Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0

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

Abstract

In this paper we propose a first empirical mapping between the RST-DT and the PDTB 3.0. We provide an original algorithm which allows the mapping of 6,510 (80.0%) explicit and implicit discourse relations between the overlapping articles of the RST-DT and PDTB 3.0 discourse annotated corpora. Results of the mapping show that while it is easier to align segments of implicit discourse relations, the mapping obtained between the aligned explicit discourse relations is more unambiguous.

1 Introduction

Different linguistic frameworks have been proposed to model the discourse relations that hold between textual segments. Two widely used frameworks are the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse Treebank (PDTB) (Miltsakaki et al., 2004; Prasad et al., 2008). Following these frameworks, several annotated corpora have been developed for a wide variety of NLP tasks, such as discourse parsing (Chi and Rudnicky, 2022), implicit discourse relation classification (Liu and Strube, 2023) and discourse generation (Stevens-Guille et al., 2022).

Since generating and manually annotating discourse corpora at the large scale required for fine-tuning large language models is prohibitively expensive and laborious, a viable alternative is to establish a mapping between already existing corpora so that they can be used seamlessly and interchangeably together. The two primary discourse annotated corpora are the RST-DT (Carlson et al., 2001) and the PDTB (PDTB 1.0, 2.0 and 3.0) (Prasad et al., 2006, 2007; Webber et al., 2019). However, since both corpora are annotated based on different frameworks, they differ in how they segment and label discourse relations. The resulting structural differences limit the extent to which they can be used together to train discourse models.

In this paper, we present a first empirical mapping between the RST-DT and the PDTB 3.0 based on the overlapping sections of the two annotated corpora. Previous work has addressed such a mapping between the RST-DT and PDTB 2.0. Sanders et al. (2021) proposed a theoretical mapping between both frameworks, while Demberg et al. (2019) established an empirical mapping based on the subset of the corpora that they share. However, to the best of our knowledge, no work has proposed a mapping between the RST-DT and the PDTB 3.0.

2 Background

The linguistic frameworks behind the RST-DT and the PDTB differ in how textual units are segmented and in how discourse relations are defined.

2.1 RST-DT

The RST-DT corpus (Carlson et al., 2001) is based on the RST theoretical framework (Mann and Thompson, 1988). In this framework, a text is first segmented into minimal non-overlapping units, referred to as elementary discourse units (EDUs). The grammatical clause is the starting point of the segmentation. After segmentation, relations between EDUs are identified using an open set of discourse relations. These relations are established recursively between adjacent EDUs until the entire text is connected, forming a single tree-like structure that encompasses multiple embedded relations (Taboada and Mann, 2006).

Consider the text in Example (1)[1] and its corresponding RST diagram in Figure 1.

(1) [There have been three days of hot, wind-swept rain,]^{edu1} [and now with the first sun we are after speckled sea trout,]^{edu2} [which with redfish provides most of the game fishing hereabouts.]^{edu3}

[1] Taken from the WSJ_1323 article in the RST-DT corpus.

Proceedings of Recent Advances in Natural Language Processing, pages 344–352
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_039

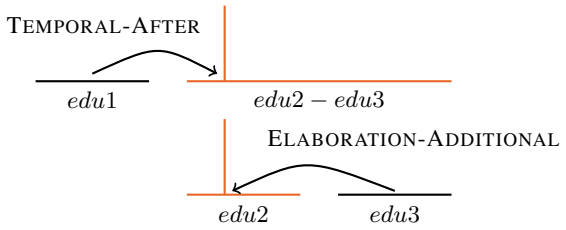

Figure 1: RST diagram of Example (1).

The leaves of the resulting RST diagram in Figure 1 correspond to the EDUs of Example (1) (i.e., $edu1$, $edu2$ and $edu3$), while the internal node of the tree correspond to multiple contiguous EDU segments (i.e., $\langle edu2 - edu3 \rangle$). Vertical lines in the diagram represent the nucleus of the discourse relation. All discourse relations in the RST framework hold between a nucleus and a satellite (mononuclear) or between two nuclei (multinuclear). The nucleus of a relation (depicted with a vertical line and shown in orange in Figure 1) represents an essential unit of information, while the satellite provides supporting information.

2.2 PDTB

The PDTB corpora (Prasad et al., 2006, 2007; Webber et al., 2019) are based on their namesake theoretical framework (Miltsakaki et al., 2004; Prasad et al., 2008). In the PDTB framework, discourse relations are annotated by first identifying discourse connectives (e.g., *but*, *however*) and then the arguments between which the relation holds. Unlike in the RST framework, in the PDTB framework arguments are not annotated for their nuclearity.

