prompt-based learning had not yet become a common practice at the time of this study, we initially focused on enhancing standard fine-tuning approaches, and later extended our exploration to prompt-learning settings. The following chapter explores a more direct and efficient approach to integrating hierarchical label relationship into model training in prompt-based learning setting. This advanced method not only enhances performance in monolingual setting but also demonstrates improvements in the cross-lingual scenario, representing a step forward in discourse relation recognition.

Chapter 4

Improving Sense Labeling in Prompt-based Learning using the PDTB Sense Hierarchy

This chapter reflects on what we learned from applying hierarchical label information to prompt-based learning for discourse relation recognition. Whereas earlier approaches in this area predominantly relied on pretraining and fine-tuning, recent work shows that prompt-based learning offers a flexible and efficient alternative. Building on this trend, we find that integrating the PDTB sense hierarchy into prompt construction improves model performance. This insight not only addresses limitations of prior prompt-based methods but also highlights the value of hierarchical structure for guiding the model's reasoning.

A key takeaway from our study is that leveraging label hierarchies helps prompt-based models generalize better, especially in scenarios with limited training data. Notably, this approach proves effective beyond English datasets—our cross-lingual experiments suggest that hierarchical prompts support more reliable zero-shot transfer to low-resource languages. These findings emphasize the importance of structured label information for improving performance in discourse understanding in both monolingual and cross-lingual settings. This chapter is based on the work of (Long and Webber, 2024).

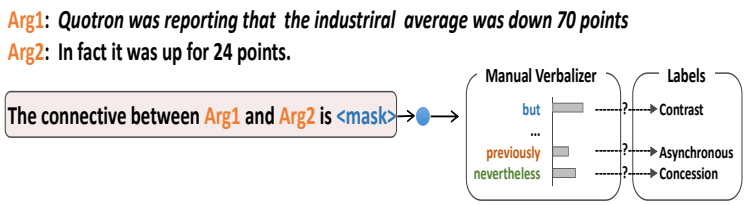


Figure 4.1: The manual verbalizer and the labels.

4.1 Motivation

For implicit discourse relation recognition, several efforts (Zhou et al., 2022; Xiang et al., 2022b; Chan et al., 2023) have applied prompt learning by transforming the task into a connective-cloze task, where the model is prompted to predict a missing discourse connective that best fits the relation between two text spans. Specifically, the input is reformulated into a masked sentence with a placeholder (e.g., [MASK]) representing the missing connective. Figure 4.1 shows the main template and the verbalizer adopted in their work. The model is required to infer the connective for the [MASK] token, and the predicted connective is then mapped to a discourse relation sense label using a manually designed verbalizer.

However, since many connectives can express more than one sense, the connectives they selected can correspond to multiple labels. In the given example depicted in Figure 4.1, the gold label of this example is "Concession". In the manual verbalizer used by Zhou et al. (2022); Chan et al. (2024), the connective "but" was connected with the label "Contrast" and "nevertheless" was mapped for the label "Concession". However, it should be noted that "but" can also be used to indicate other senses such as "Concession" and "Substitution" (see Appendix C in PDTB-3 annotation manual (Webber et al., 2019)). Likewise, "nevertheless" is not strictly limited to conveying "Concession" but can also be employed to indicate "Contrast". Similarly, the connective "since" is not exclusively associated with the label "Pragmatic Cause" but can also be used for "Cause" or "Asynchronous"; The connective "specifically" can be used for "List" and "Restatement", etc. Therefore, while previous approaches have shown improvements in task performance, the manual selection of connectives as verbalizers may not consistently serve as reliable indicators, resulting in potentially less accurate results for certain labels.

Instead of relying on manual design, we introduce an alternative verbalizer to implicit discourse relation recognition (IDRR) by leveraging prototype learning and the sense hierarchy reflecting the organizations of the labels. Our method exploits the hier-

archical organization by learning prototype vectors that capture not only the semantic features of individual relations but also their position within the broader hierarchical structure. These prototypes serve as semantic centroids for each discourse relation class, with higher-level prototypes informing the learning of more specific relation types. By incorporating both prototype learning and this multi-level hierarchical label information, our approach not only enhances monolingual IDRR performance but also enables effective zero-shot cross-lingual transfer.

4.2 Methodology

In this section, we present our approach to using the PDTB sense hierarchy in prompt learning methods. We first introduce the background concepts including prompt-based tuning and prototype learning. Then, we detail our hierarchical prototype-based verbalizer learning method for monolingual IDRR, explaining how we utilize the sense hierarchy to improve sense labeling. Finally, we extend our approach to zero-shot cross-lingual transfer learning scenarios, demonstrating how our method enables effective knowledge transfer across languages without requiring labeled data in target languages. Throughout this section, we explain the contrastive learning objectives that adjust distances between instances and prototypes based on hierarchical label relationships, as well as our template construction process for both monolingual and cross-lingual settings.

4.2.1 Prompt-based Tuning

The original prompt-based tuning approach transforms the downstream task into a cloze task with masks. An example is shown in Figure 4.2. To convert the original input x into a format suitable for masked language modeling, we wrap it with a prompt template. Specifically, we insert x into a predefined textual prompt $T(\cdot)$ of the form: "The discourse relation between Arg1 and Arg2 is [MASK]". This prompt guides the model to predict the sense label at the [MASK] position that best fits the relation between Arg1 and Arg2.

4.2.2 Prototype Learning

Prototype Learning is a method that learns class-specific representations (prototypes) to perform classification tasks (Snell et al., 2017). This approach is inspired by how hu-

mans categorize objects by comparing them to representative examples of each class. In the context of deep learning, prototype learning combines the advantages of representation learning with interpretable decision-making processes.

In prototype learning, the model learns a set of prototypes $P = \{p_1, ..., p_M\}$ where M is the number of prototypes, and each prototype $p_m \in \mathbb{R}^D$ represents a characteristic point in the feature space. The classification process involves comparing input features with these learned prototypes.

4.2.3 Hierarchical Prototype-based Verbalizer Learning for Monolingual IDRR

Figure 4.2 illustrates our framework. We formulate implicit discourse relation recognition (IDRR) as a multi-task learning problem, where the model jointly predicts the Level-1 and Level-2 senses. Each level is treated as a separate classification task, but they share the same encoder, template, and prototype-based verbalizer learning framework. The hierarchical structure is explicitly modeled by optimizing three types of contrastive objectives—instance–instance, instance–prototype, and prototype–prototype—across both levels.

We employ a template where Arg1 and Arg2, representing two discourse segments, are concatenated into a single word sequence. To provide background information, we include the list of Level-1 and Level-2 labels in the prompt before presenting Arg1 and Arg2. This allows the model to be aware of the label space when predicting the implicit discourse relation between Arg1 and Arg2 based on the template.

Following Cui et al. (2022), instead of designing a manual verbalizer or using label words as the answer set for the [MASK] token, we use the hidden states of the [MASK] token to represent instances. These hidden states are then projected into another embedding space for prototype learning. The resulting prototypes serve as the verbalizer during prediction.

Given a piece of training text x wrapped with a template, we take the last layer's hidden state of the [MASK] token, denoted as $\mathrm{h}_{\texttt{[MASK]}} \in \mathbb{R}^d$, as the initial representation of the instance. We apply a linear transformation to map it into a new embedding space:

$$v = E_\phi(x) = W\, h_{\texttt{[MASK]}} + b \tag{4.1}$$

where $W \in \mathbb{R}^{d' \times d}$ is a learnable projection matrix, $b \in \mathbb{R}^{d'}$ is a learnable bias vector,

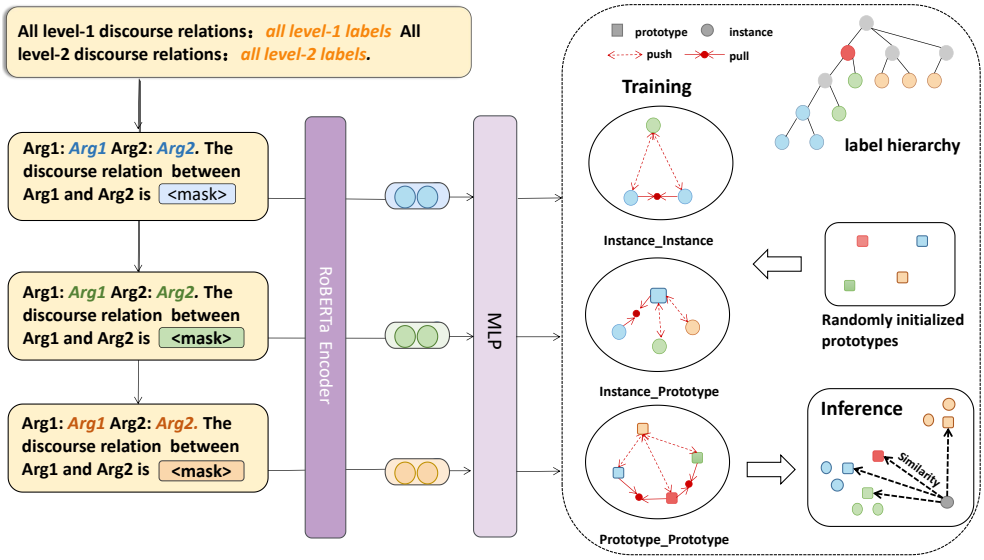


Figure 4.2: The hidden states of [MASK] token represent instances and project them to another embedding space for prototype learning. Three contrastive learning losses adjust the distances among prototypes, the distances among instances, and the distances between prototypes and instances based on the sense hierarchy. Finally, we calculate the similarity scores of query and prototypes during inference.

and $v \in \mathbb{R}^{d'}$ is the resulting instance representation used for prototype matching.

We randomly initialize a set of prototype vectors to represent the label space. Let $C = \{c_1, \ldots, c_M\}$ denote the set of M prototype vectors in the same embedding space $\mathbb{R}^{d'}$.

We use contrastive learning to adjust the distances between class prototypes, the distances between instances, and the distances between prototypes and instances in terms of the sense hierarchy.

(1) For instance to instance pairs, we follow the work described in Chapter 3, to bring representations of instances from the same type closer together, and push apart those from different types, based on the sense hierarchy at Level-2 or Level-3. To encourage the model to learn class-discriminative representations, we define a contrastive loss over instance pairs within the same batch. Specifically, we distinguish between two types of pairs: *intra-class* pairs, where both instances belong to the same class at Level-2 or Level-3, and *inter-class* pairs, where the instances belong to different classes at these levels. The objective encourages intra-class pairs to have higher

similarity scores than inter-class pairs. The instance-instance loss is defined as:

$$\mathcal{L}_{ins\_ins} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|e^i_{pos}|}\sum_{j=1}^{N} 1_{i\neq j} 1_{j\in e^i_{pos}} log\frac{e^{sim(v_j,v_i)/\tau}}{\sum_{k=1}^{N} 1_{i\neq k} e^{sim(v_k,v_i)/\tau}} \tag{4.2}$$

where e^i_{pos} represents the set of positive examples for the i-th instance in the same batch. For every instance in a batch, any one of other examples in this batch can form pairs with. Therefore, if the batch size is N, we will have N^2-1 pairs in total.

(2) For instance to prototype pairs, the purpose of the loss function is to force each prototype to lie at the center point of its instances:

$$\mathcal{L}_{ins\_pro} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{sim(v_i,c_i)/\tau}}{\sum_{j=1}^{M} e^{sim(v_i,c_j)/\tau}} \tag{4.3}$$

where c_i represents the prototype of the corresponding category of the example v_i . We create pairs by pairing each prototype of a class with examples from the same batch. For a specific class, the number of pairs will be N. For the instance-prototype pairs, each example has three levels, so each example has three prototypes. This is to narrow the distances between the example and the prototype of its class and to widen the distances between the example and prototypes of other classes.

(3) For the prototype to prototype pairs, this loss function maximizes intra-class similarity and minimizes inter-class similarity between prototypes. In terms of the sense hierarchy, a father prototype should be nearer to its children prototype than to its non-children prototypes. We pair each father prototype with each child prototype, resulting in a total number of pairs equal to $N_1 \times N_2$, where N_1 is the number of father prototypes and N_2 is the number of child prototypes.

$$\mathcal{L}_{pro\_pro} = -\frac{1}{M}\sum_{i=1}^{M} log\frac{e^{sim(c_i,c'_i)/\tau}}{\sum_{j=1}^{M'} e^{sim(c_i,c_j)/\tau}} \tag{4.4}$$

where c'_i represents the father class of c_i.

Overall, combining the instance-instance loss, instance-prototype loss and the prototype-prototype loss, the final training objective is:

$$\mathcal{L} = \mathcal{L}_{ins\_ins} + \mathcal{L}_{ins\_pro} + \mathcal{L}_{pro\_pro} \tag{4.5}$$

During inference, we calculate the similarity scores of query and prototypes. For instance i, the probability score for class k is

$$P(y_k \mid x) = \frac{e^{sim(v_i,c_i)}}{\sum_{j=1}^{M} e^{sim(v_i,c_j)}}. \tag{4.6}$$

English template:
Argument 1: *Arg1* Argument 2: *Arg2* The discourse relation between the argument 1 and the argument 2 is <mask>.

German template:
Argument 1: *Arg1* Argument 2: *Arg2* Die Diskursbeziehung wischen argument 1 und argument 2 ist <mask>.

Figure 4.3: One example on constructing our language-specific templates.

To make predictions, we use the arg max function, selecting the label y_k that maximizes the probability $P(y_k \mid x)$. We calculate the similarities between the representation of the examples and the level-1 prototypes for the level-1 senses prediction, while calculating the similarities between the representation of the examples and the level-2 prototypes for the level-2 senses prediction.

4.2.4 Hierarchical Cross-lingual Prototype Transfer for Zero-shot IDRR

This subsection describe how we extend our approach to the zero-shot cross-lingual transfer learning scenario, enabling the models to learn the language-agnostic class features. Although we refer to this as "zero-shot" for the target languages, the model is fine-tuned only on English prototypes. Prototypes for other languages are obtained via the the language-specific templates, without using target-language training data. Unlike the monolingual case, the zero-shot setting is a single-task scenario, where the model only predicts the Level-1 sense in the target language.

In order to make use of the approach described in the Section 4.2.3 for the target languages and to enhance cross-lingual representation in zero-shot scenarios, a language-specific template is constructed for each target language. Figure 4.3 displays the example of how we construct language-specific template for German.

The representations of the prototypes for the target languages are trained by using the same methodology described in the Section 4.2.3. After obtaining all class prototypes for the source and the target language, contrastive learning is employed for adjusting the distance among prototypes in the feature space for class-wise alignment. We bring together the source prototypes and target prototypes if they belong to the same class, while simultaneously pushing away the source and target prototypes if they are from different classes. This behavior is illustrated in Figure 4.4.

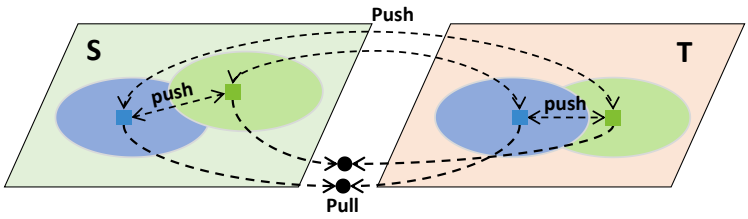


Figure 4.4: Prototype Alignment between the source and the target languages. If the source prototypes and target prototypes are from the same class, we pull them closer, otherwise we push them away.

4.3 Experimental Setting

We evaluate our approach on both monolingual and cross-lingual scenarios using established discourse relation datasets.

