Discourse beyond Units: The Role of Context in Relation Recognition

Anonymous ACL submission

Abstract

Discourse frameworks have traditionally centered on *minimal* spans of “discourse units” or arguments, as defined by annotation schemas in frameworks like PDTB or RST. While discourse relations have been understood to not be viewed in full isolation, this approach may still be limiting, as annotators typically have access to the entire context when labeling spans and relations. In this study, we empirically evaluate the inclusion of contextual information in discourse modeling. Further, we also evaluate the effect of including explicit modeling of interactions between the spans. Our findings reveal that context-inclusive models outperform non-contextual baselines in case of explicit relations, with the inclusion of context proving more beneficial than explicit interargument modeling, but not beneficial in the case of implicit relations. We observe average improvements of 10.04% for PDTB3-L1, and 16.25% for L2. This work suggests that discourse units are not as minimal as previously assumed and contributes to a more nuanced understanding of discourse structure, opening new avenues for improving NLP for discourse comprehension.

1 Introduction

Discourse is typically conceptualized as sequences of discrete semantic units, where the larger semantics is determined based on the relationships established between the units. While the true interplay of relationships between pieces of text in the document could be extensive, the frameworks implicitly suggest a constrained view – for example, Penn Discourse Tree Bank (PDTB) (Prasad et al., 2008; Liang et al., 2020) annotates for a relationship between a pair of units whereas the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) looks at the nested hierarchical structure.The utilization of discourse units in NLP have mostly been limited to just the units themselves or the connectives if they are explicitly marked, or *immediate* surrounding context. Here, we seek to explore how much contextual information outside the units inform the discourse relation recognition.

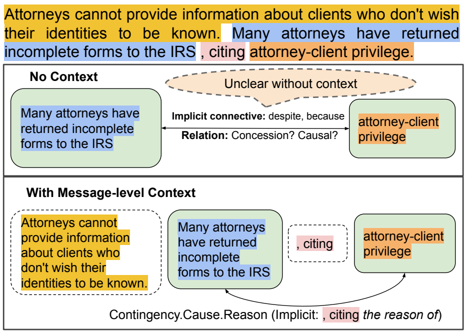

Attorneys cannot provide information about clients who don't wish their identities to be known. Many attorneys have returned incomplete forms to the IRS , citing attorney-client privilege.
No Context
Unclear without context
Many attorneys have returned incomplete forms to the IRS
Implicit connective: despite, because
Relation: Concession? Causal?
attorney-client privilege
With Message-level Context
Attorneys cannot provide information about clients who don't wish their identities to be known.
Many attorneys have returned incomplete forms to the IRS
, citing
attorney-client privilege
Contingency.Cause.Reason (Implicit: , citing the reason of)

Figure 1: While discourse relations are meant to be wholly captured by the individual units themselves, there is a lot of information in the rest of the message that provides context to distinguish the relations better. We explore how relevant this context is with respect to inferring relationship between discourse units.

Our results reveal several key insights about discourse relations and context in language models. Our main findings are: (1) That the importance of discourse context is not universal, with implicit relations and dissonance performing better without extensive context, highlighting their independence from connective elements; (2) Both explicit and implicit discourse relations exhibit attention patterns extending to distant tokens, particularly 500-1000 tokens ahead in the sequence, challenging conventional assumptions about local context dependencies; (3) Our straightforward sliding window approach proves more effective than models trained on larger contexts in discourse relation recognition tasks; and finally, (4) Despite the inherent hierarchical structure of transformers, explicit cross-attention mechanisms enhance the model’s ability to capture discourse-level interactions, suggesting that direct modeling of these relationships still provides tangible benefits.

