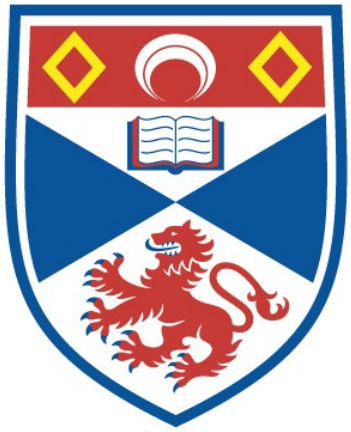
Towards integration of discourse frameworks

[Redacted version]

Yingxue Fu

A thesis submitted for the degree of PhD
at the
University of St Andrews



2025

Identifier to use to cite or link to this thesis:

DOI: https://doi.org/10.17630/sta/1181

Candidate's declaration

I, Yingxue Fu, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 51,665 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2020.

I, Yingxue Fu, received assistance in the writing of this thesis in respect of grammar, which was provided by ChatGPT.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date Aug 22, 2024 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date Aug 22, 2024 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Yingxue Fu, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Electronic copy

Embargo on part (Abstract, Chapter 4, Chapter 5) of electronic copy for a period of 2 years on the following ground(s):

- Publication would preclude future publication

Supporting statement for electronic embargo request

The Abstract contains a summary of the contents that I wish to apply embargo to, i.e. Chapter 4 and Chapter 5. Papers based on the two chapters are under review.

Title and Abstract

- I require an embargo on the abstract only.

Date Aug 22, 2024 Signature of candidate

Date Aug 22, 2024 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Yingxue Fu, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date Aug 22, 2024 Signature of candidate

ABSTRACT

Discourse coherence relates to the way that a monologue or dialogue is organized so that it is a coherent entity, instead of a random collection of clauses or sentences. As such, coherence represents an important aspect of text quality.

Existing discourse frameworks differ considerably in discourse segmentation, discourse relation taxonomies, and assumptions about structural representation. Despite superficial discrepancies, the frameworks may be related. How existing discourse frameworks are related with each other has been an open research question.

Discourse-level analysis is typically concerned with discourse relations. These relations describe the link with which two textual segments are associated with each other and they form an integral part in discourse frameworks, such as the Rhetorical Structure Theory (RST) and the Penn Discourse Treebank (PDTB). The present research begins by addressing the problem of discourse relation alignment. Existing empirical studies face the challenge of differences in discourse segmentation. To overcome the hurdle, a neural approach is proposed, which is based on label embedding techniques, and the relationship between discourse relations can be obtained automatically by comparing the learnt label embeddings.

ACKNOWLEDGEMENTS

I would like to express my profound gratitude to my supervisor, Dr Mark-Jan Nederhof, who gave me a chance to do a PhD with him, and thank him for being a kind, helpful, insightful and patient supervisor. His guidance is always timely and his notable intelligence is evident in the ingenious ideas he comes up with and his insights in questions that do not always fall into his expertise. He did not hesitate to help me when I sought help with providing references for my house and job hunts.

I would also like to extend my appreciation to my second supervisor, Dr Lei Fang, who agreed to help whenever needed in my PhD research. The books he recommended are very helpful.

I am indebted to my mentor, Dr Alice Toniolo, for her practical suggestions and guidance on various questions during my second and third years, which were particularly challenging. Even after concluding her role as my mentor, she remains kind and helpful, offering advice on conference attendance and assisting with my flight booking to Italy.

During my PhD, I was lucky to meet Prof Bonnie Webber from the University of Edinburgh. She generously provided invaluable suggestions and insights on my research topic more than once. Despite her tight schedule, she has remained responsive to my questions. Much of the literature referenced in the thesis was identified with her guidance, especially the earlier critical studies on RST and existing studies on mapping different frameworks. She showed me how to take a critical view towards these studies and the claims in existing literature about the dependency framework. My thanks also go to Prof Alex Lascarides and Dr Julie Hunter whom I came to contact with through the help of Prof Alex Lascarides. They were warm and helpful when I emailed them during the early stage of my PhD.

I would like to express my sincere appreciation to my PhD examiners for their insightful comments and thought-provoking questions.

I must acknowledge that my annual reviewers, including Dr David Harris-Birtill, Prof Saleem Bhatti and Dr Loraine Clarke, have been instrumental in identifying potential risks I was unaware of, particularly concerning ethics review.

I would like to thank the admin team and Dr Stuart Norcross for their highly

efficient work. Some senior PhD students when I started, now Dr Abd Alsattar Ardati and Dr Ryo Yanagida, are always generous to share their experience in doing a PhD.

For anyone I may have inadvertently missed, please know that your support has not gone unnoticed, and I am truly grateful.

Funding

This work was supported by the China Scholarship Council–University of St Andrews Scholarships (PhD) [202008300012].

CONTENTS

List of Figures

List of Tables

Chapter One

Introduction

1.1 Background

Discourse relates to the way that a monologue or dialogue is organized so that it is a coherent entity, instead of a random collection of isolated clauses or sentences. As such, coherence represents an important aspect of text quality:

> John took a train from Paris to Istanbul. He likes spinach.[1]
>
> Jane took a train from Paris to Istanbul. She had to attend a conference.

Although the two texts appear similar and contain sentences that are both syntactically correct and semantically sound, the first text is more difficult to comprehend than the second.

Different from conventional tasks in natural language processing (NLP), such as syntactic parsing or semantic parsing, discourse processing focuses on information beyond single sentences. The information can be utilized as a complement to sentence-level processing, such as clarifying a pronoun, or for combining sentence-level units into larger chunks (Stede, 2012), yielding a whole that is greater than the sum of its parts (Fetzer, 2014).

Discourse typically involves the interplay of multiple information sources. Position, order, adjacency and context of linguistic units all contribute to the formation of coherence (Webber and Joshi, 2012). Similar to other linguistic phenomena, ambiguity is not rare at the level of discourse. In many cases, more than one

[1] An example illustrating the role of discourse, from Hobbs (1979).

interpretation is possible for the same textual segment, as in the following example[2]:

> Small businesses say a recent trend is like a dream come true: more-affordable rates for employee-health insurance, initially at least. But then they wake up to a nightmare.[3]

With the explicitation of "but", it does not take much effort to infer that a comparison relation exists between the two sentences. However, with the word "then", a temporal relation is also perceivable.

Apart from the ambiguity of language, a text is often used for a particular purpose of communication, such as for illustration, for persuading readers, or for informing readers of the development of an event, which may influence the way that text segments are ordered and structured. One can easily notice the differences between an argumentative essay and a user manual for a product. Moreover, under different genres, information is typically unfolded in different ways in conformance with stylistic conventions. For instance, a news article tends to show the most important message at the beginning and give background information later, which stands in contrast with a novel, which generally places background information in the beginning.

Considering all these factors, discourse is a linguistic phenomenon that is intricate and challenging to capture. The elusive nature of discourse gives rise to varied assumptions and models of discourse, ranging from the model by Grosz and Sidner (1986), which is characterized by three separate but interrelated components of discourse: linguistic structure, intentional structure and attentional structure, the Linguistic Discourse Model (LDM) by Polanyi (1988), which takes a rule-based approach to discourse parsing, the lexicon-based cohesion model by Halliday and Hasan (1976), the entity-based local coherence model by Barzilay and Lapata (2008), the Centering theory (Grosz et al., 1995), to various coherence-relation-based discourse models, such as the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), the Penn Discourse Treebank (PDTB) (Webber et al., 2003)[4], and the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides,

[2]Inspired by Webber and Joshi (2012) and Webber (2019).

[3]wsj_0518, following the way of representing IDs of Wall Street Journal articles in PDTB, which will be introduced later.

[4]The name of the corpus is conventionally used to refer to the PDTB framework.

2003), and the Question Under Discussion (QUD) framework (Roberts, 2012; Onea, 2016), where discourse structure is organized based on the implicit questions answered by discourse units[5], and so on[6].

Discourse-level analysis is typically concerned with discourse relations (Rutherford and Xue, 2015). These relations describe the semantic or logical links with which two textual segments are associated with each other. Consider the example at the beginning of this chapter. In the text about Jane's activities, one can easily infer a logical connection between the two sentences. In contrast, it is far-fetched to relate the two statements "John took a train from Paris to Istanbul" and "He likes spinach", because no plausible relations seem to exist between them.

Discourse relations form an integral part of coherence-relation-based discourse models. Even though discourse models may have different assumptions and focus on different aspects of discourse, they could come to similar findings in terms of discourse relations. The following example shows RST-style annotation and PDTB-style annotation for the same text (The original text is "the acquisition should be completed by December after a definitive agreement is completed and regulatory approval is received", a segment taken from a news article):

[5]A detailed introduction to this framework is shown in Chapter 5.

[6]Chapter 2 gives a detailed account of these discourse theories, models or frameworks, which are terms used interchangeably in this thesis.

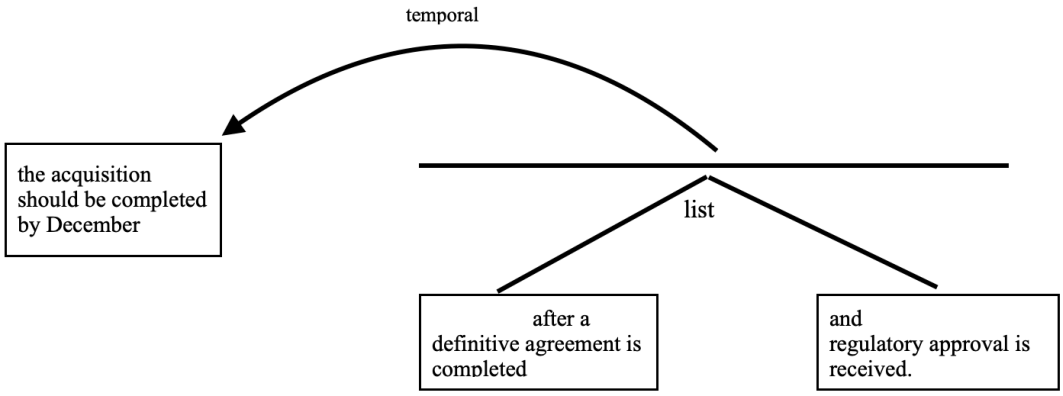


Figure 1.1: RST-style annotation for a part of the article from wsj_0635. The boxes around the textual segments are for illustrative purpose, showing the boundaries of elementary discourse units (EDUs). The horizontal line over the two lowest EDUs indicates a span/complex discourse unit (CDU) formed by two EDUs, which is further connected with an EDU (leftmost textual segment) through a ***temporal*** relation. The ***list*** relation is symmetric, which means that both EDUs are equally important, while the ***temporal*** relation is asymmetric, with the arrow head pointing to the nucleus EDU, i.e., the semantically more important EDU.

The PDTB-style annotation for the same text is[7]:

- *the acquisition should be completed by December* after **a definitive agreement is completed and regulatory approval is received.** (explicit, sense label: Temporal.Asynchronous.Succession)
- *a definitive agreement is completed* and **regulatory approval is received.** (explicit, sense label: Expansion.Conjunction)

The two relations in RST-style annotation, ***temporal*** and ***list***, resemble the two relations identified in PDTB-style annotation, ***Temporal.Asynchronous.Succession*** and ***Expansion.Conjunction***, even though the labels assigned under the two frameworks are not the same.

The intuition that different discourse frameworks are related despite surface differences has been applied to computational experiments on RST parsing (Braud

[7]Interpretation of the annotation: The label *explicit* denotes that the relation is explicitly marked. In the example, the first relation is marked by the word *after* and the second is marked by *and*. The two arguments are shown in different styles. The sense label *Temporal.Asynchronous.Succession* means that the Level-1 sense is *Temporal*, the Level-2 sense is *Asynchronous*, and the Level-3 sense is *Succession*, following the sense hierarchy in PDTB 3.0 (Webber et al., 2019).

et al., 2016) and PDTB Level-1 (L1) implicit relation classification (Liu et al., 2016). These two studies use multi-task learning to incorporate data annotated under different frameworks, so that the main task can benefit from the training signals of a related task, which represents a way of increasing training data for discourse relation classification. An issue with this approach is that the systematic correlation between relation taxonomies of different discourse frameworks remains poorly understood.

As discourse annotation is a demanding task, discourse corpora are generally small, which is challenging for developing computational means for discourse processing. Since different discourse frameworks provide distinctive but not incompatible perspectives of discourse phenomena, the interoperability and integration of different discourse models has been a topic of interest for a long time (Bunt and Prasad, 2016; Benamara and Taboada, 2015; Sanders et al., 2018; Chiarcos, 2014). Research in this direction can be used to improve understanding of discourse phenomena and discourse models. For example, in the enhanced version of RST proposed by Zeldes et al. (2024), more relations are added based on studies of signals of discourse relations and insights from PDTB. Moreover, annotations from other frameworks may provide a reference for checking the quality of annotation under a certain framework. For example, the PDTB-style annotation for two consecutive sentences from wsj_0641 is:

> *The company said it will move the storage and cross-blending operations to a site 23 miles northeast of Las Vegas to distance the operations from residential areas.* **Ammonium perchlorate is an oxidizer that is mixed with a propellant to make rocket fuel used in the space shuttle and military rockets.** (NoRel)

The label *NoRel* means that no relations can be inferred between the two sentences. However, as shown by Figure 1.2, it can be seen that there is a relation *Reason* connecting the last infinitive verb phrase of the first sentence with the second sentence in RST annotation, which is an error, because the infinitive verb phrase follows the preceding sentence more closely. The incorrect attachment of the EDU may influence the interpretation of the relation between the two sentences in RST-style annotation.

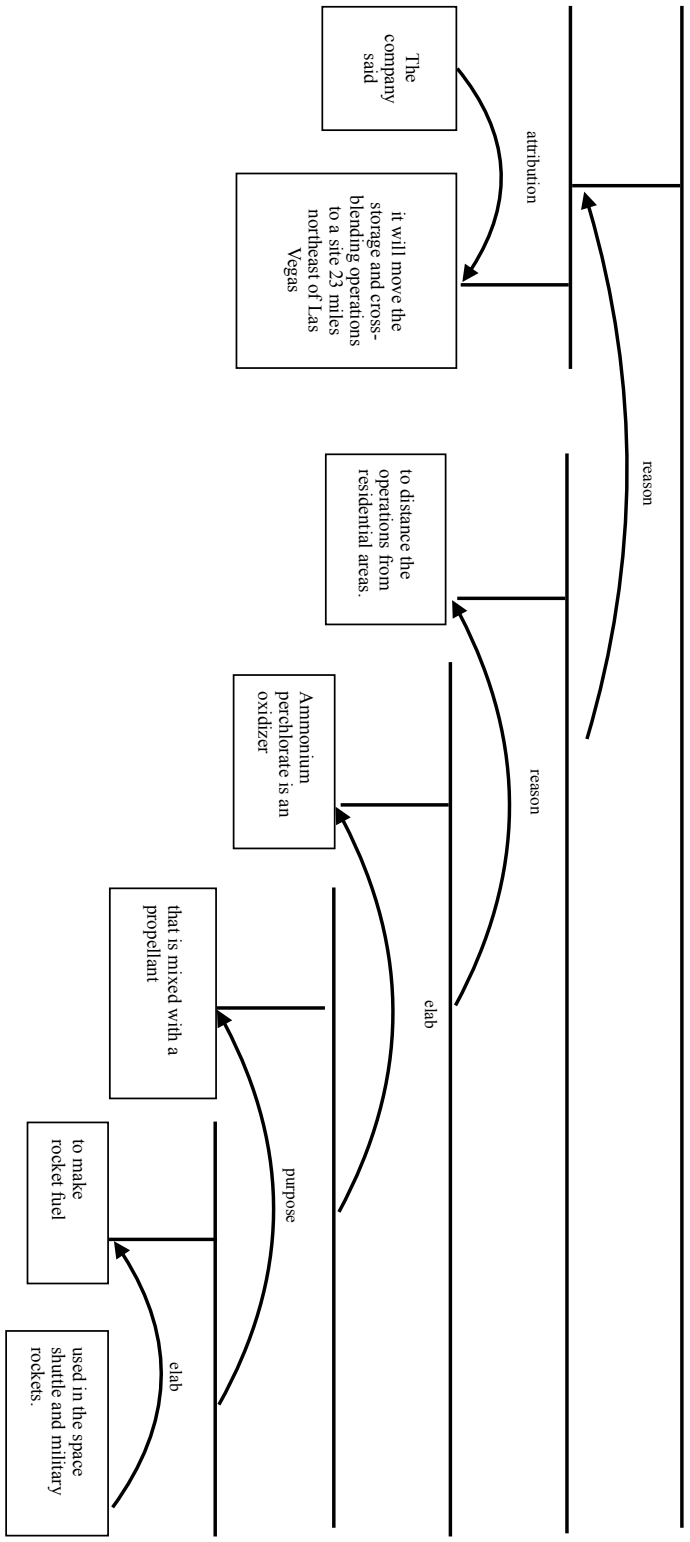


Figure 1.2: RST-style annotation for a PDTB instance of *NoRel* relation.

Moreover, if systematic relationship between discourse frameworks can be uncovered, it is possible to integrate discourse corpora effectively to increase the data amount for developing automatic systems. However, in contrast with the multi-task learning approach, most of the studies in this strand are theoretical, although it is believed that a good way to test the usefulness of the proposed methods is to merge different corpora based on the methods and apply the data in computational experiments to see whether the increased size of data improves the performance (Benamara and Taboada, 2015). Demberg et al. (2019) try to validate existing proposals for integrating discourse frameworks against annotated data. One of their research purposes is to enable joint usage of discourse corpora annotated under different frameworks for computational purposes.

This thesis continues with this research question and explores ways of using existing discourse corpora annotated under different frameworks together, with a focus on computational approaches. Integration of different discourse representations is a broad topic. Apart from the investigation of the relationships between discourse relations of different frameworks, there are other research questions, such as exploring ways to improve one framework with the insights offered by another framework. An example is the study by Zeldes et al. (2024). Motivated by the insights from the PDTB framework, the empirical study by Liu et al. (2023) and studies on relation signalling (Das and Taboada, 2018), more relations are added to RST representation, and a graph structure is constructed, which represents a theoretical development of the RST framework. Other frameworks for discourse representation can also be investigated, such as the QUD approach, which receives increasing attention over the years (Ko et al., 2022; Westera et al., 2020; Ko et al., 2023; Wu et al., 2023b, 2024) owing to its potential in converting discourse parsing into a question answering task, making it easy for large language models (LLMs) to process discourse information. Riester et al. (2021) investigate the possibility of using this framework for combining different discourse representations.

Among coherence-relation-based discourse models, RST, PDTB and SDRT are widely used for annotating discourse corpora in different languages. SDRT focuses on the role of rhetorical relations in enriching dynamic semantics, and it has been applied to annotate discourse structure of multi-party dialogues, such as the STAC corpus (Asher et al., 2016) and the Molweni corpus (Li et al., 2020), where discourse structure is too complex to be covered by trees. RST and PDTB have been primarily used to annotate monologues, represented by the RST Discourse Treebank (RST-

DT) (Carlson et al., 2001) and the Penn Discourse Treebank (PDTB) (Prasad et al., 2008, 2018), respectively. The two discourse corpora have an overlapping section of the Wall Street Journal articles from the Penn Treebank (Marcus et al., 1993). As the focus of this thesis is computational means, RST and PDTB are chosen in all the computational experiments to reduce the confounding effect of domain shift caused by differences in language and genre.

1.2 Research Questions

As outlined in the previous section, within the broader topic of integrating discourse frameworks, this thesis focuses on exploring approaches to enable the joint use of data across different frameworks and addressing some limitations identified in existing mainstream discourse frameworks based on insights from the other frameworks.

The research questions can be summarized as follows:

1. As discourse relations play a pivotal role in mainstream discourse frameworks, previous research efforts on understanding the relationships between discourse frameworks are directed towards the alignment of discourse relation taxonomies. The availability of corpora annotated in different frameworks in parallel spurs data-driven methods to investigate the research question (Demberg et al., 2019; Bourgonje and Zolotarenko, 2019; Scheffler and Stede, 2016). However, different criteria in discourse segmentation hinder systematic studies, as shown by the semi-automatic algorithm proposed by Demberg et al. (2019). Even if the *strong nuclear hypothesis* (Marcu, 2000)[8] can be used to alleviate the problem, there are still many ambiguous cases and the method works on limited data. Moreover, despite the claim of testing the methods for aligning relation taxonomies extrinsically (Benamara and Taboada, 2015), no empirical results have been reported.

   —Can the relationship between different relation taxonomies be learnt automatically? How could the alignment of discourse relations be used?

2. If relabeling data based on alignment rules and simply combining data from different frameworks does not work effectively for enabling joint use of data across different frameworks, are other approaches possible? Existing

[8] A detailed explanation is provided in section 2.6.1.4.

literature contains a number of studies that provide theoretical analysis of the relationship between relation taxonomies. Are these studies useful?

—What are efficient ways of discourse processing across different frameworks?

3. RST captures deep discourse structure, while PDTB focuses on local discourse relations. An emerging approach in discourse parsing is based on the QUD framework. Although each framework has its advantages, they also present challenges and theoretical limitations.

   —Is it possible to combine the advantages of the frameworks while addressing some of their limitations?

1.3 Research Statements and Contributions

An important strand of research on the interoperability of discourse frameworks focuses on the relationships between discourse relation taxonomies employed by different frameworks (Hovy and Maier, 1992; Bunt and Prasad, 2016; Benamara and Taboada, 2015; Chiarcos, 2014). As can be seen from the examples of RST-style and PDTB-style annotations shown in section 1.1, while the two frameworks differ considerably in structural organization[9], the relations that connect pairs of discourse units may be related.

Building on this line of work and considering the central role that discourse relations play in mainstream frameworks such as RST, PDTB and SDRT, this thesis begins by **investigating the relationships among relation taxonomies across different discourse frameworks**. Thanks to the availability of discourse corpora annotated in different frameworks in parallel, researchers have been able to validate theoretical insights on real data (Demberg et al., 2019). However, some challenges remain. For instance, these data-driven methods tend to be limited, as they often rely on string matching to align discourse units before identifying potentially related discourse relations. Different frameworks adopt varying rules for discourse segmentation; for example, RST and SDRT typically treat clauses as basic discourse units, while PDTB considers semantically motivated abstract objects (such as propositions, events, or claims) as discourse units. As a result, aligning discourse units can be challenging and is feasible only on a limited

[9]RST uses trees to represent discourse structure, while PDTB does not annotate high-level discourse structure. The differences are highlighted in Chapter 2.

amount of data. The study by Demberg et al. (2019) presents a method to address this challenge by leveraging the notion of nuclearity in RST, enabling coverage of discourse units that are not adjacent in the original text but can be connected along the nuclearity path over RST trees. However, their work also reveals numerous cases where discourse unit alignment is uncertain, and even when alignment of discourse units is achieved, manual inspection is needed to identify correct relation mappings. This is because in PDTB, more than one relation can be annotated for the same pair of discourse units, whereas RST, owing to its theoretical constraint of using trees to represent discourse structure, allows only one relation to be annotated. **This thesis tries to overcome the challenges posed by different criteria of discourse segmentation and proposes an automatic method to learn the alignment of relation taxonomies.** Experimental results show that the method achieves slightly higher performance than that based on the approach proposed by Demberg et al. (2019).

One aim of mapping discourse relation taxonomies across different frameworks is to increase the amount of data available for discourse processing. This thesis presents experimental results demonstrating how to use relation mapping for data augmentation and the performance that can be achieved based on the alignment of discourse relations. The results indicate that, even when relation taxonomies are aligned, simply relabeling data based on alignment rules and combining data from different frameworks does not yield performance gains over using data from a single framework. This finding highlights the need for research on more effective methods for data augmentation across frameworks.

In existing studies, another approach to facilitating interoperability among discourse frameworks is to decompose discourse relations into primitive dimensions, forming a platform-agnostic basis for representing discourse relations across frameworks (Sanders et al., 1992). Although different frameworks use varied relation taxonomies, this decompositional approach allows for the joint use of data from different frameworks to train these primitive dimensions. The predicted dimensions can then be applied for discourse relation classification. The UniDim proposal proposed by Sanders et al. (2018) represents a theoretical effort to expand the Cognitive approach to Coherence Relations (CCR) framework (Sanders et al., 1992, 1993) as an *interlingua* for representing discourse relations. **Following this line of thought, this thesis explores methods for applying the UniDim proposal in cross-framework discourse relation classification. Experimental**

results suggest that the UniDim proposal is effective in facilitating discourse processing across different frameworks.

While mainstream discourse frameworks are generally grounded in discourse relations and are widely used for creating corpora and developing computational systems, they are subject to theoretical limitations. For example, RST conflates intentional, semantic and textual relations into a single tree representation (Stede, 2008b), while PDTB focuses solely on local semantic relations (Webber, 2004). Compared with RST and PDTB, where discourse relations are encoded by senses such as "Contrast" and "Elaboration", the QUD framework assumes that discourse units provide answers to some implicit questions, with discourse structure shaped by the relationships between these questions. By encoding discourse relations in free-form questions, this approach allows annotators to bypass the need to familiarize themselves with fixed-term representations of discourse relations predefined by experts in discourse annotation projects. Moreover, as question answering is a widely studied NLP task (Kamalloo et al., 2023), the QUD framework has the potential to transform discourse parsing into a question-answering (QA) task, which may explain its increasing prominence in the computational linguistics community, especially with the advent of large language models (Raffel et al., 2020; Brown, 2020). However, incorporating discourse structure within the QUD framework remains challenging, as evidenced by the low inter-annotator agreement in annotating QUD trees (De Kuthy et al., 2018). **This thesis explores how the QUD framework can be leveraged to integrate different perspectives of discourse representation, probably beyond mainstream discourse frameworks, aiming to overcome some of the limitations identified in existing frameworks.** The model proposed by Van Kuppevelt (1993) is implemented, and high inter-annotator agreement is achieved on naturally occurring texts, which suggests that the model can be applied reliably. As sentences are taken as the basic discourse unit for which a topic can be identified, to develop a full QA approach for discourse processing, the method proposed by Pyatkin et al. (2020) is implemented for intra-sentential level processing.

To sum up, the main contributions of the thesis include:

1. An automatic approach for learning the alignment of discourse relation taxonomies employed by different frameworks, for which intrinsic and extrinsic evaluations are performed (Chapter 3).

2. A platform-agnostic approach for discourse relation classification, which is based on the theoretically motivated UniDim proposal (Chapter 4).

3. A new QUD model for integrating different perspectives of discourse representing, incorporating topic segmentation in discourse processing and arguably achieving the multi-level analysis advocated by Grosz and Sidner (1986) (Chapter 5).

1.4 Overview of the Dissertation

A short description of each chapter in this thesis is given below.

Chapter 1 provides the background of the research and introduces the topic of the thesis, outlining the research questions to be addressed.

Chapter 2 reviews existing studies about discourse frameworks and research on the relationship between discourse frameworks.

Chapter 3 focuses on the experiments of developing a fully automatic means of learning the alignment of relation taxonomies for RST and PDTB.

Chapter 4 presents the experiments of applying the UniDim proposal in cross-framework discourse relation classification.

Chapter 5 presents a study on combining deep discourse structure with shallow discourse annotation.

Chapter 6 provides a general discussion and concludes the thesis.

CHAPTER TWO

CHAPTER 2
LITERATURE REVIEW

2.1 Chapter Overview

As has been introduced briefly in Chapter 1, a text is not a simple collection of isolated sentences. Sentences generally appear in a certain order and are connected with each other through logical or semantic means to form a coherent whole. In recent years, information beyond the sentence level has been attracting increasing attention, and various studies have shown the benefits of incorporating discourse-level information or coherence-related training objectives in NLP tasks, such as text generation (Bosselut et al., 2018), language modelling (Iter et al., 2020; Lee et al., 2020; Stevens-Guille et al., 2022), and summarization (Xu et al., 2020).

While Chapter 1 demonstrates the existence of various discourse frameworks, the complex theoretical background of these frameworks necessitates a more detailed description to provide sufficient foundational knowledge for the research topic. This chapter presents an overview of discourse frameworks and previous studies on the relationship between different discourse frameworks. The categorization of discourse frameworks in Jurafsky and Martin (2023, Chapter 27) is adopted, but reference may be made to other studies, such as Webber (2006), which analyzes discourse frameworks from the perspective of the source of discourse relations in terms of constituency and anaphoric dependency.

The review starts with some influential discourse frameworks other than the discourse frameworks in focus in this thesis, i.e. coherence-relation-based discourse

frameworks represented by RST (section 2.6.1) and PDTB (section 2.6.2), with the purpose of fostering a better understanding of the central issues discussed in the thesis. Different from Jurafsky and Martin (2023), frameworks focusing on global coherence are excluded, because greater text segments may involve multiple discourse relations linked to items outside the textual segments, giving rise to a different structure than discourse structure. Additionally, at the global level of discourse, other factors such as genres and communicative purposes are at play (Taboada and Mann, 2006), and analysis from these perspectives is more informative than from the perspective of coherence relations.

2.2 Grosz & Sidner's Model

The theory by Grosz and Sidner (1986) is one of those theories whose linguistic claims about discourse are also computationally significant (Mann and Thompson, 1987). With this theory, it is believed that discourse structure is composed of three separate but interrelated components: linguistic structure, intentional structure and attentional structure.

The linguistic structure refers to the level of discourse segmentation. Words or clauses may work together as a single textual unit to achieve a function in discourse. These textual units are discourse segments. Two consecutive utterances can be in the same discourse segment, but it is also possible that they belong to different discourse segments, and non-consecutive utterances can belong to the same discourse segment. Moreover, discourse segments can be embedded to reflect the requirements of the intentional structure, which is to be explained below. These observations are shared in later discourse models, such as RST and SDRT, and later studies also show that discourse segmentation influences inferences and relations that can be attached (Demberg et al., 2019; Carlson et al., 2001; Asher and Lascarides, 2003).

