Towards Full Text Shallow Discourse Relation Annotation: Experiments with Cross-Paragraph Implicit Relations in the PDTB

Rashmi Prasad
University of Wisconsin-Milwaukee
prasadr@uwm.edu

Katherine Forbes Riley
University of Pittsburgh
katherineforbesriley@gmail.com

Alan Lee
University of Pennsylvania
aleewk@seas.upenn.edu

Abstract

Full text discourse parsing relies on texts comprehensively annotated with discourse relations. To this end, we address a significant gap in the inter-sentential discourse relations annotated in the Penn Discourse Treebank (PDTB), namely the class of cross-paragraph implicit relations, which account for 30% of inter-sentential relations in the corpus. We present our annotation study to explore the incidence rate of adjacent vs. non-adjacent implicit relations in cross-paragraph contexts, and the relative degree of difficulty in annotating them. Our experiments show a high incidence of non-adjacent relations that are difficult to annotate reliably, suggesting the practicality of backing off from their annotation to reduce noise for corpus-based studies. Our resulting guidelines follow the PDTB adjacency constraint for implicits while employing an underspecified representation of non-adjacent implicits, and yield 62% inter-annotator agreement on this task.

1 Introduction

Empirical approaches for modeling discourse relations rely on corpora annotated with such relations, such as the PDTB (Prasad et al., 2008), the RST-DT (Carlson et al., 2003), and the ANNODIS corpus (Afantenos et al., 2012). The PDTB is currently the largest of these annotated corpora and widely used for theoretical and empirical research on discourse relations. However, it does not provide exhaustive annotation of its source texts (Prasad et al., 2014). A critical kind of gap is found within the class of inter-sentential relations, i.e., relations with arguments in different sentences. In particular, while the PDTB provides annotations of explicit inter-sentential relations within and across paragraphs, and of implicit relations between adjacent sentences within paragraphs, it ignores cross-paragraph implicit relations. Ex. (1) illustrates the problem in a PDTB-annotated text, showing 6 sentences (S1-S6) in the first four paragraphs of a longer article. (Empty lines indicate paragraph boundaries.) While all annotation elements are not shown here, the key issue to note is that the relations of sentences S2 and S3 with the prior text are left *unannotated* because they are paragraph-initial sentences lacking any inter-sentential explicit connectives.

(1) **S1:** As competition heats up in Spain's crowded bank market, Banco Exterior de Espana is seeking to shed its image of a state-owned bank and move into new activities.

(**unannotated**)
S2: Under the direction of its new chairman, Francisco Luzon, Spain's seventh largest bank is undergoing a tough restructuring that analysts say may be the first step toward the bank's privatization.

(**unannotated**)
S3: The state-owned industrial holding company Instituto Nacional de Industria and the Bank of Spain jointly hold a 13.94% stake in Banco Exterior.
(Conjunction)
S4: The government directly owns 51.4% and Factorex, a financial services company, holds 8.42%.
(Conjunction)
S5: The rest is listed on Spanish stock exchanges.

(Contrast)
S6: Some analysts are concerned, however, that Banco Exterior may have waited too long to diversify from its traditional export-related activities.

There are more than 12K such unannotated tokens in the current version of PDTB (PDTB-2), constituting 30% of all inter-sentential discourse contexts and 87% of all cross-paragraph inter-sentential contexts. Furthermore, research on discourse parsing shows that there is value in filling these gaps. For example, Pitler et al. (2009) report improvements in implicit relation sense classification with a sequence model. And more re-

cent systems, including the best systems (Wang and Lan, 2015; Oepen et al., 2016) at the recent CONLL shared tasks on PDTB-style shallow discourse parsing (Xue et al., 2015, 2016), while not using a sequence model, still incorporate features about neighboring relations. Such systems have many applications, including summarization (Louis et al., 2010), information extraction (Huang and Riloff, 2012), question answering (Blair-Goldensohn, 2007), opinion analysis (Somasundaran et al., 2008), and argumentation (Zhang et al., 2016).

This paper describes our experiments in annotating cross-paragraph implicit relations in the PDTB (Section 2), with the goal of producing a set of guidelines (Section 3) to annotate such relations reliably (Section 4) and produce a representative dataset annotated with complete sequences of inter-sentential relations.

Our main findings from the experiments are as follows:

- The ratio of cross-paragraph implicit relations between non-adjacent sentences and between adjacent sentences is almost 1 to 1 (47% vs 51% in our experiment). This is similar to the distribution of cross-paragraph explicit relations (Prasad et al., 2010). Hence, non-adjacency is a non-trivial factor to consider when annotating cross-paragraph implicit relations.

- Inter-annotator agreement for the non-adjacent cross-paragraph implicits is substantially lower compared to their adjacent counterparts (47% versus 68%). Furthermore, the disagreements, while possible to resolve through discussion, are time-consuming and therefore prohibitive to large-scale annotation.

