

The Role of Context and Uncertainty in Shallow Discourse Parsing

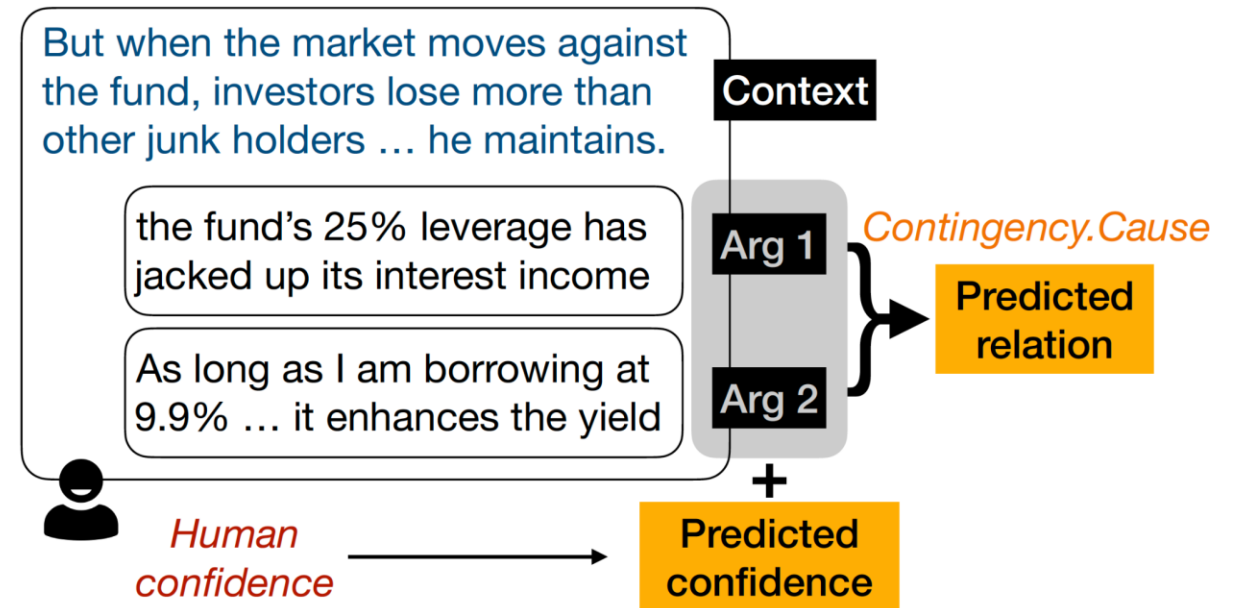
Katherine Atwell¹, Remi Choi¹, Junyi Jessy Li², Malihe Alikhani¹

¹Department of Computer Science, University of Pittsburgh

²Department of Linguistics, The University of Texas at Austin

Goals

- Determine whether shallow discourse parsers need context
- Determine whether human annotation accuracy and confidence scores can improve model accuracy and calibration



What are discourse relations?

Discourse relations represent the rhetorical relations between different units of text

- How the units of text are connected

Temporal.Asynchronous



Casey drove to the mall. She went to the Nike store.

What are discourse relations?

Discourse relations represent the rhetorical relations between different units of text

- How the units of text are connected

Contingency.Cause



Casey drove to the mall. She wanted to go to the Nike store.



Context in Discourse

The context of an utterance influences its interpretation

- Lewis (1980)
- Glanzberg (2002)
- Thompson-Schill (2003)

We hypothesize that context can influence the prediction of discourse relation labels

Context in Discourse

- Motivating this is the example below (arg1 is **bolded** and arg2 is *italicized*)

Argument pair (no context)

the fund's 25% leverage has jacked up the interest income

As long as I am borrowing at 9.9% and each (bond) yields over that, it enhances the yield

Expansion.Conjunction

Context in Discourse

- Motivating this is the example below (arg1 is **bolded** and arg2 is *italicized*)

Argument pair (with context)

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 Fund, says **the fund's 25% leverage has jacked up the 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

Contingency.Cause



Uncertainty Quantification

- *Calibration* - the distribution of error and the model's level of self-assessment, or confidence (Bella et al., 2010)
 - Neural networks tend to be poorly calibrated (Guo et al., 2017)
 - But pretrained models, even very complex ones, are typically more well-calibrated and benefit from temperature scaling (Desai and Durrett, 2020)
-



Uncertainty Quantification

- Well-calibrated models are more explainable and can improve downstream tasks that utilize estimated probabilities
 - Calibration is especially important for discourse parsing due to the difficulty of the task
 - This is the first work to study calibration for discourse parsing
-



Research Questions

How does context
affect annotation
accuracy and
confidence levels?