The intentional structure captures the purposes of discourse segments and the purpose of discourse, the former being referred to as discourse segment purposes (DSPs) and the latter being called discourse purpose (DP). In Grosz and Sidner (1986), it is believed that although discourse participants may have more than one aim in engaging in a discourse, only one of these purposes is foundational, and one intention can be specified for a discourse segment. This position is also taken in RST. Grosz and Sidner (1986) notice that this assumption may be too strong, but

they take it to be a convenient step to formulate the theory. DSPs contribute to the achievement of DP, and DSPs may be structurally related to each other: if DSP1 provides partial satisfaction of DSP2, then DSP1 contributes to DSP2 or DSP2 dominates DSP1, thus giving rise to a hierarchy of intentions. If the order in which the DSPs are satisfied is important for discourse participants, when DSP1 must be satisfied before DSP2, it forms a case that DSP1*satisfaction-precedes* DSP2. Grosz and Sidner (1986) mention explicitly that no finite lists of discourse purposes exist, but they give some broad categories of such intentions.

The attentional structure records the salience of entities, properties or relations during the development of discourse and therefore, it is a dynamic structure, with different entities gaining prominence at different periods. It is modeled by associating each discourse segment with a *focus space*, which contains the salient entities, properties, relations and DSP of that segment. A stack can be used to simulate the functioning of this structure: when the DSP for a new discourse segment contributes to the DSP of an immediately preceding segment, the focus space of the new discourse segment is pushed on the stack of the discourse, and when the DSP contributes to some DSP higher in the dominance hierarchy, for instance, DSP_n of a discourse segment n contributes to DSP_{n-2} of a discourse segment n-2, the focus spaces of the top two discourse segments need to be popped before the focus space of the new discourse segment n is pushed on the stack. From the way the attentional structure operates, it can be seen that the attentional structure is closely related to the intentional structure, but Grosz and Sidner (1986) take pains to show that these two types of structure should not be conflated in an adequate description of discourse structure.

These three aspects capture discourse phenomena in a systematic way, and other discourse theories may be related to this theory in some way. For instance, the Centering Theory (Grosz et al., 1995) (section 2.3) relates to the attentional state, and RST (section 2.6.1) deals with the intentional structure.

2.3 Centering Theory and Entity-Grid Model

At different points of discourse development, different entities come into focus and become salient, and this process influences the choice of referring expressions in linguistic realizations. The Centering Theory studies the relationships between focus of attention, choice of referring expressions and perceived coherence *within* a

discourse segment, which suggests that the Centering Theory is built on the work of Grosz and Sidner (1986) and studies the linguistic and attentional structures and how the two levels constrain interpretation at separate levels.

With this theory, centers are defined as entities that link one utterance with other utterances in the same discourse segment (Grosz et al., 1995). Each utterance has a set of forward-looking centers and a single backward-looking center. For a sequence of utterances, represented as U_0, U_1, U_2, ..., U_n, under a discourse segment, the backward-looking center of U_2 is connected with one of the forward-looking centers of U_1. The forward-looking centers are partially ordered based on their relative prominence and the most highly ranked one, if it is realized in the next utterance U_2, becomes the backward-looking center of U_2.

If the backward-looking center of an utterance U_2 is the same as that of U_1 and this center is the most highly ranked element in the forward-looking centers of U_2, it forms a case of *center continuation*, because the center will likely be the backward-looking center of U_3. However, if the backward-looking center of U_2 is the same as that of U_1 but this center is not the most highly ranked element in the forward-looking centers of U_2, it forms a case of *center retaining*, because although the backward-looking center of U_1 is the same as that of U_2, it is not the most probable candidate as the backward-looking center of U_3. The case of *center shifting* refers to the situation where the backward-looking center of U_2 is not identical to that of U_1.

As can be seen here, the ordering of forward-looking centers is important. It is found from empirical data that the grammatical role is a relevant indicator, with the subject being most likely to be the highest ranked forward-looking center of an utterance, followed by object, and then other constituents. Grosz et al. (1995) show that lower-ranked elements in forward-looking centers cannot be pronominalized before higher-ranked ones, which reflects the constraint of centers on linguistic realization.

Another important part of the theory concerns constraints on the movement of centers. Two rules are proposed in Grosz et al. (1995). The first rule stipulates that if any of the forward-looking centers of an utterance is realized with a pronoun in the next utterance, the backward-looking center of the next utterance should also be realized with a pronoun. This rule constrains the pronominalization of utterances. The second rule specifies that center continuation is preferred over

center retaining, which is preferred over center shifting. This rule reflects the intuition that smooth transition between centers is conducive to local coherence[1].

Inspired by the Centering Theory, Barzilay and Lapata (2005) propose the entity-grid model for local coherence. The model is based on similar intuitions as the Centering theory, for instance, entity distribution follows some regular patterns, the salience status of an entity is often signaled by its grammatical function, and continuation of entity is preferred over shifts of entity mentions when building discourse coherence. However, different from the Centering Theory, which studies transitions of centers inside a discourse segment, the entity-grid model captures distributions of entities across the whole discourse and computes a coherence metric for the whole discourse, even though the type of coherence it tries to capture is still **local** coherence. The idea of the model is to represent entity transitions in a text as a two-dimensional array, called entity grid, with the rows representing sentences of a text, and columns representing entities, which are classes of coreferent noun phrases. If an entity appears in a sentence, its grammatical role is recorded in the corresponding cell of the grid. Similar to Grosz et al. (1995), three grammatical roles are distinguished: subjects, objects and others, and a probability is computed for transition between different roles in the corpus based on the frequency of a role over the overall count of transitions from the role to other roles.

2.4 Linguistic Discourse Model

The Linguistic Discourse Model (LDM) proposed by Polanyi (1988) represents a formal model for discourse parsing. In this model, discourse is represented with a tree structure, which is built recursively through sequencing and embedding of discourse units, similar to RST. The basic discourse unit, or elemental discourse unit, is generally a clause or one-word utterance, such as a false start in speech. Different combinations of these elemental discourse units form varied types of *discourse constituent units* (DCUs). A structural category *discourse operator* is defined in LDM, which is a term referring to linguistic mechanisms that do not have propositional content but serve to modify the force of discourse constituents, such as connectives that link DCUs together. As LDM is intended to be a general discourse model, linguistic phenomena of dialogues are taken into account, and

[1]Local coherence can be loosely interpreted as the coherence established between consecutive or neighboring discourse segments.

some discourse operators function to accommodate special features of speech communication.

There are four types of DCUs: *Sequence*, where a DCU is created from arbitrarily many constituents of the same type; *Expansion Unit*, where a DCU is constructed by a clause and a subordinate clause that expands on it; *Binary Structure*, where a DCU is formed by joining two DCUs with an explicit connector operator, such as "because"; and *Interruption*, where one DCU is interrupted by intervening materials. These patterns of combining clauses to form DCUs resemble *schemas* in RST (section 2.6.1.2) but RST schemas are applicable to all the levels of tree nodes while the types of DCUs here only describe how elemental discourse constituents are combined to form a DCU.

With LDM parsing, these DCUs are processed incrementally, a record is kept at all times, specifying which discourse units have been processed, which have been interrupted, and which are open for attachment for incoming new discourse units. The Right Frontier Constraint (RFC) forms the structural constraint for LDM. With RFC, a new DCU is attached to the parse tree as the rightmost constituent, and the rightmost nodes at all the levels of the tree are nodes on the right frontier and open for new attachment, irrespective of whether the nodes have been attached to the tree with a coordination or subordination relation[2]. This principle restricts access to previously parsed DCUs. An illustration of RFC is shown in Figure 2.1.

For the derivation of discourse relations, each DCU is designed to have a *context frame*, which is a semantic frame that contains propositional information and context information for interpreting a DCU. It provides a mechanism for deriving relations between an incoming DCU and open nodes on the parse tree.

In addition, LDM takes communication purposes into consideration, and incorporates discourse structuring conventions for different types of communication as constraints for discourse tree construction.

[2]A coordinating relation means that the segments connected by the relation are equally important, and typical connectives include "and" and "in contrast". In a subordinating relation, one segment plays a central role while the other provides subsidiary information. A typical connective is "because".

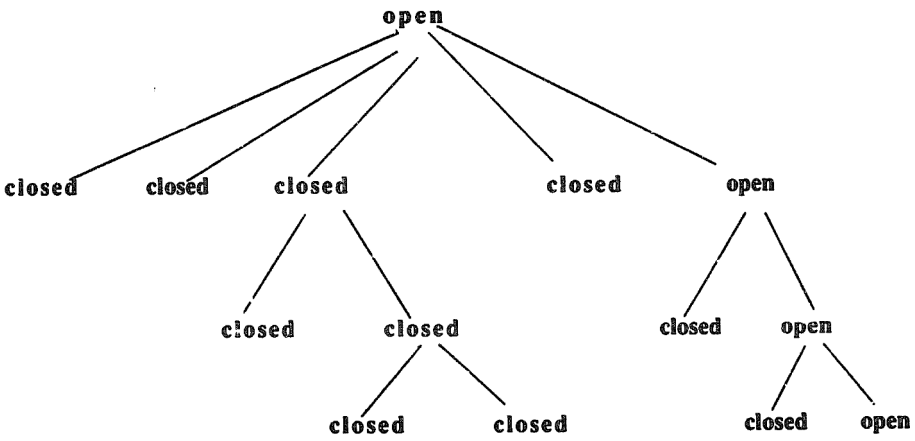


Figure 2.1: RFC illustrated, originally from Polanyi (1988). The rightmost nodes at all the tree levels are open for attachment of new discourse units and the rest are closed.

2.5 Lexicon-Based Discourse Model

The theory proposed by Halliday and Hasan (1976) focuses on how various lexical means are used to achieve cohesion[3]. These lexical means include:

1. reference: Reference is achieved by some words or phrases that cannot be interpreted without reference to something else. Based on the entities that are involved, it can be divided into different types, such as *personal* (e.g., "I", "me", "my"), or *demonstrative* (e.g., "this", "here", "the").

2. substitution: Substitution is similar to reference but the major difference is that substitution is a relation in wording, rather than in meaning. Some words are used to replace content words or clauses mentioned in the previous context to avoid repetition. Based on the grammatical categories of the replaced items, substitution may be divided into *nominal* (generally using "one", "ones", and "same" for replacing a noun/nouns or a pronoun/pronouns), *verbal* (using "do" to replace a verb, for example, "(a) Does she like the fruit? (b) I think she does.") and *clausal* (using "so", "not", etc. to replace a clause or sentence, for example, "(a) Is your sister happy? (b) I think so.").

3. ellipsis: Ellipsis means leaving a structural slot to be filled by the readers

[3]Loosely speaking, cohesion refers to links established through surface-level lexical or grammatical devices, without involving deep inference.

through inference, because the content inside is assumed to be known (Halliday and Hasan, 1976, p. 143), for example, "The first article is about politics, and the second, finance".

4. lexical cohesion: Lexical cohesion is mainly achieved through reiteration, including repetition of a lexical item, using a general word to refer to a lexical item in the preceding context, and using synonyms, near-synonyms, or superordinate terms for referring to lexicons in the preceding context.

5. conjunction: Conjunctions can be divided into additive, adversative, causal, and temporal types.

As pointed out by Webber (2006), cohesion realized through the first four lexical means is in essence anaphoric dependency, and conjunction is the only type that can be related to discourse relations. Except for the last type, the lexicon-based discourse framework remains under-explored. Studies that apply the intuitions of the framework include those by Mesgar and Strube (2016) and Lei et al. (2021).

Mesgar and Strube (2016) focus on the fourth type of lexical cohesive device. For two sentences *S1* and *S2*, *S1* followed by *S2*, they compute cosine similarities of lexical items of the two sentences and take the highest value of cosine similarity between words in *S2* and words in *S1* as the weight of semantic connection between the two sentences. In this way, a text is modeled as a directed graph, with sentences as nodes and the weights computed above as strengths of edges, and the edges are directed to represent sentence order. A threshold may be set for edge strengths to remove some edges. Larger subgraphs, i.e. subgraphs with a greater number of nodes, can better capture structural relationships between sentences, but they tend to suffer from sparsity. To tackle this challenge, the Kneser-Ney smoothing method (Kneser and Ney, 1995) is adopted in their experiments. As is pointed out by Lei et al. (2021), their method is based on cosine similarity of static word embeddings and contextual information of words is not considered.

Lei et al. (2021) focus on identifying contrastive word pairs as a way of testing a language model's capability of capturing lexical coherence. A corpus is created for this purpose. They compare the capabilities of static word embeddings and contextualized word embeddings in this task. It is shown that contextualized embeddings are more powerful. Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) yield the best performance for this task, but the overall accuracy only exceeds 70%, indicating that it is challenging to detect

contexual contrastive word pairs. Furthermore, they compare the performances on detecting contrastive word pairs that appear in the same sentence but in different clauses and identifying contrastive word pairs from two adjacent sentences. Their results indicate that the task is much easier when the word pairs are inside the same sentence, and the advantage of BERT over static embeddings is greater in this setting. Further experiments reveal that the improvement of BERT over static embeddings is mainly attributed to its ability of recognizing repetitive patterns that serve as clues for identifying such contrastive word pairs, for instance, "...lives in happiness, ...lives in misery...". The capability of BERT in modelling lexical coherence is still limited.

2.6 Coherence-Relation-Based Discourse Models

Discourse models that are based on coherence relations derive the source of coherence from the relations between textual segments.

2.6.1 RST

The main thesis of RST is presented in Mann and Thompson (1988) and Mann and Thompson (1987), which can be summarized with three components: relations, schemas, and structures.

2.6.1.1 Relations

Relations refer to links between two non-overlapping textual spans. Based on the relative semantic salience of the two spans, relations can be symmetric, if both spans are equally salient, or asymmetric, if one span provides central information while the other plays a subordinate role. For an asymmetric relation, the span of greater importance is called the *nucleus* and the other span is called the *satellite*.

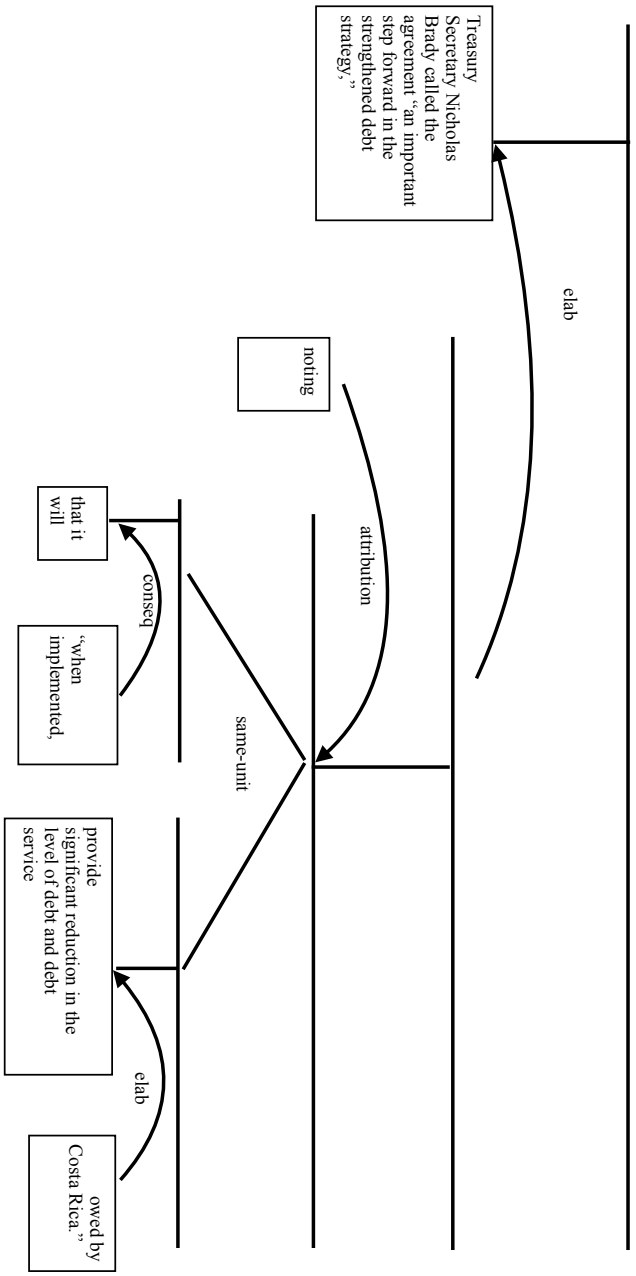


Figure 2.2: RST analysis of wsj_0624 from RST-DT. *elab* means *elaboration*, and *conseq* denotes *consequence*. These are names of RST discourse relations.

An example of RST analysis of a text is shown in Figure 2.2. The RST graphic convention introduced in Mann and Thompson (1988) is used, where the horizontal

lines indicate intermediate textual spans formed by composition of lower-level spans, and the vertical bars highlight the nuclei. The leaf nodes in the boxes are EDUs and the arcs are discourse relations, with the arrowheads pointing to the nuclei for asymmetric relations. Thus, the whole text is the span denoted by the top horizontal line, which suggests that RST analysis aims at full coverage of the text. The two spans of the relation *same-unit*[4] are equally important, hence the absence of arrowheads in the depiction of the relation.

The relations in RST are specified by four types of *constraints*: constraints on the nucleus (N), constraints on the satellite (S), constraints on the combination of the nucleus and the satellite (N+S), and constraints on the effect of the relation. When a particular relation is determined, nuclearity analysis is also performed as a natural by-product of this process, which is considered questionable in later studies (Stede, 2008b; Taboada and Mann, 2006). Mann and Thompson (1988) show that nuclearity forms a text structure organizing principle with deletion tests: if the satellite parts are deleted, the gist of the text can still be captured, while if the nuclei are deleted, the text becomes unintelligible.

Each relation must be defined in terms of constraints on the effect, which specify why each span should be included and what function is to be achieved with this relation. This focus on functions in the definition of relations forms one of the differences between RST and PDTB. Mann and Thompson (1988) provide a list of relations and their definitions with these four constraints. For instance, the definition of the relation *background* is shown in Table 2.1:

constraints on N:	R won't comprehend N sufficiently before reading text of S.
constraints on the N+S combination:	S increases the ability of R to comprehend an element in N.
the effect:	R's ability to comprehend N increases.

Table 2.1: Constraints of *background*. *R* represents the **reader** of the text, as opposed to the **writer**, which are roles deemed necessary for the analyst to make RST analysis of an text, as described in Mann and Thompson (1988).

In Mann and Thompson (1988), it is believed that relations are not determined based on morphological or syntactic signals, because no unambiguous signals can be identified, which makes another difference between RST and PDTB. The

[4]Strictly speaking, *same-unit* is not a typical discourse relation, because the purpose of introducing this relation is to link some embedded discourse units. The example here is merely for illustrating the notion of symmetric relations.

proposed RST relations are grouped based on their resemblance, and the relations can be further grouped into two broad classes: subject matter relations and presentational relations, the former concerning the semantic or ideational relations and the latter focusing on intended effects. Moore and Pollack (1992) elaborate on this division and propose parallel annotations of these two types of relations.

Although the order of N and S for the relations is not specified in the definitions, some patterns emerge, for example, for the *background* relation, S normally appears before N, while for the *purpose* relation, N typically appears before S. Mann and Thompson (1988) point out that if a non-canonical order happens to appear, i.e. N before S for *background*, the text quality can be improved by reversing the order most of the time.

The relations proposed by Mann and Thompson (1988) are not meant to be a fixed set and they can be extended and modified to suit the needs for analyzing specific data. Nevertheless, later studies tend to treat these relations as a fixed taxonomy.

2.6.1.2 Schemas

Schemas are abstract patterns defined by textual spans, relations that apply to the spans, and how the textual spans form a structure, outlining how textual spans co-occur with each other. Figure 2.3 shows the five schemas in RST. The other schemas not shown follow the *circumstance* schema, where two textual spans are connected by one relation, characterized by a nucleus and a satellite. The schemas are named based on the relations applied. The *contrast* schema has two and only two nuclei, the *joint* schema contains two or more nuclei, the *sequence* schema has two or more nuclei and a succession relation holds between adjacent nuclei, and the *motivation+enablement* schema is characterized by a shared nucleus between two sets of adjacent span pairs.

2.6.1.3 Structures

To make RST analysis, the text is first segmented into basic units. Mann and Thompson (1988) choose clauses as basic units to keep this step theory-neutral. Following this step, the analyst has to determine which relation holds, and then applies the schema. Schema application yields complex discourse units. The step of analyzing text structure is constrained by four principles: completedness, which means that the entire text should be covered in the analysis; connectedness, which specifies that except for the span covering the whole text, each span should

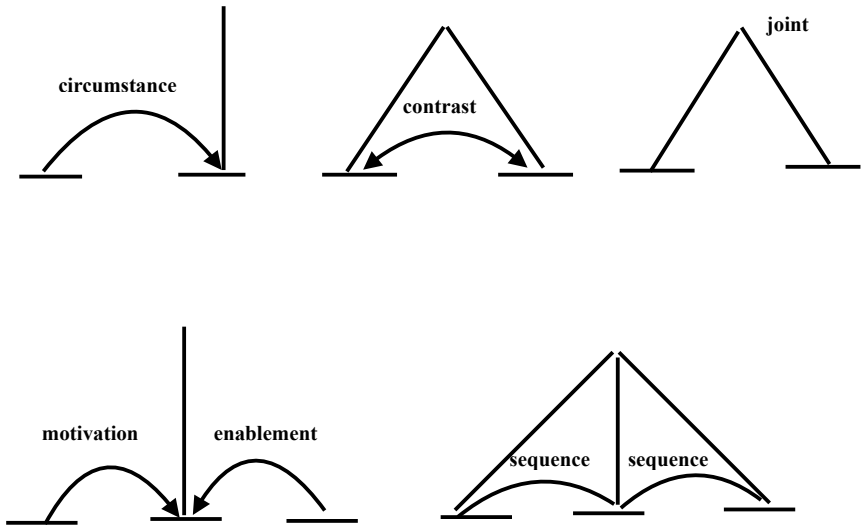


Figure 2.3: Five schemas in RST, originally from Mann and Thompson (1988).

be either a terminal or a constituent of a schema; uniqueness, which means that a single schema application is applied to a set of textual spans; and adjacency, which requires that the textual spans of each schema application form a continuous textual span. With these constraints, the final product of RST analysis is guaranteed to be a single tree.

Mann and Thompson (1988) stress that the adjacency requirement can be relaxed so that texts of different genres can be analyzed, and like the other types of linguistic analysis, multiple RST analyses of a text are possible. Although this point of view is endorsed in the development of computational means for RST parsing, only the discourse tree with the highest probability is retained (Marcu, 1997). Additionally, discourse markers are used extensively for identifying discourse relations in Marcu (1997), which deviates from the original assumption in Mann

and Thompson (1988) but yields high accuracy.

2.6.1.4 Strong Nuclearity Hypothesis

Marcu (1996) finds that the four principles for constraining text structure are not sufficient, because compositionality of textual spans is not specified, which gives rise to ambiguity at higher levels of span aggregation during the process of constructing an RST tree. Inspired by the observation of using nuclearity as a text organization principle in Mann and Thompson (1988), Marcu (1996) proposes the *strong nuclearity hypothesis*, which means that for a large span with two composite spans, the relation that holds between the two composite spans should also hold between the nuclei of the two spans. Take the example from Marcu (1996) for illustration:

> [No matter how much one wants to stay a non-smoker,]$_1$[the truth is that the pressure to smoke in junior high is greater than it will be any other time of one's life.]$_2$[We know that 3,000 teens start smoking each day,]$_3$[although it is a fact that 90% of them once thought that smoking was something that they'd never do.]$_4$

Suppose a set of relations, *R*s, which hold between spans [1]-[4], can be identified (the nuclearity of each relation is shown as (N(ucleraity): span id):

- *justification* (1, 2) (N: 2)
- *justification* (4, 2) (N: 2)
- *evidence* (3, 2) (N: 2)
- *concession* (3, 4) (N: 4)
- *restatement* (4, 1) (N: 1)

As adjacent EDUs 1 and 2 are linked by a relation, the two EDUs can be aggregated, and similarly, adjacent EDUs 2 and 3, and EDUs 3 and 4 can be aggregated into larger spans. Thus, multiple structures are possible, which are shown in Figure 2.4.

The structure in d) involves a schema *evidence+concession*, which is not possible based on Mann and Thompson (1988), hence discarded in further analysis. In a), for a span $S_{1\text{-}4}$ formed by the aggregation of two spans $S_{1\text{-}2}$ and $S_{3\text{-}4}$, if the

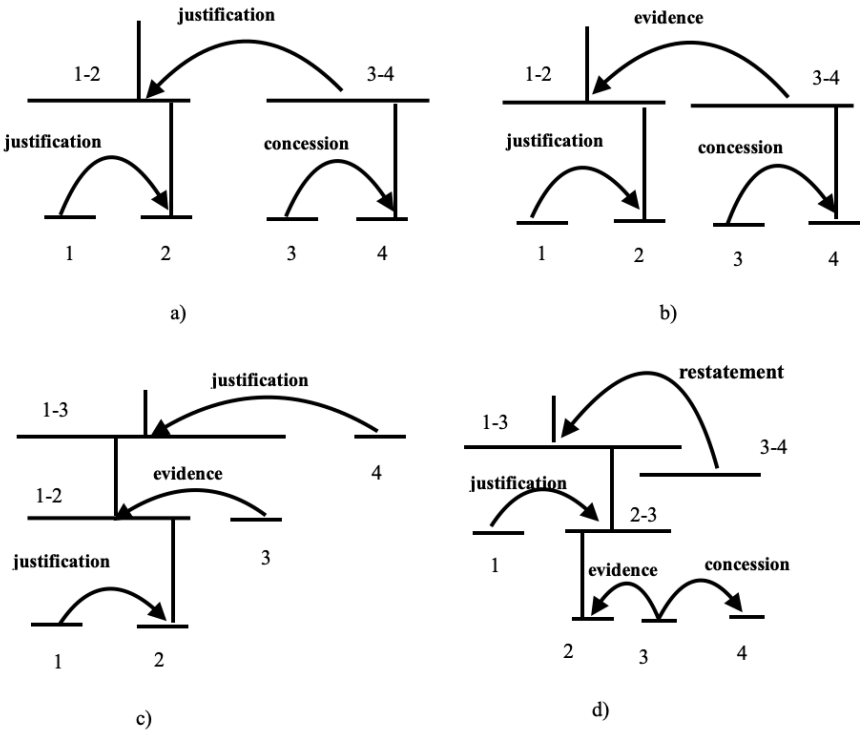


Figure 2.4: Possible structures, originally from Marcu (1996).

nucleus of $S_{1\text{-}2}$ is EDU 2 and the nucleus of $S_{3\text{-}4}$ is EDU 4, the relation that holds between $S_{1\text{-}2}$ and $S_{3\text{-}4}$ should also hold between EDU 2 and EDU 4. Therefore, the relation that links $S_{1\text{-}2}$ and $S_{3\text{-}4}$ should be *justification*, and the analysis shown in a) is correct. In b), the structure is similar to a). However, the relation *evidence* that holds between EDU 3 and EDU 2 cannot be applied when $S_{1\text{-}2}$ and $S_{3\text{-}4}$ are aggregated into $S_{1\text{-}4}$, because EDU 3 is not the nucleus of $S_{3\text{-}4}$.

When the tree is built recursively from the bottom up, the nucleus of each span is recorded with a property called *promotion set* to facilitate span aggregation at higher levels. The promotion sets of leaf nodes, i.e. EDUs, are EDUs themselves, and the promotion sets of intermediate spans are formed by the union of the

promotion sets of their immediate nuclei spans/EDUs. Hence, a span can be characterized by three properties: a promotion set, a relation, and a nuclearity status.

Span	Promotion set	Relation	Nuclearity
1	1	leaf	S
2	2	leaf	N
1-2	2	justification	N
3	3	leaf	S
1-3	2	justification	N
4	4	leaf	S

Table 2.2: Illustration of span properties with the case shown in Figure 2.4 c).

An example is shown in Table 2.2, which is based on the structure shown in Figure 2.4 c). The leaves, i.e. EDUs have themselves as promotion sets. The span 1-2 has EDU 2 as nucleus, and it is aggregated with EDU 3 to form a larger span 1-3. To determine the promotion set of span 1-3, one first identifies its nucleus, i.e. span 1-2, and takes the promotion set of this span.

The notion of strong nuclearity hypothesis is employed in later studies (Demberg et al., 2019), which is a related work in Chapter 3 and will be discussed in greater detail in section 2.7.

2.6.1.5 Controversy over Nuclearity

The strong nuclearity hypothesis is even questioned by Marcu (1998) himself when nuclei are used for summarization. Marcu (1998) shows that the nuclei alone are insufficient to create high-quality summaries and importance weighting based on tree-depth and span adjacency is needed. For some relations, satellites are as important as nuclei, and for some multi-nuclear relations, only the first nucleus is important. In some cases, the satellites are considered more important than nuclei by human judges in the extractive summarization task.

Similarly, Stede (2008b) discusses problematic cases with the notion of nuclearity. He points out that nuclearity assignment is determined not only by semantic salience but also by referential and/or thematic continuity, because when a clause introduces an entity that becomes a referent in the next clause, it is reasonable for the clause to be assigned a nucleus status. As a result, although some relations tend to have a canonical order of nucleus and satellite, this is often only a pattern, and when the linear order of textual spans is changed, because of the need for

thematic continuity, the result of nuclearity assignment may be different for the same relation. Moreover, if nuclearity assignment is deemed a must in choosing a relation, when more than one relation is possible, the annotators may choose the one whose nucleus assignment makes it easy to build the tree structure, rather than the most suitable coherence relation. Based on established observations that coherence is created from multiple sources of information (Grosz and Sidner, 1986), Stede (2008b) proposes multi-layer annotation so that a comprehensive picture of discourse can be obtained for further studies.