On the basis of these findings, we established the following guidelines for our annotation of cross-paragraph implicit relations:

- We fall back to the PDTB strategy of fully annotating only adjacent implicit relations, while also employing an underspecified marking of non-adjacent ones.

- We introduce new guidelines to (a) better represent the inter-dependency of relations in a text, (b) represent new senses we have encountered, and (c) better represent the relation of entity-based coherence. These new guidelines are discussed at various points in Section 3.

We achieve a final overall agreement of 62% with our guidelines.

We discuss related work in Section 5 and conclude in Section 6, outlining our goals for this task and future work beyond.

2 A Brief Review of PDTB

Our study is carried out within the annotation framework of the PDTB, and incorporates the most recent PDTB (PDTB-3) sense hierarchy (Webber et al., 2016), shown in Fig. 1 (with two modifications – see Section 3.2). Annotated over the ˜1 million word WSJ corpus (Marcus et al., 1993), the PDTB follows a lexically-grounded approach to the representation of discourse relations (Webber et al., 2003) while remaining theory-neutral in its annotation approach. Discourse relations are taken to hold between two abstract object arguments, named Arg1 and Arg2 using syntactic conventions, and are triggered either by *explicit* connectives (Ex. 2) or, otherwise, by adjacency between clauses and sentences. (Throughout the paper, the expression of a relation is underlined, its Arg2 is **bolded**, its Arg1 is *italicized*, and its type and sense are in parentheses.)

(2) *The Manhattan U.S. attorney's office stressed criminal cases from 1980 to 1987, averaging 43 for every 100,000 adults.*
(Explicit, Contrast)
But **the New Jersey U.S. attorney averaged 16**.

(3) *So far, the mega-issues are a hit with investors.*
(Implicit, Arg2-as-instance, For example)
Earlier this year, Tata Iron & Steel Co.'s offer of $355 million of convertible debentures was oversubscribed.

(4) *When the plant was destroyed, "I think everyone got concerned that the same thing would happen at our plant*," a KerrMcGee spokeswoman said.
(AltLex, Reason)
That prompted Kerr-McGee to consider moving the potentially volatile storage facilities and cross-blending operations away from town.

(5) *The proposed petrochemical plant would use naphtha to manufacture the petrochemicals propylene and ethylene and their resin derivatives, polypropylene and polyethylene.*
(EntRel)
These are the raw materials used in making plastic

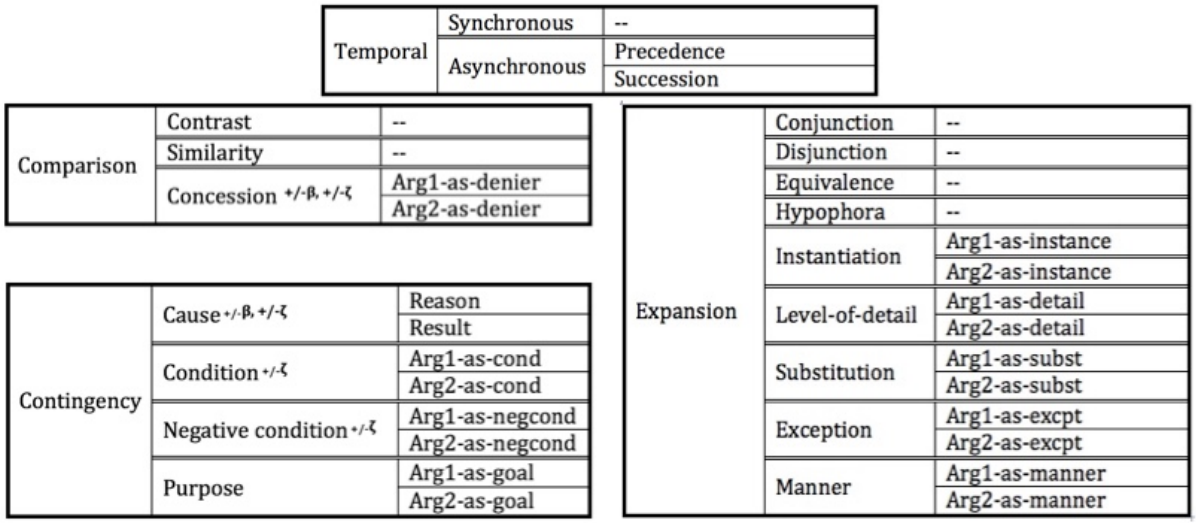

Figure 1: PDTB-3 Sense Hierarchy (Webber et al., 2016) Modified to Include Arg1/Arg2-as-instance and Hypophora. Only asymmetric relations are specified further at Level-3, to differentiate directionality of the arguments. Superscript symbols on Level-2 senses indicate features for implicit beliefs (+/-β) and speech-acts (+/-ζ) that may or may not be associated with one of the defined arguments of the relation.

