Where Are We in Discourse Relation Recognition?

Katherine Atwell
Dept. of Computer Science
University of Pittsburgh
kaa139@pitt.edu

Junyi Jessy Li
Dept. of Linguistics
The University of Texas at Austin
jessy@austin.utexas.edu

Malihe Alikhani
Dept. of Computer Science
University of Pittsburgh
malihe@pitt.edu

Abstract

Discourse parsers recognize the intentional and inferential relationships that organize extended texts. They have had a great influence on a variety of NLP tasks as well as theoretical studies in linguistics and cognitive science. However it is often difficult to achieve good results from current discourse models, largely due to the difficulty of the task, particularly recognizing implicit discourse relations. Recent developments in transformer-based models have shown great promise on these analyses, but challenges still remain. We present a position paper which provides a systematic analysis of the state of the art discourse parsers. We aim to examine the performance of current discourse parsing models via gradual domain shift: within the same corpus, on in-domain texts, and on out-of-domain texts, and discuss the differences between the transformer-based models and the previous models in predicting different types of implicit relations both inter- and intra-sentential. We conclude by describing several shortcomings of the existing models and a discussion of how future work should approach this problem.

1 Introduction

Discourse analysis is a crucial analytic level in NLP. In natural language discourse, speakers and writers often rely on implicit inference to signal the kind of contribution they are making to the conversation, as well as key relationships that justify their point of view. While early AI literature is full of case studies suggesting that this inference is complex, open-ended and knowledge-heavy (e.g., Charniak (1973); Schank and Abelson (1977)), recent work on computational discourse coherence offers a different approach. Take the following example from Pitler and Nenkova (2008):

(1) *"Alice thought the story was predictable. She found it boring."*

This discourse shows the classic pattern of implicit information. The overall point is that Alice had a negative opinion of the story: the underlying explanation is that the story was not interesting because it had no surprises. But given available lexical resources and sentiment detection methods, we can capture such inferences systematically by recognizing that they follow common general patterns, known as "discourse relations", and are guided by shallow cues.

An example of an instance in which discourse analysis can produce insights that may be missed by employing other NLP methods is this example from Taboada (2016), where without discourse relations it may be difficult to capture sentiment:

(2) *"While this book is totally different from any other book he has written to date, it did not disappoint me at all."*

This represents a *Concession* relation according to both Rhetorical Structure Theory and the Penn Discourse Treebank (where it is notated as *Comparison*.Concession), resolving the incongruity of the first clause being negative and the second clause being positive by illustrating how the negative statement in the subordinate clause is reversed by the positive one in the main clause.

The importance of discourse has led to active research based on predicting what *coherence relations* are present in text based on shallow information. The predicted relations are then used to draw inferences from the text. The value of predicting *the semantic classes of coherence relations* has been demonstrated in several applications, including sentiment analysis (Marcu, 2000; Bhatia et al., 2015), machine comprehension (Narasimhan and Barzilay, 2015), summarization (Cohan et al., 2018; Marcu, 1999; Xu et al., 2019; Kikuchi et al., 2014), and predicting instructor intervention in an online course discussion forum (Chandrasekaran

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 314–325
July 29–31, 2021. ©2021 Association for Computational Linguistics

et al., 2017). However, it is still the case that few works have so far found discourse *relations* as key features (Zhong et al., 2020). We argue that one reason for this gap between theory and empirical evidence is the quality of the parsers exacerbated by the distributional shifts in the texts they need to apply to.

The necessity of discourse research has resulted in several shared tasks (Xue et al., 2015, 2016) and corpora development in multiple languages (Zeyrek and Webber, 2008; Meyer et al., 2011; Danlos et al., 2012; Zhou et al., 2014; Zeyrek et al., 2020). Yet shallow discourse parsing is a very difficult task; more than 10 years after the introduction of the Penn Discourse Treebank (Eleni Miltsakaki, 2004), performance for English implicit discourse relation recognition has gone from 40.2 F-1 (Lin et al., 2009) to 47.8 (Lee et al., 2020), less than 8 percentage points; a similar story could be said about the relation prediction performance of RST parsers. Such performance hinders the wider application of parsers. If downstream tasks are to use predicted relation senses, the data to which the systems are applied is typically different from their training data—the Wall Street Journal (WSJ) in a 3-year window—to varying degrees. This tends to further aggravate the low performance observed. As a result, often we find that adding *parsed* discourse relations into models are unhelpful.

Although domain difference is a recognized issue in shallow discourse parsing by existing work (Braud et al., 2017; Liu et al., 2016), we still have little understanding of the types of distributional shift that matter and by how much, even within one language. This position paper seeks to shed some light on our current state in discourse parsing in English. Surprisingly, we found that parsers have some issues even within the same news source as the training set (WSJ); the differences in accuracy were not significant between in-domain and out-of-domain data for the qualitative examples that we looked at, although the distribution of errors tend to be different. This differs from other NLP tasks such as entity recognition, where training on data in the target domain increased the F1 score by over 20 points (Bamman et al., 2019).