2.6.1.6 Controversy over Single Layer Representation

As discussed above, Stede (2008b) represents one of those proponents of multi-layer discourse annotation to ameliorate the issues of RST analysis. When working on a project on creating a dialogue system, Moore and Pollack (1992) find that generating answers to follow-up questions cannot be sufficiently addressed by RST representations. They take issue with the single tree representation postulated in RST and suggest simultaneous representation of informational and intentional relations, which correspond to subject matter relations and presentational relations in RST, respectively. However, there is no fixed one-to-one mapping between these two types of information, and structures at the two levels are isomorphic. These reasons make simple fixes to RST, such as representing each relation with two types of relation labels, futile. However, as indicated in Stede (2008b), the assumption of coexistence of two types of information is too strong, and under-specification for one type of information may be needed when another type of information is clearly more pronounced. Furthermore, Stede (2008a) points out that systematic ambiguity at these two levels does not occur so frequently as to warrant full annotation at both levels.

When tackling a task of generating descriptive texts that are characterized by a sequence of entities introduced one after another and elaborated on later[5], Knott

[5]An example is given in Knott et al. (2000):

- (1) In the women's quarters the business of running the household took place.
- (2) Much of the furniture was made up of chests arranged vertically in matching pairs (. . .).
- (3) Female guests were entertained in these rooms, which often had beautifully crafted wooden toilet boxes with fold-away mirrors and sewing boxes, and folding screens, painted with birds and flowers.
- (4) Chests were used for the storage of clothes . . . .

The "chests" are introduced in (2) but elaborated on in (4).

et al. (2000) identify some cases where the properties of RST analysis, i.e. compositionality, continuous constituency and single-tree representation, cannot be supported. They attribute the problem to the *elaboration* relation, which is essentially an entity-based relation and does not allow any embedding. They find that the special case with *elaboration* can be accounted for by an entity-chain model interleaved with RST trees minus the *elaboration* relation, where the entity-chains represent the global focus structure and RST trees encode local rhetorical relations. Figure 2.5 shows the model.

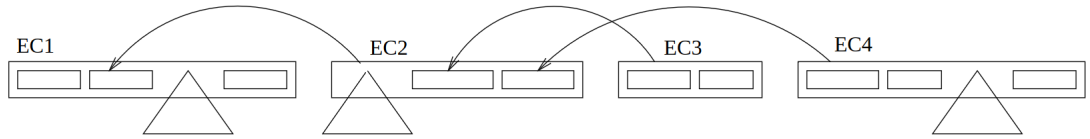


Figure 2.5: The RST model based on entity-chains proposed by Knott et al. (2000). EC*n* denotes the entity-chains. The rectangles represent atomic RST trees and the triangles represent non-atomic RST trees. If an entity is introduced first and elaborated on later, a directed arc is used to show this relationship. These arcs do not have to link adjacent segments, as shown by the arc linking EC4 and EC2, and edge crossings are allowed.

Different from Stede (2008b) and Moore and Pollack (1992), the approach taken by Knott et al. (2000) does not assume the simultaneous effects of multiple sources of information. Another distinctive feature of the model proposed by Knott et al. (2000) is that coherence at higher levels is achieved by reiteration of the proposition in focus, or normalization of the proposition, or discourse deixis, diverging from RST, which posits that the same set of relations can be applied to textual spans of arbitrary sizes.

2.6.1.7 Controversy over Tree Structure

If there is a need for multi-layer annotation, it may suggest that a tree structure is inadequate for discourse representation. Webber (2006) and Knott et al. (2000) suggest combining trees with other mechanisms, allowing some non-tree structures to be accommodated.

Wolf and Gibson (2005) propose using a chain graph for discourse representation. In this structure, directed arcs represent asymmetric discourse relations and undirected arcs represent symmetric relations. They annotate a corpus with texts collected from the Wall Street Journal (WSJ) and the AP Newswire, with a total of 135 texts, known as the Discourse Graphbank and released on Linguistic

Data Consortium[6]. They choose clauses as basic discourse units. Adjacent basic discourse units are allowed to be grouped based on two criteria: having the same attribution source, and being about the same topic. These groupings allow hierarchical structures to be formed, but compared with RST, the hierarchical structures are limited since only specific kinds of groupings are allowed.

In their experiments, the relation set proposed in Hobbs et al. (1985) is adopted, with some modification, and the relations include asymmetric relations, such as *cause–effect*, *condition*, *violated expectation*, *elaboration*, *example*, *generalization*, *attribution*, and *temporal sequence*, and symmetric relations, including *similarity*, *contrast* and *same*. When determining a relation, a procedure of explicitating relations is adopted to facilitate the process, and if no conjunctions can be found to explicitate relations, it is considered that no discourse relations are applicable. This procedure restricts the inferences that can be made. In the created corpus, they find a large number of cross-dependencies and multi-parent nodes, which, they believe, provide strong evidence that trees are not adequate for discourse representation. They examine the relations involved in these cases, and identify *elaboration* and *similarity* as the major sources of violations of tree constraints. However, they argue that if these relations are removed from the corpus, the representation is impoverished.

The evidence against using trees for discourse representation in Wolf and Gibson (2005) is questioned by Egg and Redeker (2010) and in the discussion by Marcu (2003). They argue that the *elaboration* relation in Wolf and Gibson (2005) is essentially not a coherence relation but a lexical cohesive device, which is not necessary in the representation of discourse structure, and the multi-parent structures can be done away with the strong nuclearity hypothesis in the text structuring processes. Webber (2006) attributes the case with the *elaboration* relation to a different type of discourse, i.e. coherence created from anaphoric dependency. Her account with respect to the multi-parent structures suggests that the extra complexity introduced in the model by Wolf and Gibson (2005) is not necessary.

Lee et al. (2006) present some non-tree constructions when building PDTB (Prasad et al., 2008), which will be discussed in section 2.6.2. These constructions include shared arguments, properly contained arguments, pure crossings, and partially overlapping arguments, shown in (a), (b), (c) and (d), respectively, in Figure 2.6. The structures in (a), (b) and (d) contain nodes with more than one parent and (c)

[6] `https://catalog.ldc.upenn.edu/LDC2005T08`

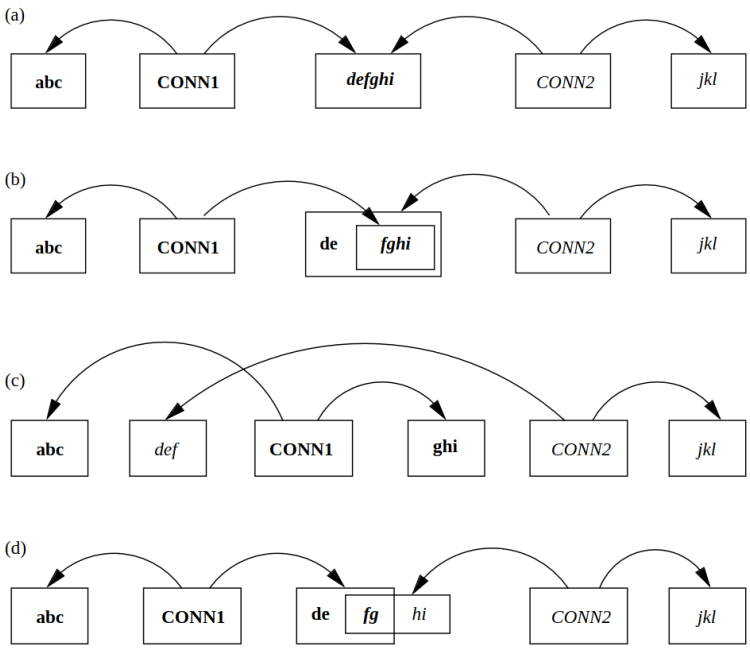


Figure 2.6: Non-tree like constructions in Lee et al. (2006). "CONN" means "connective".

shows a case of edge crossing. These structures depart from tree constraints and cannot be represented by trees.

However, they find that only shared argument structure and a subset of properly contained argument structures are valid discourse structures, while the rest can be attributed to anaphoric dependency or attribution relations, which belong to a different type of discourse information. Their approach is questioned by Egg and Redeker (2010), who suggest that the discourse framework employed by them only focuses on individual local coherence relations and it is not valid to analyze discourse structure by juxtaposing such independently annotated structures.

2.6.1.8 RST-DT

The best known English corpus annotated according to RST is RST-DT. The corpus contains 385 Wall Street Journal articles that are also included in the Penn Treebank. The corpus annotation process is presented in Carlson et al. (2001). The original theoretical formulation in Mann and Thompson (1988) specifies that no syntactic or morphological clues can be used. However, to achieve high annotation consistency, Carlson et al. (2001) allow the use of syntactic clues and adopt clauses as the basic discourse units, with the exception of clauses that act as subjects, objects or complements of main verbs. Moreover, a relation *same-unit* is used to link discourse segments that are separated by intervening segments. A few

discourse markers that introduce phrasal clauses, such as "because of", "owing to", and "as a result of", are considered as signals introducing discourse segments.

A protocol is created to help the annotators to make choices in ambiguous cases. The protocol defines an order of relations so that annotators can choose the most specific one when more than one relation is possible for connecting two segments (Marcu et al., 1999). Recall that for an RST tree, the leaves are EDUs and they are recursively aggregated to form greater spans, where only adjacent EDUs can be aggregated and no edge-crossings are allowed. The final tree forms a full-coverage of the text. Each node is characterized by a nuclearity status and the relations linking the segments can be mononuclear or multinuclear, with the former referring to relations that hold between two segments characterized by a nucleus and a satellite and the latter denoting relations that hold between multiple segments of equal salience.

There are 53 mononuclear and 25 multinuclear relations in RST-DT. These 78 relations can be grouped into 16 classes based on semantic similarity. The grouping improves inter-annotator agreement considerably and is typically employed in later studies for developing computational systems. Figure 2.7 shows the relation set used in RST-DT. Based on Carlson et al. (2001), the relations can be intentional, semantic or textual.

It is mentioned in Carlson et al. (2001) that nuclearity assignment and relation selection are straightforward at the inter-clausal level, but are challenging at the inter-sentential level and rather difficult when large spans are linked. Stede (2008a) also indicates difficulties in determining span boundaries higher up the tree when they annotate the Potsdam Commentary Corpus (PCC) (Stede, 2004), which suggests that it is difficult to define the scopes of relations at higher levels of the tree clearly. Therefore, it is questionable whether the relations at higher levels of RST trees can be reliably annotated and whether they are semantically meaningful, except for their function in maintaining a connected tree structure.

Another property of the corpus is the tendency of sentence-boundedness (Fu, 2022). Although RST does not enforce well-formed discourse sub-trees at the sentence level, it is found that 95% of the discourse trees in RST-DT have sentence-level well-formed sub-trees (Soricut and Marcu, 2003). This observation forms the basis of the research by Joty et al. (2015) and Soricut and Marcu (2003). Zhang et al. (2021a) and Guz et al. (2020) explicitly model this property in developing RST

Classes	Representative Members
Attribution	attribution, attribution-negative
Background	background, circumstance
Cause	cause, result, consequence
Comparison	comparison, preference, analogy, proportion
Condition	condition, hypothetical, contingency, otherwise
Contrast	contrast, concession, antithesis
Elaboration	elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
Enablement	purpose, enablement
Evaluation	evaluation, interpretation, conclusion, comment
Explanation	evidence, explanation-argumentative, reason
Joint	list, disjunction
Manner-Means	manner, means
Topic-Comment	problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
Summary	summary, restatement
Temporal	temporal-before, temporal-after, temporal-same-time, sequence, inverted- sequence
Topic-Change	topic-shift, topic-drift

Figure 2.7: The relation set used in RST-DT (Carlson and Marcu, 2001).

discourse parsers.

2.6.1.9 Other RST Discourse Corpora

The Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017b) features multiple layers of annotation, where one layer focuses on RST discourse annotation. The texts are selected from four genres: interviews, news articles, instructional texts, and travel guides. In determining the relation set, Zeldes (2017b) refers to relation taxonomies adopted in existing discourse corpora, discarding less frequent relations and adopting more general types when some relations are similar to each other, for instance, multi-nuclear relations including

list, *joint*, *conjunction* and *disjunction* are merged under the general *joint* relation to ease the annotation process. 20 relations in total are used. As annotators are students, the corpus does not include Wall Street Journal articles like RST and PDTB, which may be challenging for them to finish within limited time. This corpus represents a multi-genre English discourse corpus annotated based on RST and it is still growing.

The Georgetown Chinese Discourse Treebank (GCDT) (Peng et al., 2022) is a Chinese discourse corpus annotated in the RST framework. It follows the design of GUM and contains texts of different genres. The focus of this corpus is discourse annotation of documents of medium lengths and long documents. Hence, the corpus is still small, with 50 documents evenly distributed in five genres.

Similar to GUM, PCC is also a multi-layer corpus, but the texts in the corpus only cover German newspaper commentaries. In Stede (2004), no formal annotation guidelines are developed for the annotation of rhetorical structures, because such analysis in German suffers from some open problems, such as ambiguities in span boundary determination and relation selection, and the author treats this annotation process as a qualitative study of these problems. In Stede and Neumann (2014), connectives and arguments are annotated additionally, following the annotation of explicit relations in PDTB. However, implicit relations are not covered, and the connectives are restricted to a predefined set, and sense labels are not attached. Overall, the corpus features multi-layer annotation of a total of 176 German newspaper commentaries, which are short texts with a typical length of 8 to 10 sentences[7].

Redeker et al. (2012) annotate a discourse corpus of 80 Dutch texts from expository and persuasive genres. The corpus is also multi-layer, containing annotations of RST structure, genre analysis and lexical cohesion analysis.

Corpora that feature RST-style annotation also exist for other languages, such as Basque (Iruskieta et al., 2013), Bangla (Das and Stede, 2018), Russian (Toldova et al., 2017), Spanish (da Cunha et al., 2011), and so on.

2.6.1.10 Extensions of RST

Veins Theory The Veins Theory is proposed by Cristea et al. (1998), which combines RST with the Centering Theory (Grosz et al., 1995), so that the Centering

[7] `https://angcl.ling.uni-potsdam.de/resources/pcc.html`

Theory can be applied at the global level. The model defines the scope of referential accessibility for each discourse segment based on the nuclei of RST trees. It restricts the referential accessibility of entities so that the entities in two nuclei that are in a structural relation can refer to each other, and satellites can refer to entities within their respective nuclei. In this way, reference is not limited within a single discourse segment, and long-distance reference is possible. Different from the Centering Theory, which specifies the preferred movement of centers to create smooth transitions in a sequential order of utterances, the Veins Theory assigns scores to different types of center movement across referential accessibility domains. In Cristea et al. (1998), discourse relation types are considered irrelevant, while Zeldes (2017a) shows the influence of relation types on domains of referential accessibility.

RST Signaling Corpus Das and Taboada (2018) focus on how discourse relations are signaled. Apart from discourse markers, they assemble other possible signals from existing studies, such as referential, lexical, semantic, syntactic, graphical and genre features. These signals are organized into a three-level hierarchy. The top level contains three approaches of signalling: *single*, *combined* and *unsure*, where *single* means that a relation is signaled by only one feature, *combined* means that a relation is signaled by two features in combination, one feature being an independent signal and the second feature being dependent on it, and *unsure* indicates that no signals can be identified. Under each class, more specific types can be identified, for instance, the *single* class has nine types: *discourse markers*, *lexical*, *morphological*, *syntactic*, *semantic*, *genre*, *graphical*, *numerical* and *reference*. Furthermore, for each type, some more specific features can be discriminated. For example, *reference* can be categorized into *personal*, *demonstrative*, *comparative* and *propositional* reference, which can be traced to the lexicon-based discourse model by Halliday and Hasan (1976) (section 2.5). Zeldes and Liu (2020) propose an automatic approach for quantifying the signaling strength of instances.

RST and Grosz & Sidner's Model Moser and Moore (1996) investigate the relationship between RST and Grosz & Sidner's Model. They argue that the dominance relation of DSPs (note: discourse segment purposes) in the model proposed by Grosz and Sidner (1986) is similar to the notion of nuclearity in RST. Although RST does not claim explicitly that nuclearity encodes intentionality, a nucleus expresses an action or a belief that the speaker who initiates the discourse intends the hearer to take or adopt, and a satellite is intended to facilitate the

achievement of the purpose. Based on this observation, as nuclearity represents a property of intentional relations, Moser and Moore (1996) propose removing nuclearity in informational relations in RST to resolve the issue indicated in Moore and Pollack (1992) (discussed in section 2.6.1.6), where two types of discourse information, informational and intentional, compete with each other when only one relation is represented in the annotation. Furthermore, they point out that RST and Grosz & Sidner's model both make some claims on the orderings of segments. In RST, some relations tend to have a canonical order in terms of nucleus-satellite combination, while in Grosz & Sidner's model, the satisfaction-precedence relation specifies an ordering of segments but no claim is made about the ordering of segments for dominance relations. In addition, Moser and Moore (1996) indicate that the attentional structure in Grosz & Sidner's model constrains referential accessibility for pronouns, while RST does not give any account of this structure.

2.6.1.11 RST Parsing

Bottom-Up Approach For a text D formed by a sequence of tokens t_0, t_1, ..., t_n, the bottom-up approach of RST parsing includes the following tasks:

1. EDU segmentation: segment the token sequence t_0, t_1, ..., t_n into EDUs, i.e. EDU_0, EDU_1, ..., EDU_m

2. parsing: link EDU pairs in the set formed by EDU_0, EDU_1, ..., EDU_m and build trees recursively from the bottom up.

3. nuclearity assignment: For each of the EDU/(intermediate) span pairs determined, assign a nuclearity status to each component. Possible combinations include N + S, S + N, and N + N.

4. relation attachment: determine which relation holds between each EDU/span pair from a predefined relation set $R = \{r_0, r_1, ..., r_z\}$.

Marcu (1999) develops the first transition-based RST parser. The step of EDU segmentation is separated from the rest, and EDU segmentation is performed with a binary classifier to identify EDU boundaries. For the parsing step, the input is an empty stack and a list of elementary discourse trees, one elementary discourse tree for each EDU. The relation, nuclearity status and promotion set of the elementary discourse trees are initialized first. The discourse parsing

process is characterized by a sequence of SHIFT-REDUCE operations. By a SHIFT operation, the next elementary discourse tree is pushed onto the top of the stack and a REDUCE operation replaces the top two discourse trees by a single tree. Combining discourse trees modifies the values of the relations, nuclearity statuses and promotion sets of the trees. There are six types of REDUCE operation: NS, NS-below, SN, SN-below, NN, NN-below, which means that for each possible nuclearity status (task 3 above), there are two possible ways of attaching the discourse tree at the top of the stack to the discourse tree immediately below it. If the two discourse trees are to be combined as child nodes of a new binary tree, NS, SN and NN REDUCE operations will be applied, and if the discourse tree at the top is to be attached to the one below it as a child node, creating a non-binary tree, the NS-below, SN-below and NN-below REDUCE operations will be applied. To mitigate data sparsity, Marcu (1999) clusters the relation labels into 17 groups, and therefore, there are 17*6+1 operations, where 1 refers to the SHIFT operation.

Soricut and Marcu (2003) incorporate syntactic information into EDU segmentation, and employ syntactic and lexical features for identifying links between EDUs in the parsing step. Their model is based on two components, one for computing structural probabilities and one for computing probabilities of combinations of relations and nuclearity status. A dynamic programming algorithm is used to find the most probable parse tree. Their method improves the method shown in Marcu (1999) to a large margin. Sagae (2009) applies a transition-based constituent syntactic parsing algorithm to RST parsing, and their model takes a whole document for processing, and this method yields even higher performance. Hernault et al. (2010) use support vector machines (SVMs) (Cortes, 1995) for discourse segmentation and relation labelling. Similarly, they employ a binary classifier for discourse segmentation. They adopt a binary classifier for structural labelling and a multi-class classifier for determining relation and nuclearity combinations. Discourse tree construction is performed with a recursive bottom up process, where all pairs of consecutive elements are read left to right, and the pair with the highest probability is chosen and treated as child nodes of a subtree. This process is repeated until only one element is left, which is the discourse parse tree for the whole document. Similar to Marcu (1999), Subba and Di Eugenio (2009) employ a shift-reduce parser for RST parsing, but they adopt the RFC principle for determining the possible attachment points of an incoming segment. To incorporate more information for improving the accuracy of relation identification, they utilize inductive logic programming based on

first-order logic to learn rules from a variety of sources of information, such as compositional semantics, structural information, WordNet, and linguistic cues for relations. Feng and Hirst (2012) adopt the approach proposed by Hernault et al. (2010) but refine the features. Joty et al. (2013) argue that previous approaches ignore sequential dependencies between the constituents of discourse trees, and they use a conditional random field (CRF) model (Lafferty et al., 2001) as a remedy. They also notice that discourse relations are distributed differently at intra-sentential and inter-sentential levels, and implement two parsers to handle parsing at the two separate levels. Two ways of combining the results at the two levels are investigated. To find the best discourse tree, they implement a CKY-like bottom-up algorithm based on dynamic programming. Their model achieves notable performance gains over the baseline model by Hernault et al. (2010). Ji and Eisenstein (2014) propose a method for learning a lower-dimensional projection of input features for discourse parsing. To learn a classifier for shift-reduce operations, they adopt a max-margin learning objective.

Wang et al. (2017) believe that in transition-based algorithms, if an action encodes too much information simultaneously, i.e. Relation-(Shift/Reduce)-Nuclearity, data sparsity is likely to occur. Therefore, they adopt a two-stage pipeline approach. First, they use a transition-based system to construct naked discourse trees, without predicting relations. Similar to Joty et al. (2013) and Feng and Hirst (2012), they distinguish different levels of spans in relation classification. Apart from intra-sentential and inter-sentential relation classifiers, they include a paragraph-level relation classifier. Post-order traversal of the naked trees is performed and for each internal node, the system determines if its left and right subtrees are in the same sentence or not and in the same paragraph or not. The action classifier of the transition-based system and three relation classifiers are all based on SVMs.

Li et al. (2014) propose to convert hierarchical RST tree structures into dependency structures to reduce structural complexity. For a mononuclear relation, the nucleus is treated as the head, and the satellite is the dependent, and for a multinuclear relation, the leftmost nucleus is treated as the head. They apply graph-based syntactic dependency parsing techniques to discourse dependency parsing. Hirao et al. (2013) propose another method for converting RST trees into dependency structures, and their method differs from that of Li et al. (2014) in the handling of multinuclear relations. Morey et al. (2018) address the issue of ambiguity in

converting RST trees into dependency structures. Zhang et al. (2021a) adapt unsupervised syntactic dependency parsing methods to discourse dependency parsing. Following Wang et al. (2017), Zhou and Feng (2022) take a two-stage pipeline approach to discourse dependency parsing. They apply a transition-based system for discourse tree construction in the first step. Different from previous studies, they use contextualized embeddings and obtain different representations at intra-sentential and inter-sentential levels. At the step of relation classification, they use sequence labelling to incorporate greater context.

Top-Down Approach Bottom-up parsers tend to make local greedy choices in the tree construction process. There has been a move towards top-down approaches for RST parsing in recent years.

Koto et al. (2021) adopt a sequence labelling approach to top-down RST parsing. Given a sequence of EDUs, a sequence labelling model determines where to split the EDU sequence, and this process is performed iteratively, until a single binary RST tree is formed. As errors close to the root have greater influence on tree construction than errors close to the leaves, they introduce a penalty factor inversely proportional to tree depth. Moreover, to prevent errors in the segmentation step from influencing the following steps, they adopt a dynamic oracle, which compares the current structure with the gold structure and decides the next step on this basis to minimize deviation from the gold structure. Kobayashi et al. (2020) employ a deep biaffine attention model (Dozat and Manning, 2018) to learn the splitting points, but different from Koto et al. (2021), they apply the top-down splitting procedure at three granularity levels, and merge the outputs at the three levels into one parse tree.

RST Parsing with Transformer-Based Language Models Inspired by the training mechanism of BERT, Yu et al. (2022) propose an EDU-level pre-training stage to enhance RST parsing, which includes two objectives: next EDU prediction (NEP), and discourse marker prediction (DMP). For NEP, they sample continuous EDUs as positive examples and non-adjacent EDUs as negative examples and train a model to predict which are positive examples and which are negative examples. For DMP, they mask discourse connectives of EDUs and train the model to predict them.

Xiao et al. (2021), Huber and Carenini (2019) and Huber and Carenini (2022) explore extracting RST trees from attention matrices of transformer-based lan-

guage models fine-tuned on tasks that require high-level understanding, such as summarization. Built on this line of research, Huber and Carenini (2020) present an RST-style discourse corpus with nuclearity annotated but without discourse relation annotation, created using distant supervision from sentiment analysis. Using distant supervision from another NLP task for RST parsing forms an approach to addressing the data shortage problem of RST parsing (Huber et al., 2022).

Shared Task The Discourse Relation Parsing and Treebanking (DISRPT) Shared Task (Braud et al., 2023; Zeldes et al., 2021, 2019) includes EDU segmentation, connective detection, and relation classification tasks, with a focus on evaluating systems developed for these separate tasks across different languages, where differences in discourse frameworks are marginalized.

2.6.2 PDTB

PDTB is another mainstream discourse framework that has been widely used in discourse annotation projects and discourse parsing. The theoretical foundation of PDTB is the lexicalized Tree Adjoining Grammar for discourse (D-LTAG) proposed by Webber (2004).

D-LTAG is proposed as a way to extend syntactic processing to discourse. Specifically, it applies the lexicalized Tree Adjoining Grammar (LTAG) (Joshi, 1987) to discourse modelling. In an LTAG, syntactic structures are tied to specific words, for instance, the word "enjoy" can have a set of tree structures that feature a predicate of "enjoy" connecting two arguments, "enjoyer" and "enjoyee". The word "like" has a similar structure as "enjoy" but it can also function in a prepositional phrase to modify a verb or noun, such as "work **like** a horse" or "a teacher **like** Jane", which is discussed in Webber (2004).

A Tree Adjoining Grammar (TAG) can be defined as: $G = (N, T, I, A, S)$, where N is a finite set of non-terminal symbols, T is a finite set of terminal symbols, I denotes a finite set of initial trees, A denotes a finite set of auxiliary trees, and S is a distinguished non-terminal symbol (Joshi and Schabes, 1991). The union of I and A forms a set of elementary trees.

For both initial trees and auxiliary trees, interior nodes are labeled by non-terminal symbols, and frontier nodes can be labeled by terminal or non-terminal symbols. The major difference between the two types of elementary trees is that

for initial trees, non-terminal symbols on the frontier are marked for substitution, conventionally denoted by a ↓ symbol, and for auxiliary trees, non-terminal symbols on the frontier are marked for substitution, with a node as an exception, whose label is identical to that of the root node, and the node is called *foot* node, which is denoted by a ∗ symbol by convention. A lexicalized TAG means that at least one terminal symbol must appear on the frontier of all the elementary trees. In other words, each lexical item is associated with a finite number of structures for which the lexical item is the *anchor* (Joshi and Schabes, 1991).

Figure 2.8 illustrates elementary trees for the word "like". Trees shown by (a), (b) and (c) are initial trees, whose root nodes are labeled *S*, and trees shown by (d) and (e) are auxiliary trees, whose root nodes are labeled *NP* and *VP* respectively. The linguistic realizations are (a) "liker" **like** "likee"; (b) "likee" "liker" **like**; (c) "likee" **like** by "liker"; (d) NP **like** NP; (e) VP **like** NP.

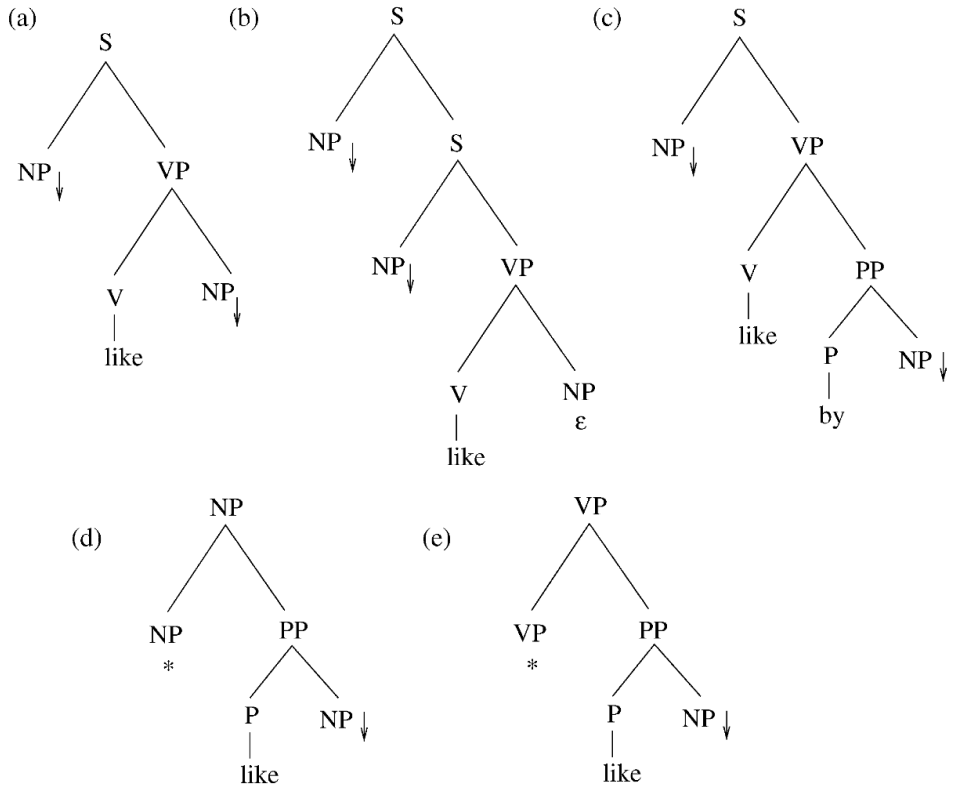


Figure 2.8: Initial and auxiliary trees that "like" can appear in, originally from Webber (2004).

