The Role of Context and Uncertainty in Shallow Discourse Parsing

Katherine Atwell[†], **Remi Choi**[†], **Junyi Jessy Li**[§], **Malihe Alikhani**[†]
[†]Department of Computer Science, University of Pittsburgh
[§]Department of Linguistics, The University of Texas at Austin
{kaa139, theresachoi, malihe}@pitt.edu,
jessy@utexas.edu

Abstract

Discourse parsing has proven to be useful for a number of NLP tasks that require complex reasoning. However, over a decade since the advent of the Penn Discourse Treebank, predicting implicit discourse relations in text remains challenging. There are several possible reasons for this, and we hypothesize that models should be exposed to more context as it plays an important role in accurate human annotation; meanwhile adding uncertainty measures can improve model accuracy and *calibration*. To thoroughly investigate this phenomenon, we perform a series of experiments to determine 1) the effects of context on human judgments, and 2) the effect of quantifying uncertainty with annotator confidence ratings on model accuracy and calibration (which we measure using the Brier score (Brier et al., 1950)). We find that including annotator accuracy and confidence improves model accuracy, and incorporating confidence in the model's temperature function can lead to models with significantly better-calibrated confidence measures. We also find some insightful qualitative results regarding human and model behavior on these datasets.

1 Introduction

The context of an utterance influences the interpretation. Linguistics, philosophy, cognitive science and neuroscience (Lewis, 1980; Glanzberg, 2002; Thompson-Schill, 2003) studies have shown the effect of context in text interpretation. In this paper, we hypothesize that context can influence the prediction of discourse relation labels. Motivating the need to discover the effects of context on annotation accuracy is the following example, where argument 1 is **bolded** and argument 2 is *italicized*:

(1) **the fund's 25% leverage has jacked up its interest income**
As long as I am borrowing at 9.9% and each {bond} yields over that, it enhances the yield

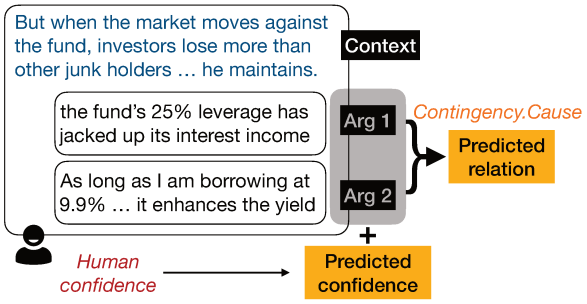

Figure 1: A diagram of our model, in which we use contextual information to predict uncertainty and pass the results into a model that predicts the discourse relation. We experiment with several different architectures and machine learning methods for utilizing uncertainty, including concatenation.

The lack of contextual information makes it difficult to determine the proper discourse relation here, and our annotator chose *Expansion.Conjunction* as a result. However, when more context is added, the passage presented to the annotator is as follows:

(2) But when the market moves against the fund, investors lose more than other junk holders because the market decline is magnified by the amount the fund is leveraged. Fund managers, for their part, defend their use of leverage. Carl Ericson, who runs the Colonial Intermediate High Income Fund, says **the fund's 25% leverage has jacked up its interest income**. "*As long as I am borrowing at 9.9% and each {bond} yields over that, it enhances the yield*," he maintains. Mr. Ericson says he tries to offset the leverage by diversifying the fund's portfolio.

This added context makes it more apparent that a causal relation holds between the two arguments, and thus our annotator chose the correct relation, *Contingency.Cause*, after being presented with this context.

Proceedings of the 29th International Conference on Computational Linguistics, pages 797–811
October 12–17, 2022.

While the annotation process of discourse corpora such as the Penn Discoruse Treebank (PDTB) exposes annotators to the full context of the document, we do not yet understand how human annotation behavior would change if context were limited. Additionally, context is limited for most of the models built for automatic discourse relation classification. We start with a basic question: is the argument pair (where arguments are defined as the minimum span a relation could be interpreted in PDTB (Prasad et al., 2008; Webber et al., 2019)) enough to determine the discourse relation?

In addition, context may affect human annotation to varying degrees and impact annotators' confidence in their judgements, how can we make sure that this information is factored into discourse parsers and model confidence? For the first time, we propose to study and measure the human annotator's confidence and incorporate it into the architecture of the deep-learning-based discourse parsers. We utilize our human-annotated confidence scores to predict human confidence, and test whether this method improves model accuracy and *calibration* (how well the predicted probabilities produced by a model reflect the true likelihood of the corresponding events to occur in the studied population). Model calibration is an important issue in modern neural networks (Guo et al., 2017), and to our knowledge we are the first to study it for discourse relation classification. Properly calibrating a model (i.e. properly quantifying uncertainty) is especially important for this task, because determining the correct discourse sense involves a large degree of uncertainty, and more correctly quantifying uncertainty allows a model to be more explainable.

Our two main research questions, as described above, are illustrated in Figure 1, and our contributions can be summarized as the following:

1. Determine the effects of added context on the discourse annotation task by increasing the context window given to the annotators and comparing the results to those of presenting annotators with only the two arguments across three different datasets. Measure the annotation accuracy and confidence under each of these conditions.

2. Perform a qualitative error analysis of these results, providing insight into cases in which adding context may improve annotation results.

3. Add annotation accuracy and confidence scores to the input of an implicit sense classifier, and measure the resulting changes in model accuracy.

4. Use confidence scores to impact the training and evaluation mechanisms of an implicit sense classifier, and use accuracy and calibration metrics as validation metrics for these models. Measure the change in accuracy and model calibration.

5. Perform a qualitative error analysis on the model results, finding cases where model performance suffers without access to context and providing explanations as to why.

Our code can be found here [1].