We further found that parsers perform differently on implicit discourse relations held within vs. across sentences. We believe these findings are strong evidence for the sensitivity of existing models to distributional shift in terms of both linguistic structure and vocabulary.

Additionally, as part of our evaluation, we asked linguists to perform manual annotation, which allowed us to evaluate the accuracy of these parsers on plain, unlabeled text, and gain some insight about the mistakes made by the parsers. During the annotation process, we uncovered information that can guide future research, including but not limited to the critical role of context for implicit discourse sense classification. We discuss this need for context, hypothesize what scenarios may cause two arguments to need additional context, and provide some examples for which this is the case. We urge future researchers to consider developing context-aware models for shallow discourse parsing moving forward. We release our dataset to facilitate further discourse analysis under domain shift. [1]

2 Related Work

There are various frameworks for studying inferential links between discourse segments, from local shallow relations between discourse segments in PDTB (Rashmi Prasad, 2008) to hierarchical constituent structures in RST (Carlson et al., 2003) or discourse graphs in Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) and the Discourse Graphbank (Wolf and Gibson, 2005).

Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) provides a hierarchical structure for analyzing text that describes relations between text spans known as elementary discourse units (EDUs). The RST Discourse Treebank (Carlson et al., 2003) contains 385 Wall Street Journal articles from the Penn Treebank (Marcus et al., 1993) which have been split into elementary discourse units and annotated according to Rhetorical Structure Theory, where discourse relations are annotated in a tree structure across the whole document. A full list of these relations can be found in Carlson and Marcu (2001).

The Penn Discourse Treebank (PDTB) (Eleni Miltsakaki, 2004; Rashmi Prasad, 2008; Prasad et al., 2018), which also uses Penn Treebank Wall Street Journal articles, contains discourse relations annotated in a shallow, non-hierarchical manner. For each relation between two arguments, each argument and the discourse connective (word or phrase that indicates the discourse relation) are labeled. The PDTB also annotates whether a

[1] Our data is located here: https://github.com/katherine-atwell/discourse-domain-shift

relation is explicit or non-explicit, the latter type of which has three subtypes: Implicit, AltLex, and EntRel. In this paper, we focus on implicit relations, where a connective can be inserted between the two arguments that indicates a discourse relation. These relations are considered extremely challenging for discourse parsers to automatically identify.

There is a need to examine the performance of the proposed discourse parsers, their representational choices, their generalizability, and interpretability both across domains, distributions, and frameworks. One recently developed framework is the PDTB-3. Since its release in 2019, several papers have evaluated the performance of implicit sense classifiers on this new corpus, which includes newly annotated intra-sentential implicit discourse relations. In addition to proposing a new evaluation framework for PDTB, Kim et al. (2020) evaluate the performance of pretrained encoders for implicit sense classification on the PDTB-2 and the PDTB-3. Liang et al. (2020) identify locating the position of relations as a new challenge in the PDTB-3, due to the significantly increased number of intra-sentential implicit relations annotated.

Techniques of discourse parsing range from supervised (Liu et al., 2019; Mabona et al., 2019; Lin et al., 2019; Zhang et al., 2020; Kobayashi et al., 2020) and weakly supervised and unsupervised approaches (Lee et al., 2020; Nishida and Nakayama, 2020; Kurfalı and Östling, 2019); recent developments such as word/contextual embeddings have improved parser performance, although not as significantly as other tasks (Shi and Demberg, 2019; Chen et al., 2019) Yet most works have made simplifying assumptions concerning the linguistic annotations for practical purposes that affect their evaluation and generality. For instance, most shallow discourse parsers use only the argument pairs to determine the discourse sense without considering further context. Additionally, in RST parsing, standard practice involves classifying only the 18 top-level RST classes (Hernault et al., 2010; Feng and Hirst, 2014; Morey et al., 2017). Thus, all *Elaboration* relations are lumped together, making it a huge class. We reveal findings about these assumptions in Section 4.

Other works evaluating discourse parsers include DiscoEval (Chen et al., 2019), a test suite of evaluation tasks that test the effectiveness of different sentence encoders for discourse parsers, and an improved evaluation protocol for the PDTB-2 (Kim et al., 2020). In contrast, our work aims to analyze and evaluate existing discourse parsers via gradual domain shift. We provide a comparative genre-based analysis on distributionally shifted text data and present a qualitative analysis of the impact of the practical choices that these models make while doing discourse parsing across frameworks.

3 Where are we in discourse parsing?

3.1 Experiments

Data. We start by focusing on possible distributional shifts in a shallow parser's application, by considering different linguistic types of implicit discourse relations (inter- vs intra-sentential) (Liang et al., 2020). To do this, we evaluate performance on the PDTB-2 and PDTB-3, as well as the intra-sentential relations in the PDTB-3 specifically.