4.3.1 Datasets

For monolingual evaluation, mirroring the setup in Chapter 3, we conduct experiments on both PDTB-2 (Prasad et al., 2008) and PDTB-3 (Webber et al., 2019). Following standard practice (Bai and Zhao, 2018; Liu et al., 2020; Xiang et al., 2022a), we used Sections 2-20 for training, Sections 0-1 for validation, and Sections 21-22 for testing. In line with (Qin et al., 2016), instances with multiple annotated labels were treated as separate examples during training, with predictions considered correct during testing if they matched any of the gold labels. We exclude the labels for both PDTB-2 and PDTB-3 which have insufficient data as described earlier in the Section 3.3.1.

For zero-shot cross-lingual evaluation, we utilize TED-MDB (Zeyrek et al., 2019) as our target language dataset. TED-MDB is a parallel corpus following the PDTB-3 framework, comprising manual annotations of six TED talks across seven languages: English, Turkish, Portuguese, Polish, German, Russian, and Lithuanian. Table 4.1 presents the distribution of top-level senses of the implicit discourse relations in the TED-MDB.

Although the texts are parallel across languages, the distribution of top-level senses differs. "Expansion" is the most frequent category in all languages, ranging from 62.60% in Lithuanian to 76.47% in Russian. "Contingency" appears more often in Turkish (24.36%) than in Polish (14.56%). "Temporal" is the least frequent, with the lowest occurrence in Russian (2.26%) and the highest in Polish (9.23%).

These differences suggest that implicit discourse relations may vary across lan-

Language	Comparison	Contingency	Expansion	Temporal	Total
German	13(10.31%)	41(19.16%)	148(69.16%)	12(5.61%)	214(100%)
Lithuanian	26(6.07%)	53(21.54%)	148(62.60%)	13(5.28%)	246(100%)
Polish	19(9.74%)	28(14.56%)	130(66.67%)	18(9.23%)	195(100%)
Portuguese	23(9.06%)	47(18.50%)	169(66.54%)	15(5.91%)	254(100%)
Russian	16(7.24%)	31(14.03%)	169(76.47%)	5(2.26%)	221(100%)
Turkish	20(9.90%)	29(24.36%)	140(69.31%)	13(6.64%)	202(100%)

Table 4.1: Distribution of top level senses of the implicit discourse relations in TED-MDB corpora

guages, even in parallel texts. One possible explanation is that discourse relations are not always preserved during translation. That is, although the content of sentence 1 and sentence 2 may remain semantically aligned in their translations, the implicit relation between them may shift due to language-specific discourse conventions, translator choices, or structural differences between languages. For instance, a causal relation in one language might be rendered as an elaboration in another, depending on how the relation is pragmatically conveyed. This cross-linguistic variation highlights the challenge of modeling implicit discourse relations in a multilingual setting.

Given the limited annotations per language (approximately 200 implicit discourse relations), we focus our evaluation on top-level senses. We use the PDTB-3 as our source language dataset, as the TED-MDB follows the same sense hierarchy.

4.3.2 Baselines

Since most previous research evaluated on either the PDTB-2 or the PDTB-3 dataset, we compare different baselines specific to each dataset. All baselines are designed for implicit discourse relation recognition (IDRR) and report results on the standard splits, allowing like-for-like evaluation. Because our work studies hierarchy-aware and prompt-based IDRR, we include baselines that inject label dependencies only at inference time (e.g., Wu et al. 2022) and prompt-based connective prediction (Xiang et al. 2022b; Zhou et al. 2022; Chan et al. 2023) to contrast with our hierarchy-aware training and prompting. All the baselines are widely cited or strong performers (often state-of-the-art at publication), so our comparisons reflect the capability frontier at the time.

The baselines for PDTB-2 are: (1) a model which incorporates external event

knowledge and coreference relations (Dai and Huang, 2019); (2) a neural model which learns better argument representations by utilizing the inserted connectives (Shi and Demberg, 2019); (3) a model used to predict the labels and connectives at the same time (Nguyen et al., 2019); (4) a working memory-driven neural networks that uses a knowledge enhancement paradigm (Guo et al., 2020); (5) a model which performs additional pre-training on text tailored to discourse classification (Kishimoto et al., 2020); (6) a RoBERTa-based model that combines a powerful contextualized representation module, a bilateral multi-perspective matching module, and a global information fusion module (Liu et al., 2020); (7) a method that recognizes the relation label and generates the target sentence containing the meaning of relations simultaneously (Jiang et al., 2021); (8) a method that develops a re-anchoring strategy by using Conditional VAE (CVAE) (Dou et al., 2021); (9) a hierarchical contrastive learning based multi-task framework (Long and Webber, 2022); (10) a transformed prompt-based Connective prediction (PCP) task by utilizing the correlation between connectives (Zhou et al., 2022);(11) a method injecting the label dependencies information via prompt tuning with aligning the representations and using connectives prediction (Chan et al., 2023).

PDTB-3 was released in 2019, and few early models were directly developed for it. We rely on baselines from Xiang et al. (2022a), who reimplemented several strong PDTB-2 models on PDTB-3. We compare the following baselines for PDTB-3: (1) a model with multi-level attention (NNMA) (Liu and Li, 2016); (2) a method which adopts a gated relevance network to capture the semantic interaction (Chen et al., 2016a); (3) a multi-task attention based neural network model through two types of representation learning (Lan et al., 2017); (4) a propagative attention learning model using a cross-coupled two-channel network (Ruan et al., 2020); (5) a multi-Attentive Neural Fusion (MANF) model to encode and fuse both semantic connection and linguistic evidence for IDRR (Xiang et al., 2022a);(6) a hierarchical contrastive learning based multi-task framework (Long and Webber, 2022);(7) a connective-cloze Prompt (ConnPrompt) to transform the relation prediction task as a connective-cloze task by designing insert-cloze Prompt and Prefix-cloze Prompt (Xiang et al., 2022b).

For zero-shot cross-lingual experiments, we compare with 2 baselines: (1) vanilla fine tuning the multilingual pre-trained language model on the PDTB-3 and tested on target language test set; (2) Kurfalı and Östling (2019), which is the first and only study about a zero-shot cross-lingual transfer learning for implicit discourse relation recognition by using the PDTB-3 and the TED-MDB . We only use PDTB-3 as the source language dataset, so we compare the results with respect to their models trained

on the PDTB-3 and tested on the TED-MDB.

4.3.3 Implementation Details

In our monolingual experiments, we use RoBERTa-base (Liu et al., 2019) as the PLM backbone. We adopt Adam (Kingma and Ba, 2015) with the learning rate of $5e{-}5$ and the batch size of 196 to update the model. The maximum training epoch is set to 10 and the wait patience for early stopping is set to 5 for all models. This setting differs from that in Section 3.3.2, primarily due to the fact that preliminary experiments showed that this configuration achieves competitive performance while reducing training cost. The temperature τ is set to 0.1. For prototype learning, we set the prototype dimension to 128. All experiments are performed with 1× 80GB NVIDIA A100 GPU. Accuracy and Macro-F1 score are used as evaluation metrics. We use the same model and parameters for PDTB-2 and PDTB-3.

For our zero-shot cross-lingual transfer learning method, we use the same parameters presented in the monolingual experiments. For all cross-lingual experiments (including ours and vanilla fine tuning), we use XLM-RoBERTa base (Conneau et al., 2020) as our multilingual backbone model.

4.4 Main Results

4.4.1 Results for Monolingual Scenario

Table 4.2 compares recent prompt-based and RoBERTa-based systems on PDTB-2/3 at both Level-1 and Level-2. Our RoBERTa-based approach is consistently best at Level-1 on both datasets, edging out strong prompt methods. On the PDTB-2, it improves over DiscoPrompt (Chan et al., 2023) by about +0.8 Acc (71.70 $\rightarrow$ 72.47) and +3.9 Macro-F1*(65.79 $\rightarrow$ 69.66). On the PDTB-3, it surpasses ConnPrompt (Xiang et al., 2022b) with +1.0 Acc (74.36 $\rightarrow$ 75.37) and +1.3 Macro-F1 (69.91 $\rightarrow$ 71.19). These gains indicate that leveraging hierarchical prototypes provides more discriminative representations than prompt-only formulations at the coarse (Level-1) sense granularity.

At Level-2, our model remains competitive and shows clear F1 advantages. On the PDTB-2, our model achieves 60.73% accuracy and 47.07% Macro-F1 on the PDTB-2, compared to 61.02% and 43.68% in DiscoPrompt (Chan et al., 2023). On the PDTB-3, our model achieves 63.53% accuracy and 52.91% Macro-F1, outperforming our results

Model	PDTB-2				PDTB-3			
	Top Level		Second Level		Top Level		Second Level	
	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1	Acc	Macro-F1
Dai and Huang (2019)	59.66	52.89	48.23	33.41	-	-	-	-
Shi and Demberg (2019)	61.42	46.40	47.83	-	-	-	-	-
Nguyen et al. (2019)	-	53.00	49.95	-	-	-	-	-
Guo et al. (2020)	57.25	47.90	-	-	-	-	-	-
Kishimoto et al. (2020)	65.26	58.48	52.34	-	-	-	-	-
Liu et al. (2020)	69.06	63.39	58.13	-	-	-	-	-
Jiang et al. (2021)	-	57.18	-	37.76	-	-	-	-
Dou et al. (2021)	70.17	65.06	-	-	-	-	-	-
Wu et al. (2022)	71.18	63.73	60.33	40.49	-	-	-	-
Long and Webber (2022)	71.70	67.85	59.19	45.54	-	-	-	-
Zhou et al. (2022)	70.84	64.95	60.54	41.55	-	-	-	-
Chan et al. (2023)(T5-base)	71.70	65.79	**61.02**	43.68	-	-	-	-
Liu and Li (2016)	-	-	-	-	57.67	46.13	-	-
Chen et al. (2016b)	-	-	-	-	57.33	45.11	-	-
Lan et al. (2017)	-	-	-	-	57.06	47.29	-	-
Ruan et al. (2020)	-	-	-	-	58.01	49.45	-	-
Xiang et al. (2022a) (BiLSTM)	-	-	-	-	60.45	53.14	-	-
Xiang et al. (2022a) (BERT)	-	-	-	-	64.04	56.63	-	-
Long and Webber (2022) (RoBERTa-base)	-	-	-	-	73.32	69.02	63.24	51.80
Xiang et al. (2022b) (RoBERTa-base)	-	-	-	-	74.36	69.91	-	-
Ours (RoBERTa-base)	**72.47**	**69.66**	60.73	**47.07**	**75.37**	**71.19**	**63.53**	**52.91**

Table 4.2: Experimental results on the PDTB-2 and the PDTB-3. We report accuracy and macro-averaged F1 scores for both top-level and second-level senses. Our model outperforms previous RoBERTa-base baselines on most metrics.

shown in Chapter 3 without the data augmentation method (63.34% and 51.80%). Together, these results suggest that injecting label hierarchy into training via prototype learning) better preserves fine-grained distinctions, translating into consistent Macro-F1 gains while maintaining strong overall accuracy.

4.4.2 Results for Cross-lingual Scenario

The results for the top-level classifications on TED-MDB are presented in Table 4.3. As observed, our method achieves better overall results for all six languages in TED-MDB. Compared with (Kurfalı and Östling, 2019), the approach improves on the average Macro-F1 by about 8% on Turkish. Better performance is achieved in the Portuguese language, with an F1 score of 43.42%. This result is nearly 10% higher than the baseline XLMR-based vanilla fine tuning. These outcomes serve as evidence of the effectiveness of our zero-shot cross-lingual learning methods.

Model	German	Lithuanian	Polish	Portuguese	Russian	Turkish
Kurfalı and Östling (2019)	39.22	39.32	37.54	39.33	35.50	33.52
XLMR-base	41.61	36.00	35.51	34.39	33.11	36.73
Ours	**45.24**	**41.97**	**40.49**	**43.32**	**37.37**	**41.12**

Table 4.3: Macro-F1 scores (%) for top level classification on 6 languages when the model is trained on PDTB-3.

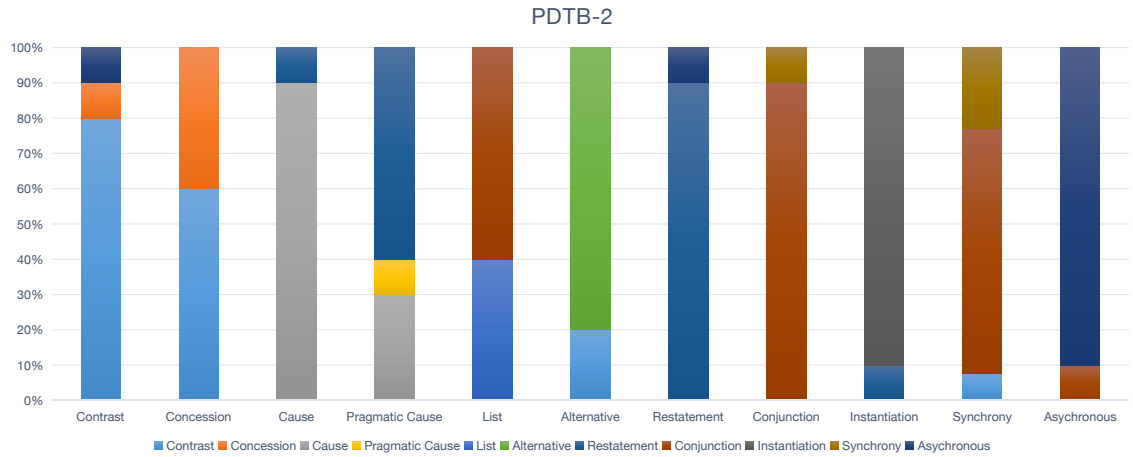
4.5 Analysis

4.5.1 Nearest Neighbors for each Learned Prototype

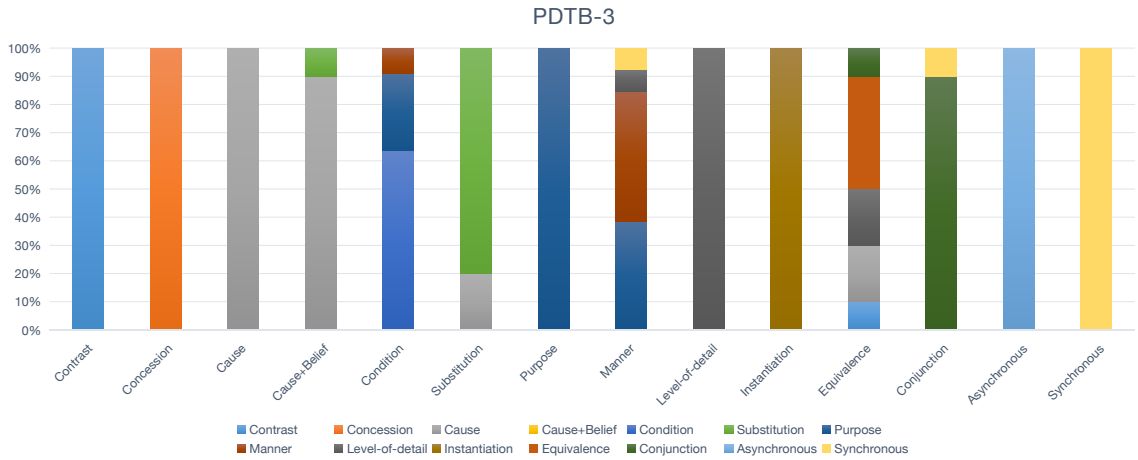
We investigate which examples are the most similar to the learned prototypes by retrieving the top ten nearest neighbors for each second-level prototype in both PDTB-2 and PDTB-3 for Monolingual Scenario. For each prototype, we examine how many of its neighbors share the same class and, if not, what alternative labels they have. As shown in Figure 4.5, a majority of the neighbors do share the same class as the prototype, supporting the effectiveness of the learned prototypes in capturing class-specific characteristics. However, for those neighbors that do not belong to the same class, we observe that they often come from semantically related discourse relations. For example, "Contrast" and "Concession" in PDTB-2. This suggests that the model may struggle to distinguish between closely related classes.