2 Related Work

The effects of context on implicit sense classification As mentioned above, implicit sense classification is a very challenging task. Further, most implicit sense classifiers (Chen et al., 2019) do not include context outside of the two arguments contained in a discourse relation,[2]despite the annotators having access to context during the PDTB annotation task, wherein the annotator inserts a connective between the two arguments and then determines the discourse relation. An example of this connective insertion from the PDTB is as follows, with Arg1 in **bold** and Arg2 in *italics*:

(3) Several leveraged funds don't want to cut the amount they borrow because it would slash the income they pay shareholders, fund officials said. But a few funds have taken other defensive steps. **Some have raised their cash positions to record levels.** Implicit = BECAUSE *High cash positions help buffer a fund when the market falls.*

However, no paper has yet studied the effect of context on the discourse annotation task, nor has a work attempted to use insights from annotators' proficiency and confidence on the model.

[1] https://github.com/katherine-atwell/DiscourseContextUncertainty

[2] Note that in the case of PDTB-3, it is possible for local (sentence-level) context to be encoded using pre-trained encoders such as BERT, when determining intra-sentential implicit discourse relations.

The closest work to ours in this space is Scholman and Demberg (2017), who use a connective insertion task to crowdsource discourse sense annotations and examine the effect of context on this task. They find that under certain conditions (such as when argument 1 refers to an entity/event in the surrounding context or the sentence after argument 2 expands on argument 2), the presence of context can improve annotator agreement. However, this annotation task is simplified and only covers 6 level-2 relations. Thus, their result is not fully representative of the PDTB annotation task. We wish to examine this question more in-depth by presenting annotations using the traditional PDTB annotation task to a trained linguist in settings with and without additional context, and comparing these annotations against ground truth data. In doing so, we believe we can gain more insight into factors that affect human understanding in more similar conditions to the ones present in the original PDTB annotation task.

Calibration of neural networks for discourse parsing *Calibration* in machine learning refers to the distribution of error and the model's level of self-assessment, or confidence (Bella et al., 2010). It was found that neural networks tend to be badly calibrated (Guo et al., 2017), which can result in poor explainability and uncertainty quantification. However, pretrained models, even very complex ones, were found to generally be more well-calibrated, and to benefit from temperature scaling (Desai and Durrett, 2020). Therefore, we use annotator confidence scores to scale the temperature according to a the example's simplicity (as perceived by the annotator). In our work, we choose to use the Brier score (Brier et al., 1950) to measure calibration because it is a proper scoring function. As far as we are aware, we are the first paper to study the calibration of implicit sense classification models and the impact of using annotator confidence measures to improve model calibration.

3 Data and Analysis

3.1 Methods

Here we describe our human annotation experiments, in which we determine whether adding context can improve annotator accuracy and confidence.

Dataset For all of our experiments, we use gold data from the Penn Discourse Treebank 2 (PDTB-2, Prasad et al. (2008)), Penn Discourse Treebank 3 (PDTB-3, Webber et al. (2019)), and the English set of the TED Multilingual Discourse Bank (TED-MDB, Zeyrek et al. (2018)), in order to test our hypothesis across different frameworks and text domains. The PDTB-2 is the most commonly used dataset for discourse parsers, while the PDTB-3 introduces intra-sentential discourse relations and an updated label schema and the TED dataset contains speeches annotated with the PDTB-3 framework.

Discourse sense	Without context	With context	Context effect
Temp.Asynchronous	-2	-1	**better**
Cont.Cause	-3	-5	*worse*
Comp.Contrast	14	10	**better**
Comp.Concession	6	5	**better**
Comp.Similarity	2	1	**better**
Exp.Conjunction	11	5	**better**
Exp.Instantiation	9	10	*worse*
Exp.Equivalence	3	4	*worse*
Exp.Level-of-detail	-16	-9	**better**

Table 1: Frequency of discourse senses with/without context in our annotated set relative to ground truth (0 indicates a perfect overlap with ground truth count). We can see that a context window usually entails a better (more ground truth-like) distribution. Moreover, the impact of the context effect is usually stronger when it is better than when it is worse. The full table including neutral relations and raw counts is in the appendix.

Annotation To test our hypothesis that adding context improves annotation performance for the implicit sense labeling task, we recruit two expert linguists to provide level 2 sense annotations for implicit discourse relations from all three corpora listed above, calculating 60% absolute agreement. This task was approved by our institution's human subjects board.

In order to attain a representative sample, we make sure that every type of implicit discourse sense contained in the PDTB-2, PDTB-3, and TED-MDB was represented at least once in this sample. Further, in order to select for relations that may need more context than the two arguments, we randomly sample a large set of relations and, from that sample, select relations whose arguments have a high portion of pronouns and a low level of specificity. Pronouns signal coreference relations that may be missing from the argument spans (Scholman and Demberg, 2017), and a low level of specificity suggests that more information may be needed to understand the full context (Li et al., 2016; Choi et al., 2021) . We use NLTK's part of

speech tagger to detect the number of pronouns contained in the arguments, and to determine specificity we use the Ko et al. (2019) specificity classifier.

Task To study the role of context, our annotators perform two respective tasks. First, they attempt to determine the discourse connective, and corresponding discourse sense, when only shown the pair of arguments. For the second task, the annotator has access to the full sentence(s) containing the arguments, as well as the two sentences before and the sentence after the arguments. The first task always comes before the second task, and the annotator is not able to edit their annotation for the first task after seeing the additional context in the second task.

We produce 498 samples of human annotations for discourse sense and confidence before and after context. Of the samples, 147 come from the PDTB-2, 199 from the PDTB-3, and the last 152 from the TED-MDB.

3.2 Analysis

To better understand the hypothetical effect of context, we investigate annotation changes with respect to the discourse relations in the arguments, as well as in the context window supplied to the annotator.

Does context improve annotation accuracy? Generally, yes. Over all three corpora, annotator accuracy increases from 0.350 to 0.414 (+6.4%) between the task without context and the task with context. Table 2 provides the breakdown of accuracy by corpora for each task. While the task accuracy is low in general due to the difficulty of discourse relation prediction, we can see consistently higher accuracy across all three corpora in the task with context. However, it should be noted that these increases are not statistically significant, likely due to the small size of our annotated set.