Discourse relations, in the PDTB, can be categorized as explicit or implicit[2]. An explicit discourse relation is marked by a discourse connective, while an implicit discourse relation holds between two arguments in the absence of a discourse connective. Explicit and implicit discourse relations are further differentiated based on their sense. Senses are organized hierarchically into three levels. The top level has four classes: TEMPORAL, CONTINGENCY, COMPARISON, and EXPANSION, which are then further refined into second and third level senses. In this work, we consider only the second level of sense granularity in our mapping.

The release of PDTB 3.0 (Webber et al., 2019) has brought important changes to its predecessor PDTB 2.0. In particular, the second and third levels of the sense hierarchy have been revised and 13,000 additional discourse relations have been annotated. Similarly, the number of intra-sentential implicit discourse relations went from 530 instances in the PDTB 2.0 to 6,234 in the PDTB 3.0. Due to these, 19% of the discourse relations annotations in the PDTB 2.0 corpus were changed. Of these, around 56% correspond to explicit discourse relations, while around 40% correspond to implicit relations.

3 Previous Work

Previous work has attempted to establish a mapping between the RST-DT and the PDTB corpora. Most recently, Demberg et al. (2019) proposed an empirical mapping between the RST-DT and the PDTB 2.0. Their approach was able to map 76% of the PDTB explicit and implicit discourse relations (senses) to an RST-DT relation based on an analysis of the overlapping sections of the two corpora.

Additionaly, Demberg et al. (2019) compare the results of their empirical mapping with the theoretically mappings proposed by Chiarcos (2014), Bunt and Prasad (2016) and Sanders et al. (2021). They found that their empirical results matched the theoretical mappings in more than 70% of the explicit relations, but only in less than 50% of the implicit relations. Another empirical mapping between the RST-DT and the PDTB 2.0 corpora was conducted by Polakova et al. (2017). They focused only on implicit discourse relations where an exact segment span matching was possible, which included a total of 472 discourse relations.

However, previous work was based exclusively on the PDTB 2.0 corpus. Given the significant changes in the PDTB 3.0, it has become necessary to develop a new mapping algorithm to accommodate the new annotation guidelines and establish a first empirical mapping between the RST-DT and the PDTB 3.0.

4 Corpora

The RST-DT corpus (Carlson et al., 2002) consists of 385 Wall Street Journal articles annotated with 20,017 discourse relations, while the PDTB[3] corpus (Prasad et al., 2019) consists of 2162 Wall Street Journal articles with 53,631 discourse relation annotations.

Both corpora overlap on 365 articles, allowing us to establish a direct mapping between the two.

[2]Other PDTB discourse relations include AltLex, AltLexC, EntRel, NoRel and hypophora.

[3]We will simply refer to PDTB 3.0 as PDTB henceforth.

Table 1 shows the total number of individual segments[4] and discourse relations in both corpora over this overlap. Due to its non-hierarchical structure, the PDTB corpus contains far less discourse relations than the RST-DT (see Table 1). Note that out of the 9,369 PDTB relations in the overlapping section of the PDTB corpus, 4,169 (44.5%) are explicit and 3,965 (42.3%) are implicit discourse relations. This corresponds to a combined total of 8,134 (86.8%) discourse relations. The remaining 1,235 (13.2%) PDTB relations include other relations such as AltLex, AltLexC, EntRel, NoRel and hypophora, which we did not take into account.

	RST-DT	**PDTB**
Text Segments	21,789	18,738
Discourse Relations	20,017	9,369

Table 1: Number of segments and discourse relations in the RST-DT and the PDTB corpora over the overlapping set of 365 Wall Street Journal articles.

5 Aligning and Mapping Relations

Similarly to Demberg et al. (2019), and given the smaller number of PDTB relations compared to RST-DT relations (see Table 1), we used the PDTB as the starting point for the alignment and mapping of discourse relations.

5.1 Segment Alignment

The purpose of the alignment is to match PDTB segments to their closest RST-DT segment. RST-DT segments can be individual EDUs (e.g., $edu1$), or contiguous EDUs (e.g., $\langle edu2-edu3\rangle$). PDTB segments can either be continuous, as in Example (2), or discontinuous, as in Example (3), where $arg2$ is discontinuous and split into two constituents: $arg2a$ and $arg2b$.

(2) **PDTB:** [We've had a good relationship with GE]^{arg1} [which is the first time you could say that]^{arg2}

(3) **PDTB:** Mr. Carpenter notes [that these types of investors]^{arg2a} also [are "sophisticated" enough not to complain about Kidder's aggressive use of program trading]^{arg2b}

[4]We will refer to PDTB arguments and to RST-DT EDU segments simply as segments for the remainder of the paper.