2 Related Work

Discourse relations have been extensively studied as logical or structural connections between segments of discourse, typically describing how two segments are related to one another in the context of their surrounding (Knott et al., 2002; Taboada, 2006; Lin et al., 2009). The Penn Discourse Tree Bank (PDTB) (Prasad et al., 2017) and Rhetorical Structure Theory (Mann and Thompson, 1988) define a framework for these relationships.While the hierarchical nature of transformers seem to capture deeper semantic roles and relationships, newer perspectives of discourse frameworks have revealed some fundamental drawbacks in transformer-based LLMs, where inducing structure might help (Chernyavskiy et al., 2021; Miletić and Walde, 2024).

Some works explored have emphasized role of context within the discourse units themselves to improve implicit relation recognition (Qin et al., 2016; Zhang et al., 2021; Atwell et al., 2021), with some in the context of social media as well (Varadarajan et al., 2023). Some have explored modeling explicit connectives, or removing them to learn implicit relations (Liu et al., 2024; Son et al., 2022). Since the absence of explicit connectives is known to make the relation recognition problem more difficult (Xiang and Wang, 2023), most of the recent efforts have focused on improving implicit relation recognition (Kim et al., 2020; Kishimoto et al., 2020; Liu et al., 2021). Prior research on leveraging broader contextual information in discourse analysis has been limited. While Zhou et al. (2020) explored the use of global context to enhance implicit relation recognition, the impact of extended context on explicit discourse relations and connective disambiguation remains largely unexplored. Our research extends beyond this by systematically investigating how distant contextual elements influence the overall process of discourse relation recognition, including both implicit and explicit relations.

3 Experiment

We describe the experimental setup to examine the effect of inclusion of context of discourse beyond explicit markers and adjacent or in-between tokens.

3.1 Data

PDTB3 dataset The PDTB3 dataset (Webber et al., 2019) represents a recent discourse framework annotated on Wall Street Journal articles, which improves upon PDTB2 (Prasad et al., 2008) in adding more implicit, intersentential relations which were missed previously. Traditional approaches examine discourse unit pairs and their immediate sentential context, primarily focusing on text between or adjacent to discourse units. Instead, our approach incorporates all tokens from the source WSJ articles. We process articles as complete sequences of discourse units, derive pairs of units from each article that have been annotated in the PDTB3 dataset. The final dataset consists of 2024 articles, each consisting of 362.5 words on an average (min: 5, max: 3135 words), with each article containing an average of ∼10 discourse unit pairs as input (not every discourse unit is annotated as part of a pair).

Twitter Dissonance dataset The Twitter Dissonance dataset introduces consonance and dissonance as two relations between phrases that state beliefs (Varadarajan et al., 2023). Both of these relations are implicit in nature. They are annotated on noisy social media posts on Twitter, with an average message length of 35 words (max: 91, min:5). This dataset is meant to challenge the model's ability to improve upon relation recognition given shorter context.

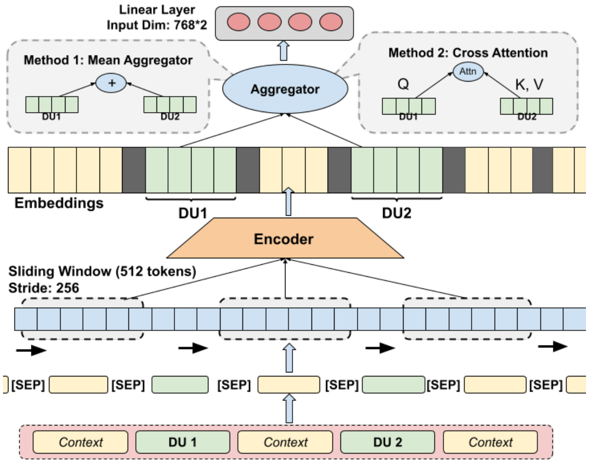


Figure 2: Description for Context Inclusion Architectures. DU stands for discourse unit.

3.2 Context Inclusion

The entire article or message that the discourse units (DUs) belong to, serves as context to those discourse units. As shown in Figure 2, each message could consist of multiple discourse unit pairs, which are fed into an encoder with separator tokens in between adjacent discourse units. The separator tokens indicate the start and end of adjacent discourse units.