Corpus	Accuracy		Confidence	
	Raw	Context	Raw	Context
PDTB2	0.306	**0.354**	3.70	**4.81**
PDTB3	0.379	**0.423**	3.56	**4.71**
TED-MDB	0.296	**0.355**	3.63	**4.84**

Table 2: Average annotator accuracy and confidence on a scale of 1-5 for each corpus before and after context. Adding context improves accuracy and confidence across all corpora, an increase which is statistically significant with respect to confidence ($p < 0.01$).

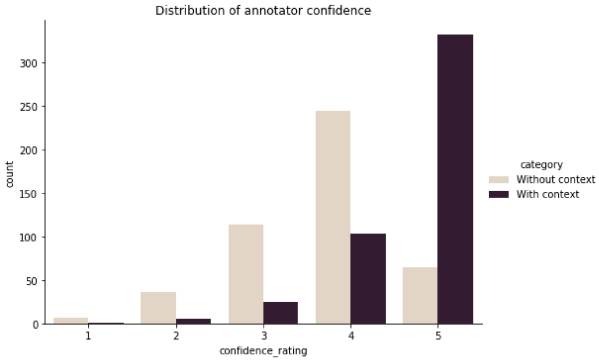

Figure 2: Distribution of the confidence scores (discrete scores from 1-5) chosen by the annotators. As this distribution shows, annotator confidence consistently increases with context.

Further, as Table 1 shows, the discourse sense distribution of annotations models the ground truth distribution more closely after context.

Does context improve annotator confidence? Annotator confidence provides information beyond simply whether an annotation is incorrect/correct. We found that over all three corpora, annotator confidence ratings improved from 3.59 to 4.77 (+1.18) after being given context. Similarly to annotator accuracy, we see all three corpora reflect this improvement in Table 2. The effect of context on annotator confidence is statistically significant across all three corpora ($p < 0.01$).

In addition, the full results can be found in Table 3, which shows the distribution of annotations that are positively or negatively impacted by context across each discourse relation. Discourse relations with no samples that change after context are excluded.

But why study annotator confidence? Annotator confidence provides information beyond simply whether an annotation is incorrect/correct. For instance, an annotator could be correct but lack confidence due to a lucky guess. Further, in our annotations, we found that highly confident answers were twice as likely to be correct (0.390) as answers rated with a lower confidence (0.200). We incorporate these annotator confidence scores into our model and experimentally validate whether incorporating these scores helps guide model decisions in a meaningful way.

Is the presence of certain discourse relations in an expanded context window beneficial? Yes. We find both instances where context helped

(an incorrect annotation before context became correct after the annotator was given context) and where context hurt (a correct annotation before context became incorrect after the annotator was given context). In particular, we find that *Comparison.Contrast*, *Contingency.Cause*, *Expansion.Conjunction*, and *Expansion.Restatement* are likely to improve annotation accuracy. Contrarily, *Expansion.List* and *Temporal.Asynchronous* seem to provide less helpful context.

The full results are reported in Table 4. We exclude relations in the context that had no effect on the annotation's accuracy, or when the annotation remained correct or incorrect before and after context.

In what instances does context actually help? Our annotator performed a qualitative analysis on the annotations where context helps (an incorrect prediction turned correct) and hurts (a correct prediction turned incorrect). From this, we found that context typically helps in cases where the arguments are very short, the discourse structure in the surrounding sentences is made clearer, and background information in the surrounding texts illustrates a relationship between pieces of the two arguments that could not be extrapolated from only the argument pair. The example in Section 1 represents the latter phenomenon, and we provide an example of the first phenomenon in Appendix A. Below we provide an example where the discourse structure makes the correct relation more apparent, with argument 1 in **bold** and argument 2 in *italics*:

(4) USA Today reported that the Rales brothers, Washington, D.C.-based investors who made an unsuccessful offer to acquire Interco last year, have bought nearly 3% of Mead's common shares. Entertainment and media stocks generally escaped the market's slide as well. Paramount Communications rose 5/8 to 58 3/4, **Time Warner climbed 1 7/8 to 138 5/8**, *Walt Disney advanced 3 1/8 to 127 1/2*, MCA rose 1 1/8 to 65 5/8 and McGraw-Hill added 1/2 to 67 1/8. The American Stock Exchange Market Value Index lost 3.11 to 379.46.

Discourse sense	Incorr.→Corr.	Corr.→Incorr.
Temp.Asynchronous	**5 (15.63%)**	1 (3.13%)
Cont.Cause	7 (6.25%)	6 (5.36%)
Cont.Purpose	1 (5.00%)	0
Comp.Contrast	**4 (13.79%)**	1 (3.45%)
Comp.Concession	3 (14.29%)	2 (9.52%)
Comp.Similarity	1 (33.33%)	0
Exp.Conjunction	**12 (12.50%)**	5 (5.21%)
Exp.Instantiation	4 (10.00%)	3 (7.50%)
Exp.Equivalence	0	**2 (25.00%)**
Exp.Level-of-detail	**12 (17.91%)**	4 (5.97%)

Table 3: Distribution of ground truth Level 2 senses by whether an annotation turns from incorrect to correct after context or vice versa. Bolded cells in each row indicate whether a sense seems to benefit or hurt from context.

Discourse Sense	Incorr.→Corr.	Corr.→Incorr.
Temp.Asynchronous	4 (3.36%)	**9 (7.56%)**
Temp.Synchronous	2 (3.45%)	3 (5.17%)
Cont.Cause	**17 (6.37%)**	11 (4.12%)
Cont.Condition	3 (4.41%)	4 (5.88%)
Comp.Concession	2 (4.17%)	2 (4.17%)
Comp.Contrast	**26 (10.70%)**	17 (7.00%)
Exp.Alternative	0	1 (4.35%)
Exp.Conjunction	**30 (8.55%)**	14 (3.99%)
Exp.Exception	0	1 (33.33%)
Exp.Instantiation	6 (7.79%)	5 (6.49%)
Exp.List	2 (3.70%)	**4 (7.41%)**
Exp.Restatement	**12 (7.79%)**	2 (1.30%)

Table 4: Distribution of Level 2 senses in sample context by whether an annotation turns from Incorrect to Correct after context or vice versa. Bolded cells in each row indicate whether a sense seems to provide beneficial or harmful context.