We then evaluate the performance of three widely used or state-of-the-art models under gradual shift of the domain of texts, noting that users who would want to use a parser will be applying it on data that varies linguistically to different degrees from the parser's training data (a fixed 3-year window of WSJ articles). The data we examine is: WSJ texts outside of the Penn Treebank , other news texts, and the GUM corpus (Zeldes, 2017). Note that none of these texts contain gold PDTB annotations, and only the GUM corpus contains gold RST annotations.

Setup. To examine the impact of changing the linguistic distribution by introducing intra-sentential discourse relations, we run the model developed by Chen et al. (2019) using the same train-test split as the authors and training/testing on discourse senses which contain 10 or more examples. To get results for the PDTB-2, we train and test the model on the PDTB-2; to get results for the PDTB-3 and intrasentential relations in the PDTB-3, we train the model on the PDTB-3 and evaluate its performance on both of these sets.

To parse plain-text documents for PDTB relations, we use the Wang and Lan (2015) parser as our end-to-end parser and the Chen et al. (2019) DiscoEval parser as our implicit sense classifier. The former is needed in order to parse unlabeled text, and the latter is a more accurate BERT-based implicit sense classifier (implicit sense classification is the most difficult PDTB parsing task). To evaluate these parsers, we look at quantitative as-

	PDTB-2	PDTB-3	PDTB-3 Intra-Sent
Base	0.4236	0.4897	0.6251
Large	0.4358	0.5094	0.6251

Table 1: Accuracy of the BERT-based model described in Chen et al. (2019) on implicit relations in the PDTB.

pects of their output (e.g. the distributions) and qualitative aspects (manual annotation and inspection of parser output).

For our RST experiments, we use the state-of-the-art (Wang et al., 2017) parser. We evaluate the performance of this parser on the standard RST Discourse Treebank test set with a 90-10 split (347 training documents and 38 test documents). We also evaluate it on the gold labels from the GUM corpus (but trained on the RST). Because GUM is annotated with 20 different discourse relations which do not precisely map to the conventional 18 types used in the Wang et al. (2017) parser, we map the ones that don't match these types or the more fine-grained relations in the following manner, following Braud et al. (2017): *preparation* to BACKGROUND, *justify* and *motivation* to EXPLANATION, and *solutionhood* to TOPIC-COMMENT.

For the plain-text news articles from outside of the PDTB corpus, we mirror the PDTB experiments on these documents by parsing them with the (Wang et al., 2017) parser, then examining the resulting distributions and manually inspecting the parser output.

3.2 Findings

Transformer-based models perform better on linguistically different intra-sentential relations than they do on inter-sentential relations. As mentioned above, we aim to examine the results of distributional shifts in both vocabulary and linguistic structure. Here, we look at shifts in linguistic structure, namely, inter- vs. intra-sentence implicit discourse relations (Hobbs, 1985). The latter was introduced in the PDTB-3 (Liang et al., 2020) from which we show the following example:

(3) ...Exxon Corp. *built the plant* **but** (Implicit=then) **closed it in 1985**

Unlike the inter-sentence relations that were annotated across adjacent sentences, implicit intra-sentence relations do not occur at well-defined positions, but rather between varied types of syntactic constituents. Additionally, they often co-occur with explicit relations.

Table 1 shows the accuracies of the base and large BERT model (Chen et al., 2019) on the implicit relations in the two versions of the PDTB. The results on the PDTB-3 are significantly better than those of the PDTB-2, and the model tested on the PDTB-3 intra-sentential relations significantly outperformed both ($p<0.01$, $t>11.172$). This mirrors the results found from running the baseline model in Liang et al. (2020) on the PDTB-2, PDTB-3, and PDTB-3 intra-sentential relations.

Figure 1 shows the accuracy of the Wang et al. (2017) parser on the inter-sentential and intra-sentential relations in the RST, respectively. For the inter-sentential relations, we sampled only the relations between two sentences to have a "fairer" comparison (it is well known that performance suffers on higher levels of the RST tree). As with the PDTB, these results show a significant improvement in performance when run on only the intra-sentential relations compared to only the inter-sentential relations.

These results drive home the influence of the linguistic and structural differences between intra- and inter-sentence implicit relations on the performance of the parsers. We initially found this surprising since intra-sentence ones contain arguments with less information than their (full-sentence) inter-sentence counterparts. However, one explanation for this is that, while looking for relations within sentence boundaries is a problem that has been very explored, and to some extent solved, in various NLP tasks (e.g. syntactic parsing), there are not as many rules regarding relations that occur across sentence boundaries. Regardless of the cause, these results illustrate that future shallow discourse parsers may benefit from accounting for such linguistic differences explicitly.