With two operations, *substitution* and *adjoining*, elementary trees can form derived trees. A substitution operation refers to substituting a non-terminal on the frontier of a tree, such as the two NP↓*s* in Figure 2.8 (a). Given the initial tree shown in (a),

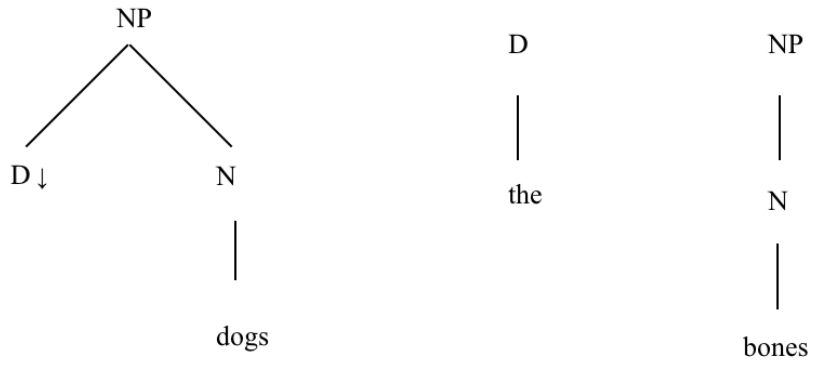


Figure 2.9: Three more trees to form a complete parse tree for the sentence "The dogs like bones".

with three more trees shown in Figure 2.9 (from left to right: $\alpha 1$, $\alpha 2$, $\alpha 3$), a parse tree can be derived for the sentence "The dogs like bones" with the following steps:

- substitute D in $\alpha 1$ with $\alpha 2$, yielding a derived tree $\gamma 1$;
- substitute the leftmost NP↓ in (a) with $\gamma 1$;
- substitute the rightmost NP↓ with $\alpha 3$.

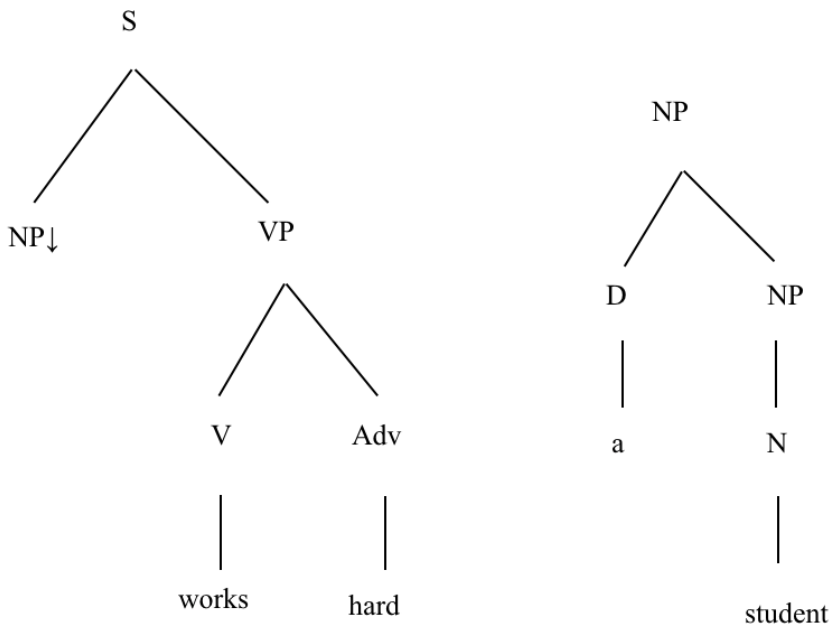


Figure 2.10: Additional trees involved in the adjoining operation.

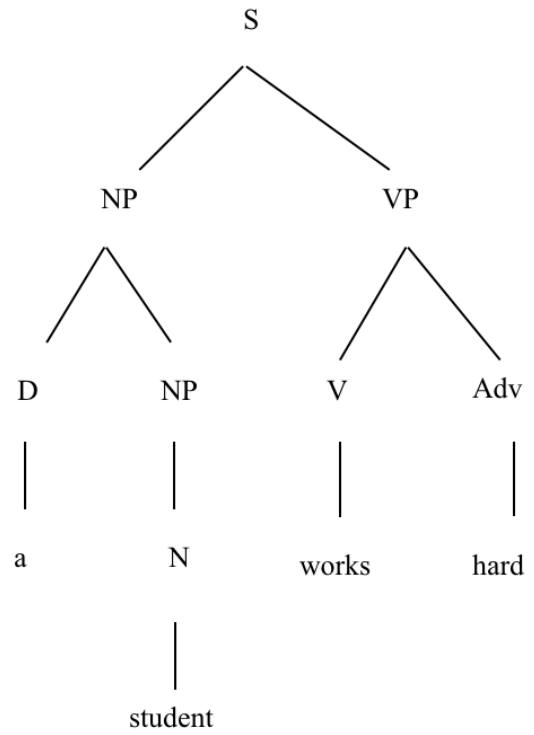


Figure 2.11: Illustration of tree γ.

An adjoining operation involves an auxiliary tree and an elementary or derived tree. Take the auxiliary tree (β) shown in Figure 2.8 (d) for example. Additional trees (left: $\alpha 1$, right: $\alpha 2$) are given in Figure 2.10 for illustration. An adjoining operation cannot be applied to a node that is marked for substitution, such as the node NP indicated by $\downarrow$ in $\alpha 1$. $\alpha 2$ can substitute this node, forming a new tree γ, which is shown in Figure 2.11. As the root of the auxiliary tree β has the label NP, it can replace the node with the same label on tree γ, yielding the tree shown in Figure 2.12.

An LTAG requires a variety of elementary trees to describe possible structures for different words. In comparison, a D-LTAG only treats conjunctions, discourse adverbials and adjacency as anchors. Owing to the limited categories of anchors, it requires only a few elementary trees. For D-LTAG, initial trees can be divided into four types. The first type concerns subordinating conjunctions, such as "since" and "when", and subordinators, which are lexical items that introduce a clause, such as "in order to", "to..." (to introduce a purpose), or "by..." (to show the means).

The possible structures are shown in Figure 2.13, where D_c denotes a discourse

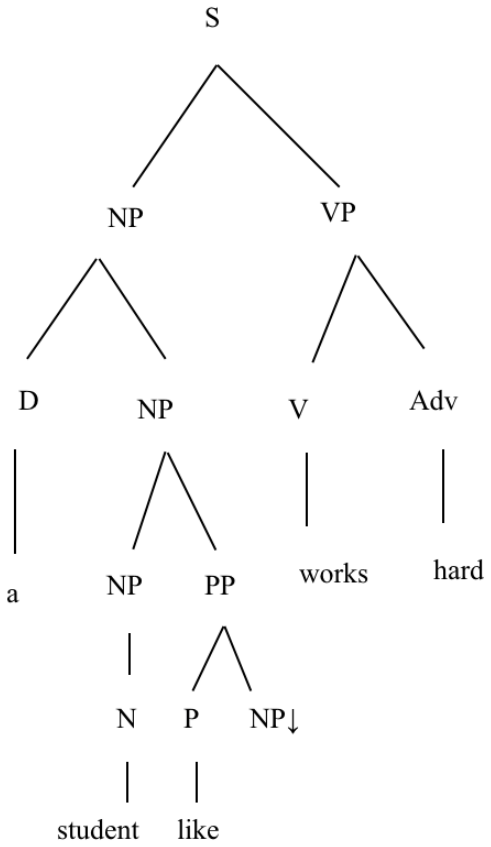


Figure 2.12: Tree formed with the adjoining operation.

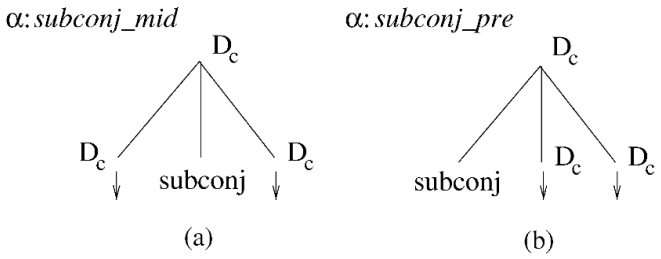


Figure 2.13: Initial trees with subordinating conjunctions as anchors, originally from Webber (2004).

clause, as in Webber (2004). In (a), they appear in the middle of two arguments, for example, "I was not caught in the rain, **because** I had an umbrella with me". In (b), they appear before the first argument, for example, "**When** it rains, the area could be flooded".

The second type of initial trees are characterized by using parallel constructions as anchors, such as "not only...but also...". Figure 2.14 shows an example using "on

the one hand...on the other hand..." as an anchor to convey a *contrast* relation.

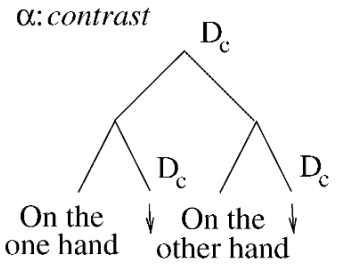


Figure 2.14: Initial trees with parallel constructions as anchors, originally from Webber (2004).

The third type of initial trees are anchored by coordinating conjunctions. Figure 2.15 shows an example with "so" as an anchor to convey a *result* relation.

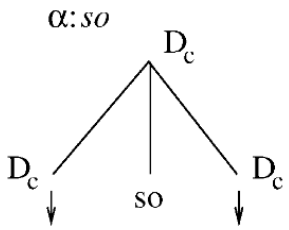


Figure 2.15: Initial trees with coordinating conjunctions as anchors, originally from Webber (2004).

Some verbs form special cases, such as the imperative "suppose". While these verbs might anchor auxiliary trees in LTAG, as shown in Figure 2.16 (a), they are anchors of initial trees in D-LTAG, as in Figure 2.16 (b). An example is: "**Suppose** you leave early$_{D_{c1}}$, you might be able to avoid a traffic jam$_{D_{c2}}$".

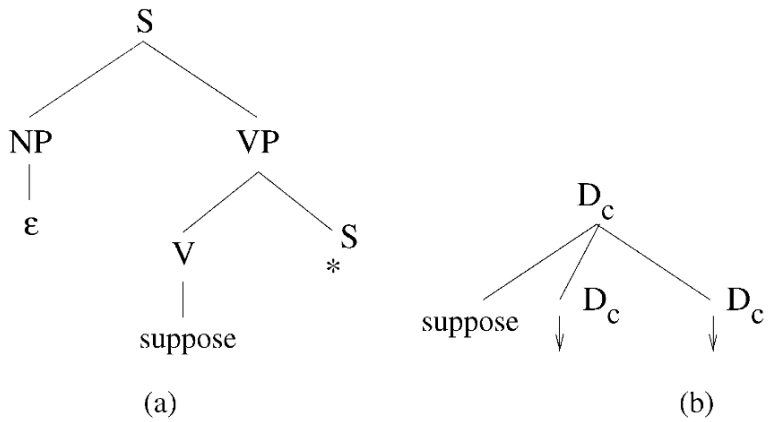


Figure 2.16: The case with the imperative "suppose", originally from Webber (2004).

Recall that auxiliary trees are used to derive more complicated structures through adjoining operations. In D-LTAG, they are used to accommodate cases where descriptions of objects, events, situations and states extend over several clauses. Figure 2.17 shows an auxiliary tree anchored by a coordinating conjunction.

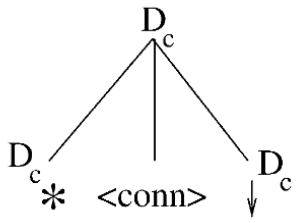


Figure 2.17: An auxiliary tree anchored by a coordinating conjunction, originally from Webber (2004).

Discourse adverbials, such as "instead" and "in contrast" are also taken as anchors of auxiliary trees. However, this type of auxiliary trees requires one discourse clause, as shown in Figure 2.18.

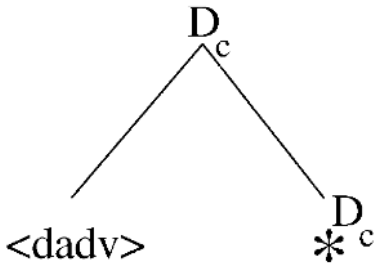


Figure 2.18: An auxiliary tree anchored by a discourse adverbial, originally from Webber (2004).

2.6.2.1 Corpora

PDTB is based on the idea that discourse relations are anchored by sentence adjacency or a finite set of discourse connectives, which are generally easy to identify. This relatively simple conceptualization lends itself to large-scale discourse annotation (Prasad et al., 2018). PDTB 3.0, which is the newest version, contains 53631 tokens of annotated relations for 2162 Wall Stree Journal articles (PDTB 2.0 contains 40600 tokens) (Webber et al., 2019). Similar to RST-DT and the Discourse Graphbank, it is distributed through the Linguistic Data Consortium[8].

[8] https://catalog.ldc.upenn.edu/LDC2019T05

PDTB and RST-DT are both built on texts chosen from the Wall Street Journal section of the Penn Treebank. The corresponding PDTB annotations for the example of wsj_0624 shown in Figure 2.2 are:

1. *the agreement "an important step forward in the strengthened debt strategy"*, **that it will "when implemented, provide significant reduction in the level of debt and debt service owed by Costa Rica."** (implicit, given, Contingency.Cause.Reason)

2. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica.*, **implemented,** (explicit, when, Temporal.Asynchronous.Succession)

3. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica.*, **implemented,** (explicit, when, Contingency.Cause.Reason)

As can be seen from this example, the annotations are independent from each other, and no higher-level structural constraints are imposed. Thus, PDTB-style annotation is shallow discourse annotation, focusing on local semantic/pragmatic relations and not all the textual elements are covered.

Argument Identification By convention, Argument 1 (Arg1) is shown in italics and Argument 2 (Arg2) is in bold. As explicit relations involve lexical anchors that can be identified first, the arguments of explicit relations do not have to be adjacent or continuous (Prasad et al., 2008). The argument spans of explicit relations are determined based on the *Minimality Principle*[9], which means that as many clauses and/or sentences should be considered as part of an argument as are minimally required but necessary for understanding the relation. As shown in the first annotation above, Arg1, i.e. "the agreement "an important step forward in the strengthened debt strategy", is taken from the original text "Treasury Secretary Nicholas Brady called the agreement 'an important step forward in the strengthened debt strategy' " and the part "Treasury Secretary Nicholas Brady called" is not covered in Arg1 because it does not contribute to the *reason* relation here. If the remaining part is relevant, it can be annotated as supplementary information to Arg1 or Arg2. Since implicit relations are anchored by positional adjacency between two sentences or clauses, the arguments are naturally the adjacent sentences or clauses.

[9]See the annotation manual of PDTB 2.0 (Prasad et al., 2008) (**https://www.cis.upenn.edu/~elenimi/pdtb-manual.pdf**).

A set of rules are defined for the assignment of Arg1 and Arg2. In PDTB 2.0, Arg2 is the argument that is syntactically related to the connective and the other one is Arg1. For implicit relations, Arg1 goes before Arg2, by linear order in the text. In PDTB 2.0, implicit relations are annotated mainly between adjacent sentences, but in PDTB 3.0, a large number of intra-sentential implicit relations are annotated. To accommodate new types of intra-sentential relations annotated in PDTB 3.0, the convention of assigning Arg1 and Arg2 is changed. For inter-sentential relations, the position of the arguments is used as the criterion, by which Arg1 is the first argument followed by Arg2. The same rule applies to intra-sentential coordinating relations. For intra-sentential subordinating relations, Arg2 is the argument that is subordinate to the other argument (Arg1).

Relation Types The annotations inside the brackets represent relation types, connectives, and sense labels, respectively. An implicit relation holds when no conjunctions or discourse adverbials that link Arg1 and Arg2 can be found in the text, and the annotators have to rely on inference to determine a sense label, as in the first annotation. The connective "given" is inserted by annotators as they think that the connective can be used to link the two arguments and explicitate the relation. When no such connectives can be inserted, there are three labels to consider: *AltLex*, *EntRel* and *NoRel*.

AltLex is applied when adding a connective makes the expressed relation redundant, as in the following example:

Profit from continuing operations has soared to $467 million from $75 million. **Mr. Hahn attributes the gains to the philosophy of concentrating on what a company knows best.** (wsj_0100, AltLex, Contingency.Cause.Reason)

Since the expression "attributes the gains to..." already implies the relation, additional connectives would cause redundancy.

EntRel is used when an entity-based relation is inferred:

An exhibition of American design and architecture opened in September in Moscow and will travel to eight other Soviet cities. **The show runs the gamut, from a blender to chairs to a model of the Citicorp building.** (wsj_0102, EntRel)

In this example, Arg2 is connected with Arg1 by the entity *the show*, which refers to *an exhibition of American design and architecture*.

When neither discourse relations or entity-based relations can be inferred, NoRel is applied:

In 1989, home purchase plans have ranged monthly from 2.9% to 3.7% of respondents. **In October, 30.6% said they will buy appliances in the coming six months, compared with 27.4% in September and 26.5% in October 1988.** (wsj_0141, NoRel)

In PDTB 3.0, two new relation types are added: *AltLexC* and *Hypophora*. AltLexC is used when lexico-syntactic expressions can be identified as signals of discourse relations:

His wife also works for the paper, **as did his father.** (wsj_0114, AltLexC, Comparison.Similarity)

The construction "as did" is treated as a signal for the relation *Comparison.Similarity*. A list of this kind of constructions and the relations they signal is given in Webber et al. (2019).

Hypophora is a relation type for Question-Answer pairs, with one argument being the question and the other providing the answer:

The target of their wrath? **Their own employer, Kidder Peabody.** (wsj_0118, Hypophora)

Sense Hierarchy In PDTB 3.0, a sense hierarchy of three levels is adopted, as shown in Figure 2.19.

The first level comprises four broad senses (L1 senses), and beneath each L1 sense, there are finer distinctions, resulting in a total of 22 L2 senses, and the third level (L3) is used to encode the directionality of asymmetric L2 senses.

Apart from English, PDTB has been applied for annotating discourse corpora in other languages, such as Chinese (Zhou and Xue, 2015), Hindi (Oza et al., 2009), Turkish (Zeyrek and Webber, 2008), French (Danlos et al., 2012), Czech (Poláková et al., 2013), German (Bourgonje and Stede, 2020) and so on. A multi-lingual parallel discourse corpus in PDTB-style, TED-MDB (Zeyrek et al., 2018), was created. Prasad et al. (2017) show how cross-paragraph implicit relations can be annotated reliably, which forms a new development of the framework.

Crowdsourcing Some studies explore the possibility of crowd-sourcing discourse annotations in PDTB style. Scholman and Demberg (2017) ask annotators to choose

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

Level-1	Level-2	Level-3
		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 2.19: Sense hierarchy in PDTB 3.0, originally from Webber et al. (2019).

from a fixed list of connectives, which are representative of some selected discourse relations in texts from the overlapping section of PDTB 2.0 and RST-DT, and these connectives can be used for inferring sense labels. Yung et al. (2019) extend their method and adopt a two-step connective insertion method to enable more relations to be covered. Annotators can first insert connectives freely. If the connectives inserted at the first step are ambiguous, the annotators can choose from a fixed list of connectives at the second step, and these connectives are less ambiguous for determining the target sense labels. With this approach, Scholman et al. (2022) create a corpus named DiscoGeM, which contains 6,505 implicit discourse relations, with data from three genres. They conclude that sense labels are better represented as probability distributions because of their ambiguous nature, and genre has considerable influence on sense label distributions.

2.6.2.2 Studies on PDTB-Style Annotation

Hierarchical and Shallow Annotation Feng et al. (2014) express doubts about PDTB-style annotation. They show that deep hierarchical discourse structures are more powerful than shallow discourse representation in sentence ordering and

essay scoring tasks.

Reasons for Explicitness/Implicitness Torabi Asr and Demberg (2012) investigate reasons for the explicit/implicit distinction of discourse relations, and test the continuity hypothesis (Segal et al., 1991; Murray, 1997)[10] and the causality-by-default hypothesis (Sanders, 2005)[11] based on the relative frequency of explicitness/implicitness of some specific relations. According to the Uniform Information Density (UID) hypothesis (Frank and Jaeger, 2008), information is expressed evenly across a text, and redundant signals are avoided. Therefore, relations that are more predictable are more likely to be expressed implicitly. This assumption is supported by the frequency distributions of different types of relations in PDTB. Westera et al. (2020) explore the relationship between the QUD framework (Roberts, 2012; Von Stutterheim and Klein, 1989) and PDTB. Implicit questions are annotated on a corpus of TED-talks, which has been annotated in the PDTB framework. They compare evoked questions and the part of questions that are actually answered later in discourse. Annotators are asked to give evaluation on the degree to which the questions are answered, which is taken as a quantitative measure of question predictability. They find that questions evoked where implicit relations are identified are more likely to be answered than questions evoked where explicit relations are annotated, which suggests that implicit relations encode more predictable relationship between consecutive textual segments. This finding is consistent with the study by Torabi Asr and Demberg (2012). Although Torabi Asr and Demberg (2012) show correlations between specific relations, which are used to test the continuity hypothesis and the causality-by-default hypothesis, and the distinction between implicit and explicit relations in PDTB, Westera et al. (2020) fail to observe strong correlation between specific question types and question predictability in their data, except for the correlation between why-questions and causal relations. This result may be partly attributed to the high variability in question formulation with the QUD framework. They also find that where the NoRel PDTB relation class is identified, the corresponding evoked questions have lower consistency in human annotations and these questions have a lower rate of being answered.

Dependency Between Labels PDTB focuses on local discourse relations, and the annotations are not assumed to be related. However, Pitler et al. (2008) show that the relations have patterns of inter-dependence, for instance, an explicit *Comparison*

[10]Readers expect consecutive sentences to be continuous in order and focus.
[11]Consecutive sentences are expected to be causally related.

relation tends to be followed by an implicit *Contingency* relation. Based on this intuition, Lin et al. (2011) add transitions of discourse relations into the entity-grid model proposed by Barzilay and Lapata (2005) (section 2.3). Entries of the entity transition matrix are enriched with roles of entities in discourse, including discourse relations they are involved in and their role as Arg1 or Arg2. In Lin et al. (2011), only PDTB L1 sense labels are used.

2.6.2.3 Shallow Discourse Parsing

The availability of large discourse corpora spurs studies on developing computational systems for discourse processing.

Lin et al. (2014) propose the first end-to-end shallow discourse parser. The pipeline consists of four components:

1. Connective classification: All the connectives of the input text are identified first, and then a discourse connective classifier is used to determine which are discourse connectives[12].

2. Argument labelling and explicit relation classification: If a connective functions as a discourse connective, the spans of its Arg1 and Arg2 will be labeled. After this step, triples of (connective, Arg1, Arg2) will be fed to an explicit relation classifier so that an explicit relation label is assigned.

3. Non-explicit relation classification: For each paragraph, if two adjacent sentences are not assigned any explicit relations, then possible labels will be EntRel, NoRel, AltLex or an implicit relation.

4. Attribution span labelling[13]: The input text is split into clauses first. If a clause is involved in a relation annotated in step 2 or step 3, then an attribution span labeler is used to determine if the clause is an attribution span or not.

This pipeline is adopted in later studies (Xue et al., 2015, 2016; Wang and Lan, 2015), although most of the time, only some subtasks are addressed and modifications may be introduced. In the shared task for the 19th Conference on Computational

[12]Not all the connectives function as discourse connectives. For instance, in the sentence "Apples and oranges are different fruit varieties.", the connective "and" is not a discourse connective, but in the sentence "I went to the theatre, and the play was very interesting.", it is a discourse connective.

[13]An attribution span is the textual span indicating an agent making a belief or assertion (Prasad et al., 2006), such as "...said" or "...argued". Attribution is annotated for explicit relations, implicit relations and AltLex relations.

Natural Language Learning (CoNLL-2015) (Xue et al., 2015), exact match is required for argument span extraction. In both the shared tasks of CoNLL-2015 and CoNLL-2016, identification of connectives focuses on the **heads** of the connectives, for instance, in the connective "three minutes before", "before" is the head. A predicted explicit relation is considered to be correct if its connective is identified correctly, its sense label is correct, and its arguments are correct both in span extraction and Arg1/Arg2 assignment. Since implicit relations are mostly identified between adjacent sentences in PDTB 2.0, which is used in the shared tasks (Xue et al., 2015, 2016), the requirement for correct identification of implicit relations is that the sense labels are correct (the same rule goes for AltLex and EntRel relations).

Biran and McKeown (2015) propose using separate classifiers for intra-sentential and inter-sentential relations. An intra-sentential tagger only focuses on explicit relations, and an inter-sentential tagger deals with all the possible relations for two adjacent sentences, including explicit, implicit, EntRel and AltLex relations. Moreover, relation parsing is implemented as a sequence labelling task with CRFs. Compared with the pipeline adopted in Lin et al. (2014), their two-tagger approach may reduce the error propagation problem and better capture dependency between discourse relations.

Discourse connectives are generally strong signals of discourse relations. For instance, "because" is a typical indicator of a causal relation. Pitler et al. (2008) show that most of the discourse connectives are unambiguous signals of discourse relations and good performance can be achieved for explicit relation classification with these connectives. Therefore, research efforts on shallow discourse parsing are directed towards implicit relation classification.

Inspired by the effectiveness of discourse connectives in explicit relation classification, researchers propose various methods to convert implicit relation classification into explicit relation classification.

Rutherford and Xue (2015) notice that simply removing discourse connectives for some instances of explicit relations changes the relations that can be inferred. Moreover, if discourse connectives are dropped, some explicit relation data still cannot be considered as instances of implicit relations and assigned counterpart implicit relation labels. Therefore, not all the explicit relation data can be used for data augmentation by simply dropping their discourse connectives. To measure

how suitable explicit relation data are for data augmentation in implicit relation classification, they identify discourse connectives and compute the percentages of instances where the discourse connectives are omitted in the corpus. Explicit relation data, which contain connectives with higher rates of omission, are considered suitable for data augmentation in implicit relation classification. As annotations of implicit relations contain manually inserted connectives, contextual similarities of the connectives in implicit relations and explicit relations are computed as a measure of closeness of the discourse connectives appearing in explicit and implicit relation data, which forms another metric of the suitability of using explicit relation data for data augmentation.

Qin et al. (2017) adopt an adversarial approach through which an implicit relation classifier is trained to resemble a classifier that is trained on connective-augmented implicit relation data. A discriminator is configured to distinguish the latent features learnt by the two modules, and the implicit relation classifier has two training objectives: maximizing the performance for implicit relation classification and reducing the accuracy of the discriminator.

Kishimoto et al. (2020) adapt the training procedure of BERT for implicit relation classification. They add a task of predicting discourse connectives of explicit relations in the pre-training stage and integrate an objective of implicit connective prediction in the fine-tuning stage.

Liu and Strube (2023) develop a model which is trained on two tasks: discourse connective prediction, and implicit relation classification based on the concatenation of predicted discourse connectives and argument pairs. They find that using predicted connectives as a part of input for relation classification is beneficial, contrary to previous studies which treat discourse connective prediction only as a subsidiary task for implicit relation classification.

Some studies focus on the influence of context in implicit relation classification. During the annotation process, annotators typically have access to the full text, but data for training classifiers of implicit relations are generally limited to argument pairs. Atwell et al. (2022) study the effect of context in implicit relation annotation. Annotation accuracy and annotator confidence with respect to different contexts are reported. They show that greater context is most of the time helpful for discourse relation annotation. They also explore how annotators' confidence scores can be added to the training and evaluation of neural models to achieve

higher performance and better model calibration. Zhang et al. (2021c) develop a graph-based model to incorporate context for implicit relation classification. The graph contains three types of edges between sentences: edges that encode linear order of sentences; edges that represent co-reference relations between two sentences; and edges that link sentences based on lexical chains. Therefore, their method tries to capture the relationship between sentences based on co-reference and lexical relatedness. Dai and Huang (2018) develop a hierarchical representation model, the lower level being per discourse unit representation and the higher level modelling sequential dependency between discourse units. To enrich the representations of discourse units, they concatenate part-of-speech (POS) tags and named entities with word embeddings. To better represent contextual information, they take a whole paragraph containing discourse units as input and employ a CRF model on the predicted relations to model dependency between discourse relations.

As large language models (LLMs), such as LLaMA (Touvron et al., 2023) and Generative Pre-trained Transformer (GPT) models (Radford, 2018), show impressive performance on various NLP tasks (Achiam et al., 2023; Brown, 2020), some studies explore how LLMs can be used for implicit relation classification.

Similar to previous supervised methods, which focus on discourse connective prediction and try to convert implicit relation classification into explicit relation classification, Xiang et al. (2022) introduce a method under the prompt learning framework. Based on the intuition that English connectives are frequently inserted between clauses or at the start of a sentence, they design different templates, which differ in where the predicted connectives are placed. Probability distributions over a set of predefined connectives can be obtained from a pre-trained language model, and predicted connectives are then mapped to sense labels. When more than one prompt template is used and multiple results are aggregated, they adopt majority voting to improve the performance. In their experiments, connectives with only one word are produced and compound connectives such as "thanks to" are not considered. Wu et al. (2023a) adopt a knowledge-distillation approach based on the method proposed in Xiang et al. (2022). They first train a teacher model to predict connectives when (Arg1, Arg2, sense label) is given, and then train a student model to predict connectives when only (Arg1, Arg2) is given. The KL-divergence loss between the probability distributions over connectives produced by the student model and the teacher model and the loss of predicting

the target sense label are to be minimized.

Chan et al. (2023) stress interaction between label information at different levels of the sense hierarchy and design a template to elicit joint probabilities of labels at different levels from T5 (Raffel et al., 2020), an encoder-decoder model pre-trained on a mixture of tasks, where the tasks are converted to a text-to-text format, and the path of sense labels with the highest joint probability along the sense hierarchy is chosen. Different from other studies that utilize the original sense hierarchy of PDTB (Long and Webber, 2022; Wu et al., 2022), Chan et al. (2023) replace the third level with discourse connectives that are representative of the discourse relations, which is shown to be beneficial for the task.