(6) The executive producer of "Saturday Night With Connie Chung," Andrew Lack, declines to discuss recreations as a practice or his show, in particular. *"I don't talk about my work," he says.*
(NoRel)
The president of CBS News, David W. Burke, didn't return numerous telephone calls.

In adjacent contexts not related by a connective, an inferred relation is annotated as either an *implicit* relation (Ex. 3) when it can be expressed by inserting a connective, or an *AltLex* (alternatively lexicalized) relation (Ex. 4) if insertion of a connective leads to a perception of relation redundancy, indicating the presence of some alternative lexico-syntactic marking of the relation. When a discourse relation is not inferred, the context is annotated as *EntRel* (Ex. 5) if an entity-based relation is perceived, and as *NoRel* (Ex. 6) otherwise. Section 3.2 discusses in further detail how the EntRel and NoRel relations are used in PDTB.

Where a relation's arguments can be annotated depends on the type of relation. The Arg2 of explicit relations is always some part of the sentence or clause containing the connective, but the Arg1 can be anywhere in the prior text. For all other relation types, Arg1 and Arg2 are only annotated when adjacent. Arguments can be extended to include additional clauses/sentences in all cases except NoRel, but a minimality constraint requires inclusion of only the minimally necessary text needed to interpret the relation.

3 The Experiment

To identify challenges and explore the feasibility of annotating cross-paragraph implicit relations on a large scale, texts from the PDTB corpus were selected to cover a range of sub-genres (Webber, 2009) and lengths. These texts contained 440 current paragraph first sentence (CPFS) tokens (excluding the first sentence in each text) not already related to the prior text by an inter-sentential explicit connective. These tokens were annotated in the PDTB Annotation Tool (Lee et al., 2016) over the three phases described below.

3.1 Phase One

Phase One involved guidelines training and developing a preliminary understanding of the task. Two expert annotators worked together to discuss and annotate 10 texts (130 tokens) with the PDTB guidelines, except we did not enforce the PDTB adjacency constraint in order to explore the full complexity of the task. Each token was annotated for its type (Implicit, EntRel or Altlex), sense (Fig. 1), and minimal argument spans. From this exercise, two observations emerged. First, while 52% of the CPFS tokens took their prior (Arg1) argument from a unit involving the prior paragraph's last sentence (PPLS), the remaining 48% of the CPFSs took their Arg1 from somewhere else in the prior discourse, i.e. formed a non-adjacent relation. This suggested that the argument distri-

bution of cross-paragraph implicits was similar to that of cross-paragraph explicits, which are also non-adjacent roughly half (51%) the time (Prasad et al., 2010). Thus, whether this would be shown more generally became a hypothesis to explore in Phase Two.

Second, it was found that working together, the two annotators could isolate and agree upon the arguments not only of the adjacent implicit relations, but also the non-adjacent ones. Therefore, and also because of the observed high incidence of non-adjacent relations, a second hypothesis to explore in Phase Two became whether both adjacent and non-adjacent Arg1s could be reliably identified and annotated. Ex. (7) shows a CPFS (Arg2) and its Arg1 in a non-adjacent Contrast relation. In this case, the intervening material is excluded because of the minimality constraint: it only provides further detail about the Arg1 eventuality and can thus be excluded without loss of interpretation.

(7) Kidder, Peabody & Co. is trying to struggle back.

Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown, racked by internal squabbles and defections. Its relationship with parent General Electric Co. had been frayed since a big Kidder insider-trading scandal two years ago. Chief executives and presidents had come and gone.

(Contrast, But)
Now, the firm says it's at a turning point. By the end of this year, 63-year-old Chairman Silas Cathcart – the former chairman of Illinois Tool Works who was derided as a "tool-and-die man" when GE brought him in to clean up Kidder in 1987 – retires to his Lake Forest, Ill., home, possibly to build a shopping mall on some land he owns.

3.2 Phase Two

Based on Phase One observations, we decided in Phase Two to fully explore the feasibility of reliably annotating adjacent *and* non-adjacent cross-paragraph implicits. To this end, a further 103 tokens (10 texts) were separately annotated by each annotator for type, sense and minimal argument spans, regardless of whether arguments were adjacent or non-adjacent.