Without the added context, the relation appears as though it could be contrastive to the annotator (where Time Warner's rise is compared to Disney's rise). However, additional context allows the annotator to see the discourse structure of the surrounding clauses in the sentence (also holding a *Conjunction* relation), as well as the sentence contain the arguments with the previous sentence (an *Instantation* relation where the sentence containing the arguments provides several similar examples to back up the first argument's claim). Thus, in this example, the discourse structure of the surrounding text is beneficial for labeling the discourse relation.

Table 1 further sheds light on ways in which context influences the chosen relations, showing that some discourse relations are predicted more than others with and without context (relative to their ground truth counts). For instance, *Expansion.Conjunction* is predicted at a much higher rate without context than with context, likely due to instances (such as the example in Section 1 where the annotator does not have enough information about the relationship between the two arguments to pick a relation with more rigidly defined semantics. We

find that *Comparison.Contrast* tends to also be chosen less often with context (the second example in this section illustrates a scenario in which the annotator changes from *Contrast* to another relation given context). *Expansion.Level-of-detail*, on the other hand, is chosen more often when more context is provided, which makes sense given that a *Level-of-detail* relation requires knowing that the semantics of argument 2 restate the semantics of argument 1, and that both arguments hold true at the same time. This information is not always available when only the two arguments are shown.

4 Modeling Insights

Above, we illustrate some insights attained from the annotated data with respect to changes in domain and access to context. In this section, we use both the annotator correctness and confidence metrics to inform our model decisions, in order to determine whether annotator performance in any way correlates with model performance. In addition to model accuracy, we evaluate model calibration scores when the model is and is not given access to annotator confidence. To influence the model's decisions when given access to annotator confidence, we adjust the training and validation mechanisms accordingly. We hypothesize that access to annotation metrics will improve model accuracy, and that changing the temperature function with respect to annotator confidence will improve model calibration. We describe the setup for each of our experiments below, and report our results in Section 4.2.

4.1 Experimental Setup

We use the Kim et al. (2020) XLNet-large baselines as our base model, for which the large XLNet (Yang et al., 2019) model is trained for a maximum of 10 epochs, but early stopping occurs if there is no improvement to the development set for 5 evaluation steps. We use sections 4-24 for training, 2-3 for development, and 0-1 for testing. Following Kim et al. (2020), we use the standard L2 classification with 12 labels for the PDTB-2, and use the 14 senses with more than 100 labels for the PDTB-3. For each result, we report the average across 3 different seeds. We first experiment with concatenating the features described below to the sentence embeddings produced by the XLNet-large model and passing the resulting embedding to a classification head to predict the discourse relation.

Exploiting annotation accuracy and confidence We first experiment with using features from our annotated data, in order to determine whether they provide any benefit to the model. The first feature we experiment with is a binary prediction (using the annotations as training data) as to whether or not the relation will be labeled correctly. To obtain this feature, we trained an SVM using bag-of-words features on our annotation data, where the label is *true* if the annotator labeled the relation correctly given only the argument pairs and *false* otherwise. We pass the argument pairs as input to the model. We used an 80/20 train/test split for this model and the classification accuracy is .838.

The second feature that we use is the confidence score for the two arguments given additional context, to determine whether using features that incorporate some contextual features help the model at all. As with the previous feature, we train an SVM using bag-of-words features on our annotation data, and again pass the two arguments as input to the model, but here we predict confidence as a regression task as opposed to a classification task. For this model, we also use an 80/20 train/test split, and report a mean square error of .180.

Reweighting using confidence annotations Beyond experimenting with adding annotation metrics as features to our model, we experiment with adjusting the training and validation mechanisms of our model using predicted annotator confidence scores. We use these scores to adjust the training weights, weighting the examples with lower predicted confidence higher and the examples with higher predicted confidence lower. For each example, we predict its corresponding confidence feature and divide 5 by this value (as 5 is the highest the confidence level can go). We then use this value as the weight for the sample, thus upsampling all examples with a predicted confidence score less than 5 out of 5.

Temperature adjustment using confidence annotations Similarly, we experiment with adjusting the temperature of the softmax function, in order to increase model confidence in proportion to predicted annotator confidence. We thus weight the examples with higher predicted confidence scores higher, and vice versa, by dividing 5 by the confidence score for each example to get our temperature (which is inversely proportional to the desired model confidence). We show this in the following

Model	PDTB-2	PDTB-3	TED
XLNet-large (cased)	.5527	.6326	**.5381**
+Correctness	**.5694***	.6452*	.5347
+Confidence	.5648*	**.6518***	.5035*
+Correctness & Confidence	.5642*	.6428*	.4931*
+Reweighting	.5665*	.6419*	.4861*

Table 5: Accuracy scores for each model evaluated on the PDTB-2 and PDTB-3, with the best performing model in bold for each metric (+Correctness for the PDTB-2, +Confidence for the PDTB-3, and the baseline for TED). * indicates statistical significance ($p < 0.05$).

Model	PDTB-2	PDTB-3	TED
XLNet-large	.5527	.6326	.5381
+ temp change	.5597	.6326	.5486

Table 6: Accuracy for baseline and model with adjusted softmax temperature. Changing the temperature appears to slightly improve performance for the PDTB-2 while not affecting performance for the PDTB-3.

equation, denoting the temperature as $\mathcal{T}$ and the confidence score as c: $\mathcal{T} = \frac{5}{c}$

4.2 Results

In order to determine the effects of adding annotator performance metrics as input to the model, we detail the results from each of the models above, in particular looking at accuracy for the models with concatenated annotation features and accuracy and calibration for the models that use annotator confidence features to influence their training mechanism. We provide several questions we wish to answer with these analyses, and the corresponding results, below. We test for significance using a two-tailed t-test for each experiment.