Continuous For each continuous PDTB segment, we find the RST-DT segment that maximizes the character overlap, while minimizing the number of additional characters in the RST-DT segment.

A PDTB segment is considered *perfectly* aligned if all of its characters overlap with the RST-DT segment, or if the extra characters in the RST-DT segment are punctuation or explicit connectives. We consider instances of the latter as perfect since PDTB segments systematically exclude terminal punctuation and explicit connectives contrary to RST-DT segments. In Example (4), $arg1$ of the PDTB relation is perfectly aligned with $edu67$ since only punctuation differs and $arg2$ is perfectly aligned with the RST-DT segment $\langle edu68 - edu69\rangle$.

(4) **PDTB:** [We've had a good relationship with GE]^{arg1} [which is the first time you could say that]^{arg2}

RST-DT: ["We've had a good relationship with GE,]^{edu67} [which is the first time]^{edu68}[you could say that]^{edu69}

On the other hand, a PDTB segment is considered *imperfectly* aligned with an RST-DT segment, if that RST-DT segment has the longest overlap with the PDTB segment among all RST-DT segments, and either the RST-DT or the PDTB segment includes extra characters beyond punctuation or explicit connectives. In Example (5), $arg1$ is *imperfectly* aligned with $edu92$ since the PDTB segment includes the additional tokens 'of the opportunity'.

(5) **PDTB:** [of the opportunity to "rebuild a franchise" at Kidder]^{arg1}

RST-DT: [to "rebuild a franchise" at Kidder.]^{edu92}

Table 2 shows statistics of the alignment of continuous PDTB segments onto RST-DT segments. As the table shows, most of the alignments found (85%) are perfect alignments and 50% consist of one PDTB argument being perfectly aligned with a single RST-DT EDU (1 : 1 alignments).

Discontinuous If PDTB segments are discontinuous, we align each of its constituents to an RST-DT segment using the same method as for continuous arguments. In Example (6), $arg2$ is discontinuous and split into two constituents: $arg2a$, which is aligned with $edu110$, and $arg2b$, which is aligned

Type	Arg : EDU	Count (%)	Total (%)
Perfect	1 : 1	7,621 (50%)	12,959 (85%)
	1 : n	5,338 (35%)	
Imperfect	1 : 1	1,705 (11%)	2,329 (15%)
	1 : n	624 (4%)	
	Total	**15,288 (100%)**	**15,288 (100%)**

Table 2: Statistics of the alignment of continuous PDTB segments onto RST-DT segments.

Type	Constituent : EDU	Count	Total
Perfect	1 : 1	762 (38%)	936 (47%)
	1 : n	174 (9%)	
Imperfect	1 : 1	818 (41%)	1053 (53%)
	1 : n	235 (12%)	
	Total	**1,989 (100%)**	**1,989 (100%)**

Table 3: Statistics of the alignment of discontinuous PDTB segment constituents onto RST-DT segments.

with the RST-DT segment $\langle edu110 - edu111 \rangle$.

(6) **PDTB:** Mr. Carpenter notes [that these types of investors]^{arg2a} also [are "sophisticated" enough not to complain about Kidder's aggressive use of program trading]^{arg2b}

RST-DT: [Mr. Carpenter notes]^{edu109} [that these types of investors also are "sophisticated" enough]^{edu110} [not to complain about Kidder's aggressive use of program trading.]^{edu111}

Table 3 shows statistics of the alignment of discontinuous PDTB segments onto RST-DT segments. As the table shows, the ratio of perfect alignments is lower than in the case of continuous arguments (47% vs 85%, see Table 2). However, $1:1$ alignments (i.e., one PDTB argument constituent being perfectly aligned to a single RST-DT EDU) are still more frequent than $1:n$ alignments.

5.2 Relation Mapping

After aligning PDTB segments onto RST-DT segments, we map the PDTB relations to their most likely RST-DT relations. To do so, we rely on the strong nuclearity principle (Marcu, 2000) and on the notion of nucleus path (Demberg et al., 2019). In the context of the RST, the strong nuclearity principle dictates that relations annotated between segments of multiple contiguous EDUs also hold between the nucleus of each of these contiguous segments. The nucleus path, in turn, identifies the single nuclear EDU that originated the entire complex segment by always following the segments annotated as nuclei. Five different mapping scenarios are considered.