Moreover, PDTB-3 consistently yields a higher proportion of same-class neighbors compared to PDTB-2 on average, 8.7 out of 10 neighbors in PDTB-3 match the prototype's class, whereas this number drops to 6.9 in PDTB-2. This difference indicates a substantial improvement in prototype quality with PDTB-3 and further highlights the advantages of using PDTB-3 for prototype-based modeling.

Upon closer examination of the figure for PDTB-2, we can see that for the prototype learned for the discourse relation "Synchrony", we can see a significant number of its closest neighbors belong to the "Conjunction" label, while the figure for PDTB-3 shows that 40% of the closest neighbors for the discourse relation "manner" is "Purpose". "manner" and "Purpose" are semantically related, as the way an action is performed often aligns with its intended goal; likewise, "Synchrony" and "Conjunction" is very likely to co-occur, since simultaneous events are often described in an additive structure. This suggests the presence of multi-label examples in the data, even though



(a) Label distribution of the top ten nearest neighbors for each second level prototype in PDTB-2. Each bar represents a prototype category, and the colors indicate the proportion of different discourse relation labels among its top ten nearest neighbors. This visualization helps assess how well the learned prototypes align with their expected categories by showing the consistency of nearest neighbors' labels.



(b) Label distribution of the top ten nearest neighbors for each second level prototype in PDTB-3. Similar to (a), this figure shows the distribution of discourse relation labels for PDTB-3 for each prototype category.

Figure 4.5: Label distribution of the top ten nearest neighbors for second level prototypes in PDTB-2 and PDTB-3.

the model is trained to predict only a single label per instance, and incorporating multi-label modeling probably could better capture the complexity of discourse relations.

4.5.2 Ablation Studies

In order to know the effects of each of the three losses and the effects of the list of labels provided in our template, we carried ablation studies on both PDTB-2 and PDTB-3.

We conduct an ablation study by evaluating different loss components: 1) only

Datasets	Model	Top Level		Second Level	
		Acc	Macro-F1	Acc	Macro-F1
PDTB-2	Ours	**72.47**	**69.66**	**60.73**	**47.07**
	w/o ins_ins	71.51	68.44	59.19	45.73
	w/o pro_pro	70.83	67.32	58.32	42.10
	w/o ins_ins & pro_pro	68.14	63.27	56.02	42.76
	w/o label information	72.18	69.07	59.86	45.33
PDTB-3	Ours	**75.37**	**71.19**	**63.13**	**52.91**
	w/o ins_ins	73.59	69.80	61.31	50.48
	w/o pro_pro	72.77	69.64	61.00	50.27
	w/o ins_ins & pro_pro	71.67	67.63	60.08	49.36
	w/o label information	74.48	70.52	62.00	51.19

Table 4.4: Ablation study on PDTB-2 and PDTB-3.

using instance-to-prototype loss, 2) combining instance-to-prototype and prototype-to-prototype losses, and 3) combining instance-to-prototype and instance-to-instance losses. From Table 4.4, all three losses contribute to performance, with prototype-to-prototype loss being more crucial than instance-to-instance loss. On the PDTB-2, removing instance-to-instance loss reduces top-level accuracy from 72.47% to 71.51% and Macro-F1 from 69.66% to 68.44%, while removing prototype-to-prototype loss decreases accuracy to 70.83% and Macro-F1 to 67.32%. The trend is similar on the PDTB-3, where removing instance-to-instance loss results in accuracy dropping from 75.37% to 73.59% and Macro-F1 from 71.19% to 69.80%.

Additionally, we analyze the impact of providing the list of Level-1 and Level-2 labels by removing them from the template. Without this information, Macro-F1 drops from 69.66% to 69.07% on PDTB-2 and from 71.19% to 70.52% on PDTB-3, confirming that providing the list of label improves model performance. Overall, combining all three losses and incorporating label information yields the best results.

4.6 Conclusion

In this chapter, we introduce a verbalizer for prompt learning that does not require manual design and employs the hierarchical label relationship reflecting the organization of discourse relations for sense labeling in the prompt learning scenarios. Extensive experiments show this method outperforms competitive baselines and effectively supports zero-shot cross-lingual learning for low-resource languages.

The studies presented in Chapter 3 and this chapter highlight the significance of the PDTB sense hierarchy in identifying discourse relations while using single-label classification frameworks, paving the way for more advanced methods to apply the hierarchical label relations into discourse relation recognition.

Additionally, through our investigations in Chapter 3 and Chapter 4, we observe that instances that might have the second labels could impact the effectiveness of our applied methods like contrastive learning and prototype learning. While these issues were not extensively explored in Chapters 3 and 4, we recognize the importance of addressing them to enhance system performance and deepen our understanding of the task's complexities. Consequently, in the next chapter, we investigate the application of multi-label classification frameworks for discourse relation learning and make use of the sense hierarchy to understand its effectiveness within this context.

Chapter 5

Enhancing Multi-label Classification with the PDTB Sense Hierarchy

The previous two chapters demonstrated that leveraging label hierarchies can improve discourse relation learning in both fine-tuning and prompt-based settings. However, these approaches treated discourse relation recognition as a single-label classification task, despite the fact that the PDTB allows multiple sense labels per instance. This simplification overlooks the complexity of real-world discourse, where multiple relations can co-occur within a single example.

Our analysis reveals that single-label frameworks struggle to capture such nuances. For instance, as shown in Section 4.5.1, the learned prototype for "Manner" shares many nearest neighbors with "Purpose"—two senses that often co-occur. This suggests that enforcing strict boundaries between labels during training may distort the natural overlap among discourse relations. Moreover, evaluating performance based on matching any single label provides a limited view, failing to reflect the model's ability to fully capture multi-label phenomena.

To address these limitations, we turn to multi-label classification frameworks. Through experimentation, we learn that such frameworks not only preserve strong performance on single-label cases but also more accurately reflect the overlapping nature of discourse relations. This shift in perspective enables a more faithful modeling of linguistic reality and supports more consistent predictions.

Building on our earlier work (Long et al., 2024), this chapter further investigates how hierarchical label structures can be integrated into multi-label classification via contrastive learning. A key insight from this study is that contrastive learning, when aligned with hierarchical label relationships, significantly enhances the model's abil-

ity to distinguish and represent multiple co-occurring discourse senses. In contrast, applying contrastive learning without such structural guidance can degrade performance, highlighting that label relationships must be explicitly considered when modeling sense interactions.

Overall, this chapter contributes to a deeper understanding of how discourse sense hierarchies interact with classification objectives. It suggests that future approaches should avoid treating all discourse relations as mutually exclusive and instead embrace their structural and semantic relationships. Doing so can lead to more accurate, interpretable, and linguistically grounded models.

5.1 Motivation

In PDTB annotation, annotators can assign multiple sense labels to an example when they believe that both hold simultaneously, as discussed in Section 2.2.3. However, for those instances with two annotated labels, all previous work on discourse relation recognition treat them as separate and different examples during training, and at test time, a prediction matching one of the gold types is taken as the correct answer.

Nevertheless, by treating instances with multiple labels as separate examples, the model may not effectively capture the inherent complexity of discourse relations. Real-world texts often contain multiple layers of meaning, and forcing the model to treat them as distinct instances may oversimplify the problem. When implementing discourse relations recognition in downstream tasks, the inability to recognize multi-labels could potentially lead to adverse effects. For example, if the model fails to identify both "Concession" and "Asynchronous" relation in one example simultaneously, it may struggle to respond to questions concerning temporal order and contingency relation concurrently.

To address these negative impacts, we explore multi-label classification as a more effective way to capture the complexity of discourse relations. This is the first study to treat implicit discourse relation recognition as a multi-label classification issue. In PDTB-3's multi-label annotations, multiple labels can correspond to either entirely implicit or explicit relations, as well as cases where implicit relations occur alongside explicit ones or explicit relations are linked with AltLex ones. However, our work only focuses on instances where all assigned labels are implicit.

We explore different multi-label classification methods for implicit discourse relation recognition, and we show that a multi-label classifier can demonstrate better

performance than a classifier trained on examples in which multiple labels were split into two distinct and unrelated examples. Additionally, it is interesting for us to know whether applying hierarchical sense relationships could facilitate multi-label classification methods for discourse relation recognition.

5.2 Methodology

This section presents the methodological approaches used in our study. We begin by introducing three different multi-label classification methods: two encoder-only approaches and one encoder-decoder approach. Next, we explore the application of contrastive learning, both with and without consideration of the hierarchical label structure, to enhance the performance of sense labeling. Additionally, we evaluate the impact of focal loss which can help handle label imbalance and examine different cross-validation strategies, comparing section-level and example-level splits to optimize model generalization.

5.2.1 Multi-label Classification Methods

Our work has explored three different multi-label classification techniques, two encoder-only methods and one encoder-decoder method. We will introduce them in this subsection.

5.2.1.1 Method 1

The output vector corresponding to the [CLS] token aggregates input features and is used for classification. We employ RoBERTa (Liu et al., 2019) for text representation learning, and add a classification head $W_c \in \mathbb{R}^{H \times |C|}$ on top of the [CLS] token to do classification. H is the dimension size of [CLS] representation and C represents the number of classes. We use $y \in \mathbb{R}^{|C|}$ to denote the ground-truth label for an example, where $y \in \{0,1\}^{|C|}$.

The model is trained using sigmoid binary cross-entropy loss. If the predicted probability of a label surpasses 0.5, it is regarded as a predicted label.

5.2.1.2 Method 2

This method resembles Method 1, with several key distinctions. Rather than employing a single classification head to handle all labels, we utilize multiple classification heads

$W_{c_i} \in \mathbb{R}^{H\times 2}$, each c_i dedicated to the i-th specific label and treating them as individual binary classification tasks. In contrast to Method 1, which utilizes sigmoid binary cross-entropy loss, we employ softmax cross-entropy for loss calculation here. The loss for each label is computed independently, and subsequently, the mean of these individual losses is used to update the model. If the predicted probability of a label is greater than 0.5, it is considered a predicted label.

5.2.1.3 Method 3

In this approach, we use a sequence-generating model that processes input text token by token, predicting labels sequentially while considering previously predicted labels. Our method is similar to the one described in Yarullin and Serdyukov (2020). We utilize RoBERTa's last transformer block to generate word vectors and use RoBERTa's [CLS] token embedding as the initial hidden state for our decoder, which is a Gated Recurrent Unit (GRU) in our case. Our model also incorporates a dot-product attention mechanism between encoder and decoder. The motivation behind this architecture is to model the dependency among multiple labels more explicitly, by generating them sequentially in a way that each predicted label conditions on the previous ones, thus capturing label co-occurrence patterns naturally.

We train the final model to minimize the cross-entropy objective loss for a given x and ground-truth labels $\{t_1^*, t_2^*, \ldots, t_k^*\} \in \mathcal{L}$:

$$\mathcal{L}_{CE}(\theta) = -\sum_{i=1}^{k} \log P_\theta(t_i^* \mid x, t_{1:i-1}^*)$$

During inference, we conduct a beam search to identify candidate sequences with the lowest objective scores among the paths that conclude with the <eos> token. The beam size is set to 4 considering that there are only 14 labels for Level-2 prediction and the multi-label examples in the dataset only have two annotated labels.

5.2.2 Integrating Hierarchical Label Relations into Multi-label Classification

We investigate whether incorporating the hierarchical label relations into the multi-label classification method can enhance the model's ability to make more accurate predictions for implicit discourse relations. For the multi-label examples, we only consider the cases whose labels are entirely implicit.

This subsection compares three methods based on the Method 2 described in section 5.2.1. Method 2 is selected as the comparison base since it achieves better performance than Method 1 and Method 3, thus providing a stronger foundation for evaluating the impact of contrastive learning. The three methods are Method 2 without contrastive learning, applying contrastive learning into Method 2 but ignoring the label sense hierarchy, applying contrastive learning which leverages the hierarchical label relations.

We begin with Method 2 described in the previous subsection as the baseline method, which does not incorporate contrastive learning. Building upon this method, we introduce contrastive learning, first without and then with explicit use of the hierarchical label structure. The methods described here is to investigate the effectiveness of hierarchical label knowledge in multi-label classification frameworks. By analyzing these methods, we aim to assess the effectiveness of applying the PDTB sense hierarchy into contrastive learning frameworks for multi-label classification.

5.2.2.1 Contrastive Learning that Ignores Hierarchical Label Relations

This method extends the baseline by incorporating contrastive learning, designed to improve the representation without explicitly leveraging the hierarchical structure of labels.

Positive Examples: Positive pairs are defined as instances that share at least one label at the second level in the hierarchy. For example, if instance A has labels [“Concession”, “Synchronous”] and instance B has either “Concession” or “Synchronous” or both, they form a positive pair. For single-label instances, examples sharing the same second-level label (e.g., both having “Concession”) are considered positive pairs.

Negative Examples: Negative pairs are those within the batch, with all remaining instances (i.e., those that do not share labels at the second level) treated as negatives at a predefined random ratio. This ratio is set to 0.5 in this experiment. We set the ratio of negative examples to 0.5 based on preliminary experiments. We observed that 0.5 consistently provided stable performance across validation runs, and thus we use it throughout our experiments.

5.2.2.2 Contrastive Learning that Considers Hierarchical Label Relations

This method further enhances baseline method by explicitly incorporating the hierarchical label relationships into the contrastive learning framework.

Positive Examples: The definition of positive pairs is the same with the above method, in which instances that share at least one label at the second level in the sense hierarchy are positive examples.

Negative Examples: The crucial distinction in this approach lies in the selection of negative pairs. The method identifies negative pairs exclusively as instances that are sister types at the second level of the hierarchy. Unlike the approach without considering the hierarchical label relationship, it excludes examples from different level-1 labels from serving as negative examples. This design choice reflects the understanding that examples from different level-1 categories may still share semantic similarities. This approach aligns with our previous findings in single-label classification (Section 3.5.1), where using only second and third-level sisters as negative examples prove more effective than including cross-level-1 examples.

5.2.3 Other Methodological Exploration

5.2.3.1 Focal Loss vs. Standard Cross-Entropy Loss

Focal Loss, an adaptation of Cross-Entropy Loss (CE), addresses class imbalance by emphasizing challenging examples. Initially designed for object detection in computer vision by (Lin et al., 2017), it has been applied in recent NLP studies (Tan et al., 2022; Wang et al., 2022). Our dataset, PDTB-3, exhibits imbalances in single-label and multi-label data. While single-label tasks often use standard cross-entropy loss for IDRR, we explore focal loss for the multi-label classification in IDRR.

Focal Loss reduces weights for well-classified instances and emphasizes challenging ones, modulating the loss for confidently predicted instances. The focal loss function is defined as:
for positive samples $(y = 1)$:

$$L_{fl}(p) = (1-p)^{\gamma}\log(p) \tag{5.1}$$

for negative samples $(y = 0)$:

$$L_{fl}(p) = p^{\gamma}\log(1-p) \tag{5.2}$$

Here, γ serves as the focusing parameter, controlling the rate at which easy instances are down-weighted. We set it to 1 for positive samples, and 4 for negative to place relatively more emphasis on hard negative cases, as the multi-label setting in PDTB-3

contains a larger proportion of easy negatives, which may otherwise dominate the loss and hinder the learning of minority positive labels.