Table 1 presents the results of the Phase Two study. As shown, the adjacency distribution of arguments in the 76% (45%+31%) tokens agreed to be adjacent (46/103) or non-adjacent (32/103) supports our hypothesis that non-adjacent cross-paragraph implicit relations occur with high frequency (32/78, 41%), approaching half of all agreed tokens. For each of these agreed tokens, we computed sense and argument agreement to obtain

Arg1-Arg2 Tokens	Count	Pct	RelPct
Agree Adjacent:	**46**	**45%**	**100%**
Exact Match	11	11%	24%
Sent-level Match	3	3%	7%
Agree Sense, Args Overlap	14	14%	30%
Disagree Sense	18	17%	39%
Agree Non-Adjacent:	**32**	**31%**	**100%**
Exact Match	7	7%	22%
Sent-level Match	5	5%	16%
Agree Sense, Args Overlap	3	3%	9%
Agree Sense, Args Disagree	3	3%	9%
Disagree Sense	14	14%	44%
Disagree Adjacent/Non	**25**	**24%**	**100%**

Table 1: Cross-Paragraph Implicit Relations, Phase Two Agreement Counts, Percentages over all Tokens (Pct) and Relative Percentages over Subgroups (RelPct). 103 Tokens, 10 Texts.

(a) 'Exact Match', i.e., fully agreed for type, sense, and argument spans, (b) 'Sent-level match', i.e., slightly relaxing the minimality constraint subsententially to include tokens agreed for type and sense whose argument boundaries only disagreed inside a sentence boundary (e.g. because one annotator included an adjunct clause the other excluded), (c) 'Agree Sense, Args Overlap', i.e., relaxing the minimality constraint supra-sententially to include tokens agreed for type and sense whose Arg1 and Arg2 boundaries overlapped but did not exactly match (e.g. because one annotator included additional sentence(s) the other considered non-minimal), (d) 'Agree Sense, Args Disagree', i.e., agreed for type and sense but unmatched in all of the aforementioned ways, which can only occur for non-adjacent relations and not adjacent relations, and (e) 'Disagree Sense', i.e., disagreed as to type or sense, although arguments may or may not have matched in some way.

As the table shows, Exact Match agreement was low at 18% (11%+7%) for both adjacent (11/103) and non-adjacent (7/103) relations, illustrating the difficulty of the task. Agreement is boosted to 26% (26/103) when including Sentence-Level matches on argument spans (3 adjacent and 5 non-adjacent) and to 43% (43/103) when including tokens that matched for type and sense and had overlapping spans (14 adjacent and 3 non-adjacent), which we also take as the overall agreement on the task, with the most relaxed metric for argument span agreement. The table also shows that with this metric, agreement was worse for non-adjacent relations ((7+5+3)/32, 47%) than adjacent relations ((11+3+14)/46, 61%).

Discussion of the disagreements showed that while it was almost always possible to reach consensus, the time and effort required was often much greater for non-adjacent relations – twice the amount of time required for adjacent relations – and therefore prohibitive to large-scale annotation. Therefore a decision was made to maintain the PDTB adjacency constraint and focus on full annotation of only adjacent relations. Tokens perceived as forming a non-adjacent implicit relation would be annotated as **NoSemRel**, as described below, providing an underspecified marking to indicate its presence.

Also based on the Phase Two findings, two further enhancements were made to the PDTB-2 guidelines. First, two new senses were introduced (Fig. 1), as illustrated in Exs. (8-9). Our texts provide evidence of both directionalities for the asymmetric Instantiation sense, and so its Level-3 labels, **Arg1-as-instance** and **Arg2-as-instance**, were introduced. Arg2-as-instance is the more common case. In addition, a **Hypophora** label was introduced as a placeholder for question-answer pairs, until further study can shed light on the appropriate senses to capture their semantics.

(8) NBC's re-creations are produced by Cosgrove-Meurer Productions, which also makes the successful prime-time NBC Entertainment series Unsolved Mysteries.

(Arg1-as-instance, More generally)
The marriage of news and theater, if not exactly inevitable, has been consummated nonetheless.

(9) *How can we turn this situation around?*

(Hypophora)
Reform starts in the Pentagon.

The second enhancement involves a refinement of the EntRel and NoRel labels. In the absence of a semantic discourse relation between adjacent sentences, the PDTB-2 labels the relation between them as follows: **(a)** as EntRel if an entity-based coherence relation holds between Arg1 and Arg2 and the discourse is expanded around that entity in Arg2, either by continuing the narrative around it or supplying background about it; **(b)** as EntRel if (a) doesn't hold but some entity co-reference exists between Arg1 and Arg2 (even if an implicit relation also holds between Arg2 and a non-adjacent sentence); **(c)** as NoRel if neither (a) nor (b) holds (even if an implicit relation also holds between Arg2 and a non-adjacent sentence); and **(d)** as NoRel if none of (a)-(c) hold, which occurs when Arg2 is not part of the discourse (e.g., bylines or the start of a new article in a single WSJ file).