Perfect Mapping If both PDTB segments are continuous and perfectly aligned with different RST-DT segments, we map the PDTB relation to the lowest RST-DT relation covering these RST-DT segments. In Figure 2, $arg1$ is perfectly aligned with $\langle edu13 - edu18 \rangle$ and $arg2$ is perfectly aligned with $\langle edu19 - edu20 \rangle$. Therefore, we map the PDTB relation between $arg1$ and $arg2$, IMPLICIT.EXPANSION, to ELABORATION-ADDITIONAL, the lowest RST-DT relation covering $\langle edu13 - edu18 \rangle$ and $\langle edu19 - edu20 \rangle$.

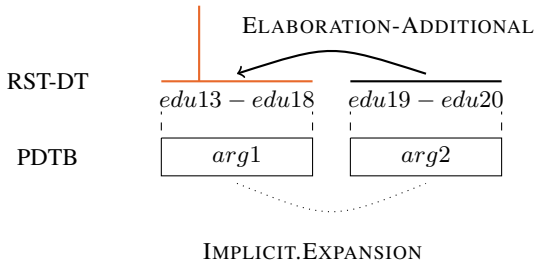

Figure 2: Example of a perfect relation mapping.

Imperfect Mapping If the nucleus paths of both RST-DT segments lead to an EDU that overlaps the aligned PDTB segment, then the potential mapping is retained. Figure 3 shows an example of an imperfect mapping. The lowest covering relation EXPLANATION-ARGUMENTATIVE, is between $\langle edu91 - edu92 \rangle$ and $\langle edu93 - edu96 \rangle$. Following the nucleus path from $\langle edu91 - edu92 \rangle$, the first nucleus found is $edu91$. Although $arg1$ is aligned with $edu92$, it overlaps with $edu91$ and is, therefore, in the nucleus path. The first nucleus in the nucleus path from $\langle edu93 - edu96 \rangle$ is $\langle edu93 - edu95 \rangle$. As $arg2$ overlaps perfectly with $\langle edu93 - edu95 \rangle$ it is also in the nucleus path. As both PDTB segments are in the nucleus path, the PDTB relation between $arg1$ and $arg2$, CONTINGENCY.CAUSE, is mapped to the RST-DT EXPLANATION-ARGUMENTATIVE relation.

Embedded Relation When both segments of a PDTB relation are aligned with the same RST-DT segment, the relation cannot be mapped. This occurs due to a difference in granularity across frameworks. In Example (7), illustrated in Figure 4, both $arg1$ and $arg2$ are aligned with $edu2$. The PDTB relation, EXPANSION.MANNER, is more fine grained and does not have an equivalent RST-DT relation. In these cases, the PDTB relation cannot be mapped.

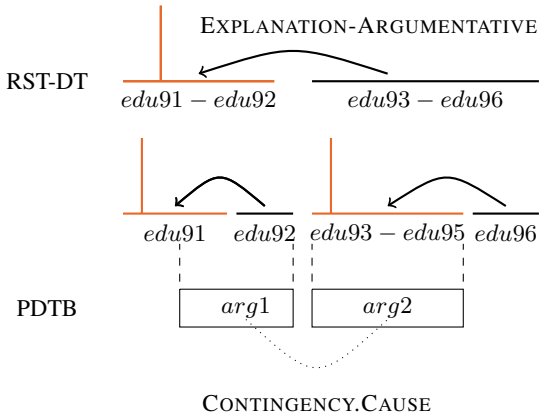

Figure 3: Example of an imperfect relation mapping.

(7) **PDTB:** [jump from murder to antitrust cases]^{arg1} [from arson to securities fraud]^{arg2}
RST-DT: [A judge must jump from murder to antitrust cases, from arson to securities fraud,]^{edu2}

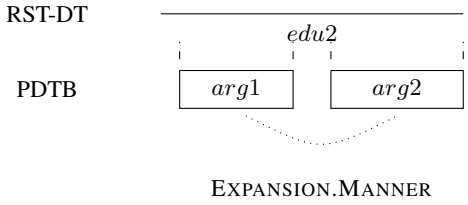

Figure 4: Example of an embedded relation which is not mapped.

If the mapping is neither perfect, imperfect or embedded, we identify the most immediate discourse relation between the aligned RST-DT segments as a potential map to the PDTB relation. We then follow the nucleus path from each of the RST-DT segments to their nuclear EDU and verify if it is included within the aligned PDTB segment. Three outcomes are possible.

Unclear Nucleus Path If at least one of the nucleus paths of the RST-DT segments leads to an EDU that does not overlap with the aligned PDTB segment, then we do not map the PDTB relation. In Figure 5, $arg1$ is aligned imperfectly with $edu101$, while $arg2$ is aligned perfectly with $edu104$. The closest covering RST-DT relation is CONSEQUENCE. As shown in Figure 5, the nucleus path from $\langle edu102 - edu104\rangle$ leads to $\langle edu102 - edu103\rangle$ which does not overlap with $arg2$. Therefore, the PDTB relation remains unmapped.