Model	Aggregator	Precision	Recall	F1 per class				AUC	F1
		(macro)	(macro)	Comparison	Contingency	Expansion	Temporal	(avg.)	(wei.)
Non-contextual Baselines									
Roberta-b	mean	0.667	0.647	0.662	0.637	0.742	0.581	0.869	0.681
Roberta-b	cross-attn	0.667	0.649	0.663	0.637	0.743	0.582	0.870	0.682
Context Inclusion Models									
Longformer	mean	0.728	0.701	0.729	0.644	0.759	0.716	0.889	0.718
nomic	mean	0.728	0.688	0.723	0.633	0.758	0.704	0.885	0.713
nomic	cross-attn	0.712	0.630	0.686	0.594	0.731	0.623	0.857	0.674
Roberta-b	mean	0.743	**0.731**	**0.784**	0.692	0.786	0.682	0.909	0.749
Roberta-b	cross-attn	**0.749**	**0.731**	0.783	**0.693**	**0.791**	**0.687**	**0.912**	**0.752**

Table 1: Effect of context inclusion on L1 relation recognition for PDTB3. While we find that inclusion of context improves relation recognition by 10% on an average. across the board, we further find that a cross attention mechanism across the contextualized representations of two discourse units under consideration can further improve modeling of the relationship between them.

Sliding Window Since articles in WSJ often exceed 512 tokens that contain multiple discourse unit pairs, we adopt a sliding window approach with a stride of 256 tokens to generate embeddings for all tokens. Token embeddings from overlapping strides are averaged to ensure all tokens are contextualized in the paragraph while allowing each token to cross-attend with more contextual tokens than usual. The two discourse units are aggregated before passing them to a linear layer for the classification task: we explore two methods as shown in Figure 2: mean-based aggregation and cross-attention-based aggregation.

Mean-aggregated model In mean-aggregated model, token representations of each discourse unit are averaged to a single discourse unit representation, and are concatenated before passing them to a linear layer.

Cross-attention model In the cross-attention-based model, the token representations of each discourse unit are singled out for performing cross-attention between the tokens of DU1 and DU2. Specifically, The embeddings of the tokens in DU1 are used as the Query (Q). The embeddings of the tokens in DU2 are used as the Key (K) and Value (V). This cross-attention mechanism enables interaction between the tokens of the two discourse units. The objective is to evaluate whether modeling interactions between the contextualized tokens improves classification performance.

Pretrained Models We employ strong encoder models to test context inclusion. To this end, we explore RoBERTa-base (Liu et al., 2019) and $\texttt{nomic}-\texttt{embed}-\texttt{text}-\texttt{v1.5}$ which is a newer BERT-based model optimized for large-context encoding representations that can handle upto 8192 tokens (Nussbaum et al., 2024). Further, since the sliding window approach is similar to the Longformer (Beltagy et al., 2020) implementation of global and local attention for large context, we also include a comparison to $\texttt{Longformer}-4096$. The results are shown in Table 1.

Aggregator	Prec. (macro)	Rec. (macro)	AUC (avg.)	F1 (wei.)
Non-contextual Baselines				
mean	0.472	0.418	0.894	0.563
cross-attn	0.492	0.437	0.890	0.569
Context Inclusion				
mean	**0.585**	0.534	**0.930**	**0.659**
cross-attn	0.563	**0.539**	**0.930**	0.657

Table 2: Effect of context inclusion on L2 relation recognition for PDTB3 for RoBERTA-base encoder. Context outperforms non-contextual baselines, with mean aggregation performing on par with the cross-attention model. Contextual models show an average improvement of 16.25% over non-contextual models.