In our implementation, we replace the cross-entropy loss in Method 2 with the aforementioned focal loss function without further modifications.

5.2.3.2 Cross-Validation Strategies: Example Level vs. Section Level

Splitting the data at section-level for PDTB means dividing the dataset based on entire sections rather than randomly splitting individual examples. This ensures that the training and test sets come from different sections, preventing better evaluating the model's generalization ability. However, this splitting method may not be optimal for multi-label classification, since the multi-label examples are limited in number and are not evenly distributed across sections. Therefore, alternative strategies can be considered. We explore an example-level method to offer a better mix of examples to train more robust models, especially when dealing with sparse labels. We first separate the multi-label data from the single-label data. Then, we divide the multi-label data into 12 portions, and the single-label data are also divided into 12 portions, with each portion having the same proportion of the number of each label. For multi-labeled data, we calculate the proportion of the pair of labels based on the frequency of the pair and ensure each portion reflects a similar distribution. Next, we combine one of the 12 portions of multi-label data with one of the 12 portions of single-label data to obtain a merged set of 12 data portions. Finally, we randomly select one portion as the test set, another portion as the validation set, and the remaining 10 portions as the training set, thus creating 12 folds of cross-validation data.

5.3 Experimental Setting

5.3.1 Dataset and Evaluation

We employ PDTB-3 (Webber et al., 2019) for our evaluation. PDTB-3 represents an advancement over PDTB-2 Prasad et al. (2008), offering a more extensive collection of annotated multi-label examples. In our study, following the work in Chapters 3 and 4, we focus exclusively on implicit discourse relations because they are more challenging to classify, requiring models to infer meaning from arguments rather than relying on explicit lexical cues. This ensures a more robust learning process. Additionally, since our single-label instances consist solely of implicit discourse relations, including

Label	Number
Cause/Level-of-detail	101
Cause/Manner	100
Purpose/Manner	378
Synchronous/Contrast	112

Table 5.1: Label counts for level-2 sense pairs with more than or equal 100 annotated instances in PDTB-3.

multi-label cases where implicit relations co-occur with explicit or AltLex relations could introduce inconsistencies in the labeling scheme. We drop all implicits that are linked to some other types of relations such as explicit or Altlex.

Additionally, following previous work, including our studies in the previous two chapters, we exclude data from Level-2 label categories with fewer than 50 annotated instances for our single-label instances in PDTB-3. The statistics of single label instances have been presented in Section 3.3.1.

We concentrate exclusively on implicit discourse relations, disregarding those with explicit connectives. About 5% of PDTB-3 implicit discourse relations receive multiple labels, which corresponds to instances with two annotated labels. We treat such instances as single examples with multi-labels during training, and during testing, predictions were considered correct only if they match the specific label. For multi-label examples, we only exclude those examples containing the labels that have been excluded for they do not have sufficient annotated data in single label examples such as "Condition+SpeechAct", "Concession+SpeechAct", and "Disjunction". We do not apply additional filtering based on the frequency of labels within the multi-label examples, meaning that low-frequency label pairs in the multi-label setting are retained as long as they were not previously excluded. This decision is made to preserve the natural distribution and diversity of label co-occurrences in the data. Table 5.1 shows the label counts of level-2 sense pairs that have at least 100 annotated instances in PDTB-3.

While previous studies (Ji and Eisenstein, 2015; Bai and Zhao, 2018; Xiang et al., 2022b) typically allocate Sections 2-20 of PDTB for training, Sections 0-1 for validation, and Sections 21-22 for testing, the limited size of the test set poses challenges, particularly for rare label and label pairs within the dataset. Acknowledging the concerns raised by Shi and Demberg (2017) regarding label sparsity, we address this issue

by employing cross-validation for Level-2 classification. In line with the methodology proposed by Kim et al. (2020), we adopted a cross-validation approach at the section level. We divide PDTB-3 into 12 folds, with each fold partitioned into 21 sections for training, two for development, and two for testing. By splitting the data at the section level, we can preserve the inherent paragraph and document structures, ensuring that data from the same sections are grouped together in the same pool.

Following the work in multi-label classification for other tasks like (Tsai and Lee, 2020), we adopt F1 scores (Manning et al., 2008) as our main evaluation metric. We report macro-averaged F1 scores because they provide a balanced evaluation across all classes, regardless of their frequency. In our data, some senses are much more frequent than others. Macro F1 ensures that performance on rare but important discourse relations is not overshadowed by frequent ones. This aligns with our goal of achieving robust performance across the full label space, especially when modeling fine-grained and hierarchical relations.

5.3.2 Implementation Details

For the experiments in Section 5.4.3, which investigate the effectiveness of the sense hierarchy, we adopt the same learning rate ($3e{-}5$) and batch size (256) as in Section 3.3.2, as both setups involve contrastive learning. For the other experiments in this chapter, we use the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $1e{-}5$ and a smaller batch size of 64. This is because contrastive learning is not applied in these settings, and thus a large batch size is not necessary. The maximum number of training epochs for all experiments is set to 20, with an early stopping patience of 10. All experiments are conducted using a single 80GB NVIDIA A100 GPU.

5.4 Results

5.4.1 Performance of Multi-Label Classification Methods

In this subsection, we examine and compare the performance of different methods for multi-label discourse relation classification. We then investigate the count of predicted labels, the predictions for multi-label examples. Furthermore, we evaluate how well each method handles multi-label instances in terms of exact and partial matches. Lastly, we provide a comparative study between single-label and multi-label predic-

Label	Method 1	Method 2	Method 3
Concession	50.98 ± 5.06	51.59 ± 4.61	50.86 ± 2.91
Contrast	50.58 ± 3.40	50.82 ± 2.99	48.49 ± 3.46
Cause	65.57 ± 1.76	65.15 ± 1.74	65.34 ± 2.19
Cause+Belief	0.00 ± 0.00	0.00 ± 0.00	3.04 ± 5.58
Condition	75.99 ± 5.95	80.97 ± 7.51	78.01 ± 10.00
Purpose	92.50 ± 2.01	92.68 ± 2.34	92.58 ± 2.21
Conjunction	62.12 ± 3.05	63.32 ± 3.34	62.11 ± 3.08
Equivalence	12.99 ± 7.56	14.55 ± 8.43	17.85 ± 6.42
Instantiation	58.76 ± 4.02	59.67 ± 5.73	58.49 ± 6.15
Level-of-detail	50.93 ± 3.97	51.80 ± 3.80	51.09 ± 2.26
Manner	58.76 ± 11.39	58.60 ± 13.47	23.23 ± 7.94
Substitution	64.11 ± 10.35	62.35 ± 7.83	54.46 ± 9.18
Asynchronous	62.46 ± 4.01	62.10 ± 3.88	61.20 ± 3.72
Synchronous	27.40 ± 9.48	30.28 ± 7.30	30.30 ± 6.96
Average	52.37 ± 1.62	53.13 ± 0.92	49.79 ± 1.12

Table 5.2: A Comparison of Macro-F1 scores across different methods by using $\text{RoBERTa}_{\text{base}}$. We use cross-validation at section level for the Level-2 classification. The standard deviations across 12 folds are reported.

tion settings, demonstrating that multi-label prediction can achieve comparable performance when using the single-label classification evaluation criteria.

5.4.1.1 Performance for Each label across Multi-label Classification Methods

Table 5.2 presents the F1 scores for each level-2 label across three methods using section-level cross-validation, where the scores are computed per label regardless of whether it appears in single-label or multi-label instances. Overall, there is no substantial differences for the average performance for all labels among the three methods.

However, we observe notable discrepancies in performance on certain labels. Most prominently, Method 3 struggles significantly with "Manner" (23.23 ± 7.94), which is more than 30 points lower than Method 2 and Method 1. Additionally, Method 3 underperforms on "Substitution", compared to Method 1 and Method 2. While both "Manner" and "Substitution" are more often observed in multi-label contexts, where they co-occur with other discourse relations. This suggests that Method 3 may struggle to accurately identify instances involving these labels, particularly when they appear alongside other relations.

	Method 1		Method 2		Method 3	
Num. gold / Num. prediction	**2**	**1**	**2**	**1**	**2**	**1**
2	395	806	405	983	379	585
1	506	17,780	498	17,614	563	19,396
0	41	1,395	39	1,383	0	0

Table 5.3: Comparative analysis of predicted label counts for instances with one and two gold labels across Method 1, Method 2, and Method 3. 'Num. Prediction" denotes the number of labels predicted by each method, while "Num. Gold" represents the number of gold-standard labels.

5.4.1.2 Count of Predicted Labels

Table 5.3 displays the distribution of the number of the predicted labels for examples with one or two gold labels across the three methods. We did not impose a limit on the number of predicted labels. However, none of the examples received more than two labels for any method, likely due to the data not containing examples with more than two labels. Analyzing the table, we find that **distinguishing one or two labels is challenging**, as over half of multi-label examples receive only one label, while more than 5% of single-label examples get two labels for three methods.

5.4.1.3 Predictions for Multi-label Examples

Here, we compare the three methods on the predictions for multi-label examples. However, the number of such examples in the dataset is small, which limits the statistical reliability of the comparison. As a result, the findings reported here should be interpreted with caution, as they may not generalize beyond this restricted subset.

For each multi-label example, a method can predict both labels correctly, only one label correctly, neither label, or make no prediction. As shown in Table 5.4, Method 2 achieves the highest proportion of exact matches (42%), indicating its strength in jointly predicting both labels when multiple relations are present. In contrast, Method 1 adopts a more conservative strategy: although its both-label accuracy is slightly lower (39%), it produces the fewest completely incorrect predictions (14%) and the most partial matches (43%). This suggests that Method 1 may be better at capturing at least one relevant discourse sense. Notably, Method 3 never produces "No Prediction" cases, as it always generates a label sequence via beam search. However, it has the highest rate of both-label errors (23%), which could be attributed to the first label

Method	Both Labels Correct	One Label Correct	Both Labels Incorrect	No Prediction
Method 1	363 (39%)	404 (43%)	134 (14%)	41 (4%)
Method 2	392 (42%)	343 (36%)	168 (18%)	39 (4%)
Method 3	382 (40%)	346 (37%)	214 (23%)	0 (0%)

Table 5.4: A comparison of methods on the predictions for multi-label examples (examples annotated with two labels).

Label	Single(base)	Multi(base)	Single(large)	Multi (large)
Concession	47.08±2.69	51.99±4.3	61.17±4.07	61.2±3.93
Contrast	49.01±2.32	52.94±2.25	57.19±4.07	59.76±2.55
Cause	66.31±1.75	66.0±1.56	70.89±1.52	70.44±1.61
Cause+Belief	4.13±5.74	6.52±9.08	8.59±6.99	10.06±11.74
Condition	78.88±8.87	80.16±7.84	84.91±10.74	84.55±10.21
Purpose	91.34±2.56	91.47±1.91	92.04±2.75	92.53±2.68
Conjunction	61.21±2.56	63.52±2.93	68.03±1.73	67.48±2.98
Equivalence	15.75±8.81	16.67±8.38	22.56±12.0	25.85±7.04
Instantiation	56.63±7.57	60.86±4.91	63.62±3.68	61.3±4.62
Level-of-detail	53.61±2.97	54.07±3.33	58.58±2.83	58.1±1.62
Manner	79.86±9.26	77.15±10.21	80.12±10.92	77.35±12.36
Substitution	60.34±13.13	65.67±8.61	70.34±6.17	71.92±8.45
Asynchronous	61.6±4.25	60.93±3.97	68.12±2.97	67.93±3.96
Synchronous	41.5±12.83	35.46±7.96	40.93±8.95	46.51±11.4
Average	54.8±1.85	55.96±0.84	60.51±1.32	61.07±1.64

Table 5.5: Comparative evaluation of cross-validation Macro-F1 scores for multi-label versus single-label prediction methods, with multi-label predictions assessed using single-label evaluation criteria. base refers to the smaller variant of the RoBERTa architecture $\text{RoBERTa}_{\text{base}}$, large means $\text{RoBERTa}_{\text{large}}$. This study provides standard deviations over 12-fold.

being incorrectly predicted, thereby causing the subsequent label to also be incorrect due to the sequential generation process.

5.4.1.4 Multi-Label vs. Single-Label Prediction: A Comparative Performance Analysis on single-label prediction

To compare multi-label and single-label prediction methods, we evaluated Method 2 under single-label criteria. In this evaluation, the highest probability label is chosen, and for multi-label examples, we consider it correct if the predicted label matches one of the gold labels. We did not evaluate the single-label prediction method in terms of multi-label criteria since our goal here is not to perform a comprehensive comparisons on multi-label evaluation, but rather to investigate whether a model trained with multi-label supervision underperform on single-label prediction. Therefore, we adopt a single-label evaluation setting for a fair comparison.

We utilize RoBERTa to obtain the [CLS] representation for each example for both single label prediction method and Method 2, but Method 2 uses separate classification heads for binary classification per class, while single-label classification employs a multi-class mapping layer, with a size of $\mathbb{R}^{h\times|C|}$. Specifically, num_class takes the value of 14 here as the number of the labels at Level-2 is 14. Training loss for both methods is softmax cross-entropy. We use both $\text{RoBERTa}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$ for comparisons. The two models differ in size and capacity: $\text{RoBERTa}_{\text{base}}$ consists of 12 transformer layers with a hidden size of 768 and approximately 125 million parameters, while $\text{RoBERTa}_{\text{large}}$ has 24 layers, a hidden size of 1,024, and roughly 355 million parameters. While it is generally expected that larger models like RoBERTa-large outperform smaller ones, we compare the multi-label prediction and the single-label prediction regardless of model scale.

The results in Table 5.5 indicate that, while the evaluation method for both single classification methods and multi-label classification methods is the same, based on the single-label evaluation criteria, the multi-label prediction method outperforms the single-label prediction method for both $\text{RoBERTa}_{\text{base}}$ and $\text{RoBERTa}_{\text{large}}$ by around 1%. This suggests that multi-label prediction does not compromise the performance of single-label prediction. Finally, It should also be noted that multi-label classification methods do not necessarily increase computational complexity. Moreover, multi-label methods are not necessarily more complicated, using no more computational resources than the single-label prediction methods in our experiments.

Label	Cross-entropy	Focal loss
Concession	50.64±3.97	52.13±4.56
Contrast	49.29±5.37	49.8±3.7
Cause	65.13±2.13	67.11±2.57
Cause+Belief	0.0±0.0	9.88±4.32
Condition	77.84±7.37	78.39±9.18
Purpose	92.41±2.4	92.34±2.26
Conjunction	63.21±3.16	64.15±3.13
Equivalence	17.55±10.12	23.69±7.16
Instantiation	59.76±5.27	58.93±4.89
Level-of-detail	52.84±3.38	54.05±1.79
Manner	58.94±10.65	58.53±12.22
Substitution	62.83±9.2	62.26±7.69
Asynchronous	62.19±3.24	60.73±4.73
Synchronous	30.3±7.85	30.33±5.35
Average	53.16±1.39	54.45±1.19

Table 5.6: Comparative analysis of F1 scores for Method 2 using $\text{RoBERTa}_{\text{base}}$: Evaluating cross-entropy and focal loss functions over 12-fold cross-validation with reported standard deviations.

5.4.2 Results of Methodological Exploration under the Multi-Label Scenario

This subsection gives the results of methodological variations on the performance of the model. We first investigate the effect of different loss functions by comparing focal loss with standard cross-entropy, highlighting the advantage of focal loss in addressing label imbalance. We then compare two cross-validation strategies, which is section-level and example-level, to assess their influence on performance. These explorations provide insights into how methodological choices can influence the effectiveness and stability of multi-label discourse relation classification models.