Is reweighting training examples using confidence scores useful? Similarly to directly adding confidence features as input to the model, reweighting the training examples based on the predicted confidence score improves upon the baseline for both corpora (Table 5). This suggests that influencing the training mechanism with confidence scores has the potential to improve model accuracy. We report level 2 results of this model in Tables 8 and 10 in the Appendix. As with the previous results, the scores on the TED dataset are not improved when these changes are made, but because of the small size of the test set, we do not believe this to be notable.

Does adjusting temperature improve accuracy, calibration, or both? To evaluate the results of the model in which temperature was adjusted, we measure both the accuracy and the *calibration* of the model (calculated using the Brier score). We find that although the model with adjusted temperature outperforms the baseline with respect to accuracy only on the PDTB-2 (Table 6), it outperforms the baseline with respect to the Brier score by a large margin for both datasets (Table 7). Thus, the model with the temperature change is more *well-calibrated* than the original model, i.e. its probabilities are more likely to reflect the actual probability of a prediction being correct given the input. A large improvement for the Brier scores is seen on the TED test set, similarly substantial to that of the other two datasets. However, although this is encouraging, we once again take caution in drawing significant conclusions from this due to the small size of our TED test set.

Model	PDTB-2	PDTB-3	TED
XLNet	.6781	.5787	.7214
XLNet + temp change	.6075*	.5295*	.6477*

Table 7: Brier scores for baseline and model with adjusted softmax temperature. For both datasets, the Brier score improves (a lower Brier score is better) when the temperature is adjusted in accordance with the predicted confidence. * indicates statistical significance ($p < 0.5$)

In addition to reporting overall calibration metrics, we visualize results on individual data points (excluding the TED dataset due to its small size). Using the XLNet-large model, we run the Data Maps tool (Swayamdipta et al., 2020) on the PDTB-2 (Figure 4) and PDTB-3 (Figure 6). We find that the PDTB-3 has more well-defined regions than the PDTB-2, with the model much more likely to get an example with high confidence and low variability correct than an example with low confidence and low variability. This suggests that the model trained on the PDTB-3 is more well-calibrated than the model trained on the PDTB-2, which is supported by the difference in the Brier scores between the two datasets (Table 7). The lack of easy-to-learn examples in the PDTB-2 also provides a possible explanation for the difficulty of the implicit sense classification task for this dataset; the results of Swayamdipta et al. (2020) indicate that easy-to-learn examples are important for optimization.

Does predicting annotation accuracy and confidence help? For both the PDTB-2 and PDTB-3, adding the features predicting annotator correct-

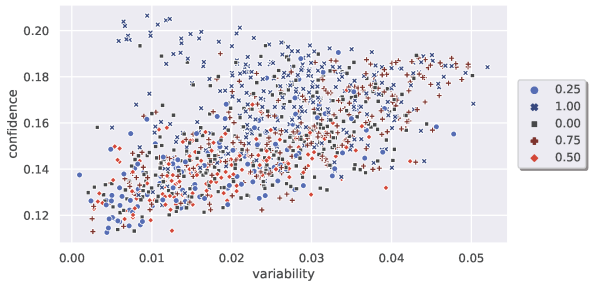

Figure 3: Data map for the baseline model trained on the PDTB-2. Here, there appear to be no distinct regions with respect to correctness, confidence, and variability.

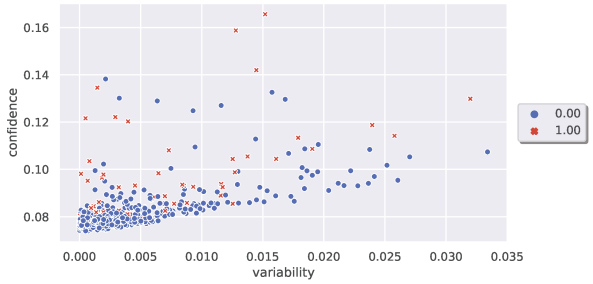

Figure 4: Data map for the PDTB-2 model with temperature changes; here, there appear to be distinct regions with respect to correctness, confidence, and variability, but similarly to the baseline PDTB-2 models, no easy-to-learn samples.

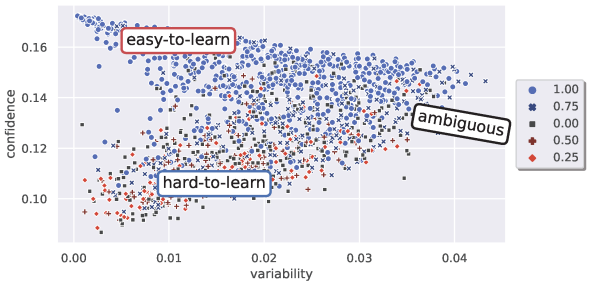

Figure 5: Data map for the baseline model trained on the PDTB-3. Here, there appear to be distinct regions with respect to ease of learning; easy to learn examples have higher confidence and lower variability, and more difficult to learn examples have lower confidence and lower variability.

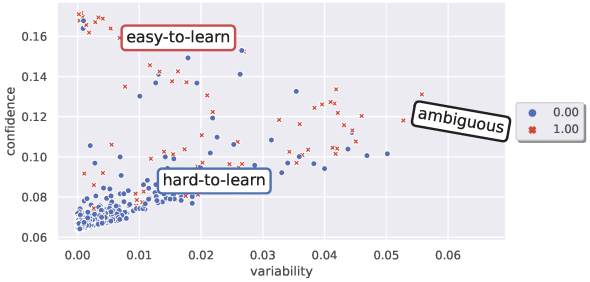

Figure 6: Data map for PDTB-3. Here, as with the PDTB-3 baseline model, there appear to be distinct regions with respect to ease of learning. However, there appears to be a clearer separation between easy-to-learn and hard-to-learn samples with respect to accuracy.

ness and confidence improved model results over the XLNet-large baseline (Table 5), with the added correctness feature yielding the best-performing model on the PDTB-2 and the confidence feature yielding the best-performing model on the PDTB-3. Therefore, it appears that correctness and ease of annotation have some impact on model correctness. We report level 2 results of this model in Tables 8 and 10 in the Appendix. We note that none of these features yielded an improvement when our model was tested on the TED set; however, due to the TED dataset's small set of implicit relations, we hesitate to draw conclusions from these numbers (our test set is comprised of 96 examples).