Parsers struggle to identify implicit relations from less frequent classes. The second distributional shift we examine is a shift in vocabulary. In order to capture this, we measure the performance across several domain shifts from the PDTB-2 using three datasets: WSJ articles from the COHA corpus (Davies, 2012), other news articles from COHA, and the GUM corpus (Zeldes, 2017). The WSJ articles are completely within the domain of the PDTB, but more shifted in timeline than the PDTB test set. The other news articles are in-domain as well, but not from the same source

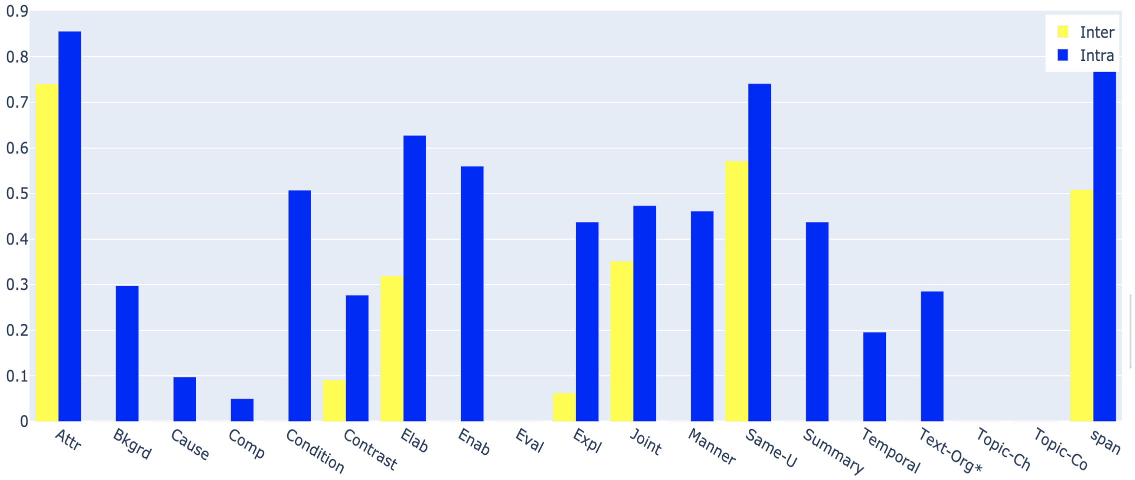

Figure 1: F-1 scores for running the Wang et al RST parser on the RST Discourse Treebank for inter-sentential (yellow) and intra-sentential (blue) relations (* denotes that this relation was not included in the set of inter-sentential relations). We can see from this graph that the performance of the parser was improved for the intra-sentential relations compared to the inter-sentential relations.

publication, and thus may be linguistically different. The GUM corpus, our out-of-domain dataset, contains data from eight domains: Academic, Bio, Fiction, Interview, News, Travel, How-to guides, and Forum Discussions. It contains gold RST annotations but no PDTB annotations.

To quantitatively evaluate the performance of these parsing models, we examine the distribution of the parser predictions and how frequently different senses are predicted. From this, we noticed that only 5 out of the 16 PDTB-2 level 2 senses were predicted at all by the Wang and Lan parser, and only 7 out of 16 were predicted by the DiscoEval parser. Of these classes, several were predicted less than 2% of the time (Table 6).

We can also see that in Tables 2 and 3, the Wang et al parser predicted at least 38.7% *Contingency*.Cause for all datasets and the DiscoEval parser predicted at least 44% *Contingency*.Cause, although these percentages were often much higher. Because only 24.9% of the total relations contained in the PDTB are *Contingency*, this overrepresentation of *Contingency*.Cause in the predictions indicates a strong bias towards *Contingency*. Indeed, many of the errors found during annotation occurred when the parser predicted *Contingency*.Cause, the most common level 2 sense, over a less represented class such as *Comparison*.Contrast; the precision for Contingency.Cause was 0.33, 0.14, and 0.33 for WSJ articles, non-WSJ news articles, and the GUM corpus respectively. This likely contributed to the low accuracy for these documents. These results show us that if PDTB parsers are run on plain text documents, whether in-domain or slightly shifted, the results are likely to be overconfident with majority classes and unlikely to predict minority classes.

Wang et al. Level-2 Predictions			
Sense	WSJ	other news articles	GUM
Expansion.Conjunction	15.2	22.7	12.1
Expansion.Instantiation	2.4	1.5	0.7
Expansion.Restatement	30.9	36.1	29.5
Comparison.Contrast	0.3	0.9	0.9
Contingency.Cause	51.3	38.7	56.7

Table 2: Percentages of Level-2 senses predicted by the Wang and Lan (2015) parser on the Penn Discourse Treebank on Wall Street Journal articles, other news articles, and the GUM corpus. All other 11 senses not included in this table were not predicted by the parser at all.