How can we quantify
the impact of context
on these metrics to
inform model
decisions?



Datasets

Penn Discourse
Treebank 2.0
(PDTB-2)

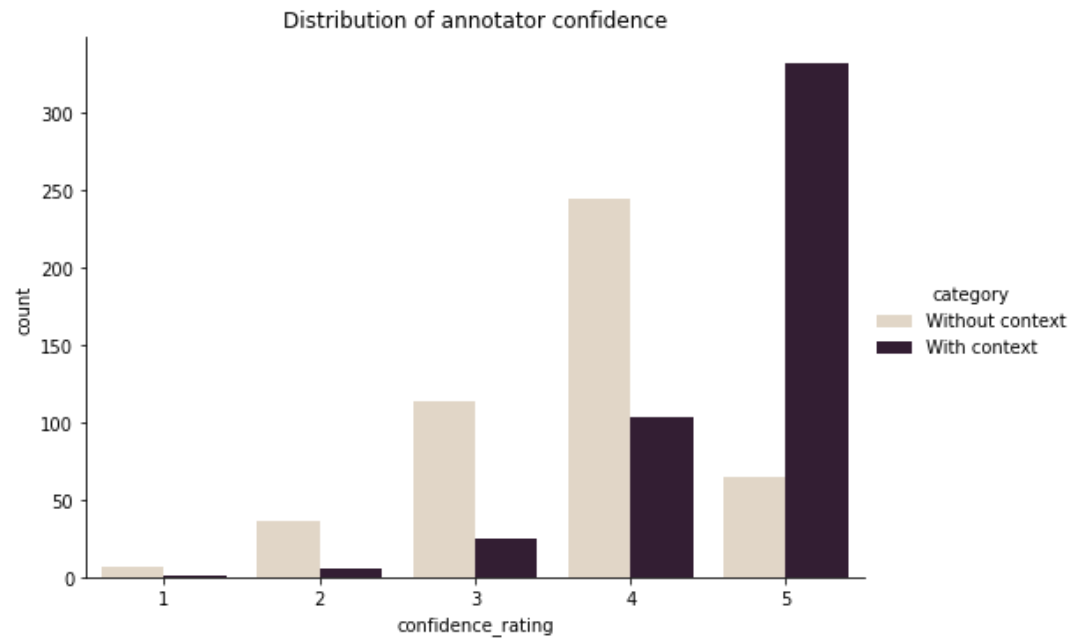
Penn Discourse
Treebank 3.0
(PDTB-3)

TED Multilingual
Discourse Bank
(TED-MDB)



Annotations

- 2 expert linguists
 - Annotate discourse connective and implicit relation under two conditions:
 - Without context
 - With context
 - Without context appeared first, and the answers to “without context” could not be changed after viewing the context
 - 498 annotated samples
 - PDTB-2: 147
 - PDTB-3: 199
 - TED-MDB: 152
-



Corpus	Accuracy		Confidence	
	Raw	Context	Raw	Context
PDTB2	.306	.354	3.70	4.81
PDTB3	.379	.423	3.56	4.71
TED-MDB	.296	.355	3.63	4.84

Annotation Results

Access to context **does** improve annotation accuracy and confidence




Modeling Questions

- Do these accuracy and confidence metrics provide helpful insights for classification models?
 - Can we improve both:
 - Accuracy (% correct)
 - Calibration (Brier score)
-




Setup

- Baseline: **XLNet-large (Kim et al, 2020)**
 - Experimental settings:
 - **Baseline + predicted accuracy/confidence**
 - **Reweighting with predicted confidence scores**
 - **Temperature adjustment with predicted confidence scores**
-



Results - Accuracy

Model	PDTB-2	PDTB-3	TED
XLNet-large (cased)	.5527	.6326	.5381
+Correctness	.5694	.6452	.5347
+Confidence	.5648	.6518	.5035
+Correctness and Confidence	.5462	.6428	.4931
+Reweighting	.5665	.6419	.4861



Results - Calibration

Model	PDTB-2	PDTB-3	TED
XLNet-large (cased)	.6781	.5787	.7214
+Temp change	.6075	.5295	.6477

Conclusion



Adding context improves **annotator accuracy and confidence**



Incorporating annotation accuracy and confidence improves **model accuracy and calibration**



Future works should continue to study calibration and uncertainty quantification for discourse models

Thank you!