Following the study by Chan et al. (2023), Yung et al. (2024) try to update the results on PDTB 3.0 with GPT-4 (Achiam et al., 2023), the fourth generation of auto-regressive GPT models. They adopt three prompting strategies. The first is based on the two-step connective insertion method proposed in Yung et al. (2019), and the second strategy breaks down the task of predicting one sense label from all the possible choices into a per-class binary yes-or-no question. With this approach, n questions are generated for each instance, n denoting the total number of sense labels. The third strategy differs from the second one in that the binary yes-or-no question is replaced by a question that is a paraphrase of the second-level sense label. For instance, instead of asking "Is this an Asynchronous relation?", the question is phrased as "Which argument describes an event occurring before the other?" and the answer is chosen from (Arg1, Arg2 or None). Their experiments show that even carefully designed prompting techniques fail to enable LLMs to achieve results comparable to existing supervised approaches.

2.6.3 SDRT

The Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) introduces rhetorical relations into traditional studies on dynamic semantics, represented by the Discourse Representation Theory (Kamp and Reyle, 1993), and focuses on the role of rhetorical relations in constraining semantic interpretation.

Under this framework, discourse structure is represented by directed acyclic graphs (DAGs). EDUs may be combined recursively to form complex discourse units (CDUs), which can be linked with other EDUs or CDUs (Asher et al., 2017). Similar to LDM, discourse structuring in SDRT follows RFC. However, greater

scope of nodes on the right frontier can be observed in SDRT, including the last rightmost constituent, complex discourse units that include the last rightmost constituent, and units super-ordinate to the last rightmost constituent through a series of subordinating relations (Afantenos and Asher, 2010; Afantenos et al., 2012).

An example of analysis in this framework is shown in Figure 2.20, from which it is clear that SDRT features full-coverage of text and a hierarchical structure of text organization, similar to RST. The vertical arrow-headed lines represent subordinating discourse relations, and the horizontal lines represent coordinating relations. The textual units in boxes are EDUs and π' and π'' represent CDUs. Discourse relations are shown in bold.

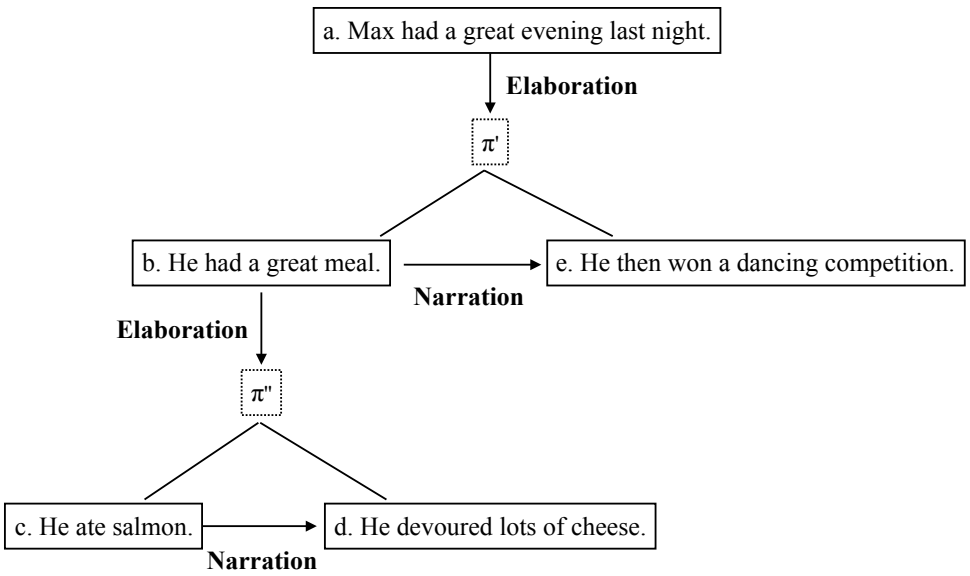


Figure 2.20: SDRT representation of the text "(a) Max had a great evening last night. (b) He had a great meal. (c) He ate salmon. (d) He devoured lots of cheese. (e) He then won a dancing competition." The example is taken from Asher and Lascarides (2003).

It is stressed in Asher and Lascarides (2003) that more than one relation is possible between two textual spans. When there are multiple possible interpretations, the

Maximize Discourse Coherence (MDC) principle can be used to choose the best interpretation of discourse. Based on this principle, the more rhetorical relations are possible between two units, the more coherent an interpretation is, and the more anaphoric expressions whose antecedents become clear, the more coherent an interpretation is.

With SDRT, the STAC corpus (Asher et al., 2016) and the Molweni corpus (Li et al., 2020) are created for analyzing multi-party dialogues. The ANNODIS corpus (Afantenos et al., 2012) is one of the few corpora built on formally written texts and annotated in SDRT style, but the texts are in French. The DISCOR project (Reese et al., 2007) aims at creating an English discourse corpus under the SDRT framework, but based on the account in Reese et al. (2007), only 60 articles taken from the MUC6 corpus (Chinchor and Sundheim, 2003) are annotated, and the corpus may be too small to be useful in computational experiments.

2.7 Integration of Discourse Frameworks

A review is given on theoretical proposals for integrating discourse frameworks, followed by a review on studies that empirically explore the relationship between different frameworks in discourse structure and in relation taxonomies.

2.7.1 Theoretical Proposals for Integrating Discourse Frameworks

Hovy (1990) is the first to try to unify discourse relations proposed by researchers from different areas who may have different perspectives on definitions of discourse relations. He suggests adopting a hierarchy of relations, with the top level being more general (categorized from functional perspective as ideational, interpersonal and textual), and no restrictions are imposed on adding fine-grained relations, as long as they can be subsumed under the existing taxonomy. The number of studies that propose a specific relation is taken as a vote of confidence for incorporating the relation into the taxonomy. Similar to Hovy (1990), Benamara and Taboada (2015) propose a hierarchy of discourse relations when they try to unify the relation taxonomies of RST and SDRT, where the top level is general and fixed, and the lowest level is more specific, accommodating variations according to genre and language.

Three systematic theoretical proposals have been made to integrate different discourse frameworks: (1) ISO DR-Core (ISO 24617-8) (Bunt and Prasad, 2016); (2) discourse extension of the Ontologies of Linguistic Annotation (OLiA) (Chiarcos, 2014); and (3) the unified dimension (UniDim) approach by Sanders et al. (2018).

ISO DR-Core The ISO DR-Core proposal is based on the understanding that a set of core relations are shared by existing frameworks. As such, the proposal is not aimed at providing a comprehensive and fixed set of discourse relations, and only a set of 20 most common relations are identified, which are based on theoretical analysis of definitions of discourse relations of different frameworks, covering written, spoken and multi-modal domains. The proposal is similar to PDTB in adopting a distinction between implicit and explicit relations and allowing connectives to be added if possible, focusing on informational relations and ignoring the intended effects of discourse relations on the reader/hearer, and being concerned solely with local discourse relations, without considering higher-level text structure. However, different from PDTB, the core set of relations is flat. This design is based on the consideration that different frameworks show divergent understandings of semantic closeness of discourse relations, which forms the criteria for grouping fine-grained discourse relations into broader classes. The set of relations are extensible and open to accommodate the results of other studies. Figure 2.21 shows the 20 discourse relations of ISO DR-Core.

OLiA The OLiA approach is to provide an ontological intermediary representation of discourse phenomena, including discourse structure and discourse relations, coreference and bridging, and information structure. As the focus of this thesis is on discourse structure and discourse relations, a brief introduction is given to these two parts of the model.

The OLiA model is characterized by a layered organization. The top-level category contains some abstract concepts, including: (1) *DiscourseCategory* for non-relational structures and entities; (2) *DiscourseRelation* for relations between DiscourseCategories; and (3) *DiscourseFeature* for annotations of DiscourseCategories and DiscourseRelations.

To generalize over structural assumptions of discourse frameworks, the OLiA model introduces a distinction between *DiscourseRelation* and *DiscourseStructuralPattern*, the former capturing relation types and the latter distinguishing discourse relations in terms of coordination/subordination. The *DiscourseSegment* concept

	ISO DRel	Symmetry	Relation and Argument-Role Definitions
1.	Cause	Asymmetric	Arg1 provides a reason for Arg2 to come about or occur.
2.	Condition	Asymmetric	Arg1 is an unrealized situation which, when realized, would lead to Arg2.
3.	Negative Condition	Asymmetric	Arg1 is an unrealized situation which, when not realized, would lead to Arg2.
4.	Purpose	Asymmetric	Arg1 enables Arg2.
5.	Manner	Asymmetric	Arg1 is a way in which Arg2 comes about or occurs.
6.	Concession	Asymmetric	An expected causal relation between Arg1 and Arg2, where Arg1 is expected to cause Arg2, is cancelled or denied by Arg2.
7.	Contrast	Symmetric	One or more differences between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
8.	Exception	Asymmetric	Arg1 evokes a set of circumstances in which the described situation holds, while Arg2 indicates one or more instances where it doesn't.
9.	Similarity	Symmetric	One or more similarities between Arg1 and Arg2 are highlighted with respect to what each predicates as a whole or to some entities they mention.
10.	Substitution	Asymmetric	Arg1 and Arg2 are alternatives, with Arg2 being the favored or chosen alternative.
11.	Conjunction	Symmetric	Arg1 and Arg2 bear the same relation to some other situation evoked in the discourse. Their conjunction indicates that they are doing the same thing with respect to that situation, or are doing it together.
12.	Disjunction	Symmetric	Arg1 and Arg2 are alternatives, with either one or both holding.
13.	Exemplification	Asymmetric	Arg1 describes a set of situations; Arg2 is an element of that set.
14.	Elaboration	Asymmetric	Arg1 and Arg2 are the same situation, but Arg2 contains more detail.
15.	Restatement	Symmetric	Arg1 and Arg2 are the same situation, but described from different perspectives.
16.	Synchrony	Symmetric	Some degree of temporal overlap exists between Arg1 and Arg2. All forms of overlap are included.
17.	Asynchrony	Asymmetric	Arg1 temporally precedes Arg2.
18.	Expansion	Asymmetric	Arg2 provides further description about some entity or entities in Arg1, expanding the narrative forward of which Arg1 is a part, or expanding on the setting relevant for interpreting Arg1. The Arg1 and Arg2 situations are distinct.
19.	Functional dependence	Asymmetric	Arg2 is a dialogue act with a responsive communicative function; Arg1 is the dialogue act(s) that Arg2 responds to.
20.	Feedback dependence	Asymmetric	Arg2 is a feedback act that provides or elicits information about the understanding or evaluation by one of the dialogue participants of Arg1, a communicative event that occurred earlier in the discourse.

Figure 2.21: 20 discourse relations considered to be common in existing discourse frameworks, originally from Prasad and Bunt (2015). The last two relations are applicable to dialogues.

denotes discourse units.

Figure 2.22 shows an example of the OLiA approach in representing discourse structure and discourse relations. The representation of coherence relations follows a hierarchical structure, with the top level being similar to L1 of the PDTB sense hierarchy: *Comparison*, *Contingency*, *Expansion*, *TemporalRelation* and *TopicContinuityRelation*. If a discourse relation is an instance of concept A, and A $\in$ B along the hierarchy, then the discourse relation is considered to be $\in$ B automatically. With this kind of operations, discourse relations in different frameworks can be compared with each other.

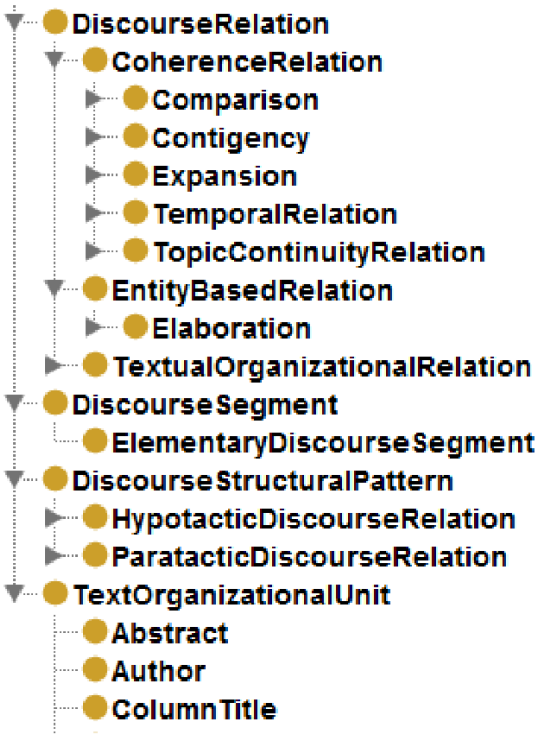


Figure 2.22: An example of ontological representation of discourse structure and discourse relations with the OLiA approach, originally from Chiarcos (2014).

UniDim Under the TextLink Action[14], which aims at unifying existing linguistic resources on discourse, Sanders et al. (2018) propose a set of unifying dimensions (henceforth UDims) as an interface for different discourse frameworks to be related with each other. The UDims originate from four cognitive primitives—*basic operation*, *source of coherence*, *order of segments* (called *implication order* in Sanders et al. (2018)) and *polarity*, which are used to define coherence relations in Sanders et al. (1992), where a different approach towards representing discourse relations is taken, forming the CCR framework. Compared with other coherence-relation based discourse frameworks such as RST and PDTB, the CCR framework treats discourse relations as cognitive entities that can be analyzed from different dimensions, and a relation is thus described from four dimensions, such as *causal, objective, basic order, positive*, rather than with a single label, such as *cause* in RST. Each of these dimensions functions as an attribute that has a number of possible values, for example, the *polarity* dimension has three values: *positive*, *negative* or *under-specified* in ambiguous cases. To make the taxonomy of CCR more expressive, additional dimensions are added, including *temporality*, and *specificity*, *lists* and *alternatives* for additive relations, and *conditionals* and *goal-orientedness* for causal

[14] http://textlink.ii.metu.edu.tr/

relations. Recall that *additive* and *causal* fall under the *source of coherence* dimension. With these UDims, discourse relations from different discourse frameworks can be decomposed and compared systematically. Roze et al. (2019) employ the UniDim proposal to identify the aspects of discourse relations that are challenging for computational systems. Additionally, they introduce a novel approach to discourse relation classification, which involves first classifying the UDims and then using the predicted UDims for classifying discourse relations. A direct mapping from predicted UDims to target relation labels is adopted, representing a potentially framework-agnostic approach to discourse relation classification, although the study focuses on PDTB. Fu (2023) shows that the upper limit for using UDims for discourse relation classification is high for RST and PDTB.

Demberg et al. (2019) compare results of relation mapping based on the three theoretical proposals and they find that different proposals agree on a part of the mappings but also show discrepancies, which arise from differences in the granularity of relation definitions, innate differences between discourse frameworks in defining discourse relations, and different interpretations of relation definitions given in the guidelines. Therefore, Demberg et al. (2019) claim that validation of theoretical proposals on annotated data is required.

2.7.2 Empirical Investigation of Integration of Discourse Structures of Different Discourse Frameworks

Stede et al. (2016) investigate the relationship between RST, SDRT and argumentation structure. The purpose is to investigate if discourse parsing can contribute to automatic argumentation analysis. The authors exclude the PDTB framework because it does not provide full discourse annotation. They add annotations of RST and SDRT to an existing corpus on argumentation structure. For the purpose of comparing the three layers of annotation, EDU segmentation in RST and SDRT is harmonized, and an "argumentatively empty" JOIN relation is introduced to address the issue that the basic unit of argumentation structure is coarser than the other two layers. The annotations are converted to a common dependency graph for calculating correlations. To transform RST trees to dependency structure, the method introduced by Li et al. (2014) is used. The RST trees are binarized and the left-most EDU is treated as the head. In the transformation of SDRT graphs to dependency structure, the CDUs are simplified with a *head replacement strategy*, which means that CDU nodes are removed from the original graphs and

any incoming or outgoing edges of these nodes are attached to the heads of the CDUs. The authors compare the dependency graphs in terms of common edges and common connected components. The relations of the argumentation structure are compared with those of RST and SDRT, respectively, through a co-occurrence matrix. Their research suggests that there is a systematic relationship between argumentation structure and the two discourse frameworks.

Yi et al. (2021) try to unify two Chinese discourse corpora annotated in PDTB style and in RST style, respectively, under the dependency framework. A small corpus annotated with discourse dependency structure is taken as reference. To convert the corpus annotated under the RST framework, the method proposed by Li et al. (2014) is used, and some of the EDUs are revised because EDU segmentation of the RST-style corpus is coarser. In the step of constructing dependency trees, it is required that each basic discourse unit should have one and only one head, and only one relation is labeled between a head and a dependent. To convert the corpus annotated in PDTB style, a pre-trained segmenter is used to re-segment the texts and the outputs are checked manually. Typical discourse connectives for each relation type are summarized and then used as clues to add some relations to form a complete dependency tree. These added relations are also checked manually. As the three corpora use different relation sets, the original relations are kept during the conversion, and then mapped to 17 predefined relations, which are derived mainly from the RST-style discourse corpus. Their experiments show that data obtained through unification of resources boost performance for parsing under the dependency framework, but without much benefit to discourse parsing under the other frameworks.

2.7.3 Empirical Investigation of Integration of Discourse Relations

Thanks to the availability of corpora annotated with different types of discourse information in parallel, the correlation between discourse relations of different frameworks can be investigated directly on data.

Scheffler and Stede (2016) focus on mapping between explicit PDTB discourse relations and RST relations. PCC, which contains annotations of both frameworks, is used. It is found that the majority of PDTB discourse connectives are associated with exactly one RST relation, and mismatches are caused by different segmentation criteria and focuses, i.e. PDTB is concerned with local/lexicalized relations

and RST focuses on global relations.

Poláková et al. (2017) try to uncover the relationship between implicit relations in PDTB 2.0 and RST relations using the RST Signaling Corpus (RST-SC) (section 2.6.1.10). The PDTB and RST-SC data are converted to the Prague Markup Language (PML) format, which makes it possible to browse, edit and query data with a specialized editor and a query system developed for treebank annotation and data processing. Annotations in RST-SC and PDTB 2.0 are thus transformed into a uniform format, and comparison can be performed based on text matching. An intersection between PDTB 2.0 and RST-SC is found, where the arguments of an implicit relation match the two discourse units of an RST relation exactly. It is found that more than half of the relations in RST-SC that have matched PDTB implicit relations are of semantic type, which originate from loosely defined lexical chains. Syntactic signals, which form the strongest signalling cue of discourse relations in RST-SC, play only a restricted role in signalling implicit relations in PDTB. With a corpus annotated with signals of discourse relations, the study provides a new perspective on interpreting underlying differences in discourse relations of different discourse frameworks.

As PCC only contains annotations of explicit relations for PDTB, Bourgonje and Zolotarenko (2019) try to induce implicit relations for PDTB from RST annotations. Since RST trees are hierarchical and PDTB annotations are shallow, RST relations that link complex discourse units are discarded. Only RST relations that are signaled by explicit discourse connectives are considered in their experiment to improve the accuracy of identifying matched relations. It is shown that differences in segmentation criteria and partially overlapping relations are two major challenges for the task.

RST-DT and PDTB have an overlapping section of annotated texts. On this basis, Demberg et al. (2019) propose a method of mapping RST and PDTB relations. Since the number of PDTB relations is much smaller than RST relations for the same text, PDTB annotations are used as the starting point for the mapping process. They aim for mapping as many relations as possible while making sure that the relations connect the same discourse units. As RST and PDTB adopt different criteria in segmentation, they use the strong nuclearity hypothesis (section 2.6.1.4) to identify alignable relations of the two frameworks, which may differ in the exact spans of discourse segments. They manage to analyse 76% of the relations in PDTB 2.0, among which 52% have corresponding arguments with RST relations, and 48%

have at least an argument mapped to multiple EDUs. Their results indicate that mapping results based on theoretical proposals diverge from those identified in annotated texts, and empirical mappings show higher consistency with theoretical mappings for explicit relations (70% of cases) than implicit relations (less than 50%). Rehbein et al. (2016) compare discourse relations of PDTB and CCR on the basis of a spoken corpus annotated in the two frameworks. They find that differences in annotation operationalisation and granularity of relation definition lead to many-to-many mappings. Similar findings are suggested in Demberg et al. (2019). Liu et al. (2024) add a layer of PDTB-style annotations on the GUM corpus and show experimental results on converting the existing layers of RST and eRST relation annotations to PDTB relations to allow cross-formalism comparison, similar to the study by Demberg et al. (2019).

CHAPTER THREE

CHAPTER 3 AUTOMATIC ALIGNMENT OF DISCOURSE RELATIONS OF DIFFERENT DISCOURSE FRAMEWORKS

3.1 Chapter Overview

As can be seen from Chapter 2, existing discourse corpora are annotated based on different frameworks, which show significant dissimilarities in definitions of arguments and relations and structural constraints. Despite surface differences, these frameworks share basic understanding of discourse relations. The relationship between these frameworks has been an open research question, especially the correlation between relation sets utilized in different frameworks. Better understanding of this question is helpful for integrating discourse theories and enabling interoperation of discourse corpora annotated under different

frameworks. However, studies that explore the relationship between discourse relation inventories are hindered by different segmentation criteria, and expert knowledge and manual examination are typically needed. Some semi-automatic methods have been proposed, but they rely on corpora annotated in multiple frameworks in parallel. In this chapter, a fully automatic approach is introduced to address the challenges. Specifically, the approach is built on the label-anchored contrastive learning method introduced by Zhang et al. (2022b) to learn label embeddings, which are then utilized to map discourse relations from different frameworks. Experimental results on RST-DT and PDTB 3.0 are shown. In addition, extrinsic evaluation is performed, and mapping results obtained with the proposed method are applied to a discourse relation classification task in comparison with results based on the state-of-the-art (SOTA) method proposed by Demberg et al. (2019).

3.2 Motivation

Discourse relations are an important means for achieving coherence. Previous studies have shown the benefits of incorporating discourse relations in downstream tasks, such as sentiment analysis (Wang et al., 2012), text summarization (Huang and Kurohashi, 2021) and machine comprehension (Narasimhan and Barzilay, 2015). Automatic discourse relation classification is an indispensable part of discourse parsing, which is performed under some frameworks, notable examples including RST and PDTB[1].

As discourse annotation has a high demand for knowledge about discourse and high-level understanding of texts, discourse corpora are costly to create. Existing discourse formalisms typically share similar understanding of discourse relations and their role in discourse construction. Thus, an option to enlarge discourse corpora is to align existing discourse corpora so that they can be used jointly. This line of work starts as early as Hovy and Maier (1992), but it remains challenging to uncover the relationship between discourse relations used in different frameworks.

An example of RST-style annotation has been shown in Figure 2.2 in section 2.6.1. It is shown again in Figure 3.1 to illustrate the challenges of mapping discourse

[1]The experiments focus on RST and PDTB because the method requires a large amount of data and these two frameworks have been applied to annotation of corpora that overlap in selected texts, thus mitigating the effect of domain shift in the results. However, the method itself does not require corpora built on the same texts.

relations of different frameworks. Recall that the textual spans in boxes are EDUs and the arrow-headed lines represent asymmetric discourse relations, pointing from satellites to nuclei. The labels *elab(oration)* and *attribution* denote discourse relations. As the two spans connected by the relation *same-unit* are equally salient, the relation is represented by undirected lines. The spans are linked recursively until full-coverage of the text is formed, as shown by the uppermost horizontal line. The vertical bars highlight the nuclei.

The text above is also included in PDTB, where the annotation is:

1. *the agreement "an important step forward in the strengthened debt strategy"*, **that it will "when implemented, provide significant reduction in the level of debt and debt service owed by Costa Rica."** (implicit, given, Contingency.Cause.Reason)

2. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica.*, **implemented,** (explicit, when, Temporal.Asynchronous.Succession)

3. *that it will provide significant reduction in the level of debt and debt service owed by Costa Rica.*, **implemented,** (explicit, when, Contingency.Cause.Reason)

where Argument 1 (Arg1) is shown in italics and Argument 2 (Arg2) is in bold. The annotations in parentheses represent *relation type*, which can be implicit, explicit or others, *connective*, which is identified or inferred by annotators to signal the relation, and *sense label*, which is delimited with dots, with the first entry showing the sense label at level 1 (L1 sense), the second entry being the sense label at level 2 (L2 sense) and so on.

One of the challenges may be attributed to distinctive assumptions about higher-level structures and discourse segmentation. PDTB focuses on semantic relations between arguments, and argument span identification is performed following the *Minimality Principle*, which means that only those parts that are necessary and minimally required for understanding a relation are annotated (Prasad et al., 2008). In comparison, EDUs in RST are typically clauses. It has been shown repeatedly that segmentation criteria affect the scope of discourse relations and influence the type of relations that can be attached (Demberg et al., 2019; Benamara and Taboada, 2015; Rehbein et al., 2016).

In the first annotation of PDTB, Arg1, i.e., *the agreement "an important step forward in the strengthened debt strategy"*, is taken from the original text "Treasury Secretary Nicholas Brady called the agreement 'an important step forward in the strengthened debt strategy' " and the part "Treasury Secretary Nicholas Brady

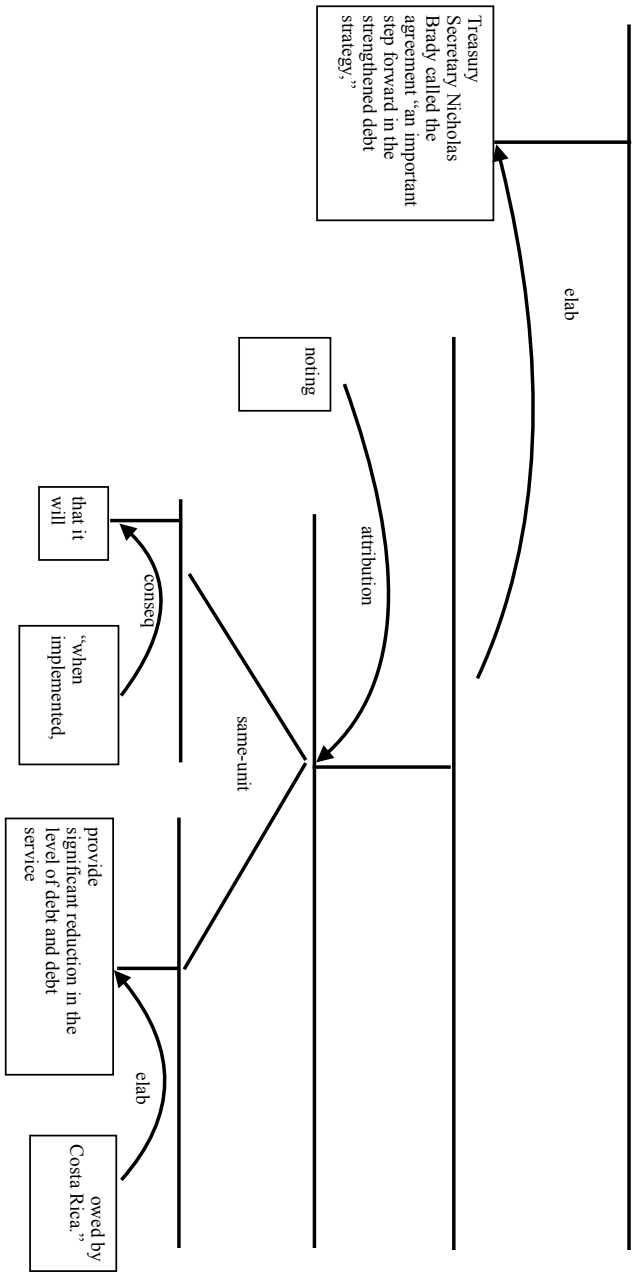


Figure 3.1: RST-style annotation (wsj_0624 in RST-DT).

called" is not covered because it does not contribute to the interpretation of the relation here. In contrast, this part is kept in an EDU in RST.

Another major difference between the two frameworks is that RST enforces a tree structure, and all the EDUs and CDUs (larger spans formed by adjacent EDUs

or CDUs) should be connected without crossings, while PDTB only focuses on local relations without commitment to any higher-level structure, as exemplified by the three independent annotations shown above. Previous studies (Lee et al., 2006, 2008) suggest that edge crossings and relations with shared arguments are common for PDTB. This distinction adds to the difficulty of exploring correlations of relations between the two frameworks, even if the two corpora are built on the same texts.

In addition, in RST-DT, an inventory of 78 relations is used, which can be grouped into 16 classes. These relations can be divided into *subject matter* relations (informational relations in Moore and Pollack (1992)), which are relations whose intended effects are to be recognized by readers, and *presentational* relations (intentional relations in Moore and Pollack (1992)), which are intended to increase some inclination in readers (Mann and Thompson, 1988). For each relation, only one sense label can be attached. In contrast, PDTB adopts a three-level sense hierarchy, and more than one sense label can be annotated for a pair of arguments. As shown in the example, items 2 and 3 represent annotations for the same argument pair, but different sense labels are assigned. In previous studies that explore the alignment of RST and PDTB discourse relations, these cases typically require manual inspection to determine the closest matching PDTB relation to an RST relation (Demberg et al., 2019). Moreover, PDTB does not take intentional relations into account but focuses on semantic and pragmatic relations.

The combination of these factors makes it challenging to investigate the relationship between discourse relations of different frameworks. Even in empirical studies that make use of corpora annotated in multiple frameworks in parallel, expert knowledge and manual examination are still required. To tackle the challenge caused by differences in discourse segmentation, Demberg et al. (2019) employ the strong nuclearity hypothesis to facilitate the string matching process of aligning PDTB arguments and RST segments. While this method alleviates the limitation of exact string matching of arguments/EDUs, it relies on a corpus annotated with multiple frameworks in parallel. Furthermore, it is conceivable that relations discarded because they violate the principle of the strong nuclearity hypothesis are not necessarily irrelevant for the goal of enabling joint usage of RST and PDTB data.

In this chapter, a fully automatic method is proposed, inspired by advances in label embedding techniques and an increasing body of research endeavors to harness

label information in representation learning, such as supervised contrastive learning (Khosla et al., 2020; Gunel et al., 2021; Suresh and Ong, 2021). Instead of relying on string matching to identify the closest matching PDTB arguments and RST segments with the aim of discovering potentially analogous relations, the method enables label embeddings of relations to be learnt and the learnt label embeddings are compared directly. Although the method is primarily used for mapping discourse relations of different frameworks, it is potentially useful for mapping labels of other types of datasets, such as datasets for human activity recognition (Ye, 2021).