We also obtained the predicted distributions of the RST relations (Table 4) on the COHA news articles; we examined these results for the set of WSJ articles as well as the other news articles. We found that relations that are highly represented in the RST Discourse Treebank such as Elaboration, Attribution, and Same Unit were predicted much more frequently than they appear in the RST. However, more minority classes were represented in

BERT Level-2 Predictions			
Sense	WSJ	other news articles	GUM
Temporal.Asynchronous	1.3	1.6	4.2
Expansion.Conjunction	16.4	20.9	19.6
Expansion.Instantiation	2.1	2.3	1.0
Expansion.List	.7	.4	2.8
Expansion.Restatement	22.9	27.2	21.8
Comparison.Contrast	2.1	3.1	1.0
Comparison.Concession	0	.02	0
Contingency.Cause	54.3	44.4	49.1

Table 3: Level-2 senses predicted by the BERT-based model described in Chen et al. (2019) on the Penn Discourse Treebank on Wall Street Journal articles, other news articles, and the GUM corpus. All other 9 senses not included in this table were not predicted by the parser at all, and thus were predicted 0% of the time.

these predictions than in the PDTB parser's.

Predicted RST Relation Percentages		
	WSJ Articles	Other News Texts
Attribution	22.02	21.38
Background	2.66	2.98
Cause	0.94	0.79
Comparison	0.90	0.49
Condition	2.96	1.93
Contrast	4.69	3.86
Elaboration	31.47	32.92
Enablement	4.58	4.20
Evaluation	0.04	0.01
Explanation	0.56	0.71
Joint	9.49	9.21
Manner-Means	1.13	1.04
Same-Unit	17.2	19.31
Temporal	1.31	1.18

Table 4: Distribution of relations predicted by running the Wang et al. (2017) parser on COHA news articles. The 4 relations not listed here were not predicted at all by the parser.

,

	WSJ Articles		Other News		GUM Corpus	
Level Correct	Wang et al	Disco Eval	Wang et al	Disco Eval	Wang et al	Disco Eval
None	46.7	60.0	35.3	41.2	43.8	23.8
Level1	20.0	6.7	29.4	44.4	21.9	28.1
Level2	33.3	33.3	35.3	29.4	34.4	28.1

Table 5: Resulting accuracies from annotating a sample of implicit PDTB relations and comparing these annotations to the output of the Wang and DiscoEval parsers

Models fail to generalize to both in-domain and out-of-domain data, and different errors are seen for different domains. We continue to analyze the effects of a change in the distribution of vocabulary by qualitatively analyzing the results of our discourse parsers through manual inspection. To qualitatively evaluate the results of the PDTB parsers across domains, we randomly selected 64 implicit relations predicted by the parsers and asked two expert linguists (a faculty member and a graduate student at a linguistics department) to annotate them. These annotations allow us to evaluate the accuracy of the parsers, since none of the documents we are looking at (Wall Street Journal articles in the COHA dataset, other news articles, and the GUM corpus) have PDTB annotations. More details about our annotation protocol are provided at the beginning of Section 4.

The annotation results are in Table 5, where the results of the parsers are compared to the ground truth labels by the annotators.

Across the three corpora, the annotators noticed that in many cases the relation type was labeled as EntRel or NoRel when it shouldn't have been, or vice versa. This led to discourse senses being predicted for relations that did not have a discourse sense and vice versa. The parsers also often had issues with argument segmentation. For the GUM corpus, segmentation was especially an issue in the travel genre, where headers or captions would be labeled as part of an argument.

As is shown in Table 5, the percentage of implicit relations that the parsers got right on the second level appeared to decrease on average as the domain shifted. However, this was a very slight decrease; they had roughly the same level of accuracy across all datasets, which was very low. In fact, for all parsing models and datasets, a larger percentage of relations was predicted completely incorrectly.

The results of running the state-of-the-art Wang et al. (2017) parser on the gold labels of the RST and GUM corpus are shown in Figure 2. These results make it clear that the RST parser performs much worse on out-of-domain data than it does on