5.4.2.1 Focal Loss vs. Standard Cross-Entropy Loss

Table 5.6 presents the results of employing focal loss and standard cross-entropy loss, demonstrating that the adoption of focal loss enhances the overall performance when applied to Method 2 in the context of IDRR.

Label	Method 1	Method 2	Method 3
Concession	45.73±3.06	48.54±4.14	49.09±4.27
Contrast	48.42±3.42	50.05±3.92	50.02±3.71
Cause	63.78±1.96	64.59±1.86	64.69±1.93
Cause+Belief	0.00±0.00	0.00±0.00	1.02±3.57
Condition	72.26±9.37	75.61±8.65	77.01±8.95
Purpose	92.51±1.99	92.79±1.84	92.7±1.98
Conjunction	61.57±2.48	62.26±2.76	62.6±2.42
Equivalence	12.94±5.24	17.2±8.71	16.08±8.08
Instantiation	58.96±4.55	59.64±4.4	59.82±4.2
Level-of-detail	51.34±3.02	51.99±2.87	51.93±3.01
Manner	57.61±5.45	59.01±5.78	46.91±18.12
Substitution	60.38±6.47	62.09±6.32	59.95±7.18
Asynchronous	60.59±3.91	61.75±3.65	61.41±3.74
Synchronous	29.04±5.91	28.67±6.86	29.29±6.96
Average	51.08±0.88	52.44±1.72	51.61±1.96

Table 5.7: Example level: Comparison of Macro-F1 scores with standard deviations detailed across three methods using $\text{RoBERTa}_{\text{base}}$. Note: the methods in Table 5.2 and this table are consistent, differing only in the application of section-level versus example-level cross-validation techniques.

5.4.2.2 Cross-Validation Strategies

Table 5.7 shows the results for the three methods where cross-validation is done at example-level. A comparison between Table 5.2 and Table 5.7, which differ only in cross-validation strategies (section-level vs. example-level), reveals that the overall performance trends remain consistent across both settings. However, the absolute scores under example-level validation are slightly lower for most methods and labels, suggesting that section-level evaluation may offer a marginal advantage in stability or data alignment.

Certain labels such as “Manner” and “Substitution” exhibit larger fluctuations between the two settings. For example, Method 3’s F1 score for “Manner” increases dramatically from 23.23 ± 7.94 under section-level validation (Table 5.2) to 46.91 ± 18.12 under example-level validation (Table 5.7), suggesting that this label is particularly sensitive to how the data is partitioned when using Method 3. These variations imply that their prediction performance of Method 3 can be strongly influenced by the distribution of examples across folds, especially when multi-label instances are involved.

5.4.3 Leveraging Hierarchical Label Relationships for Multi-label Classification

Our experimental results in Table 5.8 demonstrate that incorporating hierarchical label relationships into contrastive learning consistently yields better performance compared to the baseline and the contrastive learning approach without using the sense hierarchy. Specifically, the hierarchy-aware method achieves an overall F1 score of 55.15% $\pm$ 1.42, which is 2.6% higher than the baseline (52.51% $\pm$1.26) and 4.6% higher than contrastive learning without considering sense hierarchy (50.50% $\pm$ 1.05).

This improvement is particularly notable for several senses. Performance improves to 34.26% for "Synchronous", a gain of 7.34% over the contrastive-only method and 5.30% over the baseline. The F1 score for "Manner" rises to 57.32%, with a 7.33% improvement over the contrastive-only method and 5.53% over the baseline. These results suggest that integrating hierarchical label relations not only enhances general performance but also improves the model's ability for relations that are prone to co-occurrence with others.

In contrast, contrastive learning without leveraging the hierarchical label relations underperforms compared to the baseline. This performance drop is particularly evident for "Contrast" (50.97%), showing a decline of 2.21%, as well as for "Purpose" (91.45%), and "Asynchronous" (61.79%), each with a decrease of around 1.7 points. These labels frequently co-exist with other labels, and the performance drop suggests that treating different level-1 labels as negative examples in contrastive learning increases their distance in the embedding space, making it harder for the model to correctly classify multi-label examples.

Additionally, Table 5.9 demonstrates differences among the three methods when predicting examples with two labels. While the overall trends are similar across all three settings, contrastive learning with sense hierarchy yields better performance. It improves the rate of both-label correctness (41.97%) and reduces the proportion of completely incorrect predictions (19.49%) compared to the baseline, suggesting its effectiveness in guiding the model toward more semantically coherent decisions.

In contrast, contrastive learning without the hierarchical structure shows the lowest proportion of both-label correct predictions (31.96%) and the highest rate of both-label errors (23.41%). This decline is likely due to treating different level-1 labels as negative examples during contrastive learning, which increases their distance in the embedding space. Since more than 99% out of 986 multi-label examples in PDTB-3

Label	Contra_without_hierarchy	Contra_with_hierarchy	Without_contra
Concession	48.58 ± 4.69	50.93 ± 4.58	50.09 ± 4.90
Contrast	50.97 ± 3.32	51.63 ± 4.69	53.23 ± 3.61
Cause	62.88 ± 1.41	66.50 ± 1.41	64.67 ± 1.73
Cause+Belief	3.79 ± 1.66	3.63 ± 1.87	5.50 ± 1.83
Condition	74.55 ± 7.62	80.44 ± 8.25	76.24 ± 7.75
Purpose	91.45 ± 2.00	94.03 ± 2.45	93.22 ± 2.45
Conjunction	59.89 ± 3.00	63.92 ± 2.83	61.63 ± 3.32
Equivalence	15.45 ± 7.62	20.83 ± 7.28	17.27 ± 7.75
Instantiation	56.69 ± 4.12	59.93 ± 5.29	58.45 ± 4.36
Level-of-detail	49.76 ± 3.74	52.86 ± 2.83	51.41 ± 4.00
Manner	49.99 ± 11.87	57.32 ± 11.27	51.79 ± 11.95
Substitution	58.27 ± 9.11	63.81 ± 11.62	59.97 ± 9.22
Asynchronous	61.79 ± 4.00	64.08 ± 4.47	63.54 ± 4.24
Synchronous	26.92 ± 4.80	34.26 ± 6.32	28.76 ± 5.00
Average Macro-F1	50.50 ± 1.05	55.15 ± 1.42	52.51 ± 1.26

Table 5.8: Macro-F1 scores across different settings with and without contrastive learning, and with and without incorporating the sense hierarchy when using contrastive learning, using $\text{RoBERTa}_{\text{base}}$. Results are averaged over 12-fold cross-validation at the section level for Level-2 classification. Standard deviations across folds are reported.

Category	**No Contrastive (%)**	**Contrastive with sense (%)**	**Contrastive without sense (%)**
Both Label Correct	41.61	41.97	31.96
One Label Correct	36.51	38.54	44.63
Both Label Incorrect	21.88	19.49	23.41

Table 5.9: Proportions of both label correct, one label correct, and both label incorrect across three tables. Instances with no prediction are included in the "both incorrect" category.

contain labels from different level-1 categories, this approach inadvertently hinders the model's ability to learn effective representations for multi-label classification. This indicates that without considering the sense hierarchy, contrastive learning may have label confusion in multi-label settings.

These findings emphasize the importance of considering hierarchical label relations in contrastive learning frameworks for multi-label classification tasks. The sense hierarchy not only improves overall classification accuracy but also demonstrates more robust performance in multi-label classification scenarios.

5.5 More Analysis on Multi-label Classification Methods for IDRR

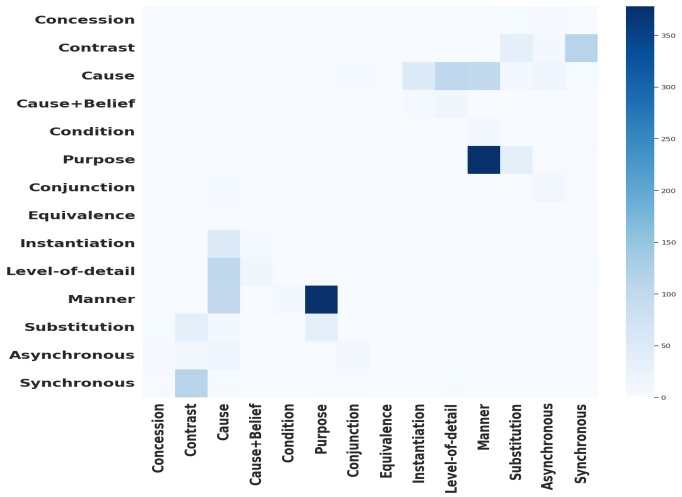
This section presents a deeper analysis of model behavior in multi-label classification for IDRR. We begin by examining whether multi-label methods can capture label co-occurrence patterns, followed by an exploration of cases where multi-label instances are only partially predicted. We also investigate the reverse situation — when the model predicts multiple labels for examples annotated with a single label.

5.5.1 Multi-label Classification Can Capture the Label Correlations

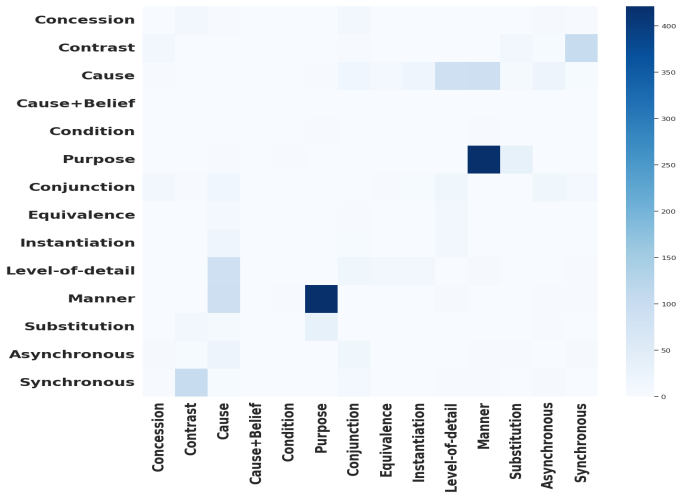
In Figure 5.1, two co-occurrence matrices visualize the joint distribution of label pairs in the dataset and in the prediction. The upper one is the co-occurrence of label pairs in the dataset and the lower one is the co-occurrence of the predicted label pairs. The gold label pairs are the multi-label data in the dataset.

For the predicted label pairs, all the predictions which have two labels on the test sets over 12-fold are used. Darker shades represent a higher frequency or probability of the label pairs co-occurring. In Method 2, for each label, the model independently predicts whether the label is the gold label by doing a binary classification. The two matrices indicate that the model has implicitly captured the correlations between the labels from the data.

However, as shown in Table 5.3, we observe that Method 2 correctly predicts both labels for only 42% of multi-label examples. This implies that the model's capacity to capture label correlations does not inherently guarantee very high accuracy in predicting multi-labels. For example, although the model can capture the correlations between "Purpose" and "Manner", it cannot distinguish which cases have both labels from those cases which only have "Purpose" or "Manner".



(a)



(b)

Figure 5.1: Co-occurrence of label pairs in the dataset and in the prediction. The upper sub-figure is for the gold label pairs, while the lower is for the predicted pairs.

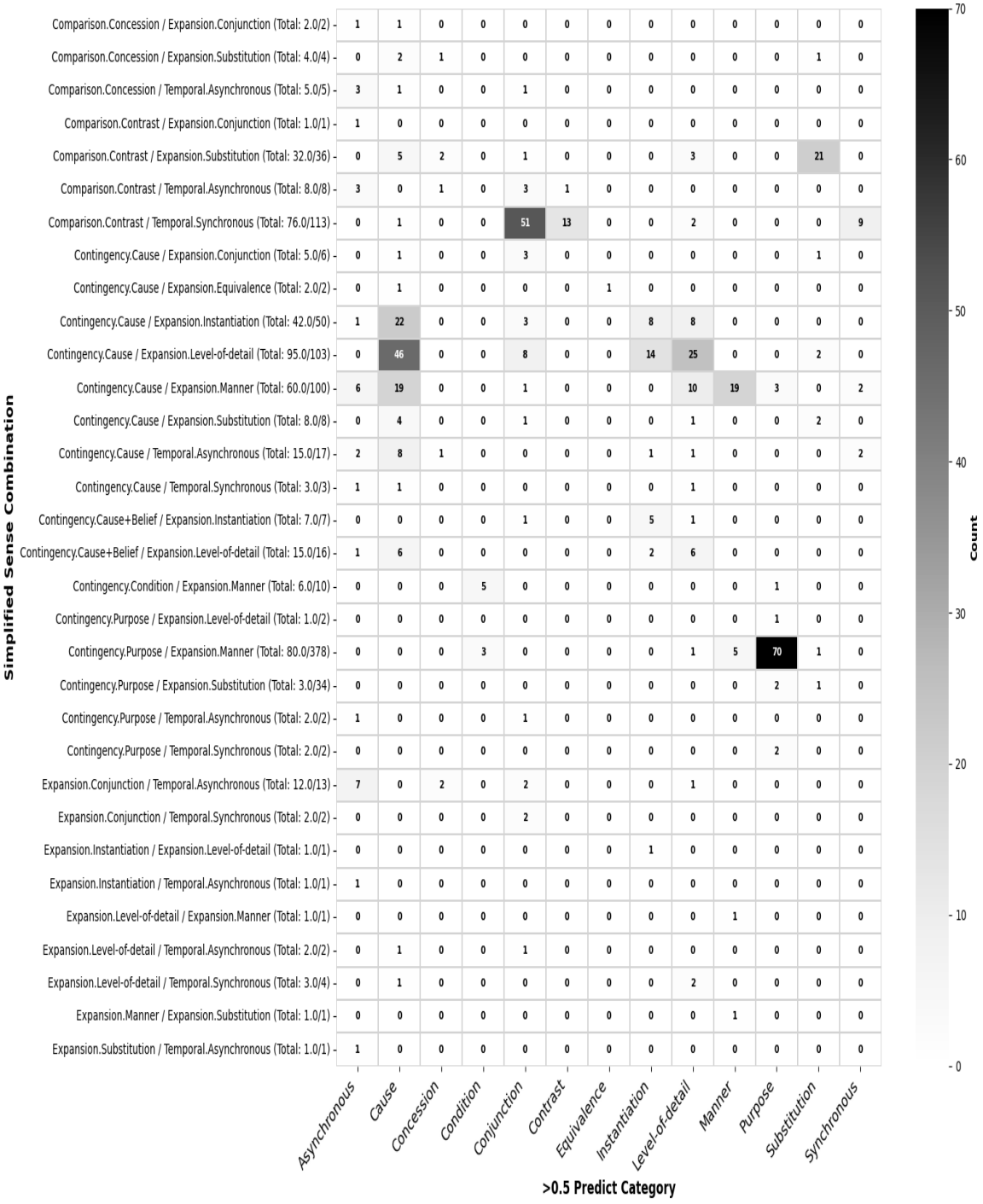


Figure 5.2: Heatmap of under-predicted multi-label instances. This figure displays the distribution of instances where two labels are annotated but only one is predicted.

5.5.2 When Multi-label Examples are Predicted as Single-label

Figure 5.2 presents a heatmap, which depicts instances where only one label is predicted for multi-label examples (with two annotated labels). The vertical axis corresponds to gold labels, while the horizontal ones correspond to predicted labels. The numbers beside the vertical axis show how many examples had two labels but received only one predicted label and the number of such label pairs in the dataset.

Figure 5.2 reveals around a quarter of instances that should have both "Purpose" and "Manner" labels are only labeled as "Purpose". Moreover, approximately one-third of instances labeled as "Purpose" are predicted as both "Purpose" and "Manner", as illustrated in Figure 5.3.