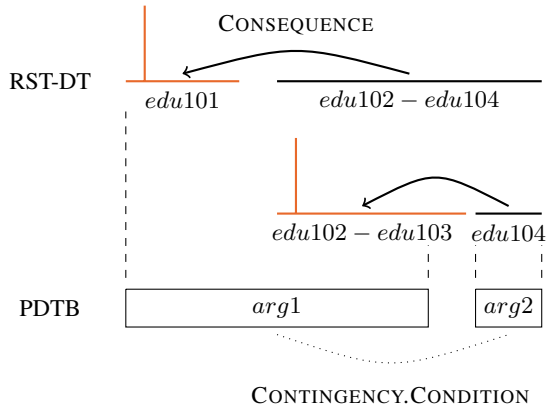

Figure 5: Example of an unclear nucleus path, which is not mapped.

Multinuclear Relation If at least one of the nucleus paths of the RST-DT segments leads to a multinuclear relation, it becomes impossible to identify a single nucleus to follow the nucleus path and we do not map the PDTB relation. In Figure 6, $arg1$ is aligned with $\langle edu139 - edu141\rangle$, while $arg2$ is aligned with $edu142$. As the figure shows, no single nucleus can be identified at the end of the nucleus path starting at $\langle edu138 - edu141\rangle$ because the following RST-DT relation, between $\langle edu138 - edu141\rangle$ and $\langle edu139 - edu141\rangle$, is a multinuclear relation and we cannot unambiguously trace it to $arg1$. As a consequence, the PDTB relation is not mapped.

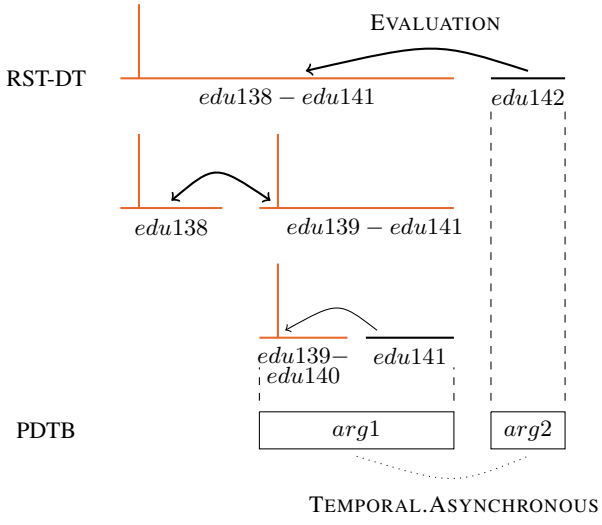

Figure 6: Example of a multinuclear relation, which is not mapped.

Discontinuous Relation If one segment of a PDTB relation is discontinuous and the other segment is embedded between its constituents, we at-

tempt to map it. To do so, we verify if the RST-DT segments aligned with the constituents are related by a SAME-UNIT relation. If so, the PDTB relation is mapped to the RST-DT relation between the RST-DT segment aligned with the continuous PDTB segment and an RST-DT segment aligned with a PDTB constituent. An example is shown in Figure 7. As shown in the figure, $\langle edu96 - edu97 \rangle$ and $edu98$ have a SAME-UNIT relation, so we map the PDTB CONDITION relation, to the RST-DT CIRCUMSTANCE relation between $edu96$ and $edu97$.

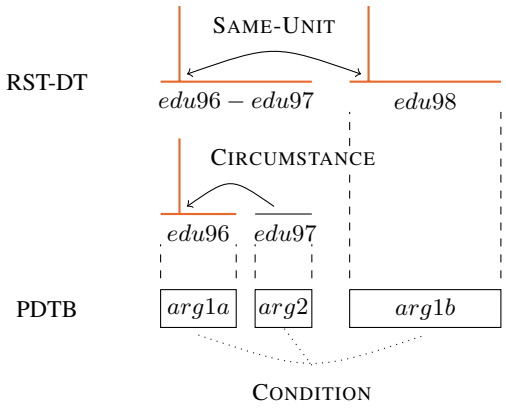

Figure 7: Example of a discontinuous relation mapping.

The five cases above illustrate how the mapping algorithm works in the different encountered scenarios. Based on it, we then established a mapping between the discourse relations that were successfully aligned in the overlapping articles of the RST-DT and the PDTB.

6 Results

We first present the results of the relation alignment (see Section 5.1) and then present the relation mapping results (see Section 5.2).