What kinds of errors do our models make? In addition to annotating discourse relations given argument pairs and additional context, our annotator performed a qualitative analysis on the results of the model. From this analysis, we found that the most common hypothesized reasons for the model guessing a relation wrong were to do with lack of context. There were also cases in which the model did not pick up on vocabulary that indicated a *Temporal* relation. For some misclassifications, the annotator could see why the predicted relation could hold, but believed that the gold annotation was the better one. There were a few cases where the annotator agreed with the model predictions as opposed to the gold labels. We provide some examples of these phenomena in Appendix D. Overall, the annotator observed more cases where errors occur because additional context is needed in the PDTB-2 and TED-MDB than in the PDTB-3.

5 Conclusion

In the previous sections, we first show the effects of presenting additional context when compared to providing only the two arguments. We find that adding context has an overall positive impact on annotator accuracy and confidence. More specifically, we find that context helps for certain discourse relations, but not all relations. We hypothesize that in cases where adding context worsens annotation, it is because context may add confusing information about argument relationship (such as whether one is a quote) or the completeness of a list of items in a

particular set (blurring the line between senses such as *Instantiation* and *Level-of-detail*). We believe this is worth exploring further.

Secondly, we find that utilizing the human performance metrics we collected in the first half of the paper yielded better model performance when compared to the baseline, suggesting that these metrics give the model some useful information about its own predictions. In particular, we find that using confidence scores to adjust the training weights improves model *accuracy*, while using them to adjust the softmax temperature improves model *calibration*, the latter of which is important for explainability and for tasks with a high degree of uncertainty (discourse relation classification being one such task). To our knowledge, ours is the first work to study calibration with respect to discourse models.

We hope that future work will continue to study the role of context in discourse relation classification, as well as model calibration for this task. We will release our annotations and model code upon the publication of this paper.

6 Ethical Considerations

Our experiments were approved by our institution's human subjects board. We acknowledge that the pretrained models we use in this paper may introduce bias.

Acknowledgements

We thank Matthew Stone, Bonnie Webber, Mert Inan, Anthony Sicilia, and the anonymous reviewers for their valuable feedback. This research was partially supported by NSF grants IIS-2145479, IIS-2107524.

References

Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. 2010. Calibration of machine learning models. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 128–146. IGI Global.

Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

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.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Michael Glanzberg. 2002. Context and discourse. *Mind & language*, 17(4):333–375.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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, Online. Association for Computational Linguistics.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *AAAI*.

Davis Lewis. 1980. Index, context, and content. In *Philosophy and grammar*, pages 79–100. Springer.

Junyi Jessy Li, Bridget O'Daniel, Yi Wu, Wenli Zhao, and Ani Nenkova. 2016. Improving the annotation of sentence specificity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3921–3927, Portorož, Slovenia. European Language Resources Association (ELRA).

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. In *LREC*. Citeseer.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics.

Sharon L Thompson-Schill. 2003. Neuroimaging studies of semantic memory: inferring "how" from "where". *Neuropsychologia*, 41(3):280–292.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Deniz Zeyrek, Amalia Mendes, and Murathan Kurfali. 2018. Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In *LREC*.

A Examples Where Context Helps

The first example represents a case where the arguments are very short and do not provide a strong clue about the relation:

(5) Currently, chips are produced by shining light through a mask to produce an image on the chip, much as a camera produces an image on film. But details on chips must now be extraordinarily fine, and the wavelengths of even ultraviolet light are long enough so that the images they draw may be too blurry – much as someone using a wide paintbrush could produce a broad line but would have trouble painting a thin one. X-rays, by contrast, **travel straighter** *and can be focused more tightly than light*. X-rays have problems, too.

Taken by themselves, **travel straighter** and *and can be focused more tightly than light* could be related in any number of ways, and our annotator guessed *Temporal.Asynchronous*. However, given the structure of the sentence the arguments are contained in as well as the information given in the previous sentences, the annotator changed to the correct relation (*Contingency.Cause*) upon seeing the additional context.

B Example Where Context is Misleading

(6) As the best opportunities for corporate restructurings are exhausted of course, at some point the market will start to reject them. But the airlines are scarcely a clear case, given anti-takeover mischief by Secretary of Transportation Skinner, who professes to believe safety will be compromised if KLM and British Airways own interests in companies that fly airplanes. **Worse, Congress has started to jump on the Skinner bandwagon.** *James Oberstar*, the Minnesota Democrat who chairs the Public Works and Transportation Committee's aviation subcommittee, *has put an anti-airline takeover bill on supersonic speed so that it would be passed in time to affect the American and United Air Lines bids*. It would give Mr. Skinner up to 50 days to "review" any bid for 15% or more of the voting stock of any U.S. carrier with revenues of $1 billion or more.

C Examples from qualitative analysis of model results

For all examples below, argument 1 is **bolded** and argument 2 is *italicized*.

C.1 Examples that need more context

PDTB-2 Below is the example without context, which the model predicted as *Contingency.Cause*:

(7) **the threat of U.S. retaliation, combined with a growing recognition that protecting intellectual property is in a country's own interest, prompted the improvements made by South Korea, Taiwan and Saudi Arabia**
What this tells us is that U.S. trade law is working

The ground truth label of this example is *Expansion.Restatement*. Without the additional context, upon inspecting the model output, the annotator concluded that "so/therefore", which signal causality, could be acceptable, but "in other words", signaling *Restatement*, could also work. Below is the example with context:

(8) They will remain on a lower-priority list that includes 17 other countries. Those countries – including Japan, Italy, Canada, Greece and Spain – are still of some concern to the U.S. but are deemed to pose less-serious problems for American patent and copyright owners than those on the "priority" list.
Gary Hoffman, a Washington lawyer specializing in intellectual-property cases, said **the threat of U.S. retaliation, combined with a growing recognition that protecting intellectual property is in a country's own interest, prompted the improvements made by South Korea, Taiwan and Saudi Arabia**. "*What this tells us is that U.S. trade law is working*," he said.