	F1 (macro))		AUC
Aggregator	Dissonance	Consonance	(macro)
Context Inclusion			
cross-attn	**0.363**	0.757	0.759
mean	**0.307**	0.667	0.718
Non-contextual			
cross-attn	0.252	**0.759**	0.719
mean	0.286	**0.779**	**0.774**

Table 3: Performance Comparison for Dissonance Classification. Dissonance is a rare class – occurring only in $\sim 10\%$ of the dataset. Inclusion of context and improves dissonance classification, but this comes at the cost of overall performance. This further bolsters our finding that context plays a tricky role in implicit relations, even in short, social media context.

3.3 Non Contextual Baselines

In the non-contextual approach, we discard the rest of context and focus only on the two discourse units as inputs concatenated together with a separator token between them. The rest of the architecture is the same as Figure 2. We explore both mean-aggregated and cross-attention models for the non-contextual setting as well.

Training The models are trained to classify Level 1 and Level 2 discourse relations of the PDTB3 dataset. For Level 1 (L1), we consider all four classes, while for Level 2 (L2), we focus on the 18 classes with the most data points. All the architectures were hyperparameter-tuned for each configuration. Following the previous works (Kim et al., 2020), we employ 12-fold cross validation across the 25 sections, considering 2 sections as development set, 2 sections as test set, and the rest 21 sections as train set such that there are no overlapping test sets. We report the performance metrics on the aggregated test sets on all folds.

4 Results

Effect of context As seen in Table 1 and 2, the context-included cross-attention architecture with RoBERTA-base performs the best with L1, and ties with mean-aggregated model for L2. We find that adding context to the models indeed helps in improving the classification performance. However, the *context* in our setting could include explicit markers of discourse relations (aka connectives) as well, where markers can indicate the relationship between two units despite not being a part of them.

Effect of token distance We examine if the model attends to tokens far away from the discourse units themselves, we capture attention weights for all the tokens in the model for the cross-attention context-inclusion model (Figure 3).

Category	Context	Intrasentential			Intersentential		
		F1 (wei.)	Prec (macro)	Rec (macro)	F1 (wei.)	Prec (macro)	Rec (macro)
Explicit	yes	0.836	0.822	0.824	**0.872**	0.843	0.854
	no	0.681	0.660	0.648	0.685	0.667	0.649
Implicit	yes	0.624	0.593	0.542	0.648	0.575	0.557
	no	**0.683**	0.669	0.655	0.679	0.664	0.648

Table 4: Performance Comparison for Explicit and Implicit Categories for the cross attention model.

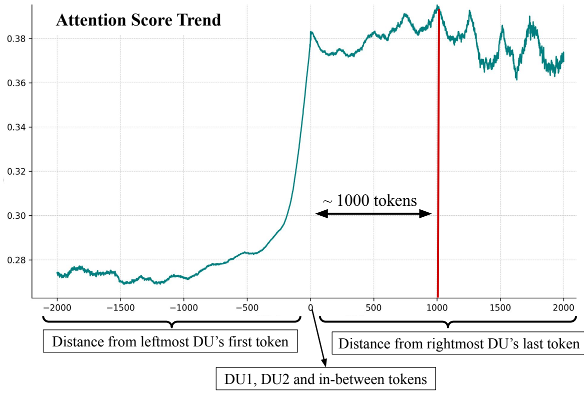


Figure 3: Plot of token attention against distance from the nearest token in discourse unit pair. All the tokens in between two discourse units, if any, are considered to be at a distance of 0. We find that the attention weights > 500-1000 tokens away still get comparable attention to that of the discourse units themselves, signaling that larger context other than connectives can be very helpful for discourse relation recognition.

Explicit and Implicit relations We observe better performance for the Explicit case in models that include context, likely since connectives are often located around discourse units – however, we also find evidence for contextual information from distant tokens (Fig 3). Conversely, for the Implicit case, we find that the inclusion of context doesn't significantly contribute to performance. Interestingly, the Non-Contextual model outperforms contextual models for the Implicit case, indicating that context doesn't provide meaningful benefits in this scenario (see §A2 for more details). This is observed across the board with PDTB3-L1, L2 and the Twitter Dissonance dataset (Tables 1, 2, 3).