The main contributions of this chapter can be summarized as follows:

- A label embedding based approach is proposed to explore correlations between relations of different discourse frameworks. The method is fully automatic and eliminates the need for matching arguments of relations.
- Extensive experiments on different ways of encoding label embeddings are conducted on RST-DT and PDTB 3.0.
- A metric for evaluating the learnt label embeddings intrinsically is proposed and experiments on extrinsic evaluation of the proposed method are performed.

3.3 Related Work

Mapping discourse relations Existing research on mapping discourse relations of different frameworks can be categorized into three types (Fu, 2022): a. identifying a set of commonly used relations across various frameworks through analysis of definitions and examples of discourse relations (Hovy and Maier, 1992; Bunt and Prasad, 2016; Benamara and Taboada, 2015); b. introducing a set of intermediary concepts for analyzing and comparing discourse relations of different frameworks (Chiarcos, 2014; Sanders et al., 2018); and c. empirical studies on mapping discourse relations based on corpora annotated in multiple frameworks in parallel (Rehbein et al., 2016). A detailed review of these studies has been given in section 2.7.

Label embeddings Label embeddings have been proven to be useful in computer vision (CV) (Akata et al., 2016; Palatucci et al., 2009; Zhang et al., 2022a) and

NLP tasks (Wang et al., 2018; Zhang et al., 2018; Miyazaki et al., 2019). In existing studies, one-hot encoding is conventionally used to represent labels, which suffers from three problems: lack of robustness to noisy labels (Gunel et al., 2021), high possibility of overfitting (Sun et al., 2017) and failure to capture semantic correlations between labels. Incorporating label representations into model training is helpful for mitigating these problems and semantics of labels can be used as additional information to improve model performance. It is shown that label embeddings can help to overcome challenges in data-imbalanced settings and zero-shot learning (Zhang et al., 2022b).

Label embeddings can be representations initialized from BERT representations (Xiong et al., 2021) or randomly initialized (Zhang et al., 2022b). Another approach of obtaining label embeddings is to learn label embeddings during model training. Akata et al. (2016) propose a method of learning label embeddings from label attributes in a classification task. Wang et al. (2018) introduce an attention mechanism that measures the compatibility of embeddings of input and labels, with which input-label-joint embedding is learnt for improving the performance on a text classification task. Additional information can be incorporated for learning label embeddings, such as label hierarchy (Chatterjee et al., 2021; Zhang et al., 2022a; Miyazaki et al., 2019) and textual description of labels (Zhang et al., 2023).

3.4 Proposed Method

Problem statement Given a corpus annotated in one discourse framework $D_1 = \{X_m^{(1)}, Y_m^{(1)}\}$ and another corpus annotated in a different framework $D_2 = \{X_m^{(2)}, Y_m^{(2)}\}$, where each X denotes an input sequence formed by a pair of arguments of the form $A_1^{(1)} \ldots A_a^{(1)}, A_1^{(2)} \ldots A_b^{(2)}$, and Y represents relation label sets of the two frameworks, $Y_1 = \{y_1^{(1)}, y_1^{(2)}, \ldots, y_1^{(k)}\}$ and $Y_2 = \{y_2^{(1)}, y_2^{(2)}, \ldots, y_2^{(c)}\}$. The objective is to learn a correlation matrix R between Y_1 and Y_2, which is a $2d$ matrix of shape $k \times c$. To achieve this purpose, the label embedding technique is used to learn the embeddings for the members of Y_1 and Y_2, and the widely used cosine similarity metric can be used as a measure of distance between embedding vectors. The label embedding learning method is the same for D_1 and D_2, and D_1 is taken as an example for illustration in the following.

The vanilla model for label-anchored contrastive learning in Zhang et al. (2022b) is utilized as the backbone. For an input sequence X, a pre-trained language model

can be used as an input encoder f_{InEnc}. Without losing generality, the popular *bert-base-uncased* model from the Huggingface transformers library (Wolf et al., 2020) is employed in implementation. For X, which is pre-processed as $[CLS]\ A_1^{(1)}\ ...\ A_a^{(1)}\ [SEP]\ A_1^{(2)}\ ...\ A_b^{(2)}\ [SEP]$, the representation of the $[CLS]$ token is used as the representation of the input sequence:

$$\mathbf{E}_X = f_{InEnc}(X) \tag{3.1}$$

where the input sequence representation $\mathbf{E}_X$ is of shape $(a+b+3) \times dim$, where dim is the dimension of the output from the pre-trained language model and a and b are the maximum lengths that the argument pairs are padded to. It is found empirically that removing the non-linear transformation to $\mathbf{E}_X$ in Zhang et al. (2022b) yields better performance for the task of this chapter.

There are several possible options of configuring label encoders, including: using representations from a BERT model (Devlin et al., 2019) (*LbEncBert*); using a RoBERTa model (Liu et al., 2020), which is trained without the objective of next sentence prediction (*LbEncRoberta*); randomly initializing from a uniform distribution (*LbEncRand*); adding textual description of the labels (*LbEncDesc*), where the label and the description are arranged in the form $[CLS]label[SEP]description[SEP]$, and the representation of $[CLS]$ is used as the label representation; and adding sense hierarchy information, where the hierarchical contrastive loss proposed by Zhang et al. (2022a) can be used and different penalty strengths are applied to losses at different levels (*LbEncHier*). As language models or trainable layers are used as label encoders, the label embeddings are learnable.

With a label encoder g_{LbEnc}, and k being the number of relations, a table T of shape $k \times lbDim$ can be obtained, where $lbDim$ is the output dimension of the label encoder. Thus, for a label y_l, its label embedding vector $\mathbf{E}_{y_l}$ is the l^{th} row of T, if the starting index is 1.

Instance-centered contrastive loss The method proposed in Zhang et al. (2022b) is applied to compute instance-centered contrastive loss $\mathscr{L}_{ICL}$:

$$\mathscr{L}_{ICL} = -\frac{1}{N} \sum_{X_i, Y_i} \log \frac{e^{\Phi(\mathbf{E}_X, \mathbf{E}_Y)/\tau}}{\sum_{1 \le l \le K} e^{\Phi(\mathbf{E}_X, \mathbf{E}_{Y_l})/\tau}} \tag{3.2}$$

where N denotes batch size, X_i is an instance in a batch, and Y_i is its label, Φ

represents a distance metric between the representations of the input and label embeddings, and cosine similarity is used in the experiment. τ denotes the temperature hyper-parameter, and lower values of τ increase the influence of hard-to-separate examples in the learning process (Zhang et al., 2021b). By minimizing this loss, distance between instance representations and label embeddings of the corresponding label is reduced, resulting in label embeddings that are compatible with input representations.

Label-centered contrastive loss The purpose of this loss is to reduce the distance between instances that have the same labels. For a batch with a set of unique classes C, c represents a member, P_c denotes the set of instances in a batch that have the label c and N_c represent the set of negative examples for c. A member in P_c is represented by X_p and a member in N_c is denoted by X_n. The label-centered contrastive loss $\mathscr{L}_{LCL}$ can be computed with:

$$\mathscr{L}_{LCL} = -\frac{1}{C}\sum_{c\in C}\sum_{X_p\in P_c}\log\frac{e^{\Phi(\mathbf{E}_{X_p},\mathbf{E}_c)/\tau}}{\sum_{X_n\in N_c}e^{\Phi(\mathbf{E}_{X_n},\mathbf{E}_c)/\tau}} \tag{3.3}$$

As indicated in Zhang et al. (2022b), $\mathscr{L}_{ICL}$ and $\mathscr{L}_{LCL}$ mitigate issues with a small batch size typical in other types of contrastive learning, which makes them suitable for scenarios with limited computational resources.

The following two supervised losses are incorporated in the training objective, which is shown to be effective empirically.

Label-embedding-based cross-entropy loss As shown in equation 3.4, a softmax function is applied to the k label embeddings in T, yielding a probability distribution over the k classes:

$$p(y_l) = \frac{e^{\mathbf{E}_{y_l}}}{\sum_{l=1}^{K} e^{\mathbf{E}_{y_l}}} \tag{3.4}$$

Let t_{y_l} denote the categorical encoding of the target y_l. The cross-entropy loss of classification based on label embeddings, denoted by $\mathscr{L}_{LEC}$, can be obtained with:

$$\mathscr{L}_{LEC} = -\sum_{l=1}^{K} t_{y_l}\log p(y_l) \tag{3.5}$$

The purpose of adding this loss is to make the label embeddings better separated from each other.

Canonical multi-class cross-entropy loss A canonical cross-entropy loss is used, which amounts to multi-class classification loss with input representations:

$$\mathscr{L}_{ICE} = -\sum_{i=1}^{N}\sum_{l=1}^{K} c_l^i \log p(c_l^i) \tag{3.6}$$

where N is the batch size, K is the total number of classes, and $p(c_l^i)$ is the probability predicted for a class c_l^i. With this loss, the input representations are learnt to be effective for the classification task.

The total loss $\mathscr{L}_{total}$ is the sum of the four losses:

$$\mathscr{L}_{total} = \mathscr{L}_{ICL} + \mathscr{L}_{LCL} + \mathscr{L}_{LEC} + \mathscr{L}_{ICE} \tag{3.7}$$

During inference, only vector matching between the representation of an input sequence $\mathbf{E}_X$ and the k learnt embeddings $\mathbf{E}_{y_l}$ is needed, with cosine similarity as a distance metric:

$$\hat{y} = \underset{1\leq l\leq k}{\operatorname{argmax}}(\Phi(\mathbf{E}_X, \mathbf{E}_{y_l})) \tag{3.8}$$

Baseline for relation classification The *BertForSequenceClassification* model from the Transformers library is used as the baseline for discourse relation classification, which is trained with cross-entropy loss only, i.e. equation 3.6.

Baseline for label embedding learning Label embeddings are generally used for improving model performance in classification tasks in previous studies (Wang et al., 2018; Zhang et al., 2018; Xiong et al., 2021; Zhang et al., 2022b). To compare with a method targeted at learning meaningful label embeddings, a baseline method is implemented, which is a combination of equation 3.4 and 3.5, but the softmax function is applied over the cosine similarities of an input $\mathbf{E}_X$ and each label embedding $\mathbf{E}_{y_l}$ in T here, similar to the approach adopted in Zhang et al. (2018) and Wang et al. (2018).

Metric After the model training stage, as the representations of the input sequences

have been learnt for the relation classification task, the average of the representations of input sequences X that belong to a class y_l can be considered as a proxy for the class representation, denoted by $\mathbf{H}_{y_l}$:

$$\mathbf{H}_{y_l} = \frac{1}{C}\sum_{i=1}^{C}\mathbf{E}_X \tag{3.9}$$

where C represents the number of instances in X.

Due to inevitable data variance, the learnt label embeddings $\mathbf{E}_{y_l}$ for a class y_l may not be the same as $\mathbf{H}_{y_l}$, but it should have a higher correlation with $\mathbf{H}_{y_l}$ than label embeddings of the other classes. Hence, the correlation matrix M between the k learnt label embeddings $\mathbf{E}_{y_j}$ and the k class representation proxies $\mathbf{H}_{y_i}$ is computed, where $0 \leq j,i \leq k-1$, with cosine similarity as the metric of correlation:

$$M_{ij} = \Phi(\mathbf{H}_{y_i}, \mathbf{E}_{y_j}) \tag{3.10}$$

For each class representation proxy, its correlation scores with the k learnt label embeddings are normalized to a range of [0, 1]. The average of values at the main diagonal of M is adopted as an overall measure of the quality of the learnt label embeddings:

$$\mathscr{LEQ} = \frac{1}{K}\sum_{i=0}^{K-1}\widetilde{M}_{ii} \tag{3.11}$$

Figure 3.2 shows the above method of intrinsic quality estimation for learnt label emebeddings.

	$\mathbf{E_1}$	$\mathbf{E_2}$	$\mathbf{E_k}$
$\mathbf{H_1}$	$\cos(E_1, H_1)$	$\cos(E_2, H_1)$	$\cos(E_k, H_1)$
$\mathbf{H_2}$	$\cos(E_1, H_2)$	$\cos(E_2, H_2)$	$\cos(E_k, H_2)$
$\mathbf{H_k}$	$\cos(E_1, H_k)$	$\cos(E_2, H_k)$	$\cos(E_k, H_k)$

Figure 3.2: Illustration of the correlation matrix M. $\mathbf{E}_{1...k}$ represents the k learnt label embeddings and $\mathbf{H}_{1...k}$ denotes the k class representation proxies. After normalization, the average of the values at the diagonal (colored) is the overall measure of the quality of the learnt label embeddings.

3.5 Experiments

3.5.1 Data Preprocessing

For the purpose of the research, it would be ideal to learn label embeddings for all the relations. However, label embeddings are trained together with input representations in a multi-class classification task, and data imbalance poses a challenge. Therefore, the experiments focus on 16 relations for RST and PDTB L2 senses with more than 100 instances, following the suggestions of Kim et al. (2020).

RST trees in RST-DT are binarized based on the procedure in Ji and Eisenstein (2014) and span pairs and relations are extracted. The 78 relations are mapped to 16 classes based on the processing step in Braud et al. (2016)[2]. 20% of the training set of RST-DT is taken for validation purposes.

For PDTB, sections 2-20 are used as the training set, sections 0-1 as the development set, and sections 21-22 as the test set, following Ji and Eisenstein (2015).

3.5.2 Hyperparameters and Training

Each model is run three times with different random seeds, and the mean and standard deviation of the results are reported. The AdamW optimizer (Loshchilov and Hutter, 2019) is used and L2 norm of gradients is clipped to 1.0. The learning

[2] https://bitbucket.org/chloebt/discourse/src/master/preprocess_rst/code/src/relationSet.py

rate is set to $1e-5$, and the batch size is set to the maximum that the GPU device can accommodate. The total training epoch is set to 10 and early stopping is adopted, with patience of 6 epochs on performance improvement on the validation set.

The temperature τ for instance-centered contrastive loss and label-centered contrastive loss is set to 0.1 based on empirical results. For the experiment with *LbEncHier* label encoder, the penalty factor is $2^{1/2}$ for L1 loss and 2 for L2 loss, which is set based on layers on the sense hierarchy.

The learning rate for the baseline *BertForSequenceClassification* model is set to $5e-5$ based on manual tuning.

The implementation is based on the PyTorch framework (Paszke et al., 2019) and a single 12GB RTX3060 GPU is used for all the experiments.

3.5.3 Details for Label Encoder Configuration

As has been noted in section 3.4, different ways of setting up label encoders are explored in the experiments. The approach with *LbEncDesc* requires textual description of the sense labels, and the details for PDTB are shown in Table 3.1.

Labels	**Description**
Synchronous	temporal overlap.
Asynchronous	one event preceding the other.
Cause	causally influenced but are not in a conditional relation.
Cause+Belief	evidence is provided to cause the hearer to believe a claim.
Cause+SpeechAct	a reason is provided for the speaker uttering a speech act.
Condition	one argument presents a situation as unrealized (the antecedent), which (when realized) would lead to the situation described by the other argument (the consequent).
Condition+SpeechAct	the consequent is an implicit speech act.

Negative-condition	one argument (the antecedent) describes a situation presented as unrealized, which if it doesn't occur, would lead to the situation described by the other argument (the consequent).
Purpose	one argument presents an action that an AGENT undertakes with the purpose of the GOAL conveyed by the other argument being achieved.
Concession	an expected causal relation is cancelled or denied by the situation described in one of the arguments.
Concession+SpeechAct	the speech act associated with one argument is cancelled or denied by the other argument or its speech act.
Contrast	at least two differences between Arg1 and Arg2 are highlighted.
Similarity	one or more similarities between Arg1 and Arg2 are highlighted with respect to what each argument predicates as a whole or to some entities it mentions.
Conjunction	when both arguments bear the same relation to some other situation evoked in the discourse.
Disjunction	the two arguments are presented as alternatives, with either one or both holding.
Equivalence	both arguments are taken to describe the same situation, but from different perspectives.
Exception	one argument evokes a set of circumstances in which the described situation holds, and the other argument indicates one or more instances where it doesn't.
Instantiation	one argument describes a situation as holding in a set of circumstances, while the other argument describes one or more of those circumstances.
Level-of-detail	both arguments describe the same situation, but in less or more detail.
Manner	presents the manner in which the situation described by other argument has happened or been done.
Substitution	exclusive alternatives, with one being ruled out.

Table 3.1: Textual description of PDTB L2 labels, which is mostly taken from the annotation manual of PDTB 3.0 (Webber et al., 2019). The labels and their descriptions are prepared in the form *[CLS] synchronous [SEP] temporal overlap [SEP]*, for instance, before being fed to the label encoder.

Details of textual descriptions for RST labels are shown in Table 3.2.

Labels	**Description**
Background	or circumstance
Cause	cause, result, consequence
Comparison	comparison, preference, analogy, proportion
Condition	condition, hypothetical, contingency, otherwise
Contrast	contrast, concession, antithesis
Elaboration	additional, general-specific, part-whole, process-step, object-attribute, set-member, example, definition
Enablement	purpose, enablement
Evaluation	interpretation, conclusion, comment
Explanation	evidence, explanation-argumentative, reason
Joint	list, disjunction
Manner-Means	manner, means
Topic-Comment	problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
Summary	summary, restatement
Temporal	before, after, same-time, sequence, inverted-sequence
Topic-Change	topic-shift, topic-drift
Textual-organization	used to link elements of the structure of the text

Table 3.2: Textual description of the 16 classes in RST-DT, which is mainly formed by the more detailed relations that each broad class contains. The labels and their textual descriptions are prepared in the same format as PDTB. However, since a broad class may encompass multiple fine-grained relations that differ in certain ways, providing descriptions for broad classes is more challenging compared to PDTB L2 senses, for which definitions are available from the annotation guidelines (Webber et al., 2019). In comparison, in the annotation guidelines of RST-DT (Carlson and Marcu, 2001), definitions of discourse relations are provided for the 78 fine-grained relations. As labels of these fine-grained relations are largely descriptive, detailed explanations, similar to those given for PDTB, are not provided for the 16 classes.

For the label encoder *LbEncHier*, sense hierarchy is required. Different from PDTB, RST does not have a three-level sense hierarchy. In the experiments, due to the small size of RST-DT, using the 78 end labels may aggravate data sparsity, and therefore, L1 and L2 sense labels are used in this set of experiments. As indicated in Chiarcos (2014), the top level of the PDTB sense hierarchy is the most coarse-grained classification of coherence relations applied to annotated data. Therefore,

PDTB L1 sense labels may be applicable to annotate RST data. Table 3.3 shows the corresponding PDTB L1 sense labels for the 16 RST relations.

RST Relations	**PDTB L1 Labels**
Contrast	Comparison
Textual-Organization	RST-Specific
Manner-Means	Expansion
Cause	Contingency
Explanation	Expansion
Evaluation	Expansion
Background	Expansion
Condition	Contingency
Elaboration	Expansion
Enablement	Contingency
Summary	Expansion
Topic-Comment	Expansion
Topic-Change	Comparison
Joint	Expansion
Temporal	Temporal
Comparison	Comparison

Table 3.3: PDTB L1 labels for the 16 RST relations. The PDTB L1 labels for RST-DT relations are determined with reference to Pu et al. (2023). As *Textual-Organization* is specific to RST, the L1 sense is denoted by *RST-Specific*.

3.5.4 Results

Since only small discrepancies are observed in data distributions between the training and test sets, the test set is chosen for generating class representation proxies necessary for the computation of the metric.

Table 3.4 shows the experimental results for PDTB and RST, including the classification accuracy (Acc.) and F1 scores of discourse relations, as well as the scores of the corresponding label embeddings (Label emb.) for different options of label encoders (Label enc.), computed based on the method shown in equation 3.11. Even though the focus is on label embedding learning, the performance on the classification task may provide useful insights for interpreting the results. Note that PDTB explicit and implicit relation data are combined. After the preprocessing step, 16 relations remain for both PDTB and RST.

It can be observed that the performance of label embedding learning on RST is lower than PDTB. Moreover, adding label embeddings generally lowers F1 compared with training with cross-entropy loss only. The decrease in F1 might

Data	Label enc.	Acc.	F1	Label emb.
PDTB total	*LbEncBert*	69.45(± 0.18)	57.80(± 0.85)	93.84(± 0.37)
	LbEncRoberta	69.34(± 0.46)	58.10(± 0.15)	**94.23(± 0.74)**
	LbEncRand	**69.87(± 0.80)**	**59.00(± 0.62)**	89.32(± 0.01)
	LbEncDesc	69.16(± 0.26)	57.53(± 0.14)	93.58(± 0.42)
	LbEncHier	69.21(± 0.45)	56.70(± 0.14)	93.67(± 0.23)
	Baseline	69.42(± 0.46)	58.73(± 0.78)	79.15(± 2.06)
RST	*LbEncBert*	64.62(± 0.90)	44.86(± 1.85)	**78.64(± 1.02)**
	LbEncRoberta	**65.20(± 0.07)**	45.39(± 0.60)	76.56(± 0.85)
	LbEncRand	65.09(± 0.70)	45.53(± 4.82)	69.98(± 3.10)
	LbEncDesc	64.62(± 0.21)	43.69(± 1.20)	74.18(± 0.91)
	LbEncHier	63.66(± 0.50)	41.30(± 0.39)	74.54(± 0.77)
	Baseline	63.55(± 0.23)	**48.57(± 0.73)**	48.21(± 1.27)

Table 3.4: Results over three runs are collected. The Pearson correlation coefficient between classification accuracy and label embedding scores is 0.5814 and it is 0.8187 between f1 and label embedding scores, both with $p < 0.05$, which shows that the learnt label embeddings are closely related to F1 scores.

be related to data sparsity when more learning objectives are added but the data amount is the same, which is visible when supplementary information of labels is added, as shown by the cases of *LbEncDesc* and *LbEncHier*. This phenomenon is rather pronounced for RST, which has a much smaller data amount. Additionally, although the label encoder *LbEncRand* works best for the classification task, the learnt label embeddings rank the lowest among the different setups. Further examination suggests that with this approach, the label embeddings of different classes are not close to the class representation proxies. It may be the case that during training, the label embeddings are mainly used as anchors, as in Zhang et al. (2022b), but the input representations are better learnt, hence the higher classification accuracy and F1 score. Zhang et al. (2022b) did not report results using other options of label encoders besides random initialization, and their focus was on classification accuracy.

3.5.5 Data Augmentation for RST

To improve the performance on RST, data augmentation is performed. Back translation (BT) is a well-studied data augmentation technique in NLP (Yaseen and Langer, 2021; Feng et al., 2021). To create more training data, texts in one language are translated into another language and then translated back into the original language. With an intermediate translation step, linguistic variation is introduced, yielding new instances that preserve the original context and meaning (Beddiar et al., 2021). If machine translation (MT) systems are well-trained on the language

pair involved in the intermediate translation step, the quality of the generated data may be high. As French is widely used by international organizations, the Europarl corpus (Koehn, 2005), a large parallel corpus commonly used for training MT systems, contains a large amount of French data. Therefore, French is chosen as the intermediate language. However, other languages are potentially valid choices, which is subject to empirical investigation. All the files containing EDUs in the training set (only) are translated into French and the French texts are translated back into English, with Google Translate[3]. As machine translation is performed at the EDU level, which is influenced by syntactic rules, syntactic differences between English and French make errors inevitable in the BT step. Future research may be conducted to propose more sophisticated data augmentation techniques, such as combining paraphrasing with BT to introduce greater linguistic variations (Beddiar et al., 2021). The choice of the intermediate language and MT system does not preclude the other valid options. Note that data augmentation is not performed for *Elaboration* and *Joint*, which are the two largest classes in RST-DT, to achieve a more balanced data distribution.

Based on the results shown in Table 3.4, *LbEncRoberta* is chosen in the following experiments because of its good performance but results with *LbEncBert* are comparable.

Table 3.5 shows the results. The F1 scores and label embedding scores are improved to a large margin. As back translation is performed at the EDU level, it is unavoidable that errors are introduced, and given that data augmentation is not performed for the two largest classes, their influence on the results is reduced, hence the lower classification accuracy.

	Acc.	**F1**	**Label emb.**
+aug.	62.75(± 0.79)	50.76(± 0.94)	92.96(± 0.90)
-aug.	65.20(± 0.07)	45.39(± 0.60)	76.56(± 0.85)

Table 3.5: Results for RST with data augmentation (+aug) and without data augmentation (-aug).

Figure 3.3 and Figure 3.4 show the T-SNE visualization plots of the learnt label embeddings together with the class representation proxies for the test set of RST-DT. The label embeddings learnt with data augmentation are shown in Figure 3.3 in comparison with Figure 3.4, where no data augmentation is performed. It is visible

[3] https://translate.google.com/

that in Figure 3.3, more label embeddings fit into the class representation proxies while in Figure 3.4, label embeddings of only six classes are close to the class representation proxies, and the rest form a nebula, which suggests that the label embeddings cannot be distinguished clearly from each other. In Figure 3.3, label embeddings for five relations including *Explanation*, *Textual-Organization*, *Topic-Comment*, *Evaluation* and *Topic-Change* show such behavior. *Textual-Organization*, *Topic-Comment*, and *Topic-Change* are classes with a small amount of data and it is difficult to obtain good performance on these classes in a classification task. The reasons for *Explanation* and *Evaluation* are not clear.

3.5.6 Separate Experiments on PDTB Explicit and Implicit Relations

Previous studies (Demberg et al., 2019; Sanders et al., 2018) indicate that it is much easier to obtain consistent results on aligning PDTB explicit relations with relations from the other frameworks, while implicit relations are generally ambiguous and the consistency is much lower. Therefore, experiments are also performed on PDTB explicit and implicit relations separately. *LbEncRoberta* is used in this set of experiments. After the data preprocessing step outlined in section 3.5.1, 12 explicit relations and 14 implicit relations remain in the experiments.

Data	Acc.	F1	Label emb.
explicit	88.98(± 0.41)	79.19(± 0.64)	99.15(± 0.60)
implicit	56.05(± 0.56)	40.56(± 0.81)	82.21(± 0.85)

Table 3.6: Experimental results on PDTB explicit relations and implicit relations.

The classification results and label embedding learning results indicate that the learnt label embeddings for PDTB explicit relations are representative of the classes while the performance on implicit relations is sub-optimal.

3.5.7 Ablation Study

In this set of experiments, *LbEncRoberta* is chosen, and ablation studies are performed on the combination of PDTB explicit and implicit relation data, similar to the experimental settings in Table 3.4. The impact of each loss can be seen in Table 3.7.

As shown, the label-centered contrastive loss ($\mathscr{L}_{LCL}$) is of paramount importance for the model's performance, followed by the instance-centered contrastive loss

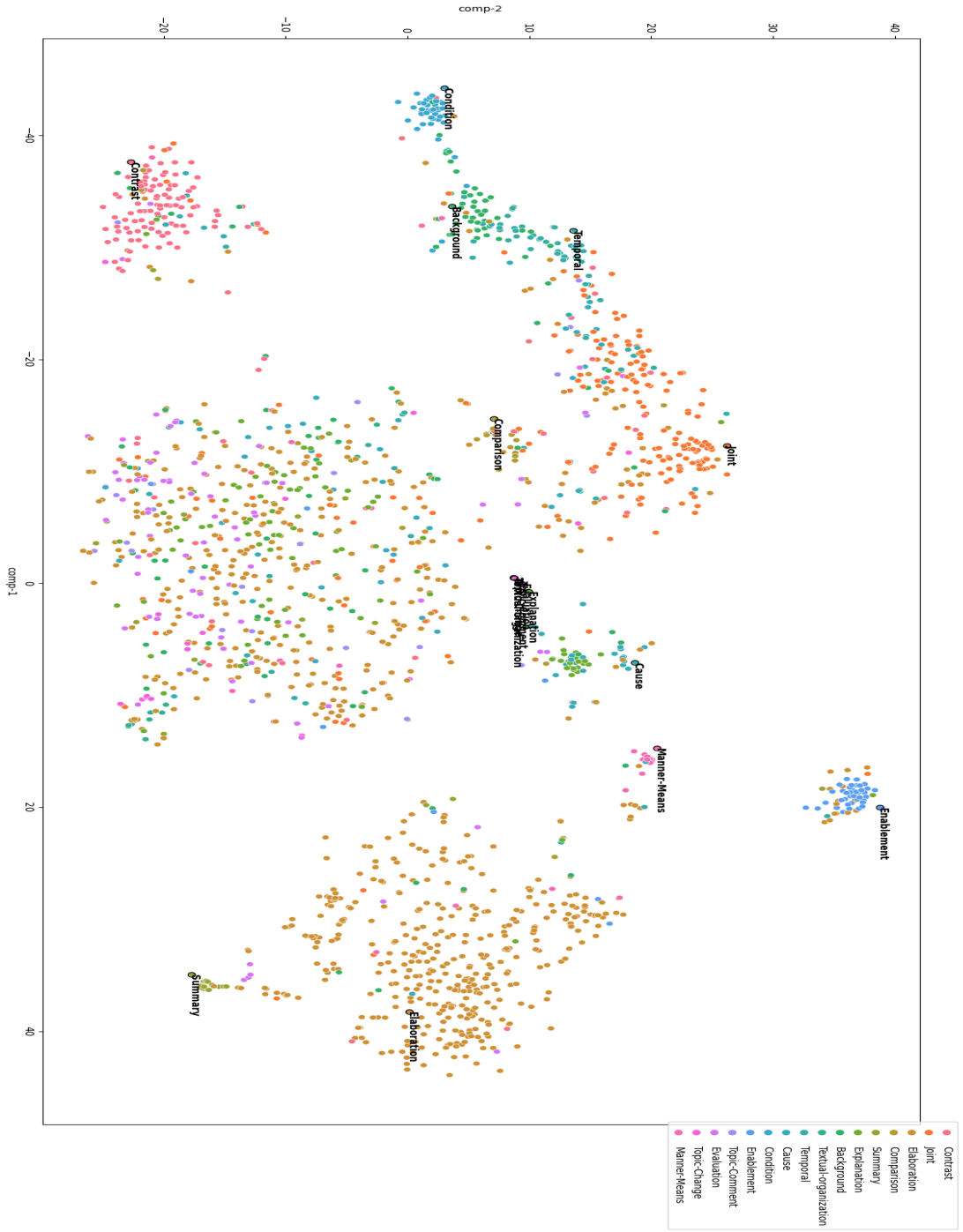


Figure 3.3: Label embeddings learnt with data augmentation.