This context makes it clearer that the proper relation to annotate here is *Expansion.Restatement*.

PDTB-3 Below is the example without context, which the model predicted as *Temporal.Asynchronous*:

(9) **In 1976, for example, dividends on the stocks in Standard & Poor's 500-stock index soared 10%, following much**

slower growth the year before. *The S&P index started sliding in price in September 1976,*

The ground truth label of this example is *Comparison.Contrast*. Without the additional context, upon inspecting the model output, the annotator concluded that more context was needed, but without the additional context they could understand how either a temporal or contrastive relation could be held. Below is the example with context:

(10) Indeed, analysts say that payouts have sometimes risen most sharply when prices were already on their way down from cyclical peaks. **In 1976, for example, dividends on the stocks in Standard & Poor's 500-stock index soared 10%, following much slower growth the year before.** *The S&P index started sliding in price in September 1976,* and fell 12% in 1977 – despite a 15% expansion in dividends that year.

This context makes it more clear why *Comparison.Contrast was chosen*. However, a *Temporal.Synchronous* relation also holds between the two arguments even with the surrounding context. Thus, though the model predicted this relation incorrectly, it predicted a relation that was close to another relation that holds between the two but was not annotated in the gold label set.

TED-MDB Below is the example without context, which the model predicted as *Expansion.Level-of-detail*:

(11) **I want to show you a new kind of map.** *This is not a geographic map.*

The ground truth label of this example is *Comparison.Concession*. Without the additional context, the annotator understood how *Level-of-detail* could be inferred, as the speaker seems to be elaborating on the type of map. However, the example below, with added context, clarifies this:

(12) When we think about mapping cities, we tend to think about roads and streets and buildings, and the settlement narrative that led to their creation, or you might think about the bold vision of an urban designer, but there's other ways to think about mapping cities and how they got to be made. Today, **I want to show you a new kind of map.** *This is not a geographic map.* This is a map of the relationships between people in my hometown of Baltimore, Maryland, and what you can see here is that each dot represents a person, each line represents a relationship between those people, and each color represents a community within the network.

With the addition of the sentence before argument 1, it is a lot more clear why the correct label is *Comparison.Concession* and not *Expansion.Level-of-detail*.

D Examples from qualitative analysis of model results

For all examples below, argument 1 is **bolded** and argument 2 is *italicized*.

D.1 Examples where annotator disagrees with ground truth

PDTB-2 Below is the example without context, which the model predicted as *Contingency.Cause*:

(13) **Pro-forma balance sheets clearly show why Cray Research favored the spinoff.** *Without the Cray-3 research and development expenses, the company would have been able to report a profit of $19.3 million for the first half of 1989 rather than the $5.9 million it posted.*

The ground truth label of this example is *Expansion.Restatement*. When inspecting the model output without context, our annotator questioned the reason for this, as they believed there was a stronger causal relation. Below is the example with context:

(14) Analysts calculate Cray Computer's initial book value at about $4.75 a share. Along with the note, Cray Research is transferring about $53 million in assets, primarily those related to the Cray-3 development, which has been a drain on Cray Research's earnings.
Pro-forma balance sheets clearly show why Cray Research favored the spinoff. *Without the Cray-3 research and development expenses, the company would have been able to report a profit of $19.3 million for the first half of 1989 rather than the $5.9 million it posted.*

Here, context does not have much of an impact on the meaning of the relation. Thus, the opinion remained that *Contingency.Cause* is more correct than *Expansion.Restatement*, and thus the model did not commit an error here.

PDTB-3 This example illustrates the common ambiguity between *Expansion.Instantiation* and *Expansion.Restatement*, and represents a case in which our annotator disagreed with the ground truth label. Below is the example without context, which the model predicted as *Expansion.Instantiation*:

(15) **The competition has cultivated a much savvier consumer.**
The average household will spread 19 accounts over a dozen financial institutions,

The ground truth label of this example is *Expansion.Instantiation*. Below is the example with context:

(16) "Today, a banker is worrying about local, regional and money-center banks, as well as thrifts and credit unions," says Ms. Moore at Synergistics Research. "So people who weren't even thinking about targeting 10 years ago are scrambling to define their customer base." **The competition has cultivated a much savvier consumer.** "*The average household will spread 19 accounts over a dozen financial institutions,*" says Michael P. Sullivan, who runs his own bank consulting firm in Charlotte, N.C. "This much fragmentation makes attracting and keeping today's rate-sensitive customers costly."

Though this context sheds light on the fact that the focus of this passage is on customers' behavior with respect to banking, it is unclear whether argument 2 represents the only way in which the customer has become more savvy as a result of the competition. Thus, it is still ambiguous as to which relation holds here, and the model's decision to predict *Expansion.Instantiation* is close to if not the correct choice.

TED-MDB This example represents a case in which the ground truth connective appears to make the most sense, but our annotator did not agree with the ground truth sense label. Below is the example without context, which the model predicted as *Expansion.Conjunction*:

(17) **the balance of power to really influence sustainability rests with institutional investors, the large investors like pension funds, foundations and endowments.**
I believe that sustainable investing is less complicated than you think, better-performing than you believe, and more important than we can imagine.