For classwise, and intersentential error analysis of the contextual model, see Appendix§A.5, A.3.

5 Conclusions

We found that the inclusion of context does not universally enhance discourse relation classification, but most of the time there is a lot to gain from inclusion of distant context, suggesting that future approaches to this task can benefit from incorporating contextual information when necessary. Additionally, cross attention at the discourse-level despite using transformer-based LMs for incorporating contextual information improves the model further, suggesting that higher order semantic interactions might be present that a standard word-context transformer cannot capture. This approach may also be applicable to other tasks, such as identifying dissonance. Our experiments indicate that developing a more effective method to model context could yield better results, and designing improved fine-tuning objectives for encoders with context could further discourse comprehension in NLP.

6 Limitations

Our experiments were limited to PDTB3 and Twitter Dissonance datasets alone, both of which are based on the PDTB framework of annotating discourse segments and relations. Both of these datasets rely on high annotator agreement for determining ground truth values, where some of these relations can be largely subjective. Further, there might be other datasets where these results might not be observed, either due to the different conceptualizations of discourse or domain differences. To some extent, we have addressed this by exploring discourse relations of two distinct domains.

We run all of our experiments on an NVIDIA-RTX-A6000 with 50 GB of memory in an internal server, on open-sourced models from HuggingFace and wrote our code with PyTorch. The experiments took about 150 hours to run.

Our work is also limited to the English documents, and these results might not be reproducible in a different languages, especially ones with low resources.

References

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 314–325, Singapore and Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. Preprint, arXiv:2004.05150.

Alexander Chernyavskiy, Dmitry Ilvovsky, and Boris Galitsky. 2021. Correcting texts generated by transformers using discourse features and web mining. In Proceedings of the Student Research Workshop Associated with RANLP 2021, pages 36–43, Online. INCOMA Ltd.

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 Twelfth Language Resources and Evaluation Conference, pages 1152–1158.

Alistair Knott, Ted Sanders, and Jon Oberlander. 2002. Levels of representation in discourse relations.

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, Online. Association for Computational Linguistics.

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.

Wei Liu, Stephen Wan, and Michael Strube. 2024. What causes the failure of explicit to implicit discourse relation recognition? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2021. On the importance of word and sentence representation learning in implicit discourse relation classification. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Preprint, arXiv:1907.11692.

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

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. Transactions of the Association for Computational Linguistics, 12:593–612.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. Preprint, arXiv:2402.01613.

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

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. Handbook of linguistic annotation, pages 1197–1217.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1914–1924, Osaka, Japan. The COLING 2016 Organizing Committee.

Youngseo Son, Vasudha Varadarajan, and H. Andrew Schwartz. 2022. Discourse relation embeddings: Representing the relations between discourse segments in social media. In Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS), pages 45–55, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maite Taboada. 2006. Discourse markers as signals (or not) of rhetorical relations. Journal of pragmatics, 38(4):567–592.

Vasudha Varadarajan, Swanie Juhng, Syeda Mahwish, Xiaoran Liu, Jonah Luby, Christian Luhmann, and H Andrew Schwartz. 2023. Transfer and active learning for dissonance detection: Addressing the rare-class challenge. arXiv preprint arXiv:2305.02459.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. Philadelphia, University of Pennsylvania, 35:108.

Wei Xiang and Bang Wang. 2023. A survey of implicit discourse relation recognition. ACM Computing Surveys, 55(12):1–34.

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, Online. Association for Computational Linguistics.

Meilin Zhou, Qi Liang, Peng Zhang, Lu Ma, Dan Luo, and Bin Wang. 2020. Global context-aware representation for implicit discourse relation recognition. In 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), pages 458–465. IEEE.