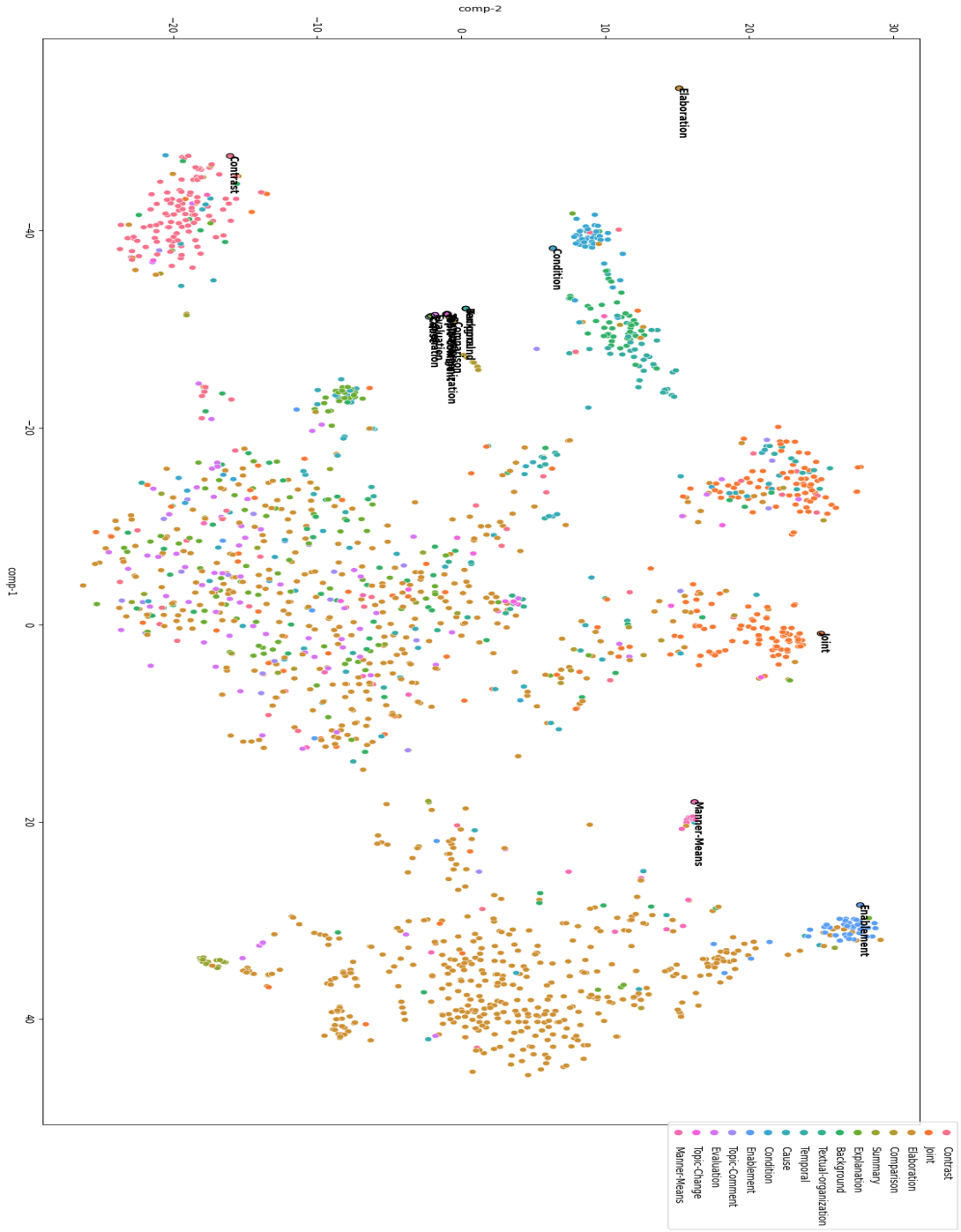


Figure 3.4: Label embeddings learnt without data augmentation. For visualization, the label embeddings with the highest score from the three runs are selected.

Loss	Acc.	F1	Label emb.
-$\mathscr{L}_{ICL}$	68.22(± 0.44)	53.65(± 1.13)	91.36(± 0.73)
-$\mathscr{L}_{LCL}$	65.02(± 0.47)	51.23(± 1.62)	80.37(± 1.42)
-$\mathscr{L}_{LEC}$	69.32(± 0.30)	57.57(± 0.87)	**94.36(± 0.37)**
-$\mathscr{L}_{ICE}$	**69.88(± 0.09)**	56.94(± 0.36)	90.79(± 0.76)
Total	69.34(± 0.46)	**58.10(± 0.15)**	94.23(± 0.74)

Table 3.7: Effect of each loss on model performance.

($\mathscr{L}_{ICL}$) and canonical cross-entropy loss ($\mathscr{L}_{ICE}$). This differs from the findings in Zhang et al. (2022b), where $\mathscr{L}_{ICL}$ is the primary contributing factor to their results, indicating the distinct nature of their task and the task of this chapter. $\mathscr{L}_{LEC}$ has some effect on F1 score of the classification task.

3.6 RST-PDTB Relation Mapping

3.6.1 Mapping Results

Table 3.8 shows the mapping results of 11 RST relations, with the five relations discussed in section 3.5.5 excluded, and 12 PDTB explicit relations discussed in section 3.5.6. Two relations with the highest values in cosine similarity (greater than 0.10) are presented.

The table on the left shows the mapping results from RST's perspective. For most of the RST relations, a PDTB relation can be identified as having a much higher value than the others. However, it is noticeable that the differences between the matched PDTB relations are small for RST relations *Contrast* and *Comparison*. RST's *Contrast* relation class subsumes *contrast*, *concession* and *antithesis*, which are also mapped mainly to these two PDTB relations in the statistics shown in Demberg et al. (2019, Table 3). RST's *Comparison* is defined as not involving contrastive elements in the annotation manual (Carlson and Marcu, 2001). Except for *Similarity*, no PDTB relations correspond to this relation[4]. These kinds of relations are innately difficult to align. In terms of definition, the RST *Summary* relation can be mapped to PDTB's *Level-of-Detail*, but in the results, the PDTB *Contrast* relation ranks the highest. Since PDTB's *Contrast* is meant to highlight at least two differences between contents presented in the arguments, it might be difficult for the model to capture its nature and RST's *Summary* relation generally spans multiple EDUs.

[4]*Similarity* is not included in the experiments on PDTB explicit relations because of small data amount.

RST	Relations in PDTB
Contrast	Concession(0.25), Contrast(0.24)
Manner-Means	Manner(0.30), Purpose(0.25)
Cause	Cause(0.40), Level-of-Detail(0.17)
Background	Synchronous(0.23), Manner(0.16)
Condition	Condition(0.39), Purpose(0.18)
Elaboration	Concession(0.19), Disjunction(0.14)
Enablement	Manner(0.24), Purpose(0.18)
Summary	Contrast(0.35), Level-of-Detail(0.23)
Joint	Disjunction(0.25), Synchronous(0.20)
Temporal	Asynchronous(0.24), Purpose(0.20)
Comparison	Purpose(0.17), Level-of-Detail(0.16)

PDTB	Relations in RST
Conjunction	Contrast(0.22), Elaboration(0.13)
Concession	Contrast(0.25), Elaboration(0.19)
Cause	Cause(0.40), Manner-Means(0.20)
Level-of-Detail	Manner-Means(0.25), Summary(0.23)
Synchronous	Background(0.23), Joint(0.20)
Disjunction	Joint(0.25), Temporal (0.16)
Manner	Manner-Means(0.30), Enablement(0.24)
Condition	Condition(0.39), Summary(0.15)
Substitution	Manner-Means(0.17), Summary(0.17)
Asynchronous	Temporal(0.24), Joint(0.19)
Contrast	Summary(0.35), Background(0.13)
Purpose	Manner-means(0.25), Temporal(0.20)

Table 3.8: Mapping between 11 RST relations and 12 PDTB explicit relations. The values in brackets represent cosine similarity scores.

The table on the right shows the mapping results from PDTB's perspective. As relation distributions are different, it is understandable that perspectives from RST and PDTB are not symmetric. For the alignment of PDTB relations onto RST relations, similar patterns can be found, but the differences between matched RST relations for *Level-of-Detail*, *Synchronous* and *Substitution* are not large, indicating ambiguity in the results. RST's *Manner* is less "goal-oriented", and it describes the way in which something is performed (Carlson and Marcu, 2001), which may be the reason for the higher similarity between PDTB's *Level-of-Detail* and RST's *Manner-Means*. The commonly used discourse connectives such as "as" and "when" may cause RST's *Background* to be similar to PDTB's *Synchronous* relation in the results, while RST's *Temporal* relation is mainly mapped to PDTB's *Asynchronous* relation. The PDTB's *Substitution* relation has no directly corresponding RST relations and similar to the case of RST's *Comparison*, it is difficult to align this relation with RST relations.

3.6.2 Extrinsic Evaluation

As indicated in Benamara and Taboada (2015), a way to test the proposed mapping method between frameworks is to merge annotated data based on the mapping results and check if the increased data amount leads to performance improvement. Thus, the obtained mapping results are compared with those provided by Costa et al. (2023), which is the most recent work on this topic, where the approach proposed in Demberg et al. (2019) is adopted but results are updated to PDTB 3.0. Since label embeddings learnt for PDTB explicit relations are more reliable, as shown in section 3.5.6, this set of experiments focus on the mapping between PDTB explicit relations and RST relations. Based on Table 3.8, PDTB's *Substitution* relation is excluded in the experiments, because no RST relations with higher similarity are observed, and 11 PDTB explicit relations are relabeled with RST labels based on Table 3.9. While the corresponding RST labels are chosen mostly based on the cosine similarity values shown in Table 3.8, distribution of relations is taken into account. For example, PDTB's *Conjunction* relation is not mapped to RST's *Contrast* relation but to *Elaboration*, because *Conjunction* is a large class in PDTB, similar to *Elaboration* in RST, and relabelling in this way may keep the label distribution of the training set close to the test set. Meanwhile, in preliminary experiments, mapping PDTB's *Contrast* relation to RST's *Summary* relation yields poor performance. Therefore, PDTB's *Contrast* is relabeled as RST's *Contrast* relation based on the results from RST's perspective.

Similarly, PDTB explicit relations are to be relabeled based on the mapping results shown in Costa et al. (2023, Table 5). As their results provide a mapping of 12 fine-grained RST relations (the taxonomy of 78 relations) and seven PDTB L2 relations, their results are not directly comparable. To overcome this issue, for a PDTB relation, if there are multiple mapped RST relations that fall under a broad class (based on the taxonomy of 16 relations), the corresponding RST relation from the 16 classes is chosen, and the average of the percentages for the mapped classes is taken as mapping strength, similar to cosine similarity used in the proposed method in this chapter. For instance, in their results, PDTB's *Concession* relation is mapped to *Contrast* (61.0%), *Antithesis* (84.0%), and *Concession* (88.0%), which are fine-grained relations under RST's *Contrast* relation, and the mapping strength is the average of the three percentages, i.e., 0.78.

Based on the results of the method proposed in this chapter, 14964 instances of PDTB explicit relations are relabeled, and with the results in Costa et al. (2023),

Original PDTB —Sense Labels	**RST Labels** —Proposed method	**RST Labels** —Costa et al. (2023)
Concession	Contrast (0.25)	Contrast (0.78)
Contrast	Contrast (0.24)	Contrast (0.26)
Conjunction	Elaboration (0.13)	Joint (0.84)
Manner	Manner-Means (0.30)	—
Cause	Cause (0.40)	Explanation (0.69)
Synchronous	Background (0.23)	Temporal (0.98)
Condition	Condition (0.39)	Condition (0.84)
Disjunction	Joint (0.25)	—
Asynchronous	Temporal (0.24)	Temporal (0.94)
Level-of-Detail	Manner-Means (0.25)	—
Purpose	Manner-Means (0.25)	—

Table 3.9: Relabelling rules of PDTB explicit relations. Similarity scores are shown in brackets.

13905 PDTB instances are relabeled.

Adding PDTB data to RST data causes a marked performance drop. The best result is obtained using an ensemble model, which is formed by a model trained with a target of minimizing a supervised contrastive loss, a model trained to minimize a label embedding loss, the label embeddings being randomly initialized, and a model that takes the input for relation classification. The outputs of the three models are averaged and used for model prediction, and a cross-entropy loss is to be minimized in addition to the supervised contrastive loss and label embedding loss. Figure 3.5 shows the model architecture. The losses in orange are to be minimized through model training. The input representations from the encoder are scaled by the similarity scores shown in Table 3.9 before being passed to the three modules. In the module for computing supervised contrastive loss, the input representations are fed to a linear layer and transformed to a lower dimensional space, and in the module for computing label embedding loss, the input representations are fed to a feed-forward network with LeakyReLU activation function, and the third module is formed by a simple linear layer.

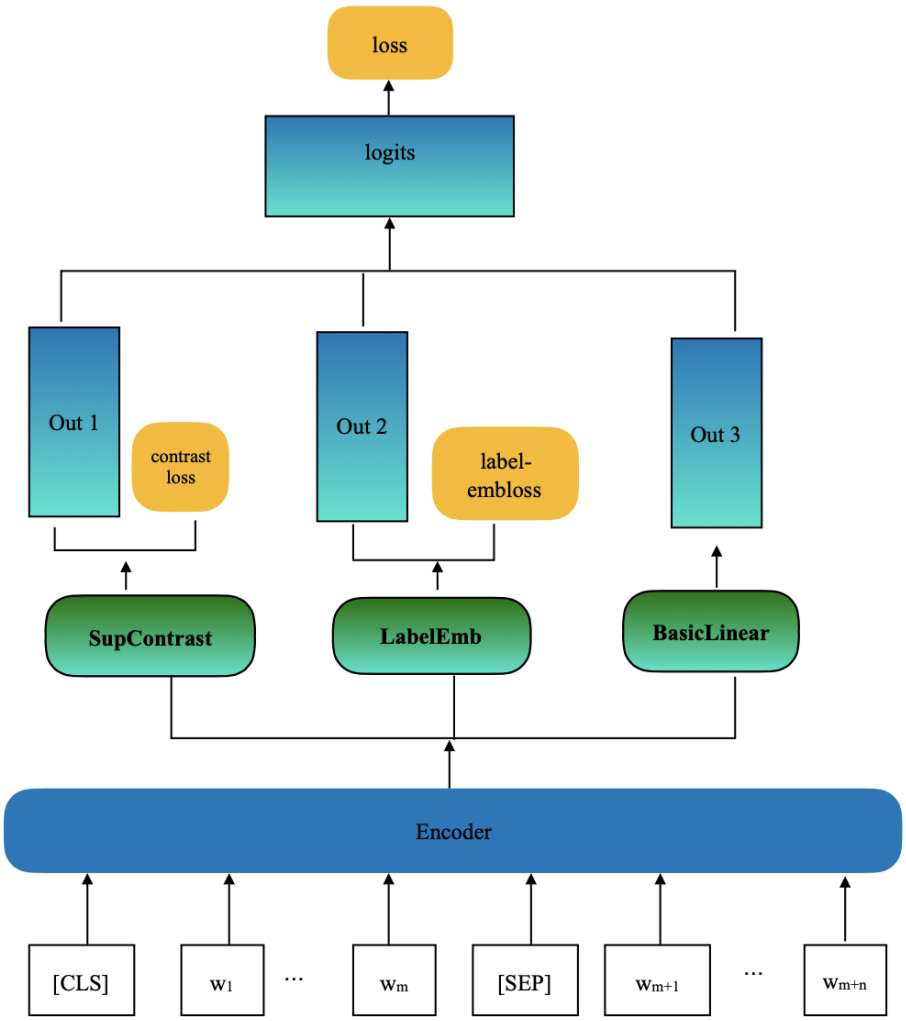


Figure 3.5: The ensemble model.

As shown in Table 3.10, the performance with the proposed method is slightly higher, although it is also noticeable that the standard deviation across three runs is larger than the results with Costa et al. (2023). Less PDTB data is used in the experiments with Costa et al. (2023). At higher levels of RST trees, spans can be rather large, which are typically not covered in PDTB-style annotation. This could be a reason for the performance drop compared to the case without using data augmentation with PDTB data.

	Acc.	F1
Costa et al. (2023)	62.13 ± 0.34	46.96±0.43
Proposed method	63.13 ± 1.12	47.95± 1.07
-PDTB aug.	63.82± 1.07	48.72± 0.11

Table 3.10: Results of extrinsic evaluation.

3.7 Interim Summary

A method for automatically aligning discourse relations of different frameworks is proposed in this chapter. By employing label embeddings that are learned concurrently with input representations during a classification task, it is possible to circumvent the challenges posed by segmentation differences, a significant hurdle encountered in prior studies. Intrinsic and extrinsic evaluations are performed. Similar to the other empirical studies, the proposed method is sensitive to the amount of data, and some relations are excluded because there is not enough training data for learning reliable label embeddings. The method may extend beyond mapping discourse relations of different frameworks to alignment of any label sets, leaving the possibility of application to a variety of scenarios, which is subject to further investigation in future work.

Chapter Four

*****

4.1 Chapter Overview

4.2 Interim Summary

Chapter Five

******

5.1 Chapter Overview

5.2 Interim Summary

CHAPTER SIX

CONCLUSIONS AND FUTURE WORK

6.1 Summary of Contributions

This section outlines the hypotheses, reiterates the research questions and discusses the approaches proposed to address them.

6.1.1 Hypotheses

The research of the thesis is based on the following hypotheses:

1. Despite superficial dissimilarities, discourse relations of different frameworks are related, making the relation taxonomies used by these frameworks alignable. By aligning the relation taxonomies effectively, data from different frameworks can be used together, offering a method to alleviate data scarcity in computational discourse processing.

2. Existing theoretical proposals for mapping discourse frameworks are useful for computational discourse processing, enabling data augmentation that goes beyond merely aligning relation taxonomies and relabeling data based on the alignment rules.

3. ******

To investigate these hypotheses, three research questions are proposed and presented in section 1.2. The first question concerns whether discourse relation

taxonomies employed by different frameworks can be aligned automatically. Existing studies primarily rely on theoretical analyses of discourse relations. Although Demberg et al. (2019) introduced a semi-automatic approach, differences in discourse segmentation continue to pose a challenge. A neural approach is not yet developed. To test the effectiveness of the alignment, data can be relabeled according to the alignment rules and then combined. If increasing the amount of data improves performance in computational experiments, the alignment is considered effective. Different methods for aligning relation taxonomies can be compared in this way.

The second research question deals with how existing theoretical proposals for aligning relation taxonomies across different frameworks can be leveraged in computational experiments, potentially leading to platform-agnostic approaches for discourse relation classification and novel methods for data augmentation across frameworks. The application of this line of research remains under-explored.

The third research question focuses on whether a new scheme for integrating hierarchical and local discourse representations is feasible while addressing some limitations of existing mainstream frameworks.

6.1.2 Proposed Approaches

In order to answer the research questions, three studies are performed.

6.1.2.1 A Neural Approach for Aligning Discourse Relations Across Different Frameworks

This approach is discussed in detail in Chapter 3. The method is based on label embedding techniques, where label embeddings are learned for relation sets adopted across different frameworks. The similarity of discourse relations is then automatically computed using cosine similarity scores of these embeddings. Unlike previous studies that employ label embeddings to enhance model performance in discourse relation classification, this method focuses on learning label embeddings that accurately represent discourse relations and encode the distances between them. Two contrastive learning objects are incorporated in model training for learning label embeddings: one aims to separate instances with different target labels, and the other seeks to minimize the distances between

instance representations and their corresponding label embeddings. The label-centered contrastive loss is more important for this task, different from existing studies that focus on discourse relation classification. While randomly initialized label embeddings are shown to enhance model performance on discourse relation classification in previous studies, the experimental results indicate that initializing label embeddings with pre-trained language models achieves superior performance in learning label embeddings.

A metric is proposed to measure the quality of the learnt label embeddings: class representation proxies, obtained by averaging instance representations for each class after model training, are compared with the learned label embeddings for those classes. The cosine similarity of the class representation proxy and the class label embeddings should be the highest among all.

The effectiveness of label embedding learning is influenced by the amount of data available. Thus, while the method does not require parallel corpora annotated with different frameworks, it necessitates large corpora to achieve optimal performance.

Through extrinsic evaluation, the method demonstrates a modest improvement over the state-of-the-art approach based on Demberg et al. (2019). It is worth mentioning that adding PDTB data to RST causes a performance drop. With the proposed approach, the step of aligning discourse segments can be bypassed, and thus, more discourse relations are aligned. This could be a reason for the small increase compared with the SOTA method. Further investigation is needed to explore more effective usage of the learned alignment.

Moreover, discourse relation classification is influenced by the context (Liu and Zeldes, 2023). Similar to existing studies on this question, this aspect is not considered in the research and the results represent general patterns of discourse relation alignment across different frameworks. Admittedly, more fine-grained results on relation alignment may improve the performance in extrinsic evaluation, but it is foreseeable that to learn such patterns automatically, a larger amount of data is needed.

6.1.2.2 A Computational Approach for Applying a Theoretical Proposal for Mapping Discourse Relations Across Different Frameworks

6.1.2.3 ******

6.2 Outstanding Issues and Future Work

Based on the summary above, certain challenges in existing studies and proposed approaches are readily identifiable. The following discussion may reiterate some of the points mentioned earlier.

Alignment of Discourse Relations The proposed method still requires a large amount of data to obtain reliable results. An even greater amount of data is needed to learn fine-grained alignment of discourse relations across different frameworks. Moreover, compared with methods based on string matching, the results are not straightforward to interpret or verify, which is common for neural approaches. In future work, synthetic data generated from LLMs may be used to increase the data amount, although other challenges may exist with this approach. Meanwhile, more research is needed to investigate efficient ways of applying the results of discourse relation alignment.

Platform-Agnostic Discourse Relation Classification Since different frameworks have varying assumptions about discourse structure, a common task across these frameworks is discourse relation classification, which now has a benchmark provided by Braud et al. (2024). In future work, studies can be performed on this benchmark in order to be comparable.

As the dimensions in the UniDim proposal are conceptually simpler, large-scale annotation of these dimensions is potentially feasible with LLMs. If sufficient data can be collected, more experiments can be conducted on building a universal classifier for discourse relations across different frameworks. However, the success of this approach calls for further studies on the CCR framework so that the dimensions can be mapped to discourse relations unambiguously.

Apart from the UniDim proposal, experiments can be conducted on the application of other theoretical proposals. Although the results of the UniDim proposal are closer to the empirical findings reported by Demberg et al. (2019), comparing the major theoretical proposals in terms of their utility in computational settings may be a beneficial supplement to existing research.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Afantenos, S. and Asher, N. (2010). Testing SDRT's right frontier. In Huang, C.-R. and Jurafsky, D., editors, *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1–9, Beijing, China. Coling 2010 Organizing Committee.

Afantenos, S., Asher, N., Benamara, F., Bras, M., Fabre, C., Ho-dac, M., Draoulec, A. L., Muller, P., Péry-Woodley, M.-P., Prévot, L., Rebeyrolles, J., Tanguy, L., Vergez-Couret, M., and Vieu, L. (2012). An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2727–2734, Istanbul, Turkey. European Language Resources Association (ELRA).

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016). Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438.

Asher, N., Hunter, J., Morey, M., Farah, B., and Afantenos, S. (2016). Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Asher, N. and Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Asher, N., Muller, P., Bras, M., Ho-Dac, L. M., Benamara, F., Afantenos, S., and Vieu, L. (2017). ANNODIS and related projects: case studies on the annotation of discourse structure. In *Handbook of Linguistic Annotation*, pages 1241–1264. Springer.

Atwell, K., Choi, R., Li, J. J., and Alikhani, M. (2022). The role of context and uncertainty in shallow discourse parsing. 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 797–811, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Barzilay, R. and Lapata, M. (2005). Modeling local coherence: An entity-based approach. In Knight, K., Ng, H. T., and Oflazer, K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Beddiar, D. R., Jahan, M. S., and Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Benamara, F. and Taboada, M. (2015). Mapping different rhetorical relation annotations: A proposal. In Palmer, M., Boleda, G., and Rosso, P., editors, *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, Colorado. 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.

Bosselut, A., Celikyilmaz, A., He, X., Gao, J., Huang, P.-S., and Choi, Y. (2018). Discourse-aware neural rewards for coherent text generation. In Walker, M.,

Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Bourgonje, P. and Stede, M. (2020). The Potsdam commentary corpus 2.2: Extending annotations for shallow discourse parsing. 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 1061–1066, Marseille, France. European Language Resources Association.

Bourgonje, P. and Zolotarenko, O. (2019). Toward cross-theory discourse relation annotation. 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 7–11, Minneapolis, MN. Association for Computational Linguistics.

Braud, C., Liu, Y. J., Metheniti, E., Muller, P., Rivière, L., Rutherford, A., and Zeldes, A. (2023). The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Braud, C., Liu, Y. J., Metheniti, E., Muller, P., Rivière, L., Rutherford, A., and Zeldes, A., editors, *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Braud, C., Plank, B., and Søgaard, A. (2016). Multi-view and multi-task training of RST discourse parsers. In Matsumoto, Y. and Prasad, R., editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Braud, C., Zeldes, A., Rivière, L., Liu, Y. J., Muller, P., Sileo, D., and Aoyama, T. (2024). DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. 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 4990–5005, Torino, Italia. ELRA and ICCL.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bunt, H. and Prasad, R. (2016). ISO DR-Core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54(2001):56.

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*.

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.

Chatterjee, S., Maheshwari, A., Ramakrishnan, G., and Jagarlapudi, S. N. (2021). Joint learning of hyperbolic label embeddings for hierarchical multi-label classification. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2829–2841, Online. Association for Computational Linguistics.

Chiarcos, C. (2014). Towards interoperable discourse annotation. discourse features in the ontologies of linguistic annotation. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chinchor, N. and Sundheim, B. (2003). Message understanding conference (muc) 6 ldc2003t13. https://catalog.ldc.upenn.edu/LDC2003T13.

Cortes, C. (1995). Support-vector networks. *Machine Learning*.

Costa, N. F., Sheikh, N., and Kosseim, L. (2023). Mapping explicit and implicit discourse relations between the RST-DT and the PDTB 3.0. In Mitkov, R. and Angelova, G., editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 344–352, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Cristea, D., Ide, N., and Romary, L. (1998). Veins theory: A model of global discourse cohesion and coherence. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 281–285, Montreal, Quebec, Canada. 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 Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Danlos, L., Antolinos-Basso, D., Braud, C., and Roze, C. (2012). Vers le FDTB : French discourse tree bank (towards the FDTB : French discourse tree bank) [in French]. In Antoniadis, G., Blanchon, H., and Sérasset, G., editors, *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 471–478, Grenoble, France. ATALA/AFCP.

Das, D. and Stede, M. (2018). Developing the Bangla RST Discourse Treebank. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Das, D. and Taboada, M. (2018). Rst signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149–184.

De Kuthy, K., Reiter, N., and Riester, A. (2018). QUD-based annotation of discourse structure and information structure: Tool and evaluation. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources*

and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Demberg, V., Scholman, M. C., and Asr, F. T. (2019). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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.

Dozat, T. and Manning, C. D. (2018). Simpler but more accurate semantic dependency parsing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Egg, M. and Redeker, G. (2010). How complex is discourse structure? 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).

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for NLP. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Feng, V. W. and Hirst, G. (2012). Text-level discourse parsing with rich linguistic features. In Li, H., Lin, C.-Y., Osborne, M., Lee, G. G., and Park, J. C., editors, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68, Jeju Island, Korea. Association for Computational Linguistics.

Feng, V. W., Lin, Z., and Hirst, G. (2014). The impact of deep hierarchical discourse structures in the evaluation of text coherence. In Tsujii, J. and

Hajic, J., editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Fetzer, A. (2014). Conceptualising discourse. *Pragmatics of discourse*, pages 35–61.

Frank, A. F. and Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30.

Fu, Y. (2022). Towards unification of discourse annotation frameworks. In Louvan, S., Madotto, A., and Madureira, B., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. Association for Computational Linguistics.

Fu, Y. (2023). Discourse relations classification and cross-framework discourse relation classification through the lens of cognitive dimensions: An empirical investigation. In Abbas, M. and Freihat, A. A., editors, *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 21–42, Online. Association for Computational Linguistics.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

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*.

Guz, G., Huber, P., and Carenini, G. (2020). Unleashing the power of neural discourse parsers - a context and structure aware approach using large scale pretraining. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English (1st ed.)*. Routledge.

Hernault, H., Prendinger, H., du Verle, D. A., and Ishizuka, M. (2010). HILDA: a discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., and Nagata, M. (2013). Single-document summarization as a tree knapsack problem. 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 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

Hobbs, J. R. (1979). Coherence and coreference. *Cognitive science*, 3(1):67–90.

Hobbs, J. R. et al. (1985). *On the coherence and structure of discourse*, volume 208. CSLI Stanford, CA.

Hovy, E. H. (1990). Parsimonious and profligate approaches to the question of discourse structure relations. In McKeown, K. R., Moore, J. D., and Nirenburg, S., editors, *Proceedings of the Fifth International Workshop on Natural Language Generation*, Linden Hall Conference Center, Dawson, Pennsylvania. Association for Computational Linguistics.