However, given our goal to encode the presence of non-adjacent implicit relations, the manner in which these labels are currently assigned is a problem because this information is spread across both labels, by way of scenarios (b) and (c) above. Further, (a) and (b) confound the presence of a semantic coherence relation with the presence of coreference. Both of these considerations therefore led us to create two new labels for our task: **SemEntRel** (Semantic EntRel) for scenario (a), to unambiguously identify cases of entity-based coherence relations, and **NoSemRel** for scenarios (b) and (c), to unambiguously identify cases of non-adjacent implicit relations. To maintain consistency with the PDTB-2 corpus, the EntRel label for (b) was noted as a comment feature where relevant. Scenario (d) continued to be labeled as NoRel.

A SemEntRel relation is shown in Ex. (10), where Arg2 provides background about the "humanitarian assistance" conceptual entity in Arg1. Though not yet applied to the rest of PDTB-2, we find Semantic Entrels occur quite frequently in cross-paragraph contexts (see Section 4). An example of a NoSemRel relation is the underspecified annotation of the non-adjacent relation of Ex. (7), shown below as Ex. (11).

(10) *And important U.S. lawmakers must decide at the end of November if the Contras are to receive the rest of the $49 million in so-called humanitarian assistance under a bipartisan agreement reached with the Bush administration in March.*

(SemEntRel)
The humanitarian assistance, which pays for supplies such as food and clothing for the rebels amassed along the Nicaraguan border with Honduras, replaced the military aid cut off by Congress in February 1988.

(11) Only a few months ago, the 124-year-old securities firm seemed to be on the verge of a meltdown, racked by internal squabbles and defections. Its relationship with parent General Electric Co. had been frayed since a big Kidder insider-trading scandal two years ago. *Chief executives and presidents had come and gone.*

(NoSemRel)
Now, the firm says it's at a turning point. By the end of this year, 63-year-old Chairman Silas Cathcart – the former chairman of Illinois Tool Works who was derided as a "tool-and-die man" when GE brought him in to clean up Kidder in 1987 – retires to his Lake Forest, Ill., home, possibly to build a shopping mall on some land he owns.

3.3 Phase Three

Employing the enhancements to the PDTB-2 guidelines developed during Phase Two, 207

CPFS-PPLS implicit relation tokens from 34 texts were separately annotated by the two annotators in Phase Three for type, sense and minimal argument spans. However, prior to initiating the Phase Three annotation, all Phase One and Phase Two texts were reannotated by the two annotators according to the enhanced guidelines, and a close analysis of the disagreements was performed. This yielded three recurring patterns of disagreements as well as procedures for resolving them via careful application of the guidelines, detailed below.

a) Multi-sentential or discontinuous arguments may exclude supporting relations. Minimality requires that all and only the semantic material minimally needed to interpret a relation be specified by its arguments. Therefore, relations that support Arg1 and Arg2 but aren't necessary for their interpretation should be excluded from those arguments' boundaries. Common supporting relations typically excluded include Arg2-as-Instance, Arg2-as-Detail, and Reason, as well as Semantic Entrel or Temporal relations that supply background information. Ex. (12) shows supporting sentences after the CPFS that are excluded from Arg2 for minimality.

(12) *Although bullish dollar sentiment has fizzled*, many currency analysts say *a massive sell-off probably won't occur in the near future*.

(Implicit, Reason, because)
While Wall Street's tough times and lower U.S. interest rates continue to undermine the dollar, weakness in the pound and the yen is expected to offset those factors. "By default," the dollar probably will be able to hold up pretty well in coming days, says Francoise Soares-Kemp, a foreign-exchange adviser at Credit Suisse. "We're close to the bottom" of the near-term ranges, she contends.

b) A CPFS may appear to relate to both an adjacent and a non-adjacent unit. Often, however, the adjacent unit will be providing supporting content to the non-adjacent unit, rather than continuing the more global narrative flow. The stronger relation in this case will be the non-adjacent one. E.g., in Ex. (13), Arg2 creates an Instantiation relation regarding the names of specific judges to be included. Some annotators may perceive this relation as capable of being formed with the prior adjacent sentence or the non-adjacent italicized one. However, the prior adjacent sentence itself provides supporting detail on the italicized one, concerning the number of judges to be included. Thus, the adjacent sentence and the bolded sentence are neither directly related themselves, nor advancing the more global narrative flow. Therefore, this token is labeled NoSemRel.

(13) Several organizations, including the Industrial Biotechnical Association and the Pharmaceutical Manufacturers Association, have asked the White House and Justice Department to name candidates with both patent and scientific backgrounds. *The associations would like the court to include between three and six judges with specialized training.*

(NoSemRel)
Some of the associations have recommended Dr. Alan D. Lourie, 54, a former patent agent with a doctorate in organic chemistry who now is associate general counsel with SmithKline Beckman Corp. in Philadelphia.

c) Multiple tokens can relate differently to the same sentence. Often in the PDTB, texts begin with a single complex sentence followed by other sentences or paragraphs each discussing some aspect of it. By minimality, tokens should only be grouped into a single Arg2 if they share the same relation to the same Arg1 unit. The text in Ex. 7 provides an illustration of this. The italicized and bolded CPFSs together form the Arg2 of an Arg2-as-detail relation with the first sentence, providing detail on the eventuality of the company trying to struggle back. In contrast, in Ex. (14), the bolded Arg2 in the first CPFS provides detail on the trade deficit worsening in the first sentence. The bolded Arg2 in the second CPFS, on the other hand, displays entity coreference with the first bolded unit, but more generally and strongly, continues the global narrative flow about the Treasury Department's statement, that is, it is in a SemEntRel relation with the non-adjacent Arg1 (in italics). Given the new guidelines for Phase Three, the relation is thus labeled NoSemRel.