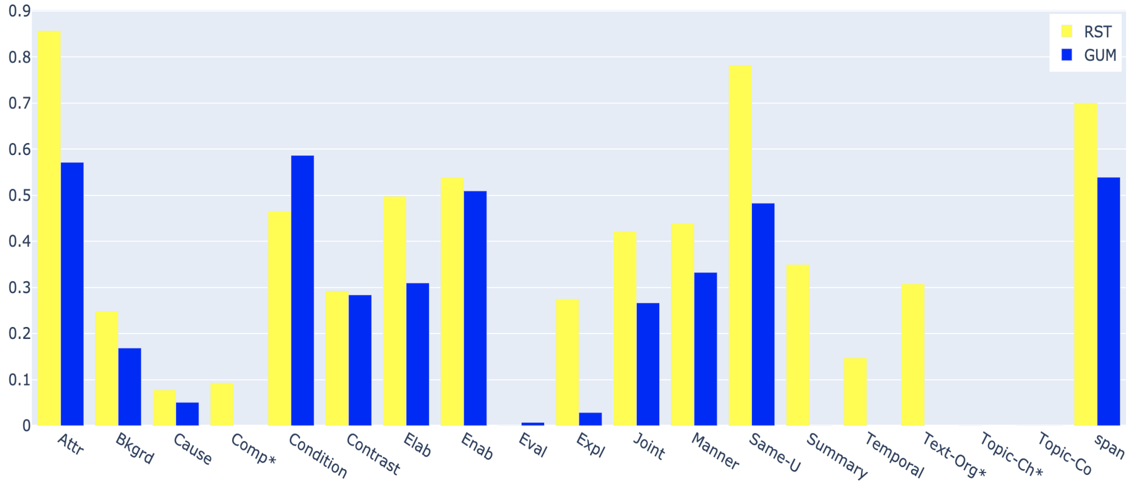

Figure 2: F-1 scores for running the Wang et al RST parser on the RST Discourse Treebank (* indicates that a relation was not annotated on the GUM corpus). For most relations, we see that the parser performed much better on the RST test set than on the GUM articles.

	Wang and Lan			BERT		
	WSJ	Other	GUM	WSJ	Other	GUM
Max	51.3	38.7	56.7	54.3	44.4	49.1
Std.dev	14.1	13.0	15.4	14.0	12.6	12.9
0%	11	11	11	9	8	8
0-2%	1	2	2	2	3	3
2-5%	1	0	0	2	2	2
>5%	3	3	3	3	3	3

Table 6: Summary stats for running the Wang and Lan parser and BERT parser on WSJ articles, other news articles, and GUM. We study the % of predicted Level 2 PDTB relations, reporting the maximum, the standard deviation, and # of sense types that were predicted 0% of the time, 0-2%, etc.

RST corpus data. This is expected; it unsurprisingly does not generalize as well for text outside of its domain as for the news text contained within the corpus test set due to a change in vocabulary. However, in order for discourse parsers to be useful for applications outside of the news domain, models that can more easily adapt to the target domain must be developed.

4 Insights for model development

While inspecting the results of the annotations, we found several helpful phenomena for developing future models, including observations regarding the role of context in shallow discourse parsing and errors that current RST parsers are making.

4.1 Annotation Details

For the qualitative analysis, we ask two annotators (a faculty member and a graduate student from linguistics departments) to provide annotations for the data, as none of the texts contain gold PDTB labels and only the GUM corpus contains gold RST labels. The annotators were trained on, and provided with, the PDTB 2.0 annotation manual (Prasad et al., 2007).

In order for the annotators to annotate this corpus, discourse relations were randomly chosen from Wall Street Journal articles, other news articles, and the GUM corpus. 64 of these discourse relations were implicit, and are the only ones reported in this paper. The annotators were given the sentence(s) containing both arguments, with the arguments labeled, and they also had access to the article text if they ever needed to reference back to it. To assess the inter-rater agreement, we determine Cohen's κ value (Cohen, 1960). We randomly selected 25 samples from the PDTB and assigned each to the annotators. We obtained a Cohen's κ of 0.88, which indicates almost perfect agreement.

4.2 Findings

More context than the two arguments is needed to determine the correct discourse relation in many cases One potential way to mitigate the impact of domain shift on the performance of shallow discourse parsers is to incorporate context. With a few exceptions (Dai and Huang, 2018; Shi and Demberg, 2019; Zhang et al., 2021), existing models for shallow discourse parsing mostly do not

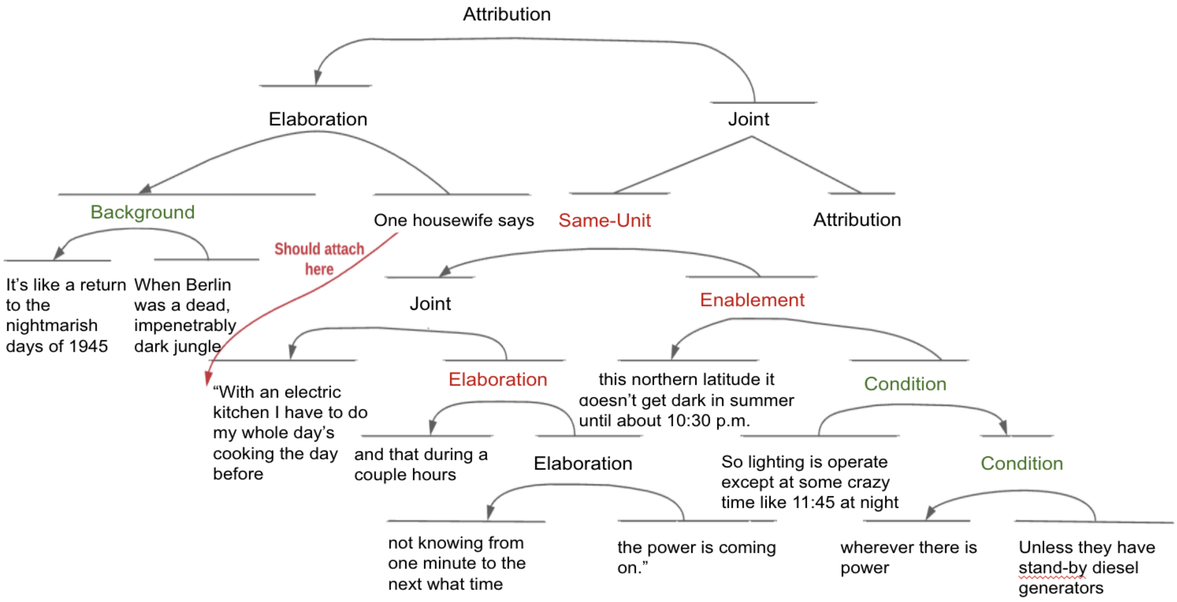
Attribution
Elaboration
Joint
Background
One housewife says
Same-Unit
Attribution
Should attach here
It's like a return to the nightmarish days of 1945
When Berlin was a dead, impenetrably dark jungle
Joint
Enablement
"With an electric kitchen I have to do my whole day's cooking the day before
Elaboration
this northern latitude it doesn't get dark in summer until about 10:30 p.m.
Condition
and that during a couple hours
Elaboration
So lighting is operate except at some crazy time like 11:45 at night
Condition
not knowing from one minute to the next what time
the power is coming on."
wherever there is power
Unless they have stand-by diesel generators