6.1 Relation Alignment

Table 4 shows the results of the relation alignment. Recall that to align a relation across frameworks both segments of the relation need to be aligned. As Table 4 shows, the approach was able to align 6,510 (80.0%) of the 8,134 explicit and implicit PDTB discourse relations in the overlapping articles of the RST-DT and the PDTB corpus. More precisely, our proposed algorithm was able to align 3,073 (73.7%) of the 4,169 explicit discourse relations and 3,437 (86.7%) of the 3,965 implicit relations.

As Table 4 shows, implicit relations have more successful alignments than explicit relations - 3,437 (86.7%) out of 3,965 vs 3,073 (73.7%) out of 4,169, respectively. This is because of the significantly higher number of discontinuous PDTB segments in explicit relations. In fact, 729 (17.5%) of all explicit discourse relations were impossible to align because at least one of the segments in the PDTB was discontinuous and no matching SAME-UNIT label was found in the RST-DT for the same segment spans. Whereas this only happened to 214 (5.4%) of all implicit discourse relations.

The higher number of discontinuous PDTB segments in explicit relations also comes as a consequence of the annotation style of the PDTB corpus. Because explicit relations are annotated based only on the presence of a connective, they are more permissive on the location and extent of their arguments. This creates a challenge when aligning the relations onto the RST-DT, where all adjacent text segments are connected. Conversely, for the implicit relations, given their more subjective interpretation, the PDTB only annotates instances where both arguments are adjacent to each other. Thus, leading to a clearer agreement with the annotation style of the RST-DT.

Another interesting result shown in Table 4 is the higher number of imperfect alignments among explicit relations (836/3,073) compared to implicit relations (375/3,437). A manual analysis shows that most of these imperfect alignments correspond to PDTB relations where the segments are not adjacent. This led to instances where the corresponding RST-DT text segments are made of multiple contiguous segments that do not exactly match the span of the PDTB segments. This, however, does not happen for implicit relations as they are only annotated in the PDTB between adjacent segments.

6.2 Relation Mapping

Once the relation segments were aligned, we mapped the relation labels (see Section 5.2). Table 5 shows the mapping of the 3,073 aligned explicit discourse relations, while Table 6 shows the mapping of the 3,437 aligned implicit discouse relations. To keep both tables readable, we show only discourse relations for which at least one mapping was found with at least 30 instances. Percentages and color gradients are calculated row-wise.

As Tables 5 and 6 show, and similarly to what Demberg et al. (2019) found, we obtain a clearer mapping for explicit discourse relations when compared to implicit discourse relations. If we consider relations that appear in both tables, such as

Relation Mapping	Discourse Relation	Type	Count	Sub-Total	Total
Possible	Explicit	Perfect Mapping	2,237 (28%)	3,073 (38%)	6,510 (80%)
		Imperfect Mapping	836 (10%)		
	Implicit	Perfect Mapping	3,062 (38%)	3,437 (42%)	
		Imperfect Mapping	375 (5%)		
Impossible	Explicit	Embedded Relation	106 (1%)	1,096 (14%)	1,624 (20%)
		Unclear Nucleus Path	64 (1%)		
		Multinuclear Relation	197 (2%)		
		Discontinuous Relation	729 (9%)		
	Implicit	Embedded Relation	50 (1%)	528 (6%)	
		Unclear Nucleus Path	81 (1%)		
		Multinuclear Relation	183 (2%)		
		Discontinuous Relation	214 (3%)		
Total			8,134 (100%)	8,134 (100%)	8,134 (100%)

Table 4: Alignment results between relations in the overlapping articles of the RST-DT and the PDTB corpus.

the RST-DT LIST relation, we observe a more predominant mapping to single explicit PDTB relations than what we observe for implicit relations. For instance, 664 (95.0%) out of the 699 RST-DT LIST relations in Table 5 are mapped to the PDTB EXPANSION.CONJUNCTION relation. On the other hand, in Table 6, only 302 (63.0%) out of 479 LIST relations are mapped to the PDTB EXPANSION.CONJUNCTION, while 92 (19.2%) are mapped to CONTINGENCY.CAUSE and 45 (9.4%) are mapped to TEMPORAL.ASYNCHRONOUS. The same is true for other discourse relations occurring in both tables.

Compared to the results obtained by Demberg et al. (2019), we observe other similar patterns. For instance, the PDTB TEMPORAL class in Table 5 shows very clear mappings between the RST-DT TEMPORAL-SAME-TIME and TEMPORAL-AFTER to the PDTB explicit TEMPORAL.SYNCHRONOUS and TEMPORAL.ASYNCHRONOUS, respectively. In addition, the explicit discourse relations in the PDTB COMPARISON and CONTIGENCY classes are harder to unambiguously map to individual RST-DT relations. Finally, for the discourse relations in the PDTB EXPANSION class in Table 6, we observe the same difficulties in establishing a mapping to their RST-DT counterparts.