A Appendix

A.1 Effect of Cross-Attention

We look to explore how the model performs when we focus on two discourse units using a cross attention layer. From Table A1 and 4 we find that Cross attention with Context models are the best performing models in both the L1 and L2 case. Even when all the encoder layers are frozen, in the L1 Case cross-attention model does better than the rest of the models even without the context. In the L2 case cross attention does the best with the context. When unfrozen, cross attention model further does better consistently. Cross attention model only performs slightly better than the mean models. While it helps in improved performance but not a significant improvement.

A.2 Ablation of frozen models

We train these models with a frozen encoder to evaluate the performance of pretrained models and compare it to fine-tuning the encoder. When the encoder is unfrozen, the model consistently outperforms the frozen case, clearly indicating that fine-tuning the encoder specifically for discourse classification is beneficial. A t-test further confirms that unfreezing the layers leads to significant(**) improvements under all conditions. The results are shown in Table A1.

A.3 Intrasential and intersentential relations

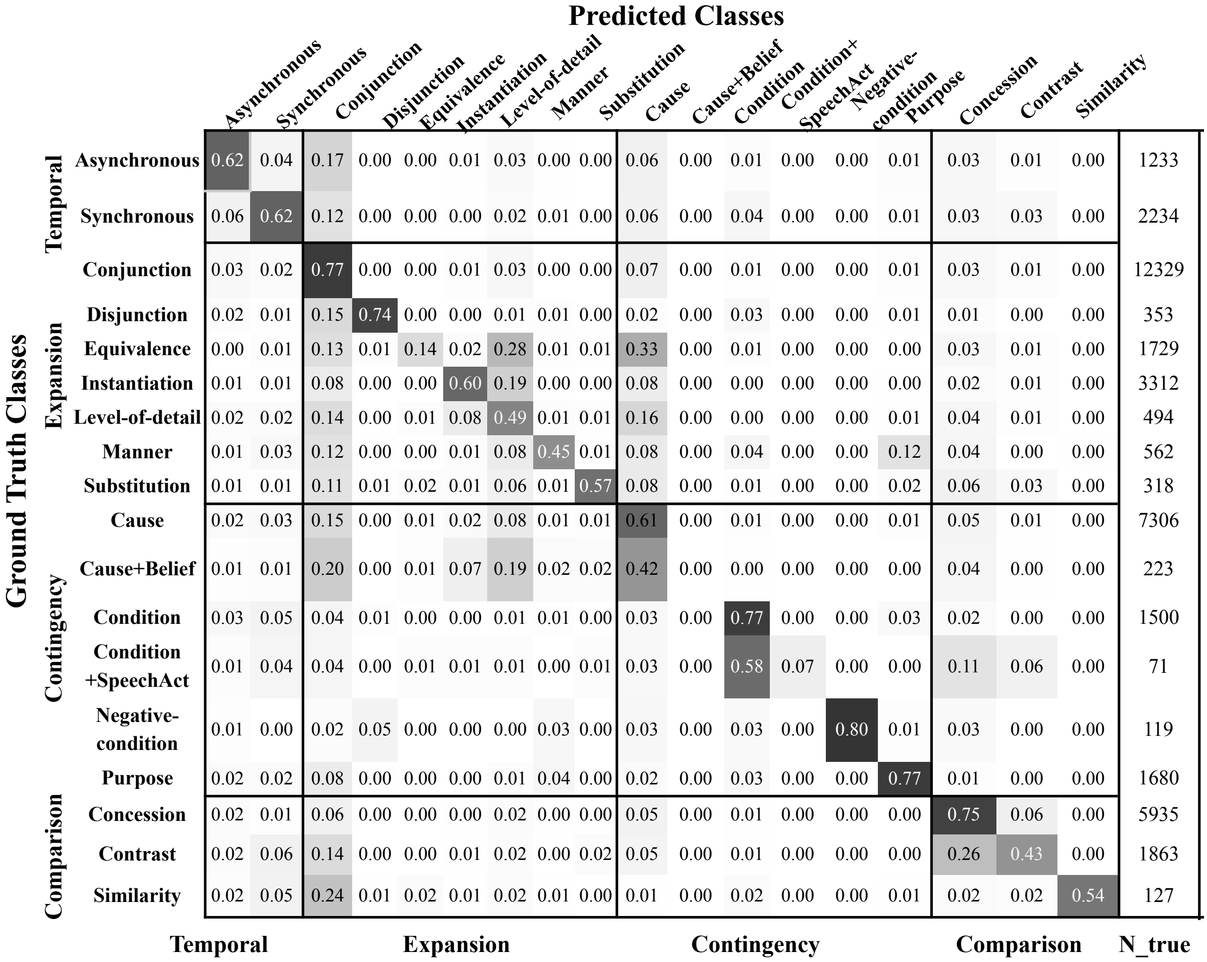
PDTB3 introduces $\sim$ 300 intersentential discourse units (Prasad et al., 2008), which are considered harder to classify as compared to the intrasentential relationsWe analyse how the best-perfoming model does on cases where the discourse Units are Intrasentential where two discourse units are part of the same sentence and intrasentential where two discourse units are part of the different sentence. The model performance in both the cases remains almost similar, as seen in the Table 4