Figure 3: RST parse tree containing a segment of the relations that were examined in the qualitative analysis. The discourse sense labels on this tree that were examined in our analysis are marked red and green, where green is correct and red is incorrect

use input beyond the two adjacent sentences that comprise the arguments of the relation (Kishimoto et al., 2020; Chen et al., 2019). We found that only considering these two sentences is not sufficient even for our expert linguist annotators. Specifically, while annotating the PDTB, the annotators found several examples where, when they looked at the larger context behind the arguments and the sentences where the arguments were contained, their annotations changed. Below, we describe a few examples that demonstrate the mistakes that can be made without the full context and their implications:

(4) *In this northern latitude it does n't get dark in summer until about 10:30 p.m. so lighting is operate except at some crazy time like 11:45 at night , whenever there is power , unless they have stand-by diesel generators.* **There 's a year 's supply of diesel oil here.**

This example is from the Wall Street Journal. At first glimpse, one would think to annotate this as *Contingency*.Factual present condition, but this does not capture the full context, which is shown below:

(5) *One housewife says : " With an electric kitchen I have to do my whole day 's cooking the day before – and that during a couple of hours , not knowing from one minute to the next what time the power is coming on. " In this northern latitude it does n't get dark in summer until about 10:30 p.m. so lighting is operate except at some crazy time like 11:45 at night , whenever there is power , unless they have stand-by diesel generators. There 's a year 's supply of diesel oil here.*

The additional context, that people in the country described are dealing with electricity issues despite there being a year's worth of diesel supply, is now made clear in this passage. Thus we can conclude that the correct relation here is *Comparison*.Contrast. Without getting this context and just seeing the two sentences in which the arguments are contained, it is difficult to discern this as an annotator. This shows that by just getting exposure to the two arguments, without additional context, the sense may be marked incorrectly. The Wang and Lan (2015) parser and the DiscoEval parser both predicted this incorrectly, with the Wang and Lan (2015) parser predicting it as *Contingency*.Cause and the BERT parser predicting it as *Expansion*.Conjunction.

Similarly, the following example, also contained in this passage, has a different true annotation than one would think from only seeing the arguments:

(6) *One housewife says : " With an electric kitchen I have to do my whole day 's cooking the day before – and that during a couple of hours , not knowing from one minute to the next what time the power is coming on . "* **In this northern latitude it does n't get dark in summer until about 10:30 p.m. so lighting is operate except at some crazy time like 11:45 at night , whenever there is power , unless they have stand-by diesel generators .**

The relation may be deemed as *Expansion*. Instantiation. However, by reading the full text, it is clear that it should be labeled as *Contingency*.Cause. Like the last example, a clearer view of the full text is needed to determine the proper annotation, not simply the two arguments.

These observations provide insights as to why contextual embeddings *with* document context such as the next sentence prediction task helps with implicit discourse relation classification (Shi and Demberg, 2019). More generally, we believe future work on discourse parsing should look beyond only the arguments of a relation because of the different interpretations one would give when taking the relation in vs. out of context. We believe that argument pairs with low specificity and one or more pronouns may be especially in need of this extra context, but more experimentation will have to be done to confirm this hypothesis.

Attachment issues tend to occur throughout the RST parse tree, and relations are often misclassified as *Same-Unit* and *Elaboration*. Regarding insights for the RST Discourse Treebank, a piece of the RST tree for this paragraph can be seen in 3. Here, the EDU "One housewife says" should attach to the EDU after it, "With an electric kitchen I have to do my whole day's cooking the day before". However, it instead attaches to EDUs from the preceding sentences, which is incorrect, as these two sentences do not contain what the housewife says. We saw several other attachment issues in the text, including a couple where the attachment should go up/down by several levels. We also saw several instances of the relation being incorrectly tagged as *Same-Unit* or *Elaboration*, some of which can be seen in the diagram.