The clearer mapping between explicit relations compared to implicit relations, contrasts with the alignment results presented in Section 6.1. However, this was expected, since the presence of an explicit discourse connective allows for a more objective interpretation of the discourse relation that holds between the text segments.

7 Conclusion

In this paper we have presented a first empirical mapping between the RST-DT and the PDTB 3.0 annotated corpora. Following our proposed algorithms we were able to map 6,510 (80.0%) of the explicit and implicit discourse relations in the 365 Wall Street Journal articles overlapping the RST-DT and the PDTB 3.0 corpora. Compared to the 76% successfully mapped relations obtained by Demberg et al. (2019) in their empirical mapping between the RST-DT and the PDTB 2.0, we were able to achieve a 4% improvement in mapping coverage.

Our alignment results show a clearer correspondence between segments of implicit discourse relations when compared to segments of explicit relations. This is a consequence of the difference in annotation between the two corpora. Since the RST-DT establishes discourse relations between all adjacent text segments, the PDTB often establishes explicit relations between text segments which are not adjacent. This creates a challenge for the alignment algorithm. However, when an alignment was found, we observed a clearer mapping between explicit discourse relations than between implicit discourse relations. This stems from the presence of discourse connectives which allow for a more objective interpretation of the relations.

8 Limitations and Future Work

The empirical mapping proposed was based exclusively on the 365 overlapping articles of both

RST-DT \ PDTB	Comparison		Contingency		Expansion	Temporal		Total
	Concession	Contrast	Cause	Condition	Conjunction	Asynchronous	Synchronous	
Contrast	61.0% (138)	26.0% (59)	0.0% (0)	0.0% (1)	9.0% (21)	0.0% (0)	4.0% (9)	**100% (228)**
List	2.0% (17)	0.0% (2)	0.0% (1)	0.0% (0)	95.0% (664)	0.0% (2)	2.0% (13)	**100% (699)**
Sequence	2.0% (2)	0.0% (0)	0.0% (0)	0.0% (0)	72.0% (62)	23.0% (20)	2.0% (2)	**100% (86)**
Antithesis	84.0% (207)	7.0% (18)	0.0% (0)	0.0% (1)	3.0% (7)	1.0% (3)	4.0% (11)	**100% (247)**
Circumstance	7.0% (20)	0.0% (1)	8.0% (22)	7.0% (18)	5.0% (15)	31.0% (86)	41.0% (112)	**100% (274)**
Concession	88.0% (170)	6.0% (11)	0.0% (0)	0.0% (0)	2.0% (4)	2.0% (3)	3.0% (6)	**100% (194)**
Condition	3.0% (4)	1.0% (2)	0.0% (0)	84.0% (127)	0.0% (0)	9.0% (13)	3.0% (5)	**100% (151)**
Elaboration-Additional	30.0% (54)	5.0% (9)	2.0% (4)	1.0% (1)	56.0% (101)	4.0% (7)	3.0% (5)	**100% (181)**
Explanation-Argumentative	19.0% (11)	0.0% (0)	66.0% (38)	0.0% (0)	0.0% (0)	2.0% (1)	14.0% (8)	**100% (58)**
Reason	0.0% (0)	1.0% (1)	71.0% (54)	0.0% (0)	8.0% (6)	7.0% (5)	0.0% (0)	**100% (76)**
Temporal-After	2.0% (1)	0.0% (0)	0.0% (0)	0.0% (0)	4.0% (2)	94.0% (50)	0.0% (0)	**100% (53)**
Temporal-Same-Time	0.0% (0)	0.0% (0)	2.0% (1)	0.0% (0)	0.0% (0)	0.0% (0)	98.0% (44)	**100% (45)**
Total	**(624)**	**(103)**	**(130)**	**(148)**	**(882)**	**(190)**	**(215)**	**(2292)**

Table 5: Mapping results for the aligned explicit PDTB discourse relations. The table shows only discourse relations for which there was at least one mapping with a total of at least 30 instances (i.e., 2292 relations instead of 3073). The percentages and the color grading were calculated row-wise.