(14) The Treasury Department said *the U.S. trade deficit may worsen next year, after two years of significant improvement.*

(Implicit=Arg2-as-detail)
In its report to Congress on international economic policies, the Treasury said **that any improvement in the broadest measure of trade, known as the current account, "is likely at best to be very modest," and "the possibility of deterioration in the current account next year cannot be excluded."**

(NoSemRel)
The statement was the U.S. government's first acknowledgement of what other groups, such as the International Monetary Fund, have been predicting for months.

Arg1-Arg2 Pairs	Count	Pct	RelPct
Agree Adjacent:	**95**	**46%**	**100%**
Exact Match	40	19%	42%
Sent-level Match	13	7%	14%
Agree Sense, Args Overlap	12	6%	13%
Disagree Sense	30	14%	31%
Agreed Non-Adjacent:	**63**	**30%**	**100%**
Disagreed Adjacent/Non	**49**	**24%**	**100%**

Table 2: Cross-Paragraph Implicit Relations, Phase Three Agreement, 207 Tokens, 34 Texts.

4 Results and Discussion

Table 2 presents the Phase Three inter-annotator agreement results. As shown, agreement on whether a relation was adjacent (95) or non-adjacent (63) was approximately the same as in Phase Two, at 76% (46%+30%), Furthermore, over these 158 (95+63) tokens, the proportion of non-adjacent tokens (63/158, 40%) was similar to Phase Two, again supporting our hypothesis about their high frequency. Because of the backoff to annotating only adjacent cross-paragraph implicit relations, overall agreement with the most relaxed metric on argument spans[1] is higher in Phase Three (62%) than in Phase Two (43%). However, there is also substantial improvement in the sense annotation of the adjacent discourse relations, from 61% in Phase Two to 69% (42%+14%+13%) in this phase,[2] which we attribute partly to our enhanced guidelines for annotating SemEntRel. The increase in tokens agreed on sense also more accurately represents the agreement on arguments. Exactly matched arguments show an increase to 42% from 24% in Phase Two and there are fewer disagreements due to supra-sentential overlapping spans, which have reduced to 13% from 30% in Phase Two. The number of sentence-level disagreements increased to 14% from 7% in Phase Two, but most of these reflect minor syntactic differences (e.g., inclusion/exclusion of adjuncts or attributions) rather than semantic ones.

Following Phase Three, gold standard annotations were produced through consensus labeling over all phases. Table 3 shows the counts and percentages for each token type. Of the 440 tokens, 207 (47%) conveyed a non-adjacent relation and thus the adjacent relation was labeled NoSemRel, confirming our initial hypothesis of an almost equal distribution of cross-paragraph adjacent and non-adjacent implicit relations. Among the remaining 233 (53%) tokens, 153 (35%) were of the Implicit type in that a connective could be inserted to express the relation, while 8 (2%) conveyed the relation through an AltLex. 62 (14%) tokens were annotated as SemEntRels, and 11 (3%) were annotated as NoRels. Table 4 presents the counts and percentages for the Implicit and AltLex gold-labeled senses. As shown, Arg2-as-Detail occurs most frequently but still accounts for only 40% of the relations. Six other senses occurring with 5% or greater frequency account for 45% of the tokens, and include Conjunction (12%), Arg2-as-instance (9%), Reason (7%), Result (6%), Arg2-as-denier (6%) and Contrast (5%). The remaining 15% of the tokens occurring with less than 5% frequency are spread across nine different senses.

	Implct	AltLex	SemEnt	NoSmRel	NoRel
Ct	152	8	62	207	11
Pct	35%	2%	14%	47%	3%

Table 3: Gold Cross-Paragraph Implicit Relation Counts and Percentages Across All Phases, 440 Tokens, 54 Texts.

5 Related Work

Given that the end goal of this research is to produce full-text annotation of discourse relations, in this section we compare our work with two related approaches to full text discourse relation annotation, focusing on how they handle non-adjacent discourse relations, or in other words, long-distance discourse relation dependencies.