The ground truth label of this example is *Expansion.Level-of-detail*. Upon seeing the model output, the annotator concluded that they would have also likely chosen *Conjunction* over *Level-of-detail*. The additional context, as seen below, does not appear to contradict this assessment:

(18) And by sustainability, I mean the really juicy things, like environmental and social issues and corporate governance. I think it's reckless to ignore these things, because doing so can jeopardize future long-term returns. And here's something that may surprise you: **the balance of power to really influence sustainability rests with institutional investors, the large investors like pension funds, foundations and endowments.** *I believe that sustainable investing is less complicated than you think, better-performing than you believe, and more important than we can imagine.*

Because the added context does make the two statements seem any more parallel, *Expansion.Conjunction* appears to be the best choice, despite the fact that *in fact* makes the most sense as a connective. Indeed, given that *Expansion.Conjunction* is the second-most-common annotation for *in fact* per the PDTB 3.0 Annotation Manual, the connective *in fact* being correct here does not preclude the possibility of *Expansion.Conjunction* being the correct label, but may have influenced the annotators of the TED dataset in the direction of *Expansion.Restatement.*

E Level 2 Recall and Annotation Distributions

Discourse Sense	XLNet	+ correctness	+ confidence	+ corr. & confidence	+ temp change	+ reweighting
Temp.Asynchronous	0.4333	0.44	0.3933	0.3867	0.4067	0.4867
Temp.Synchrony	0.1795	0.2051	0.1795	0.1795	0.1538	0.1795
Cont.Cause	0.6749	0.6737	0.6725	0.6655	0.6432	0.6573
Cont.Pragmatic cause	0	0	0	0	0	0
Comp.Contrast	0.5029	0.5478	0.5789	0.5731	0.5497	0.5439
Comp.Concession	0.0444	0.0444	0.0889	0	0	0.0444
Exp.Conjunction	0.5505	0.5745	0.5328	0.5694	0.5265	0.5366
Exp.Instantiation	0.5741	0.5802	0.6080	0.5957	0.5833	0.6420
Exp.Restatement	0.4772	0.4871	0.4859	0.4797	0.5412	0.4686
Exp.Alternative	0.2333	0.4	0.3333	0.2333	0.2	0.2667
Exp.List	0.2667	0.2667	0.3333	0.1667	0.2333	0.2667

Table 8: Recall on Level 2 senses for the PDTB-2, excluding labels that did not appear in the test set (note that the temp change model uses the Brier score as a validation metric)

Discourse Sense	XLNet	+ correctness	+ confidence	+ corr. & confidence	+ temp change	+ reweighting
Temp.Asynchronous	0.5841	0.5810	0.5810	0.5683	0.5950	0.5841
Temp.Synchronous	0.2424	0.2525	0.3030	0.2727	0.2338	0.2626
Cont.Cause	0.7506	0.7409	0.7513	0.7350	0.7439	0.7587
Cont.Cause+Belief	0	0	0.0256	0	0	0
Cont.Condition	0.7407	0.9074	0.9259	0.7222	0.7407	0.8333
Cont.Purpose	0.9271	0.9236	0.9306	0.9444	0.9256	0.9097
Comp.Contrast	0.4505	0.4835	0.4505	0.4689	0.4584	0.4762
Comp.Concession	0.6254	0.6222	0.6127	0.6894	0.6227	0.5683
Exp.Conjunction	0.6176	0.6656	0.6399	0.6522	0.6262	0.6577
Exp.Instantiation	0.6751	0.6554	0.6582	0.6638	0.6715	0.6102
Exp.Equivalence	0.1333	0.2533	0.0933	0.2533	0.1276	0.1200
Exp.Level-of-detail	0.4635	0.4793	0.4927	0.4562	0.4630	0.4818
Exp.Manner	0.1905	0.2381	0.2381	0.2143	0.1905	0.2738
Exp.Substitution	0.5938	0.5625	0.5938	0.6875	0.5938	0.6979

Table 9: Recall on Level 2 senses for the PDTB-3 (note that the temp change model uses the Brier score as a validation metric)

Discourse Sense	XLNet	+ correctness	+ confidence	+ corr. & confidence	+ temp change	+ reweighting
Temp.Asynchronous	0.5	0.5	0.5417	0.5417	0.5417	0.3333
Cont.Cause	0.7857	0.8571	0.8095	0.7143	0.8571	0.7857
Cont.Cause+Belief	0.1111	0	0.1111	0	0	0
Cont.Purpose	0.9333	1	0.9333	0.8667	1	1
Comp.Contrast	0.2222	0.3333	0.2222	0.3333	0.3333	0.2222
Comp.Concession	0.3333	0.3333	0.2222	0.2222	0.2222	0.2222
Exp.Conjunction	0.5	0.4487	0.4744	0.4359	0.4872	0.4487
Exp.Instantiation	0.5333	0.6	0.5333	0.6	0.5333	0.4667
Exp.Equivalence	0.4	0.4	0.2667	0.3333	0.3333	0.2
Exp.Level-of-detail	0.4697	0.4394	0.3636	0.4091	0.4848	0.4545
Exp.Substitution	0.6667	0.6667	0.6667	0.6667	0.6667	0.5556

Table 10: Recall on Level 2 senses for the TED-MDB, excluding labels that did not appear in the test set (note that the temp change model uses the Brier score as a validation metric)

Discourse sense	Ground Truth	Without context	With context	Context effect
Temp.Asynchronous	6	4	5	**better**
Temp.Synchronous	0	1	1	neutral
Cont.Cause	14	11	9	*worse*
Cont.Cause+Belief	2	3	3	neutral
Cont.Cause+SpeechAct	2	1	1	neutral
Cont.Purpose	4	0	0	neutral
Comp.Contrast	3	17	13	**better**
Comp.Concession	8	2	3	**better**
Comp.Concession+SpeechAct	1	0	0	neutral
Comp.Similarity	2	0	1	**better**
Exp.Conjunction	20	31	25	**better**
Exp.Instantiation	5	14	15	*worse*
Exp.Equivalence	4	7	8	*worse*
Exp.Exception	1	0	0	neutral
Exp.Level-of-detail	30	14	21	**better**
Exp.Manner	0	1	1	neutral
Exp.Substitution	4	0	0	neutral

Table 11: Frequency of discourse senses in our annotated set with respect to ground truth, annotations without context, and annotations with context. We can see that at a label distribution level, a context window usually adds a better or neutral effect.