The following are two examples for "Purpose&Manner" vs. "Purpose" :

(11) [they exercise]$_1$ [to lose weight.]$_2$. Labels: Purpose&Manner

In Example (11), the purpose of exercising is to "lose weight", while the manner in which weight loss is to be achieved is through exercising. Thus, both "Purpose" and "Manner" are appropriate sense labels.

(12) [Mr. Achenbaum will work with clients]$_1$ [to determine the mix of promotion, merchandising, publicity and other marketing outlets, and to integrate those services]$_2$. Label: Purpose

In contrast, in Example (12), while the purpose of working with clients is to determine their service needs, the annotators appear to have decided that working with clients is not the manner by which their service needs are determined. As such, only "Purpose" was annotated as a sense label.

While one might disagree with the annotators' labeling decisions here, the value of multi-sense prediction is to highlight potentially questionable cases that might well justify further review (Klie et al., 2023).

The two examples indicate the challenges the model faces in distinguishing between "Purpose and Manner" and "Purpose" in certain cases. More work is needed to determine when an example demonstrates both "Purpose and Manner" versus just "Purpose".

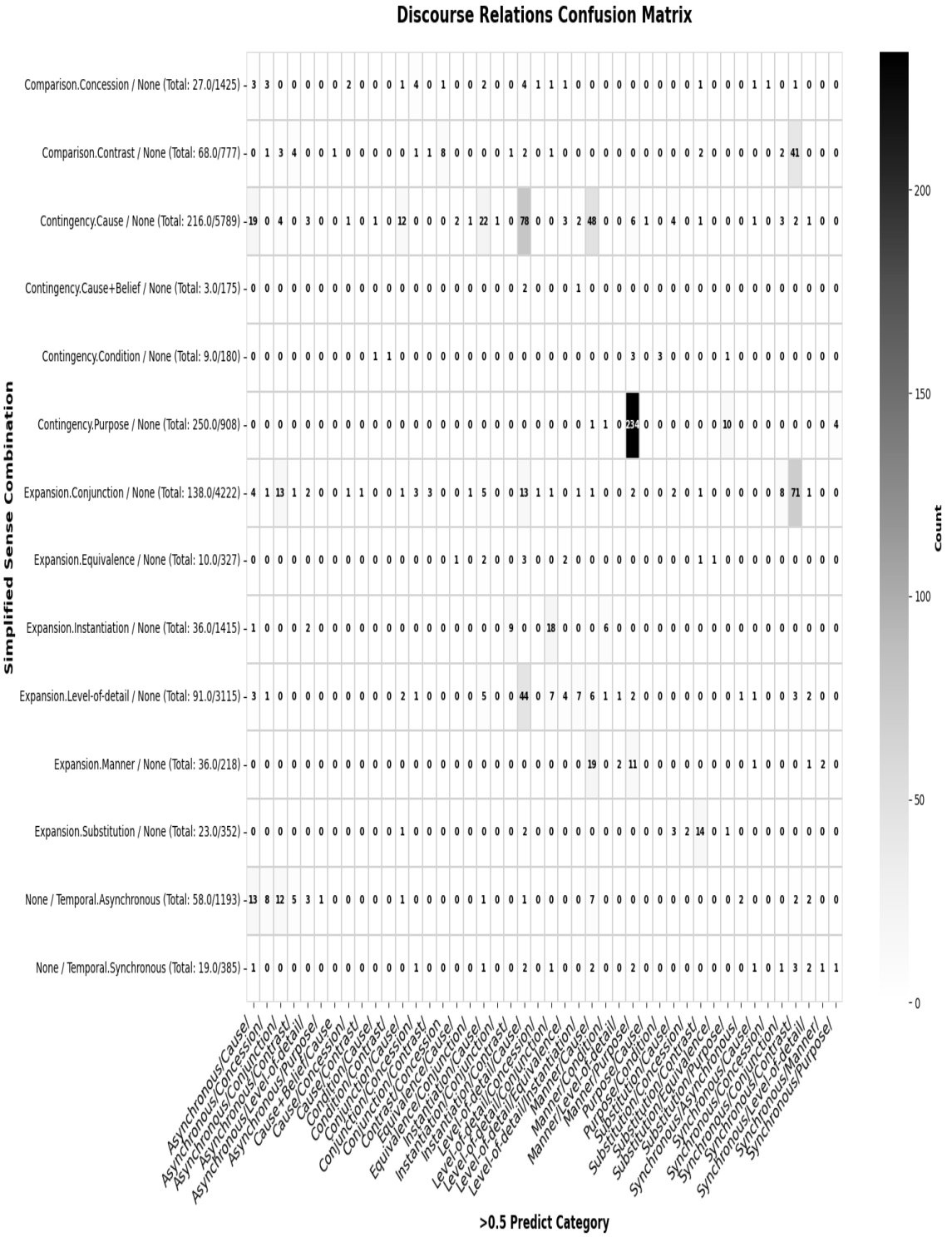


Figure 5.3: Heatmap of over-predicted single label instances. This figure displays the distribution of instances where single-label is annotated but are given two labels by the model.

5.5.3 When Two Labels are Given to the Single-label Examples

Figure 5.3 illustrates the predictions of two labels for examples manually labeled with a single label. The vertical axis corresponds to gold labels, while the horizontal one corresponds to predicted labels. The numbers beside the vertical axis indicate the number of examples with that label and how many were predicted to have additional labels.

In addition to showing that about one third of the instances labeled as "Purpose" are labeled as both "Purpose" and "Manner", Figure 5.3 also indicates that, among the 218 examples whose label is "Manner", the model gives the additional label "Cause" or "Purpose" for 30 examples. Moreover, when the model gives additional labels to those examples that are only annotated with "Cause", the labels are often under the "Expansion" category.

These observations echo our earlier analysis, suggesting that distinguishing between single and dual labels poses challenges for models, particularly concerning "Purpose"&"Manner' and "Purpose". Additionally, models occasionally predict both elaborative and argumentative relations simultaneously although only one relation (elaborative or argumentative) is annotated.

5.6 Conclusion

This chapter advances implicit discourse relation recognition by treating it as a multi-label classification task, addressing fundamental limitations in traditional single-label approaches. Our investigation demonstrates that multi-label frameworks can effectively capture the inherent complexity of discourse relations while maintaining strong performance on single-label instances. This work extends our research (Long et al., 2024) by investigating the integration of sense hierarchy into multi-label classification through contrastive learning frameworks.

Our investigation also reveals that incorporating hierarchical label relationships can enhance the effectiveness of contrastive learning in multi-label classification. The results demonstrate that the awareness of label hierarchical relations help contrastive learning achieve superior performance in capturing discourse relations, while contrastive learning without this awareness shows degraded performance compared to the baseline approach, and it particularly affects the learning of multi-label examples, as it increases the distances between the labels, resulting in much lower accuracy in pre-

dicting both labels correctly compared to the baseline. These findings provide strong evidence for the importance of leveraging discourse relation hierarchies in multi-label classification models.

This research not only advances our understanding of discourse relations but also provides concrete evidence that better modeling of label relationships is crucial for effective discourse relation recognition. These insights lay a foundation for future research in developing more effective approaches to discourse analysis that better reflect the inherent complexity of discourse relations.

Chapter 6

Limitations and Future Work

This chapter discusses the key limitations of the thesis, emphasizing that while these do not undermine the findings, they are critical for improving the methods and planning future research. It also outlines potential directions for future work to address these limitations and build upon the presented research.

6.1 Limitations

This section provides an in-depth discussion of the constraints encountered in our research, examining limitations related to both the data used and the methodological approaches employed. It highlights challenges stemming from the scope, diversity, and representativeness of datasets, as well as issues inherent in the design and application of the methods, such as scalability, adaptability, and their generalizability to other contexts or frameworks.

6.1.1 Data Limitations

We now discuss several limitations related to the data used in this study, focusing on three key aspects: the reliance on a single discourse annotation framework, constraints in genre and language coverage, and the limited range of discourse relation types considered. These factors may influence the generalizability and interpretability of our results across different discourse settings.

6.1.1.1 Limits to a Single Discourse Annotation Framework

Our evaluation relies solely on the Penn Discourse TreeBank (PDTB) for all experiments and TED-MDB, a PDTB-style dataset, in our zero-shot cross-lingual transfer learning experiments in Section 4.4.2. These experiments indicate the hierarchical relations can be used for enhancing the sense labeling. However, this exclusive reliance limits the scope of validation to a single annotation framework. In Chapter 2, we have shown that other discourse frameworks, such as enhanced Rhetorical Structure Theory (eRST) and Rhetorical Structure Theory (RST), do not adopt a three-level sense hierarchy like PDTB, but instead organize senses of discourse relations into coarse-grained and fine-grained categories. The relations of subclasses within the same broader category often exhibit significant semantic similarity, making them more difficult for models to distinguish. For example, in the original RST label hierarchy, "Evidence" and "Justify" are grouped under the same class. Despite their distinct rhetorical functions, their contextual similarities can blur the boundaries between them, increasing the difficulty of accurate classification.

This raises important questions about how our approach would perform when applied to these alternative discourse frameworks. The method we propose, which incorporates hierarchical label relations, could potentially enhance relation recognition in other frameworks by improving the model's ability to differentiate between fine-grained discourse relations that share similar functions. However, its effectiveness may vary due to differences in relation definitions and annotation schemes across frameworks.

6.1.1.2 Genre and Language Constraints

Zufferey and Degand (2024) explore how discourse relations and connectives vary across languages and genres. Their findings suggest that while core discourse relations such as "Cause", "Contrast", and "Elaboration" are largely shared across languages, the discourse relation hierarchies originally proposed for English often undergo addition, removal, or redefinition when adapted to other languages. Furthermore, the distribution and frequency of discourse relations differ across both languages and genres.

This research primarily evaluates methods on English-language data from the PDTB. Although we use TED-MDB, a multilingual PDTB style dataset for our cross-lingual transfer learning experiments, we do not discuss the challenges raised by languages

other than English. However, discourse relations exhibit significant language-specific variations. For example, Xu et al. (2019) point out that, due to the paratactic nature of Chinese and its tendency to omit explicit connectives, implicit discourse relations are significantly more prevalent in Chinese. Specifically, 75.2% of all discourse relations in the CDTB corpus (Zhou and Xue, 2014), a Chinese discourse corpus, are implicit, compared to about 40% of all discourse relations in the English PDTB (39.5% in PDTB-2 and 40.0% in PDTB-3), while both CDTB and PDTB are based on written texts. [1]

Additionally, PDTB primarily consists of news discourse, reviews, summaries, etc., while it has been noted that the text genre significantly impacts the distribution of discourse relations (Scholman et al., 2022). By confining the dataset to written texts, there is a missed opportunity to comprehensively understand and model discourse relations and label associations across a wider range of text types.

6.1.1.3 Limited Data Type Coverage

The study focuses exclusively on implicit discourse relations - in PDTB, excluding explicit relations, Entity-Based Relations (EntRel), and Alternative Lexicalizations (AltLex). While implicit relations are inherently challenging due to the absence of explicit connectives, excluding other relation types narrows the scope of evaluation. Moreover, while implementing the multi-label classification methods, we ignored those implicit relations linked with explicit or AltLex relations where where multiple relation types such as implicit-explicit or implicit-AltLex coexist, as illustrated in Section 2.2.3. This also reduces the diversity of training instances, thereby constraining the model's ability to generalize to real-world discourse scenarios where multiple relation types often coexist.

6.1.2 Methodological Limitations

In this section, we outline several methodological limitations of our study. These include simplifications in how multi-label data are handled, the substantial data requirements imposed by contrastive learning, the challenge of data imbalance, and the assumption of gold-standard argument boundaries. We also discuss overlooked factors

[1] These percentages refer to all discourse relations, including both intra-sentential and inter-sentential cases. However, CDTB only annotates comma-delimited intra-sentential discourse relations, while discourse relations can exist within a single sentence or clause without requiring punctuation in Chinese. This may limit the coverage and make the comparison not entirely parallel.

such as label co-occurrence patterns and the absence of large language models (LLMs) in our experiments. Together, these limitations highlight areas where our current modeling approach may fall short in addressing the full complexity of discourse relation classification.

6.1.2.1 Multi-label Handling

In our single-label classification methods introduced in Chapter 3 and Chapter 4, we over-simplify the presence of multiple labels by following Qin et al. (2017) in treating each label as a separate example and did not consider the second label. We treated them as separate and different examples during training. At test time, a prediction matching one of the gold types is taken as the correct answer. Thus, our approaches in Chapter 3 and Chapter 4 are inadequate for dealing with the actual distribution of the data and can be extended or modified.

6.1.2.2 Data Requirements of Contrastive Learning

Our approaches rely heavily on contrastive learning to establish meaningful relationships between instances. While this approach has shown good results, it comes with certain data requirements that should be acknowledged. Contrastive learning typically requires a substantial amount of training data to effectively learn discriminative features and establish robust representations. The performance increase observed with our data augmentation approach presented in Chapter 3 highlights the data requirements.

For low-resource languages or specific domains with limited discourse-annotated corpora, this presents a significant challenge. To achieve good performance with this method, much more training data are required.

6.1.2.3 Failure to Address Data Imbalance Issue

In the PDTB datasets, discourse relations exhibit significant imbalance, with certain sense types occurring substantially more frequently than others. This imbalanced distribution can introduce bias toward majority classes, resulting in models that better on common relations but perform inadequately on minority senses. Although our multi-label classification method replaces the standard cross-entropy loss with focal loss to address this imbalance by assigning higher weights to difficult and underrepresented examples (see Section 5.4.2), rare discourse relations remain a substantial challenge.

The limited number of examples for minority relations makes it difficult to learn robust representations that effectively capture the distinctive features of these relations. This limitation is particularly pronounced in fine-grained classification scenarios where some specific relation subtypes may have very few training instances, potentially compromising the ability of our proposed method to generalize to these rare categories. Therefore, our methods cannot significantly improve the performance for the discourse relations with the limited number of instances like "Expansion. Equivalence", "Contingency.Condition" and "Contingency.Cause+belief".

6.1.2.4 Without Identifying Arguments

One limitation of our approaches is the assumption that discourse arguments are already identified. While this allows us to focus solely on the recognition of discourse relations, it abstracts away the complexity of argument identification, which is a non-trivial and error-prone step in practical discourse parsing pipelines. In real-world discourse parsing scenarios, argument boundaries are often ambiguous and must be predicted automatically. As a result, the performance of our method may be overestimated compared to real-world settings where automatic argument identification is required.

6.1.2.5 Without Exploiting Co-occurrence Label Relationship

Yung et al. (2022) analyse the co-occurrences of relations in DiscoGem (Scholman et al., 2022) in which there are multiple annotations from different annotators for each example and they incorporate these information on multi-label distributions in the data can help improve implicit relation classifiers.

Our research does not use co-occurrence label relationship for the improvement of sense labeling. Our approach focuses primarily on the hierarchical structure of discourse relations but overlooks potentially valuable information about how different relation types tend to co-occur within texts. For instance, "Contingency.Purpose" often co-occurs with "Expansion.Manner" in PDTB-3, approximately 30% of instances labeled as "Purpose" in PDTB-3 have another label "Expansion.Manner". Our Method 3 in Section 5.2.1 may implicitly capture dependencies between labels through its sequential generation process, where the prediction of each label is conditioned on the previously generated ones. However, we do not explicitly leverage co-occurrence statistics or the co-occurrence patterns, thus the potential benefits of such information for relation classification remain underexplored in our current framework.