In the RST-based (Mann and Thompson, 1988) RST-DT corpus (Carlson et al., 2003), texts are first segmented into elementary discourse units (EDUs) and relations are then built recursively (i.e., as trees) between increasingly complex adjacent structures. Long-distance dependencies come about when the "nuclear" elements within a pair of complex adjacent structures are not adjacent in the text. In this approach, then, long-distance dependencies fall out as a function of the theory and its implementation in the annotation procedure. A disadvantage of such an approach, however, is that it tends to undervalue the evaluation and intuition of annotators with regards to such dependencies (Stede, 2012). As illustration, in the RST-DT tree (Fig. 2) for Ex. (15), the Antithesis relation clearly

[1] Exact Match + Sent-Level Match + Agree Sense, Args Overlap + Agreed Non-Adjacent

[2] The sense agreement for this task is on par with the agreement for intra-paragraph implicit relations reported in Miltsakaki et al. (2004).

Types	Senses (Count/Relative Percent of 160)							
	Detail2	Conjunction	Instance2	Reason	Result	Denier2	Contrast	Precedence
Implicit	62/39%	18/11%	13/8%	11/7%	10/6%	9/6%	8/5%	7/4%
AltLex	2/1%	1/<1%	2/1%	0	0	0	0	0
	Equivalence	Reason+β	Detail1	Instance1	Synchronous	Hypophora	Result+β	Succession
Implicit	3/2%	3/2%	2/1%	2/1%	1/<1%	1/<1%	1/<1%	1/<1%
AltLex	0	1/<1%	0	0	1/<1%	0	0	1/<1%

Table 4: Gold Cross-Paragraph Adjacent Implicit and AltLex Sense Counts and Relative Percentages Across All Phases, 160 Tokens. Detail(1/2) = Arg(1/2)-as-detail; Instance(1/2) = Arg(1/2)-as-instance; Denier2 = Arg2-as-denier.

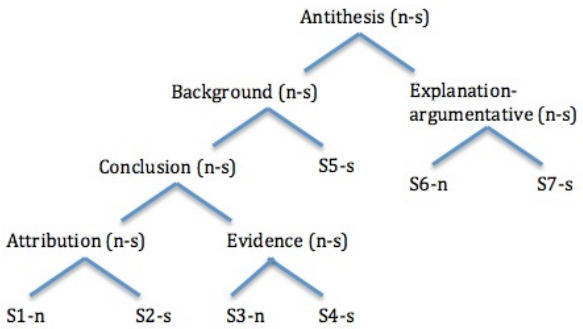

Figure 2: RST Structure for Ex. (15). Intra-sentential relations are not shown. Nodes are labeled with RST mononuclear (n-s) or multinuclear (n-n) relations and leaves are anchored by sentences IDs marked with their nuclearity status.

seems to hold between S3 and S6, but this does not fall out from the RST-DT annotation, where S1 is promoted as the nucleus of the S1-S5 complex, not S3.

(15) **S1:** FEDERAL PROSECUTORS are concluding fewer criminal cases with trials.

S2: That's a finding of a new study of the Justice Department by researchers at Syracuse University.
S3: David Burnham, one of the authors, says fewer trials probably means a growing number of plea bargains.
S4: In 1980, 18% of federal prosecutions concluded at trial; in 1987, only 9% did.

S5: The study covered 11 major U.S. attorneys' offices – including those in Manhattan and Brooklyn, N.Y., and New Jersey – from 1980 to 1987.

S6: The Justice Department rejected the implication that its prosecutors are currently more willing to plea bargain.
S7: "Our felony caseloads have been consistent for 20 years," with about 15% of all prosecutions going to trial, a department spokeswoman said.

Like the RST-DT corpus, The SDRT-based (Asher and Lascarides, 2003) ANNODIS corpus (Afantenos et al., 2012) also constructs hierarchical structures - termed complex discourse units (CDUs) - out of EDUs. A structure like Fig. 2 is thus possible in that corpus. However, CDUs are explicitly distinguished from EDUs in ANNODIS and there is at present no analogous concept of nuclearity within the theory that would promote some EDU(s) to become the prominent nucleus of the complex. The problem of identifying minimal arguments in long-distance dependencies is therefore sidestepped in the corpus; instead, the whole CDU serves as the argument. Nevertheless, identifying minimal arguments based on some principle, whether through annotation guidelines such as PDTB's "minimality constraint" or through theoretical mechanisms such as RST-DT's "nuclearity principle", is important in eliminating noise from the arguments. For example, a learning algorithm extracting features from non-minimal argument spans for sense labeling would wind up with a lot of extraneous or conflicting data. It is also an open question as to whether the speaker/hearer retains or requires such hierarchically-structured non-minimal complex units when establishing/interpreting discourse relations in speech/text. In many other respects, however, the ANNODIS approach is on par with the one addressed in this paper. Relations are defined in semantic terms, and long-distance relations are annotated regardless of whether or not they may lead to crossing dependencies in the *emergent* composite discourse structures.