Attachment issues are a particular problem for RST parsing due to its hierarchical nature; one attachment issue can lead to error propagation where the accuracy of the attachments further in the tree is impacted by that of the current one. Reducing this error is of the utmost importance for future parsers.

5 Conclusion and future work

Discourse parsing for text has seen a recent surge in experimental approaches. In this work we presented a detailed analysis of the performance of the state of the art discourse parsers and analysed their weaknesses and strength. The conclusions drawn above from these experiments make it clear that discourse parsing, though it has come a long way in the past decade or so, still has a long way to go, particularly with respect to parsing on out-of-domain texts and addressing issues of class imbalances, although the BERT-based model has made some improvements in this area. Additionally, we investigated how and when PDTB-3 can help in improving the prediction of intra-sentential implicit relations.

There are several promising future directions for the area of discourse parsing. A model that detects intra-sentential implicit relations is necessary in order to be able to parse on the PDTB-3. Exploring new neural parsing strategies is also a must. We observed that neural parsers are ignorant about what they do not know and overconfident when they make uninformed predictions. Quantifying prediction uncertainty directly by training the model to output high uncertainty for the data samples close to class boundaries can results in parsers that can make better decisions. One takeaway of our empirical analysis was the importance of the role of context in identifying the correct discourse relations. This observation suggests the need for new computational experiments that can identify the right context window that is required for the model to accurately predict relations.

Another useful direction is designing models that can learn discourse relations on their own without the help of annotated corpora. There are several unsupervised models (Kobayashi et al., 2019; Nishida and Nakayama, 2020) that are used for determining the *structure* of discourse parse trees but few that infer the relations themselves.

Acknowledgements

We would like to thank the reviewers, Diane Litman and Matthew Stone for providing helpful feedback for this work.

References

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Muthu Kumar Chandrasekaran, Carrie Epp, Min-Yen Kan, and Diane Litman. 2017. Using discourse signals for robust instructor intervention prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Eugene Charniak. 1973. Jack and Janet in search of a theory of knowledge. In *IJCAI*, pages 337–343.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 649–662.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. 2012. Vers le fdtb: French discourse tree bank. In *TALN 2012: 19ème conférence sur le Traitement Automatique des Langues Naturelles*, volume 2, pages 471–478. ATALA/AFCP.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.

Aravind Joshi Bonnie Webber Eleni Miltsakaki, Rashmi Prasad. 2004. The Penn Discourse Treebank. *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical report, Center for the Study of Language and Information, Stanford University.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1152–1158.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down RST parsing utilizing granularity levels in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8099–8106.

Naoki Kobayashi, Tsutomu Hirao, Kengo Nakamura, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2019. Split or merge: Which is better for unsupervised RST parsing? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5801–5806.

Murathan Kurfalı and Robert Östling. 2019. Zero-shot transfer for implicit discourse relation classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 226–231.

Haejun Lee, Drew A Hudson, Kangwook Lee, and Christopher D Manning. 2020. SLM: Learning a discourse language representation with sentence unshuffling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1551–1562.

Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending implicit discourse relation recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351.

Linlin Liu, Xiang Lin, Shafiq Joty, Simeng Han, and Lidong Bing. 2019. Hierarchical pointer net parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1006–1016.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2284–2295.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: a theory of text organization. Technical Report RS-87-190, USC/Information Sciences Institute. Reprint series.

Daniel Marcu. 1999. Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, 293:123–136.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT press.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank.

Thomas Meyer, Andrei Popescu-Belis, Sandrine Zufferey, and Bruno Cartoni. 2011. Multilingual annotation and disambiguation of discourse connectives for machine translation. In *Association for Computational Linguistics-Proceedings of 12th SIGdial Meeting on Discourse and Dialogue*, CONF.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *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.

Noriki Nishida and Hideki Nakayama. 2020. Unsupervised discourse constituency parsing using Viterbi EM. *Transactions of the Association for Computational Linguistics*, 8:215–230.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 186–195.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, and Aravind Joshi. 2007. The Penn Discourse Treebank 2.0 annotation manual.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alan Lee Eleni Miltsakaki Livio Robaldo Aravind Joshi Bonnie Webber Rashmi Prasad, Nikhil Dinesh. 2008. The Penn Discourse Treebank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).*

RC Schank and RP Abelson. 1977. Plans, goals aid understanding: An inquiry into human knowledge structures.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5794–5800.

Maite Taboada. 2016. Sentiment analysis: An overview from linguistics.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 17–24.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188.

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

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Discourse-aware neural extractive text summarization. *arXiv preprint arXiv:1910.14142*.

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 Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

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

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54:587–613.

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th workshop on Asian language resources*.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395.

Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese Discourse Treebank 0.5 ldc2014t21. *Web Download. Philadelphia: Linguistic Data Consortium.*