6.1.2.6 The Exploration of LLMs

With the increase of computational resources and available text corpora, the rapid evolution of large language models (e.g., ChatGPT (OpenAI, 2023a), GPT-4 (OpenAI, 2023b), LLaMA (Grattafiori et al., 2024) has demonstrated the benefits brought by scaling up model sizes. Recent studies have evaluated LLM's performance on numerous language understanding and reasoning tasks, revealing that it outperforms other models in zero-shot settings (Bubeck et al., 2023; Bang et al., 2023; Jiao et al., 2023; Liu et al., 2024).

For discourse relation recognition, which we consider a subtask of discourse parsing, Fan et al. (2024) investigates ChatGPT's capabilities in two discourse analysis tasks: topic segmentation and discourse parsing. They show that ChatGPT still faces great challenges in identifying infrequent relations. Chan et al. (2024) find that ChatGPT exhibits strong performance in detecting and reasoning about causal relations, while it may not be proficient in identifying the temporal order between two events. Eichin et al. (2025) introduce a cross-linguistic discourse relation label set by mapping labels from diverse annotation frameworks (e.g., RST, PDTB, SDRT) into a unified set of 17 core relations. They use this unified set to benchmark 23 large language models (LLMs) of varying sizes and multilingual capacities on discourse relation classification. While their work highlights that multilingual LLMs can generalize discourse information across languages and frameworks, their results also reveal notable variation in performance across languages and sense labels of discourse relations, particularly for infrequent or difficult labels. Overall, these works suggest that large language models demonstrate good capabilities in discourse relation recognition, but they exhibit performance differences across languages and labels. They continue to face substantial challenges in handling infrequent labels and hard-to-distinguish examples.

However, none of the above works directly explore how LLMs leverage the organization of senses during discourse relation recognition. This represents a significant limitation of our work, in that we do not include LLMs in our evaluation to assess whether our hierarchical sense structure could also benefit LLMs. Without such comparison, it remains an open question whether the sense hierarchy offers complementary value for LLMs. Additionally, our approach does not benefit from the rich world knowledge and linguistic patterns captured in LLMs that might be valuable for resolving ambiguous discourse relations. As the field increasingly moves toward LLM-based solutions, the practical applicability of specialized models like ours may become limited.

6.2 Future Work

Building upon the current findings and the limitations, this section proposes promising directions for future research.

6.2.1 Extending the Methods to More Datasets and Other Types of Data

Following Long and Webber (2022), subsequent studies such as Jiang et al. (2023); Lian et al. (2024); Wu et al. (2024) have explored similar approaches to incorporate the hierarchical relationships between sense labels. However, all these methods (including our own) remain restricted to implicit discourse relation instances within PDTB-2 and PDTB-3. Given the data limitations discussed earlier, an important direction for future research is to extend these approaches to a wider range of datasets from different annotation frameworks and to other discourse relation types, in order to evaluate their generalizability and applicability in more diverse discourse settings.

Cross-Framework Evaluation A critical extension would involve applying our approach to datasets annotated with alternative frameworks, such as RST and SDRT. While these frameworks do not have the same three-level sense hierarchy as PDTB, they do contain coarse and fine-grained relations that might benefit from our hierarchical modeling approach. Cross-framework evaluation can indicate whether applying the label hierarchy for other frameworks could be helpful and whether there should be differences when we use the label hierarchy for other frameworks. It would also reveal which aspects are framework-specific and which are more universal.

Genre Diversity In our work, we primary use PDTB as our evaluation dataset, despite the scarcity of datasets annotating multiple discourse relations. However, some other datasets such as GUM (Zeldes, 2017), and DiscoGem (Scholman et al., 2022) can be considered for our evaluation and analysis. GUM offer diverse genre coverage across interviews, news, academic writing, and other text types, while the texts of DiscoGem are from political speeches, novels, and Wikipedia articles. Testing our methods on these resources would assess their adaptability to genre-specific discourse patterns and reveal potential strengths or weaknesses in handling multi-sense annotations across different text styles.

Additionally, DiscoGem (Scholman et al., 2022) can be used to see whether the annotation of this corpus either agrees with or contradicts multi-sense PDTB-3 sense annotation, although DiscoGeM is not inherently a multi-label dataset but rather a collection with diverse annotations stemming from having multiple annotators and recording all their decisions. Recently, Costa and Kosseim (2024) have evaluated their multi-label classification model by training it on DiscoGeM and testing it on both PDTB-3 and DiscoGeM. Their study highlights the difficulty of generalizing discourse relation models across datasets, as performance drops significantly when a model trained on DiscoGeM is evaluated on PDTB-3; however, they do not investigate the reverse — how a model trained on PDTB-3 performs when evaluated on DiscoGeM.

Furthermore, while our work primarily focuses on written discourse, it is worth noting that PDTB-style annotation schemes have also been extended to spoken discourse corpora. For instance, Tonelli et al. (2010) annotated Italian dialogues with PDTB-style annotation, and Long et al. (2020) constructed similar annotations on Chinese TED-Talks. Similarly, Scheffler et al. (2019) applied PDTB-style annotations to Twitter conversations. These resources, although beyond the scope of our current evaluation, could be valuable for future studies aiming to extend methods to spoken or conversational domains.

Multilingual Evaluation Extending our framework to multilingual datasets represents another important direction. Several PDTB-style datasets exist for languages other than English:

- **Chinese TED Discourse Bank Corpus** (Long et al., 2020): Offers annotations tailored to Mandarin's structural and syntactic characteristics.
- **Turkish Discourse Tree Bank (TDTB)** (Zeyrek and Er, 2022): Features annotations for Turkish, with its distinctive agglutinative morphology.
- **Italian Discourse Treebank** (Pareti and Prodanof, 2010): Provides insights into Romance language-specific discourse patterns.

Additionally, other corpora, such as the Prague Dependency Treebank (PDT) (Hajič et al., 2020), introduce alternative frameworks with hierarchical sense annotation and dependency-based discourse structures. Recent work on the Prague Discourse Treebank 3.0 (PDiT 3.0) (Synková et al., 2024) makes this resource more comparable to

the Penn Discourse Treebank 3.0 by providing a revised annotation of discourse relations, offering all annotations in both the original Prague format and the PDTB-3 format and the sense taxonomy. These updates enable cross-linguistic comparison and make the resource more suitable for exploring the use of the PDTB-3 sense hierarchy in Czech.

These resources could facilitate investigations into sense hierarchy modeling and negative examples selection strategies if using contrastive learning in languages with varied syntactic and discourse structures.

Other Types of Discourse Relations Future work could explore how applying the sense hierarchy consistently across implicit, explicit, and AltLex relations can improve discourse relation recognition. This includes developing unified models that learn from all relation types while incorporating the hierarchical label relations. Additionally, rather than only using the multi-label examples which are entirely implicit, including those multi-label examples containing explicit or AltLex relations could enhance model robustness. Analyzing whether there are any differences among these relations when we select the negative examples for contrastive learning may further help in our methods to utilize the sense hierarchy and improving generalization.

6.2.2 Methodological Improvements

This subsection outlines potential methodological improvements, focusing on enhancing multi-label classification, addressing data insufficiency and imbalance, and leveraging Large Language Models (LLMs) for better discourse understanding.

Enhancing Multi-label Classification In our work, we explore three multi-label classification methods for discourse relation prediction, emphasizing the challenges of distinguishing multi-label examples from single-label ones and accurately identifying multiple labels within each multi-label instance. Subsequently, Costa and Kosseim (2024) proposed a multi-task classification model that jointly learns both multi-label and single-label representations of discourse relations. While these approaches represent important progress, further research is needed to more effectively model label dependencies in discourse.

One direction might be the use of graph-based models, where discourse relations are represented as nodes and their co-occurrence patterns form the edges. Such a

framework could naturally capture both hierarchical and concurrent relationships between labels, enabling a more structured and context-aware classification process. Incorporating these relational structures may lead to more robust predictions, especially for complex discourse relations.

Addressing Data Imbalance or Insufficiency The insufficient data for rare categories and the limited availability of annotated multi-discourse relations present challenges for current work including our approach. Although improving performance on rare categories has limited impact on overall accuracy, their proper recognition contributes to the comprehensiveness and generalizability of discourse parsing systems. To address data insufficiency and imbalance, on the one hand, we can leverage Large Language Models (LLMs) to generate additional annotated discourse relation data. Omura et al. (2024) have generated large-scale synthetic data for error-prone discourse relations and incorporated it into training, achieving state-of-the-art macro-F1 without compromising micro-F1, particularly improving recognition of infrequent relations. Since LLMs have been trained on vast amounts of text, they can be fine-tuned or prompted to generate synthetic discourse relation examples that align with existing annotation frameworks, thereby expanding training datasets in a controlled manner. On the other hand, more advanced modeling techniques or data augmentation strategies can be developed.

Jointly Model Argument Identification and Relation Recognition Future work could extend our framework to jointly model argument identification and relation recognition for more end-to-end applicability. Integrating argument identification into the model would not only improve its practicality but also test its robustness under noisier, more realistic conditions. Moreover, a joint modeling framework could potentially capture interactions between argument boundaries and relation semantics, leading to improved overall performance.

Exploring Large Language Models (LLMs) A promising direction for future work involves leveraging sense hierarchies with large language models (LLMs). As LLMs continue to advance in their accuracy in prediction or generation, incorporating structured hierarchical knowledge about discourse relations has the potential to enhance their discourse sensitivity in text generation tasks and improve their performance on discourse-related tasks. Specifically, we can finetune LLMs using our approach to better capture the relationships between discourse senses at different levels of granularity.

Additionally, discourse-aware prompting strategies which explicitly incorporate hierarchical sense information could be developed, guiding LLMs to consider and utilize the hierarchical label relations.

Furthermore, the powerful capabilities of LLMs could advance our understanding of discourse relations. Yung et al. (2024) propose a per-class prompting framework where the model is asked about each relation independently (e.g., "Is this a causal relation?"). This allows GPT-4 to predict multiple labels per instance for the 1252 multi-label examples in their data, achieving a soft-match accuracy of 92.81%, indicating that in most cases at least one predicted label overlaps with the gold labels, and an average per-item F1 of 50.63%, which measures how well the full set of predicted labels matches the gold labels. These results demonstrate that large language models can effectively identify multiple co-occurring discourse relations, and further analysis of discrepancies between model predictions and gold labels could yield valuable insights. By examining instances where LLM predictions diverge from human annotations, we can identify patterns in how machines versus humans interpret discourse relations. Additionally, analyzing these discrepancies could help improve both annotation guidelines for humans and prompting strategies for LLMs, ultimately leading to a more nuanced understanding of how discourse relations co-exist and interact in natural language.

Bibliography

Bai, H. and Zhao, H. (2018). Deep Enhanced Representation for Implicit Discourse Relation Recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Park, J. C., Arase, Y., Hu, B., Lu, W., Wijaya, D., Purwarianti, A., and Krisnadhi, A. A., editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Biran, O. and McKeown, K. (2015). PDTB Discourse Parsing as a Tagging Task: The Two Taggers Approach. In Koller, A., Skantze, G., Jurcicek, F., Araki, M., and Rose, C. P., editors, *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104, Prague, Czech Republic. Association for Computational Linguistics.

Bourgonje, P. and Demberg, V. (2024). Generalizing across languages and domains for discourse relation classification. In Kawahara, T., Demberg, V., Ultes, S., Inoue, K., Mehri, S., Howcroft, D., and Komatani, K., editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 554–565, Kyoto, Japan. Association for Computational Linguistics.

Bourgonje, P., Hoek, J., Evers-Vermeul, J., Redeker, G., Sanders, T., and Stede, M. (2018). Constructing a Lexicon of Dutch Discourse Connectives. *Computational Linguistics in the Netherlands Journal*, 8:163–175.

Bourgonje, P. and Stede, M. (2020). Exploiting a lexical resource for discourse connective disambiguation in German. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5737–5748, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv*, abs/2303.12712.

Carlson, L., Marcu, D., and Okurovsky, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Carlson, L., Marcu, D., and Okurowski, M. E. (2002). RST Discourse Treebank. Technical Report LDC2002T07, Linguistic Data Consortium.

Chan, C., Jiayang, C., Wang, W., Jiang, Y., Fang, T., Liu, X., and Song, Y. (2024). Exploring the Potential of ChatGPT on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations. In Graham, Y. and Purver, M., editors, *Findings of the European Association for Computational Linguistics: European ACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Chan, C., Liu, X., Cheng, J., Li, Z., Song, Y., Wong, G., and See, S. (2023). DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Chen, J., Zhang, Q., Liu, P., and Huang, X. (2016a). Discourse relations detection via a mixed generative-discriminative framework. *Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI)*, 30(1).

Chen, J., Zhang, Q., Liu, P., Qiu, X., and Huang, X. (2016b). Implicit Discourse Relation Detection via a Deep Architecture with Gated Relevance Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), ACL 2016, pages 1726–1735, Berlin, Germany. Association for Computational Linguistics.

Chen, R., Wang, J., Yu, L.-C., and Zhang, X. (2023). Learning to Memorize Entailment and Discourse Relations for Persona-Consistent Dialogues. *Proceedings of the The Association for the Advancement of Artificial Intelligence (AAAI)*, 37(11):12653–12661.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Chi, T.-C. and Rudnicky, A. (2022). Structured Dialogue Discourse Parsing. In Lemon, O., Hakkani-Tur, D., Li, J. J., Ashrafzadeh, A., Garcia, D. H., Alikhani, M., Vandyke, D., and Dušek, O., editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2022)*, pages 325–335, Edinburgh, UK. Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Costa, N. F. and Kosseim, L. (2024). A Multi-Task and Multi-Label Classification Model for Implicit Discourse Relation Recognition. *arXiv*, abs/2408.08971.

Cui, G., Hu, S., Ding, N., Huang, L., and Liu, Z. (2022). Prototypical Verbalizer for Prompt-based Few-shot Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022) (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland. Association for Computational Linguistics.

da Cunha, I., Torres-Moreno, J.-M., and Sierra, G. (2011). On the Development of the RST Spanish Treebank. In Ide, N., Meyers, A., Pradhan, S., and Tomanek, K., editors, *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA. Association for Computational Linguistics.

Dai, Z. and Huang, R. (2018). Improving Implicit Discourse Relation Classification by Modeling Inter-dependencies of Discourse Units in a Paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT 2018) (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Dai, Z. and Huang, R. (2019). A Regularization Approach for Incorporating Event Knowledge and Coreference Relations into Neural Discourse Parsing. 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 2019)*, pages 2976–2987, Hong Kong, China. Association for Computational Linguistics.

Das, D., Scheffler, T., Bourgonje, P., and Stede, M. (2018). Constructing a Lexicon of English Discourse Connectives. In Komatani, K., Litman, D., Yu, K., Papangelis, A., Cavedon, L., and Nakano, M., editors, *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2018)*, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

Das, D. and Taboada, M. (2017). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149 – 184.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dou, Z., Hong, Y., Sun, Y., and Zhou, G. (2021). CVAE-based Re-anchoring for Implicit Discourse Relation Classification. In *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing (ENNLP 2021)*, pages 1275–1283, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Du, H., Feng, Y., Li, C., Li, Y., Lan, Y., and Zhao, D. (2023). Structure-Discourse Hierarchical Graph for Conditional Question Answering on Long Documents. In

Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6282–6293, Toronto, Canada. Association for Computational Linguistics.

Eichin, F., Liu, Y. J., Plank, B., and Hedderich, M. A. (2025). Probing LLMs for multilingual discourse generalization through a unified label set. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18665–18684, Vienna, Austria. Association for Computational Linguistics.

Fan, Y., Jiang, F., Li, P., and Li, H. (2024). Uncovering the Potential of ChatGPT for Discourse Analysis in Dialogue: An Empirical Study. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16998–17010, Torino, Italia. ELRA and ICCL.