6 Conclusion and Future Work

In sum, our study shows that adjacent implicit discourse relations across paragraphs can be annotated reliably. Furthermore, the gold-standard sense distributions found in our study, together with the frequency of Semantic EntRels, suggest that cross-paragraph implicit relations carry varied semantic content in substantial proportions and are therefore worth annotating. Given this, one goal

of our future work is to annotate ˜200 texts of the PDTB corpus with adjacent cross-paragraph implicit relations, following the enhanced guidelines developed here, and publicly distribute the annotations via github.[3] The subset of texts to be annotated contain approximately 700 tokens of cross-paragraph implicit relations, which we have estimated (from our Phase1 to Phase3 annotations) to require 3 minutes per token on average, i.e., approximately 35 hours of annotation time per annotator. Once this corpus is completed, we can then study the distribution of senses and patterns of senses in the texts, along the lines of Pitler et al. (2008), but now over full text relation sequences. In addition, the high incidence of the underspecified implicit non-adjacent relations found in this study suggests the value of developing guidelines for their more difficult annotation to ensure it can be done reliably, and thus, this is a goal of our future work as well.

More generally, our study is the first to quantitatively assess the difficulty of annotating long-distance discourse relation dependencies. We find that annotating non-adjacent cross-paragraph implicit relations is difficult and time-consuming. Another future goal is, therefore, to develop more effective tools and methodologies to increase annotation ease, speed and reliability. These include enhancements to the PDTB annotation tool to better allow simultaneous visualization of inter-sentential relations and their arguments in a text. In addition, a two-pass annotation methodology would allow the more difficult cross-paragraph non-adjacent implicit relations to be annotated in a second pass. Sequences of inter-sentential relations from the first pass could then reveal systematic structures to inform the second pass.

Acknowledgments

This work was partially supported by NSF grant IIS-1421067.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Ccile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pry-Woodley, Laurent Prevot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*. Istanbul, Turkey, pages 2727–2734.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Sasha J. Blair-Goldensohn. 2007. *Long-Answer Question Answering and Rhetorical-Semantic Relations*. Ph.D. thesis, Columbia University.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In Jan van Kuppevelt and Ronnie Smith, editors, *Current Directions in Discourse and Dialogue*, Kluwer Academic Publishers, pages 85–112.

Ruihong Huang and Ellen Riloff. 2012. Modeling textual cohesion for event extraction. In *Proceedings of the 26th Conference on Artificial Intelligence*. Toronto, Canada, pages 1664–1670.

Alan Lee, Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2016. Annotating discourse relations with the PDTB annotator. In *Proceedings of the 26th International Conference on Computational Linguistics: System Demonstrations*. Osaka, Japan, pages 121–125.

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Tokyo, Japan, pages 147–156.

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

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn TreeBank. *Computational Linguistics* 19(2):313–330.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. Annotating discourse connectives and their arguments. In *Proceedings of the HLT/NAACL Workshop on Frontiers in Corpus Annotation*. Boston, MA, pages 9–16.

Stephan Oepen, Jonathon Read, Tatjana Scheffler, Uladzimir Sidarenka, Manfred Stede, Erik Velldal, and Lilja Øvrelid. 2016. OPT: Oslo-Potsdam-Teesside pipelining rules, rankers, and classifier ensembles for shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task on Multilingual Shallow Discourse Parsing*. Berlin, Germany, pages 20–26.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Suntec, Singapore, pages 683–691.

[3] http://www.github.com

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics*. Manchester, UK, pages 87–90.

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 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, pages 2961–2968.

Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2010. Exploiting scope for shallow discourse parsing. In *Proceedings of the 7th International Conference on Language Resources and their Evaluation*. Valletta, Malta, pages 2076–2083.

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

Swapna Somasundaran, Janyce Wiebe, and Josef Ruppenhofer. 2008. Discourse level opinion interpretation. In *Proceedings of the 22nd International Conference on Computational Linguistics, Volume 1*. Manchester, U.K., pages 801–808.

Manfred Stede. 2012. *Discourse processing*. Synthesis Lectures on Human Language Technologies (series editor, Graeme Hirst). Morgan & Claypool Publishers.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the CoNLL-2015 Shared Task*. Beijing, China, pages 17–24.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Suntec, Singapore, page 674–682.

Bonnie Webber, Aravind Joshi, Matthew Stone, and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics* 29(4):545–587.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop*. ACL, Berlin, Germany, pages 22–31.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the CoNLL-15 shared task*. Beijing, China, pages 1–16.

Nianwen Xue, Hwee Tou Ng, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*. Berlin, Germany, pages 1–19.

Fan Zhang, Diane Litman, and Katherine Forbes Riley. 2016. Inferring discourse relations from PDTB-style discourse labels for argumentative revision classification. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pages 2615–2624.