A.4 Explicit and Implicit Attention scores

Figure A2 shows the attention score trend for the Contextual Cross-Attention model trained on L1 classes. Averaged over the PDTB3 dataset, the trend highlights that in the Explicit case, the model assigns higher attention to distant tokens, likely connectives. Notably, the Explicit trend shows higher average attention scores at a distance compared to the Implicit case, underscoring the greater importance of context in the Explicit case.

A.5 Class-level analysis

As observed in Figure A1,In the L2 case, we observe that the model confuses certain classes with others, and this confusion appears to be intuitive due to the inherently ambiguous nature of the data. For instance, the model often confuses Cause+Belief with Cause and Condition+SpeechAct with Condition. This confusion predominantly occurs in cases where the number of data points is very limited. Specifically, Cause+Belief has only 223 data points in the dataset, while Condition+SpeechAct has just 71.



Predicted Classes

Ground Truth Classes		Temporal		Expansion							Contingency						Comparison			N_true
		Asynchronous	Synchronous	Conjunction	Disjunction	Equivalence	Instantiation	Level-of-detail	Manner	Substitution	Cause	Cause+Belief	Condition	Condition+ SpeechAct	Negative- condition	Purpose	Concession	Contrast	Similarity	
Temporal	Asynchronous	0.62	0.04	0.17	0.00	0.00	0.01	0.03	0.00	0.00	0.06	0.00	0.01	0.00	0.00	0.01	0.03	0.01	0.00	1233
	Synchronous	0.06	0.62	0.12	0.00	0.00	0.00	0.02	0.01	0.00	0.06	0.00	0.04	0.00	0.00	0.01	0.03	0.03	0.00	2234
Expansion	Conjunction	0.03	0.02	0.77	0.00	0.00	0.01	0.03	0.00	0.00	0.07	0.00	0.01	0.00	0.00	0.01	0.03	0.01	0.00	12329
	Disjunction	0.02	0.01	0.15	0.74	0.00	0.00	0.01	0.01	0.00	0.02	0.00	0.03	0.00	0.00	0.01	0.01	0.00	0.00	353
	Equivalence	0.00	0.01	0.13	0.01	0.14	0.02	0.28	0.01	0.01	0.33	0.00	0.01	0.00	0.00	0.00	0.03	0.01	0.00	1729
	Instantiation	0.01	0.01	0.08	0.00	0.00	0.60	0.19	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.00	3312
	Level-of-detail	0.02	0.02	0.14	0.00	0.01	0.08	0.49	0.01	0.01	0.16	0.00	0.00	0.00	0.00	0.01	0.04	0.01	0.00	494
	Manner	0.01	0.03	0.12	0.00	0.00	0.01	0.08	0.45	0.01	0.08	0.00	0.04	0.00	0.00	0.12	0.04	0.00	0.00	562
	Substitution	0.01	0.01	0.11	0.01	0.02	0.01	0.06	0.01	0.57	0.08	0.00	0.01	0.00	0.00	0.02	0.06	0.03	0.00	318
Contingency	Cause	0.02	0.03	0.15	0.00	0.01	0.02	0.08	0.01	0.01	0.61	0.00	0.01	0.00	0.00	0.01	0.05	0.01	0.00	7306
	Cause+Belief	0.01	0.01	0.20	0.00	0.01	0.07	0.19	0.02	0.02	0.42	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	223
	Condition	0.03	0.05	0.04	0.01	0.00	0.00	0.01	0.01	0.00	0.03	0.00	0.77	0.00	0.00	0.03	0.02	0.00	0.00	1500
	Condition +SpeechAct	0.01	0.04	0.04	0.00	0.01	0.01	0.01	0.00	0.01	0.03	0.00	0.58	0.07	0.00	0.00	0.11	0.06	0.00	71
	Negative- condition	0.01	0.00	0.02	0.05	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.03	0.00	0.80	0.01	0.03	0.00	0.00	119
	Purpose	0.02	0.02	0.08	0.00	0.00	0.00	0.01	0.04	0.00	0.02	0.00	0.03	0.00	0.00	0.77	0.01	0.00	0.00	1680
Comparison	Concession	0.02	0.01	0.06	0.00	0.00	0.00	0.02	0.00	0.00	0.05	0.00	0.01	0.00	0.00	0.00	0.75	0.06	0.00	5935
	Contrast	0.02	0.06	0.14	0.00	0.00	0.01	0.02	0.00	0.02	0.05	0.00	0.01	0.00	0.00	0.00	0.26	0.43	0.00	1863
	Similarity	0.02	0.05	0.24	0.01	0.02	0.01	0.02	0.01	0.00	0.01	0.00	0.02	0.00	0.00	0.01	0.02	0.02	0.54	127