Forbes-Riley, K., Miltsakaki, E., Prasad, R., Sarkar, A., Joshi, A. K., and Webber, B. L. (2003). D-LTAG System: Discourse Parsing with a Lexicalized Tree-Adjoining Grammar. *Journal of Logic, Language and Information*, 12:261–279.

Garnham, A., Oakhill, J. V., and Johnson-Laird, P. N. (1982). Referential Continuity and the Coherence of Discourse. *Cognition*, 11:29–46.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., and etc, A. Z. (2024). The Llama 3 Herd of Models. *arXiv*, abs/2407.21783.

Grindrod, J. and Borg, E. (2019). Questions under discussion and the semantics/pragmatics divide. *Philosophical Quarterly*, 69(275):418–426.

Gunel, B., Du, J., Conneau, A., and Stoyanov, V. (2021). Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. In *International Conference on Learning Representations*.

Guo, F., He, R., Dang, J., and Wang, J. (2020). Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation

recognition. *Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI)*, 34(05):7822–7829.

Hajič, J., Bejček, E., Hlavacova, J., Mikulová, M., Straka, M., Štěpánek, J., and Štěpánková, B. (2020). Prague Dependency Treebank - Consolidated 1.0. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.

Huang, Y. J. and Kurohashi, S. (2021). Extractive Summarization Considering Discourse and Coreference Relations based on Heterogeneous Graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052, Online. Association for Computational Linguistics.

Ji, Y. and Eisenstein, J. (2015). One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Jiang, F., Fan, Y., Chu, X., Li, P., and Zhu, Q. (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.

Jiang, Y., Zhang, L., and Wang, W. (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.

Jiao, W., Wang, W., Huang, J.-T., Wang, X., and Tu, Z. (2023). Is ChatGPT A Good Translator? A Preliminary Study. *ArXiv*, abs/2301.08745.

Joshua, R., Ching-Yao, C., Suvrit, S., and Stefanie, J. (2021). Contrastive Learning with Hard Negative Samples. *International Conference on Learning Representations*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Kim, N., Feng, S., Gunasekara, C., and Lastras, L. (2020). Implicit Discourse Relation Classification: We Need to Talk about Evaluation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Kim, T., Yoo, K. M., and Lee, S.-g. (2021). Self-Guided Contrastive Learning for BERT Sentence Representations. 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 2528–2540, Online. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kishimoto, Y., Murawaki, Y., and Kurohashi, S. (2020). Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

Klie, J.-C., Webber, B., and Gurevych, I. (2023). Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1):157–198.

Ko, W.-J., Dalton, C., Simmons, M., Fisher, E., Durrett, G., and Li, J. J. (2022). Discourse Comprehension: A Question Answering Framework to Represent Sentence Connections. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages

11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ko, W.-J., Wu, Y., Dalton, C., Srinivas, D., Durrett, G., and Li, J. J. (2023). Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

Kong, F., Ng, H. T., and Zhou, G. (2014). A Constituent-Based Approach to Argument Labeling with Joint Inference in Discourse Parsing. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 68–77, Doha, Qatar. Association for Computational Linguistics.

Kuppevelt, J. V. (1995). Discourse Structure, Topicality and Questioning. *Journal of Linguistics*, 31(1):109–147.

Kurfalı, M. and Östling, R. (2019). Zero-shot Transfer for Implicit Discourse Relation Classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 226–231, Stockholm, Sweden. Association for Computational Linguistics.

Lan, M., Wang, J., Wu, Y., Niu, Z.-Y., and Wang, H. (2017). Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Li, M. and Huang, R. (2023). RST-style Discourse Parsing Guided by Document-level Content Structures. *arXiv*, abs/2309.04141.

Lian, R., Sethares, W., and Hu, J. (2024). Learning Label Hierarchy with Supervised Contrastive Learning. In Graham, Y. and Purver, M., editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1569–1581, St. Julian's, Malta. Association for Computational Linguistics.

Liang, L., Zhao, Z., and Webber, B. (2020). Extending Implicit Discourse Relation Recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational*

Approaches to Discourse, pages 135–147, Online. Association for Computational Linguistics.

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Liu, S., Li, Y., Li, J., Yang, S., and Lan, Y. (2024). Unleashing the Power of Large Language Models in Zero-shot Relation Extraction via Self-Prompting. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13147–13161, Miami, Florida, USA. Association for Computational Linguistics.

Liu, X., Ou, J., Song, Y., and Jiang, X. (2020). On the Importance of Word and Sentence Representation Learning in Implicit Discourse Relation Classification. In Bessiere, C., editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-2020*, pages 3830–3836. International Joint Conferences on Artificial Intelligence Organization. Main track.

Liu, Y. and Li, S. (2016). Recognizing Implicit Discourse Relations via Repeated Reading: Neural Networks with Multi-Level Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1224–1233, Austin, Texas. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Liu, Y. J., Aoyama, T., and Zeldes, A. (2023). What's Hard in English RST Parsing? Predictive Models for Error Analysis. In Stoyanchev, S., Joty, S., Schlangen, D., Dusek, O., Kennington, C., and Alikhani, M., editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

Long, W., Narayanaswamy, S., and Webber, B. (2024). Multi-Label Classification for Implicit Discourse Relation Recognition. In Ku, L.-W., Martins, A., and Srikumar,

V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics.

Long, W. and Webber, B. (2022). Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long, W. and Webber, B. (2024). Leveraging Hierarchical Prototypes as the Verbalizer for Implicit Discourse Relation Recognition. *arXiv*, abs/2411.14880.

Long, W., Webber, B., and Xiong, D. (2020). TED-CDB: A Large-Scale Chinese Discourse Relation Dataset on TED Talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803, Online. Association for Computational Linguistics.

Mann, W., Matthiessen, C., and Thompson, S. (1989). Rhetorical Structure Theory and Text Analysis. *Discourse Description: Diverse Linguistic Analyses of a Fund Raising Text*, page 66.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text & Talk*, 8:243 – 281.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval: Scoring, term weighting, and the vector space model*. Cambridge University Press.

Metheniti, E. L., Muller, P., Braud, C., and Casas, M. H. (2024). Zero-shot Learning for Multilingual Discourse Relation Classification. In *International Conference on Language Resources and Evaluation*.

Moens, M. and Steedman, M. (1988). Temporal Ontology and Temporal Reference. *Computational Linguistics*, 14(2):15–28.

Morris, J. and Hirst, G. (1991). Lexical Cohesion Computed by Thesaural relations as an indicator of the structure of text. *Comput. Linguistics*, 17:21–48.

Nguyen, L. T., Van Ngo, L., Than, K., and Nguyen, T. H. (2019). Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via

Label Embeddings. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207, Florence, Italy. Association for Computational Linguistics.

Omura, K., Cheng, F., and Kurohashi, S. (2024). An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

OpenAI (2023a). ChatGPT Mar 14 version. *https://chat.openai.com/chat*.

OpenAI (2023b). GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Pareti, S. and Prodanof, I. (2010). Annotating Attribution Relations: Towards an Italian Discourse Treebank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Patterson, G. and Kehler, A. (2013). Predicting the Presence of Discourse Connectives. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 914–923, Seattle, Washington, USA. Association for Computational Linguistics.

Peng, S., Liu, Y. J., and Zeldes, A. (2022). GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. In He, Y., Ji, H., Li, S., Liu, Y., and Chang, C.-H., editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 382–391, Online only. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (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).

Prasad, R., Joshi, A., and Webber, B. (2010). Realization of Discourse Relations by Other Means: Alternative Lexicalizations. In Huang, C.-R. and Jurafsky, D., editors, *International Conference on Computational Linguistics (COLING 2010)*, pages 1023–1031, Beijing, China.

Pu, D., Wang, Y., and Demberg, V. (2023). Incorporating Distributions of Discourse Structure for Long Document Abstractive Summarization. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Qin, L., Zhang, Z., and Zhao, H. (2016). Shallow Discourse Parsing Using Convolutional Neural Network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77, Berlin, Germany. Association for Computational Linguistics.

Qin, L., Zhang, Z., Zhao, H., Hu, Z., and Xing, E. (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.

Rennard, V., Shang, G., Vazirgiannis, M., and Hunter, J. (2024). Leveraging Discourse Structure for Extractive Meeting Summarization. *ArXiv*, abs/2405.11055.

Rohde, H., Johnson, A., Schneider, N., and Webber, B. (2018). Discourse Coherence: Concurrent Explicit and Implicit Relations. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267, Melbourne, Australia. Association for Computational Linguistics.

Ruan, H., Hong, Y., Xu, Y., Huang, Z., Zhou, G., and Zhang, M. (2020). Interactively-Propagative Attention Learning for Implicit Discourse Relation Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages

3168–3178, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sanders, T. J. M., Spooren, W., and Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Saveleva, E., Petukhova, V., Mosbach, M., and Klakow, D. (2021). Discourse-based Argument Segmentation and Annotation. In Bunt, H., editor, *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 41–53, Groningen, The Netherlands (online). Association for Computational Linguistics.

Scheffler, T., Aktaş, B., Das, D., and Stede, M. (2019). Annotating Shallow Discourse Relations in Twitter Conversations. In Zeldes, A., Das, D., Galani, E. M., Antonio, J. D., and Iruskieta, M., editors, *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 50–55, Minneapolis, MN. Association for Computational Linguistics.

Scholman, M., Dong, T., Yung, F., and Demberg, V. (2022). DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3281–3290, Marseille, France. European Language Resources Association.

Sediqin, M. and Argamon, S. E. (2025). Esurf: Simple and effective EDU segmentation. *arXiv*, abs/2501.07723.

Shi, W. and Demberg, V. (2017). On the Need of Cross Validation for Discourse Relation Classification. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 150–156, Valencia, Spain. Association for Computational Linguistics.

Shi, W. and Demberg, V. (2019). Learning to Explicitate Connectives with Seq2Seq Network for Implicit Discourse Relation Classification. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 188–199, Gothenburg, Sweden. Association for Computational Linguistics.

Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4080–4090, Red Hook, NY, USA. Curran Associates Inc.

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

Stede, M., Scheffler, T., and Mendes, A. (2019). Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*.

Suresh, V. and Ong, D. (2021). Not All Negatives are Equal: Label-Aware Contrastive Loss for Fine-grained Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4381–4394, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Synková, P., Mírovský, J., Poláková, L., and Rysová, M. (2024). Announcing the Prague Discourse Treebank 3.0. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1270–1279, Torino, Italia. ELRA and ICCL.

Taboada, M. and Mann, W. C. (2006). Rhetorical Structure Theory: looking back and moving ahead. *Discourse Studies*, 8:423 – 459.

Tan, Q., He, R., Bing, L., and Ng, H. T. (2022). Document-Level Relation Extraction with Adaptive Focal Loss and Knowledge Distillation. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Tang, J., Lin, H., Liao, M., Lu, Y., Han, X., Sun, L., Xie, W., and Xu, J. (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.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogs. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tsai, C.-P. and Lee, H.-Y. (2020). Order-free learning alleviating exposure bias in multi-label classification. *Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI)*, 34(04):6038–6045.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Varachkina, H. and Pannach, F. (2021). A Unified Approach to Discourse Relation Classification in nine Languages. *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*.

Wan, S., Bourgonje, P., Xiao, H., and Ho, C. W. C. (2024). Chinese-DiMLex: A lexicon of Chinese discourse connectives. *Language Resources and Evaluation, LREC 2024*.

Wang, C., Balazs, J., Szarvas, G., Ernst, P., Poddar, L., and Danchenko, P. (2022). Calibrating Imbalanced Classifiers with Focal Loss: An Empirical Study. In Li, Y. and Lazaridou, A., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 145–153, Abu Dhabi, UAE. Association for Computational Linguistics.

Wang, D., Ding, N., Li, P., and Zheng, H. (2021). CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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 2332–2342, Online. Association for Computational Linguistics.

Webber, B., Prasad, R., Lee, A., and Joshi, A. (2019). The Penn Discourse Treebank 3.0 Annotation Manual. Technical report, University of Pennsylvania, Philadelphia, PA.

Webber, B. L. and Joshi, A. K. (1998). Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse. In *Discourse Relations and Discourse Markers*.

Wu, C., Cao, L., Ge, Y., Liu, Y., Zhang, M., and Su, J. (2022). A Label Dependence-Aware Sequence Generation Model for Multi-Level Implicit Discourse Relation Recognition. *Proceedings of The Association for the Advancement of Artificial Intelligence (AAAI)*, 36(10):11486–11494.

Wu, C., Hu, C., Li, R., Lin, H., and Su, J. (2020). Hierarchical multi-task learning with CRF for implicit discourse relation recognition. *Knowl. Based Syst.*, 195:105637.

Wu, Y., Li, J., and Zhu, M. (2024). Constrained multi-layer contrastive learning for implicit discourse relationship recognition. *arXiv*, 2409.13716.

Xiang, W., Wang, B., Dai, L., and Mo, Y. (2022a). Encoding and Fusing Semantic Connection and Linguistic Evidence for Implicit Discourse Relation Recognition. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3247–3257, Dublin, Ireland. Association for Computational Linguistics.

Xiang, W., Wang, Z., Dai, L., and Wang, B. (2022b). ConnPrompt: Connective-cloze Prompt Learning for Implicit Discourse Relation Recognition. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xu, S., Li, P., Kong, F., Zhu, Q., and Zhou, G. (2019). Topic Tensor Network for Implicit Discourse Relation Recognition in Chinese. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 608–618, Florence, Italy. Association for Computational Linguistics.

Yan, Y., Li, R., Wang, S., Zhang, F., Wu, W., and Xu, W. (2021). ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Yarullin, R. and Serdyukov, P. (2020). BERT for Sequence-to-Sequence Multi-label Text Classification. In *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers*, page 187–198, Berlin, Heidelberg. Springer-Verlag.

Yung, F., Ahmad, M., Scholman, M., and Demberg, V. (2024). Prompting implicit discourse relation annotation. In Henning, S. and Stede, M., editors, *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Yung, F., Anuranjana, K., Scholman, M., and Demberg, V. (2022). Label distributions help implicit discourse relation classification. In Braud, C., Hardmeier, C., Li, J. J., Loaiciga, S., Strube, M., and Zeldes, A., editors, *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeldes, A., Aoyama, T., Liu, Y. J., Peng, S., Das, D., and Gessler, L. (2025). eRST: A Signaled Graph Theory of Discourse Relations and Organization. *Computational Linguistics*, 51(1):23–72.

Zeyrek, D. and Er, M. E. (2022). A Description of Turkish Discourse Bank 1.2 and An Examination of Common Dependencies in Turkish Discourse. In *The International Conference on Agglutinative Language Technologies as a challenge of Natural Language Processing, ALTNLP'22*, pages 30–41.

Zeyrek, D., Mendes, A., Grishina, Y., Kurfali, M., Gibbon, S., and Ogrodniczuk, M. (2019). TED Multilingual Discourse Bank (TED-MDB): A Parallel Corpus Annotated in the PDTB Style. *Language Resources and Evaluation*, pages 1–38.

Zhang, D., Li, S.-W., Xiao, W., Zhu, H., Nallapati, R., Arnold, A. O., and Xiang, B. (2021). Pairwise Supervised Contrastive Learning of Sentence Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhou, H., Lan, M., Wu, Y., Chen, Y., and Ma, M. (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.

Zhou, Y. and Xue, N. (2014). The Chinese Discourse TreeBank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49:397 – 431.

Zufferey, S. and Degand, L. (2024). *Discourse Relations and Connectives across Languages and Genres*, page 142–162. Key Topics in Semantics and Pragmatics. Cambridge University Press.