Hovy, E. H. and Maier, E. (1992). Parsimonious or profligate: how many and which discourse structure relations? Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST.

Huang, Y. J. and Kurohashi, S. (2021). Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *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.

Huber, P. and Carenini, G. (2019). Predicting discourse structure using distant supervision from sentiment. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Huber, P. and Carenini, G. (2020). MEGA RST discourse treebanks with structure and nuclearity from scalable distant sentiment supervision. In Webber, B.,

Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7442–7457, Online. Association for Computational Linguistics.

Huber, P. and Carenini, G. (2022). Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2376–2394, Seattle, United States. Association for Computational Linguistics.

Huber, P., Xing, L., and Carenini, G. (2022). Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10794–10802.

Iruskieta, M., Aranzabe, M. J., de Ilarraza, A. D., Gonzalez, I., Lersundi, M., and de Lacalle, O. L. (2013). The RST basque treebank: an online search interface to check rhetorical relations. In *4th Workshop RST and Discourse Studies*, pages 40–49.

Iter, D., Guu, K., Lansing, L., and Jurafsky, D. (2020). Pretraining with contrastive sentence objectives improves discourse performance of language models. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4859–4870, Online. Association for Computational Linguistics.

Ji, Y. and Eisenstein, J. (2014). Representation learning for text-level discourse parsing. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. 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.

Joshi, A. K. (1987). An introduction to tree adjoining grammars. *Mathematics of Language*, 1:87–115.

Joshi, A. K. and Schabes, Y. (1991). Tree-adjoining grammars and lexicalized grammars.

Joty, S., Carenini, G., Ng, R., and Mehdad, Y. (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In Schuetze, H., Fung, P., and Poesio, M., editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria. Association for Computational Linguistics.

Joty, S., Carenini, G., and Ng, R. T. (2015). CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Jurafsky, D. and Martin, J. H. (2023). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Kamalloo, E., Dziri, N., Clarke, C., and Rafiei, D. (2023). Evaluating open-domain question answering in the era of large language models. 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 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic: Introduction to Model theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Dordrecht: Kluwer Academic Publishers.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

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.

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.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *1995 international conference on acoustics, speech, and signal processing*, volume 1, pages 181–184. IEEE.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2000). Beyond elaboration: The interaction of relations and focus in coherent text. In *Text Representation: Linguistic and Psycholinguistic Aspects, chapter 7*, pages 181–196. John Benjamins.

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.

Kobayashi, N., Hirao, T., Kamigaito, H., Okumura, M., and Nagata, M. (2020). Top-down RST parsing utilizing granularity levels in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8099–8106.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Koto, F., Lau, J. H., and Baldwin, T. (2021). Top-down discourse parsing via sequence labelling. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.

Lafferty, J., McCallum, A., Pereira, F., et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, volume 1, page 3. Williamstown, MA.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., and Webber, B. (2006). Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax. In *Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories*, pages 12–23. Citeseer.

Lee, A., Prasad, R., Joshi, A., and Webber, B. (2008). Departures from tree structures in discourse: Shared arguments in the penn discourse treebank. In *Proceedings of the constraints in discourse iii workshop*, pages 61–68.

Lee, H., Hudson, D. A., Lee, K., and Manning, C. D. (2020). SLM: Learning a discourse language representation with sentence unshuffling. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562, Online. Association for Computational Linguistics.

Lei, W., Miao, Y., Xie, R., Webber, B., Liu, M., Chua, T.-S., and Chen, N. F. (2021). Have we solved the hard problem? it's not easy! contextual lexical contrast as a means to probe neural coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13208–13216.

Li, J., Liu, M., Kan, M.-Y., Zheng, Z., Wang, Z., Lei, W., Liu, T., and Qin, B. (2020). Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Li, S., Wang, L., Cao, Z., and Li, W. (2014). Text-level discourse dependency parsing. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2011). Automatically evaluating text coherence using discourse relations. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006, Portland, Oregon, USA. Association for Computational Linguistics.

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, W. and Strube, M. (2023). Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. 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 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Liu, Y., Li, S., Zhang, X., and Sui, Z. (2016). Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2020). RoBERTa: A robustly optimized BERT pretraining approach.

Liu, Y. J., Aoyama, T., Scivetti, W., Zhu, Y., Behzad, S., Levine, L. E., Lin, J., Tiwari, D., and Zeldes, A. (2024). GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

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 Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

Liu, Y. J. and Zeldes, A. (2023). Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. 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.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: Description and construction of text structures. In *Natural language generation: New results in artificial intelligence, psychology and linguistics*, pages 85–95. Springer.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, D. (1996). Building up rhetorical structure trees. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.

Marcu, D. (1997). The rhetorical parsing of unrestricted natural language texts. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–103, Madrid, Spain. Association for Computational Linguistics.

Marcu, D. (1998). To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 365–372, College Park, Maryland, USA. Association for Computational Linguistics.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. MIT press.

Marcu, D. (2003). Discourse structures: trees or graphs? **https://www.isi.edu/~marcu/discourse/Discourse%20structures.htm**. Accessed: 2024, Jan 08.

Marcu, D., Amorrortu, E., and Romera, M. (1999). Experiments in constructing a corpus of discourse trees. In *Towards Standards and Tools for Discourse Tagging*.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mesgar, M. and Strube, M. (2016). Lexical coherence graph modeling using word embeddings. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, pages 1414–1423, San Diego, California. Association for Computational Linguistics.

Miyazaki, T., Makino, K., Takei, Y., Okamoto, H., and Goto, J. (2019). Label embedding using hierarchical structure of labels for Twitter classification. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6317–6322, Hong Kong, China. Association for Computational Linguistics.

Moore, J. D. and Pollack, M. E. (1992). A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Morey, M., Muller, P., and Asher, N. (2018). A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Moser, M. and Moore, J. D. (1996). Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Murray, J. D. (1997). Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25:227–236.

Narasimhan, K. and Barzilay, R. (2015). Machine comprehension with discourse relations. In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262, Beijing, China. Association for Computational Linguistics.

Onea, E. (2016). Potential questions at the semantics-pragmatics interface. In *Potential Questions at the Semantics-Pragmatics Interface*. Brill.

Oza, U., Prasad, R., Kolachina, S., Misra Sharma, D., and Joshi, A. (2009). The Hindi discourse relation bank. In Stede, M., Huang, C.-R., Ide, N., and Meyers, A., editors, *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161, Suntec, Singapore. Association for Computational Linguistics.

Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

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.

Pitler, E., Raghupathy, M., Mehta, H., Nenkova, A., Lee, A., and Joshi, A. (2008). Easily identifiable discourse relations. In Scott, D. and Uszkoreit, H., editors, *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Poláková, L., Mírovský, J., Nedoluzhko, A., Jínová, P., Zikánová, Š., and Hajičová, E. (2013). Introducing the Prague discourse treebank 1.0. In Mitkov, R. and Park, J. C., editors, *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 91–99, Nagoya, Japan. Asian Federation of Natural Language Processing.

Poláková, L., Mírovskỳ, J., and Synková, P. (2017). Signalling implicit relations: A PDTB-RST comparison. *Dialogue & Discourse*, 8(2):225–248.

Polanyi, L. (1988). A formal model of the structure of discourse. *Journal of pragmatics*, 12(5-6):601–638.

Prasad, R. and Bunt, H. (2015). Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. 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 Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Prasad, R., Forbes Riley, K., and Lee, A. (2017). Towards full text shallow discourse relation annotation: Experiments with cross-paragraph implicit relations in the PDTB. In Jokinen, K., Stede, M., DeVault, D., and Louis, A., editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 7–16, Saarbrücken, Germany. Association for Computational Linguistics.

Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A. K., Robaldo, L., and Webber, B. L. (2006). The Penn Discourse Treebank 2.0 annotation manual.

Prasad, R., Webber, B., and Lee, A. (2018). Discourse annotation in the PDTB: The next generation. In Bunt, H., editor, *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pu, D., Wang, Y., and Demberg, V. (2023). Incorporating distributions of discourse structure for long document abstractive summarization. *arXiv preprint arXiv:2305.16784*.

Pyatkin, V., Klein, A., Tsarfaty, R., and Dagan, I. (2020). QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. 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 Barzilay, R. and Kan, M.-Y., editors, *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.

Radford, A. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Redeker, G., Berzlánovich, I., van der Vliet, N., Bouma, G., and Egg, M. (2012). Multi-layer discourse annotation of a dutch text corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2820–2825.

Reese, B., Hunter, J., Asher, N., Denis, P., and Baldridge, J. (2007). *Reference manual for the analysis and annotation of rhetorical structure*. PhD thesis, University of Texas at Austin.

Rehbein, I., Scholman, M., and Demberg, V. (2016). Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1039–1046, Portorož, Slovenia. European Language Resources Association (ELRA).

Riester, A., Nápoles, A. C., and Hoek, J. (2021). Combined discourse representations: Coherence relations and questions under discussion. In De Kuthy, K. and Meurers, D., editors, *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 26–30, Tübingen, Germany. Association for Computational Linguistics.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Roze, C., Braud, C., and Muller, P. (2019). Which aspects of discourse relations are hard to learn? primitive decomposition for discourse relation classification. In Nakamura, S., Gasic, M., Zuckerman, I., Skantze, G., Nakano, M., Papangelis, A., Ultes, S., and Yoshino, K., editors, *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 432–441, Stockholm, Sweden. Association for Computational Linguistics.

Rutherford, A. and Xue, N. (2015). Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In Mihalcea, R., Chai, J., and Sarkar, A., editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Sagae, K. (2009). Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In Bunt, H. and Villemonte de la Clergerie, É., editors, *Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09)*, pages 81–84, Paris, France. Association for Computational Linguistics.

Sanders, T. (2005). Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, pages 105–114. University of Toulouse-le-Mirail Toulouse.

Sanders, T. J., Demberg, V., Hoek, J., Scholman, M. C., Asr, F. T., Zufferey, S., and Evers-Vermeul, J. (2018). Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*.

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

Sanders, T. J., Spooren, W. P., and Noordman, L. G. (1993). Coherence relations in a cognitive theory of discourse representation.

Scheffler, T. and Stede, M. (2016). Mapping PDTB-style connective annotation to RST-style discourse annotation. In *Proceedings of the 13th Conference on Natural Language Processing*, pages 242–247.

Scholman, M. and Demberg, V. (2017). Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In Schneider, N. and Xue, N., editors, *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. 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.

Segal, E. M., Duchan, J. F., and Scott, P. J. (1991). The role of interclausal connectives in narrative structuring: Evidence from adults' interpretations of simple stories. *Discourse processes*, 14(1):27–54.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Stede, M. (2004). The Potsdam commentary corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain. Association for Computational Linguistics.

Stede, M. (2008a). Disambiguating rhetorical structure. *Research on Language and Computation*, 6:311–332.

Stede, M. (2008b). Disentangling nuclearity. *'Subordination' versus 'Coordination' in Sentence and Text: A cross-linguistic perspective*, 98:33.

Stede, M. (2012). *Discourse processing*, volume 15. Morgan & Claypool Publishers.

Stede, M., Afantenos, S., Peldszus, A., Asher, N., and Perret, J. (2016). Parallel discourse annotations on a corpus of short texts. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1051–1058, Portorož, Slovenia. European Language Resources Association (ELRA).

Stede, M. and Neumann, A. (2014). Potsdam commentary corpus 2.0: Annotation for discourse research. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Stevens-Guille, S., Maskharashvili, A., Li, X., and White, M. (2022). Generating discourse connectives with pre-trained language models: Conditioning on discourse relations helps reconstruct the PDTB. 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*, pages 500–515, Edinburgh, UK. Association for Computational Linguistics.

Subba, R. and Di Eugenio, B. (2009). An effective discourse parser that uses rich linguistic information. In Ostendorf, M., Collins, M., Narayanan, S., Oard, D. W., and Vanderwende, L., editors, *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics, pages 566–574, Boulder, Colorado. Association for Computational Linguistics.

Sun, X., Wei, B., Ren, X., and Ma, S. (2017). Label embedding network: Learning label representation for soft training of deep networks. *arXiv preprint arXiv:1710.10393*.

Suresh, V. and Ong, D. (2021). Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

Taboada, M. and Mann, W. C. (2006). Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Toldova, S., Pisarevskaya, D., Ananyeva, M., Kobozeva, M., Nasedkin, A., Nikiforova, S., Pavlova, I., and Shelepov, A. (2017). Rhetorical relations markers in Russian RST treebank. In Taboada, M., da Cunha, I., Maziero, E., Cardoso, P., Antonio, J., and Iruskieta, M., editors, *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics.

Torabi Asr, F. and Demberg, V. (2012). Implicitness of discourse relations. In Kay, M. and Boitet, C., editors, *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Van Kuppevelt, J. (1993). Intentionality in a topical approach of discourse structure. In *Intentionality and Structure in Discourse Relations*.

Von Stutterheim, C. and Klein, W. (1989). Referential movement in descriptive and narrative discourse. In *North-Holland Linguistic Series: Linguistic Variations*, volume 54, pages 39–76. Elsevier.

Wang, F., Wu, Y., and Qiu, L. (2012). Exploiting discourse relations for sentiment analysis. In Kay, M. and Boitet, C., editors, *Proceedings of COLING 2012: Posters*, pages 1311–1320, Mumbai, India. The COLING 2012 Organizing Committee.

Wang, G., Li, C., Wang, W., Zhang, Y., Shen, D., Zhang, X., Henao, R., and Carin, L. (2018). Joint embedding of words and labels for text classification. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331, Melbourne, Australia. Association for Computational Linguistics.

Wang, J. and Lan, M. (2015). A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 17–24, Beijing, China. Association for Computational Linguistics.

Wang, Y., Li, S., and Wang, H. (2017). A two-stage parsing method for text-level discourse analysis. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Webber, B. (2004). D-LTAG: extending lexicalized tag to discourse. *Cognitive Science*, 28(5):751–779.

Webber, B. (2006). Accounting for discourse relations: constituency and dependency. *Intelligent linguistic architectures*, pages 339–360.

Webber, B. (2019). Discourse processing for text analysis: Recent successes, current challenges. In *BIRNDL@ SIGIR*, pages 8–14.

Webber, B. and Joshi, A. (2012). Discourse structure and computation: Past, present and future. In Banchs, R. E., editor, *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

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

Webber, B., Stone, M., Joshi, A., and Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Westera, M., Mayol, L., and Rohde, H. (2020). TED-Q: TED talks and the questions they evoke. 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 1118–1127, Marseille, France. European Language Resources Association.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11486–11494.

Wu, H., Zhou, H., Lan, M., Wu, Y., and Zhang, Y. (2023a). Connective prediction for implicit discourse relation recognition via knowledge distillation. 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 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Wu, Y., Mangla, R., Dimakis, A. G., Durrett, G., and Li, J. J. (2024). Which questions should I answer? salience prediction of inquisitive questions. *arXiv preprint arXiv:2404.10917*.

Wu, Y., Mangla, R., Durrett, G., and Li, J. J. (2023b). QUDeval: The evaluation of questions under discussion discourse parsing. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Xiang, W., Wang, Z., Dai, L., and Wang, B. (2022). 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.

Xiao, W., Huber, P., and Carenini, G. (2021). Predicting discourse trees from transformer-based neural summarizers. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152, Online. Association for Computational Linguistics.

Xiong, Y., Feng, Y., Wu, H., Kamigaito, H., and Okumura, M. (2021). Fusing label embedding into BERT: An efficient improvement for text classification. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750, Online. Association for Computational Linguistics.

Xu, J., Gan, Z., Cheng, Y., and Liu, J. (2020). Discourse-aware neural extractive text summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Xue, N., Ng, H. T., Pradhan, S., Prasad, R., Bryant, C., and Rutherford, A. (2015). The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

Xue, N., Ng, H. T., Pradhan, S., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). CoNLL 2016 shared task on multilingual shallow discourse parsing. In Xue, N., editor, *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Yaseen, U. and Langer, S. (2021). Data augmentation for low-resource named entity recognition using backtranslation. In Bandyopadhyay, S., Devi, S. L., and Bhattacharyya, P., editors, *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 352–358, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Ye, J. (2021). Shared learning activity labels across heterogeneous datasets. *J. Ambient Intell. Smart Environ.*, 13(2):77–94.

Yi, C., Sujian, L., and Yueyuan, L. (2021). Unifying discourse resources with dependency framework. In Li, S., Sun, M., Liu, Y., Wu, H., Liu, K., Che, W., He, S., and Rao, G., editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1058–1065, Huhhot, China. Chinese Information Processing Society of China.

Yu, N., Zhang, M., Fu, G., and Zhang, M. (2022). RST discourse parsing with second-stage EDU-level pre-training. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280, Dublin, Ireland. Association for Computational Linguistics.

Yung, F., Ahmad, M., Scholman, M., and Demberg, V. (2024). Prompting implicit discourse relation annotation. *arXiv preprint arXiv:2402.04918*.

Yung, F., Demberg, V., and Scholman, M. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. In Friedrich, A., Zeyrek, D., and Hoek, J., editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics.

Zeldes, A. (2017a). A distributional view of discourse encapsulation: Multifactorial prediction of coreference density in RST. In Taboada, M., da Cunha, I., Maziero, E., Cardoso, P., Antonio, J., and Iruskieta, M., editors, *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 20–28, Santiago de Compostela, Spain. Association for Computational Linguistics.

Zeldes, A. (2017b). 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. (2024). erst: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, pages 1–47.

Zeldes, A., Das, D., Maziero, E. G., Antonio, J., and Iruskieta, M. (2019). The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. 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 97–104, Minneapolis, MN. Association for Computational Linguistics.

Zeldes, A. and Liu, Y. (2020). A neural approach to discourse relation signal detection. *Dialogue & Discourse*, 11(2):1–33.

Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S. (2021). The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In Zeldes, A., Liu, Y. J., Iruskieta, M., Muller, P., Braud, C., and Badene, S., editors, *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeyrek, D., Mendes, A., and Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zeyrek, D. and Webber, B. (2008). A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Zhang, H., Xiao, L., Chen, W., Wang, Y., and Jin, Y. (2018). Multi-task label embedding for text classification. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4545–4553, Brussels, Belgium. Association for Computational Linguistics.

Zhang, K., Wu, L., Lv, G., Chen, E., Ruan, S., Liu, J., Zhang, Z., Zhou, J., and Wang, M. (2023). Description-enhanced label embedding contrastive learning for text classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhang, L., Wang, G., Han, W., and Tu, K. (2021a). Adapting unsupervised syntactic parsing methodology for discourse dependency parsing. 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 5782–5794, Online. Association for Computational Linguistics.

Zhang, O., Wu, M., Bayrooti, J., and Goodman, N. (2021b). Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*.

Zhang, S., Xu, R., Xiong, C., and Ramaiah, C. (2022a). Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669.

Zhang, Y., Meng, F., Li, P., Jian, P., and Zhou, J. (2021c). Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599, Online. Association for Computational Linguistics.

Zhang, Z., Zhao, Y., Chen, M., and He, X. (2022b). Label anchored contrastive learning for language understanding. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1449, Seattle, United States. Association for Computational Linguistics.

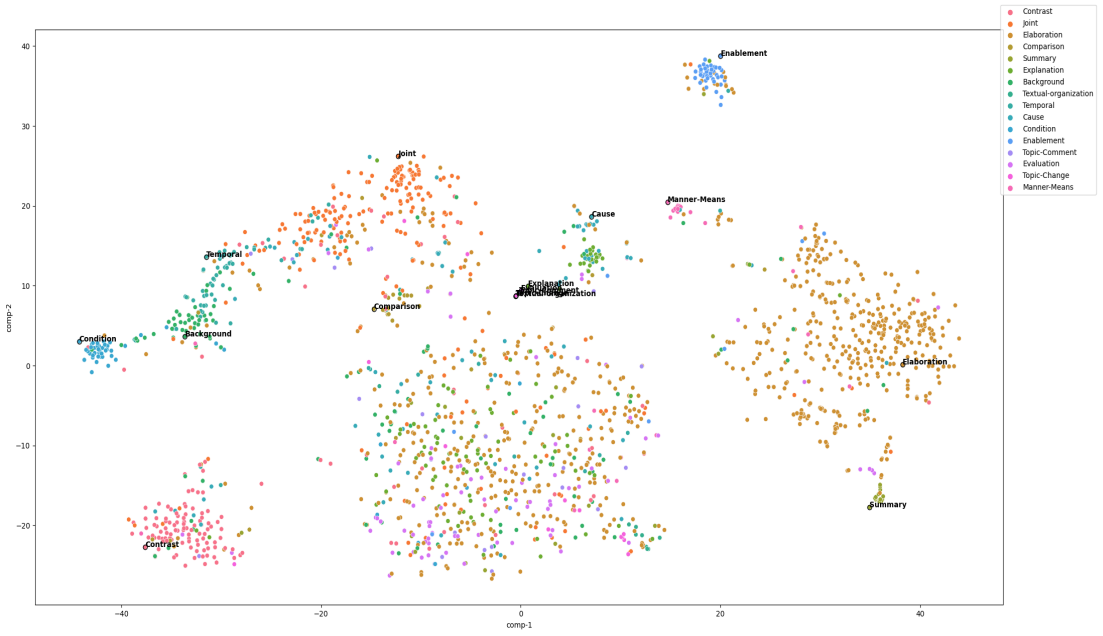
Zhou, Y. and Feng, Y. (2022). Improve discourse dependency parsing with contextualized representations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2250–2261, Seattle, United States. Association for Computational Linguistics.

Zhou, Y. and Xue, N. (2015). The chinese discourse treebank: A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49:397–431.

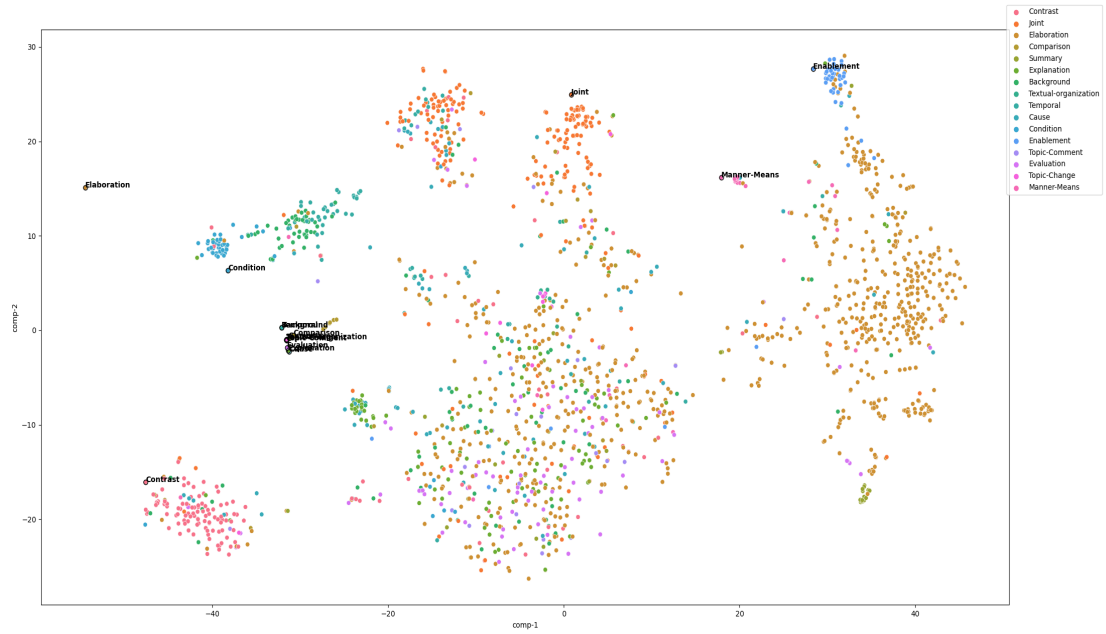
Appendix A

Appendices for Chapter 3

A.1 T-SNE Visualization Plot for RST-DT



(a)

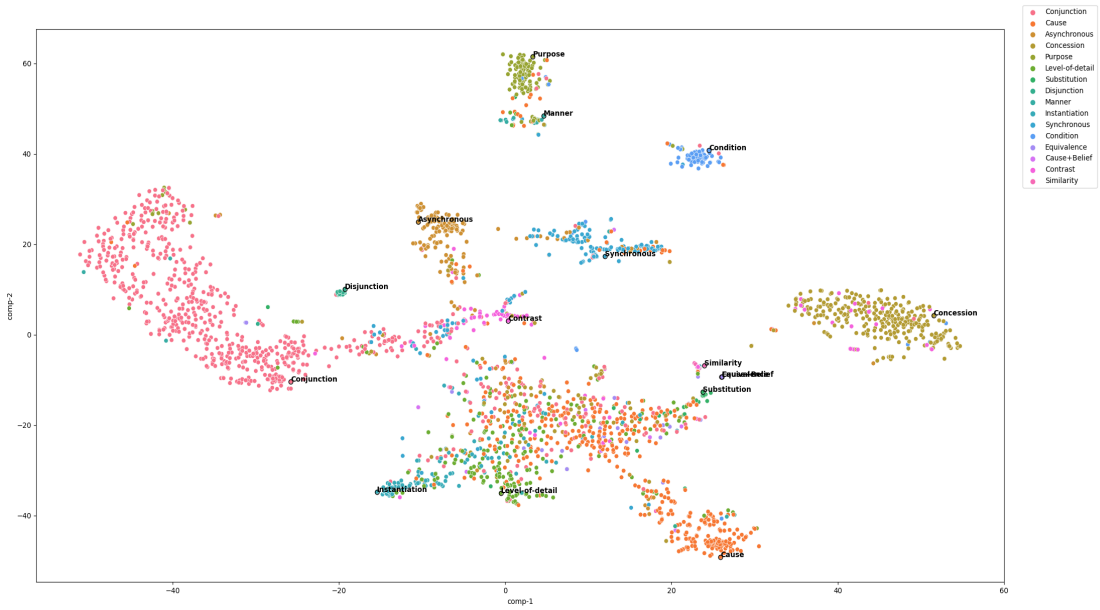


(b)

Figure A.1: (a) Label embeddings learnt with data augmentation. (b) Label embeddings learnt without data augmentation.

A.2 Appendix: T-SNE Visualization Plot for PDTB

(a)



(b)

Figure A.2: (a) Label embeddings of PDTB explicit relations. (b) Label embeddings of PDTB implicit and explicit relations combined in the training process.

A.3 Appendix: Alignment of RST-DT relations and PDTB Explicit Relations

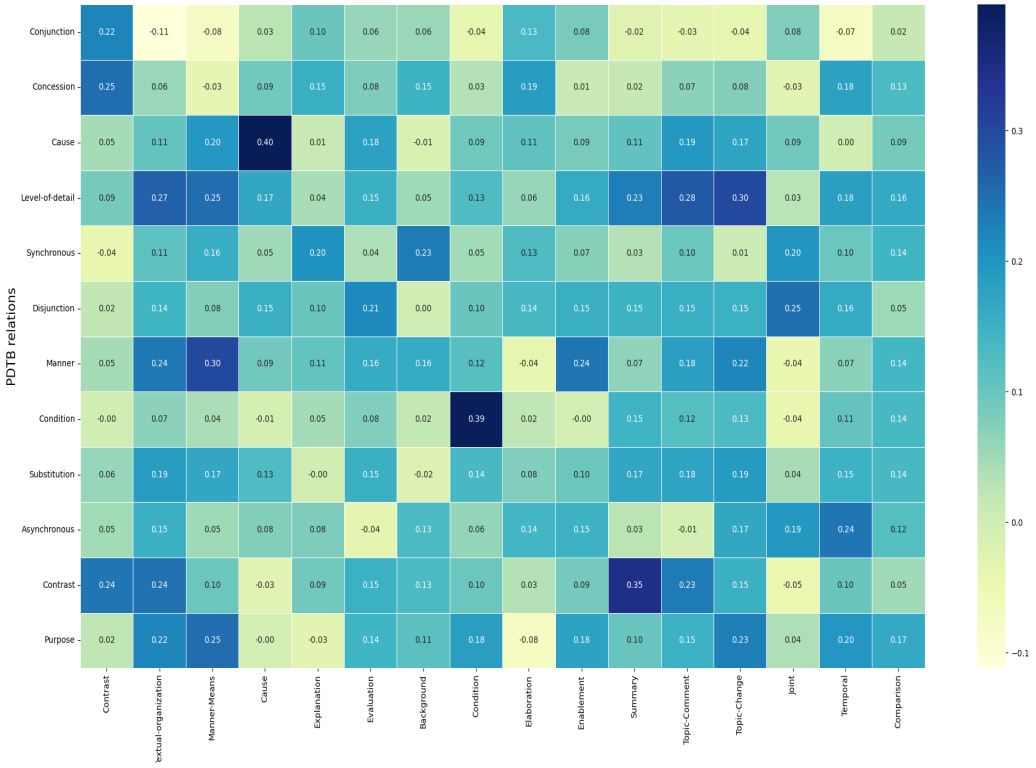


Figure A.3: Heatmap for full results of mapping between RST-DT and PDTB explicit relations.

Appendix B

List of Publications

Some of the work described in this dissertation has been published previously or is under review for publication. The following list gives an overview of these publications.

Articles in Peer-Reviewed Conferences

1. Yingxue Fu. 2024. Automatic Alignment of Discourse Relations of Different Discourse Annotation Frameworks. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 27–38, Torino, Italia. ELRA and ICCL.

2. Yingxue Fu. 2023. Discourse Relations Classification and Cross-Framework Discourse Relation Classification Through the Lens of Cognitive Dimensions: An Empirical Investigation. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 21–42, Online. Association for Computational Linguistics.

3. Yingxue Fu. 2022. Towards Unification of Discourse Annotation Frameworks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. Association for Computational Linguistics.

4. Yingxue Fu and Mark-Jan Nederhof. 2021. Automatic Classification of Human Translation and Machine Translation: A Study from the Perspective of Lexical Diversity. In *Proceedings for the First Workshop on Modelling Translation: Transla-*

tology in the Digital Age, pages 91–99, online. Association for Computational Linguistics.

Articles Under Review

- ****