Figure A1: Confusion matrix for classification L2 – each cell represents the recall rate of that class

Aggregator	Context	Level	Frozen			Unfrozen		
			F1 (wei.)	Prec (macro)	Rec (macro)	F1 (wei.)	Prec (macro)	Rec (macro)
mean	no	L1	0.538	0.524	0.478	**0.681**	0.665	0.648
cross-attn	no	L1	0.613	0.594	0.576	**0.682**	0.664	0.649
mean	yes	L1	0.588	0.580	0.535	**0.749**	0.742	0.731
cross-attn	yes	L1	0.603	0.589	0.560	**0.752**	0.748	0.731
mean	no	L2	0.394	0.291	0.198	**0.556**	0.472	0.419
cross-attn	no	L2	0.480	0.405	0.299	**0.569**	0.490	0.435
mean	yes	L2	0.450	0.352	0.237	**0.659**	0.589	0.533
cross-attn	yes	L2	0.511	0.475	0.333	**0.657**	0.577	0.540

Table A1: Performance comparison of frozen and unfrozen models for different contexts and levels. Unfreezing the layers lead to significant ** improvement under all conditions.

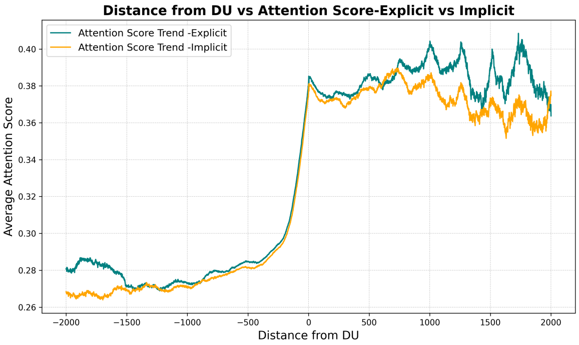


Figure A2: In the Explicit case, we observe higher attention scores at a distance from the DUs compared to the Implicit case. The model has learned to pay attention to connectives farther away from the DUs.