RST-DT \ PDTB	Comparison	Contingency		Expansion			Temporal	Total
	Concession	Cause	Purpose	Conjunction	Instantiation	Level-of-Detail	Asynchronous	
List	4.0% (18)	19.0% (92)	0.0% (1)	63.0% (302)	2.0% (9)	0.0% (1)	9.0% (45)	**100% (479)**
Sequence	8.0% (6)	7.0% (5)	0.0% (0)	12.0% (9)	0.0% (0)	5.0% (4)	67.0% (49)	**100% (73)**
Consequence	7.0% (6)	51.0% (41)	5.0% (4)	19.0% (15)	4.0% (3)	5.0% (4)	10.0% (8)	**100% (81)**
Elaboration-Additional	9.0% (77)	27.0% (236)	0.0% (4)	35.0% (311)	5.0% (40)	19.0% (169)	5.0% (42)	**100% (879)**
Elaboration-General-Specific	1.0% (1)	15.0% (15)	0.0% (0)	13.0% (13)	18.0% (17)	52.0% (50)	1.0% (1)	**100% (97)**
Evidence	2.0% (2)	14.0% (12)	0.0% (0)	13.0% (11)	40.0% (35)	31.0% (27)	1.0% (1)	**100% (88)**
Example	0.0% (0)	12.0% (13)	0.0% (0)	8.0% (9)	63.0% (68)	16.0% (17)	1.0% (1)	**100% (108)**
Explanation-Argumentative	6.0% (14)	53.0% (132)	0.0% (0)	7.0% (18)	13.0% (31)	20.0% (50)	1.0% (2)	**100% (247)**
Purpose	0.0% (0)	3.0% (8)	96.0% (222)	0.0% (1)	0.0% (0)	0.0% (0)	0.0% (0)	**100% (231)**
Reason	0.0% (0)	73.0% (35)	13.0% (6)	6.0% (3)	0.0% (0)	6.0% (3)	2.0% (1)	**100% (48)**
Total	**(124)**	**(589)**	**(237)**	**(69)**	**(203)**	**(336)**	**(150)**	**(2331)**

Table 6: Mapping results for the aligned PDTB implicit discourse relations. The table shows only discourse relations for which there was at least one mapping with a total of at least 30 instances (i.e., 2,331 relations instead of 3,437). The percentages and the color grading were calculated row-wise.

annotated corpora. We did not consider the remaining non-overlapping articles in our mapping as we would not be able to find a correspondence to the existing discourse relations on the other corpora. Based on our findings we could extrapolate our mapping to the remaining articles within a certain degree of accuracy, but a such a mapping could not be afterwards used to attest the robustness of our approach. Therefore, we preferred to focus only on the articles for which an objective correspondence could be established between both corpora.

As future work, we would like to extend the work to include AltLex and AltLexC discourse relations to have a more complete mapping between both corpora. We would also like to develop automatic segmentation and discourse relation classifiers based on our results to then establish a mapping between the remaining Wall Street Journal articles that do not currently overlap the RST-DT and the PDTB 3.0. This would allow us to generate a more comprehensive set of discourse annotated data following two of the most widely used discourse frameworks for the fine-tuning of large language models.

Reproducibility

We used the Gate Embedded API and Java for the implementation. Our code can be found on GitHub[5].

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback on the previous version of this paper. They would also like to thank the Linguistics Data Consortium (LDC) for providing the necessary corpora for this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

[5] https://github.com/CLaC-Lab/Mapping-Discourse-Relations

References

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA'16)*, pages 45–54, Portorož, Slovenia. European Language Resources Association (ELRA).

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial'01)*.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank. LDC2002T07. Web Download. Philadelphia: Linguistic Data Consortium.

Ta-Chung Chi and Alexander Rudnicky. 2022. Structured Dialogue Discourse Parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial'22)*, pages 325–335, Edinburgh, UK. Association for Computational Linguistics.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

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

Wei Liu and Michael Strube. 2023. Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 15696–15712, Toronto, Ontario, Canada. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA, USA.

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

Lucie Polakova, Jiří Mírovský, and Pavlína Synková. 2017. Signalling implicit relations: A PDTB - RST comparison. *Dialogue & Discourse*, 8(2):225–248.

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

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, University of Pennsylvania.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, and Bonnie Webber. 2006. The Penn Discourse Treebank 1.0 Annotation Manual. Technical report, University of Pennsylvania.

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

Ted JM Sanders, Vera Demberg, Jet Hoek, Merel CJ Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. 2021. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71.

Symon Stevens-Guille, Aleksandre Maskharashvili, Xintong Li, and Michael White. 2022. Generating Discourse Connectives with Pre-trained Language Models: Conditioning on Discourse Relations Helps Reconstruct the PDTB. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial'22)*, pages 500–515, Edinburgh, UK. Association for Computational Linguistics.

Maite Taboada and William C Mann. 2006. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459.

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