Enhancing Discourse Relation Classification with Attention Mechanisms on Genre-Diverse Data

Darja Jepifanova
D.Jepifanova@campus.lmu.de

Marco Flöß
Ma.Floess@campus.lmu.de

Abstract

In this work, we investigate discourse relation classification within the context of the Rhetorical Structure Theory (RST) framework, using the newly expanded genre-diverse GUM corpus (version 10.2.0). We introduce a data-driven clustering approach to group genres into five clusters, revealing genre-specific patterns according to the extracted discourse related information. Inspired by previous work on the varying strength of discourse signals, we explore the potential of transformer-based models for discourse relation classification. We compare a bidirectional LSTM baseline (following Zeldes and Liu 2020) with a transformer-based T5-small model with both fine-grained and coarse labels. Our results show that the T5 model consistently outperforms the baseline, particularly in cases where explicit discourse markers are present, and also demonstrates better generalization to less frequent relations. A potential direction for future research is to analyze attention weights in transformer models to assess whether they align with known discourse signals.

1 Introduction

Discourse models make use of various annotation frameworks to identify the structural composition of spans of text within a document. These frameworks describe how different spans relate to each other and classify those as instantiations of semantic-pragmatic discourse relations (Braud et al., 2023). Thereby a formalization of semantics beyond the sentence boundary is achieved. Over the years a multitude of annotation frameworks have been developed with Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) being among the most influential. Essentially, RST aims at producing a hierarchical constituent tree. Its components are spans of text which consist of either a single elementary discourse unit (EDU) or a set of several, related EDUs (Kobayashi et al., 2022). These constituents are connected through discourse relations such as `purpose-goal` or `restatement-repetition` and possess a nuclearity status being either satellite or nucleus.

While the state of the art discourse models for EDU-segmentation demonstrate high performance, the task of discourse relation classification is remaining challenging with less performative results (Braud et al., 2023). The task of discourse relation classification concerns itself with predicting a relation between two spans of text and is the main focus of this paper. Our experimental setup was inspired by Zeldes and Liu (2020) in which the discourse relation classifier is presented EDU satellite-nucleus pairs with special tokens indicating the nuclearity and EDU separation. While the original paper used this discourse relation classifier as a distantly supervised model to investigate the signaling strength of tokens, its setup inspired us to extend this discourse relation classifier by attention mechanisms. This was motivated by Zeldes and Liu (2020) showing that the signaling strength of tokens is context dependent and by the fact that attention mechanisms show performance increases in a variety of domains by leveraging a weighting to certain parts of the input. In this paper we therefore compare the performance of the original setup of Zeldes and Liu (2020) to the T5-small transformer-based model to investigate the effect of attention mechanisms on the task of discourse relation classification. For data we chose the most recent version of the Georgetown University Multilayer Corpus (GUM, version 10.2.0) (Zeldes, 2017). Its genre diversity is an important prerequisite to provide the discourse relation classifier with some level of generalizability. This was motivated by Liu et al. (2023b) presenting that different genres often exhibit dissimilar pattern of discourse related information (e.g. distribution of discourse signals and discourse relations). To further investigate these dissimilarities between genres we performed a data-driven genre clustering and evaluated the performance of our discourse relation classifiers on those as separate test sets. Our code will be publicly available on GitHub[1].

2 Data

Our experiments are solely based on version 10 of the Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017). We chose the GUM corpus because it is an open-source collection of texts from various genres, continuously expanded by students at Georgetown University as part of their curriculum. As our primary focus is predicting discourse relations between pairs of EDUs, the provided rsd-file format [2] is straightforward to preprocess to arrive at input-label pairs as detailed at a later point (Zeldes et al., 2024). The discourse relation between two EDUs is based on the dependency representation as described in Li et al. (2014). The multitude of genres covered in this corpus enables the discourse relation classification models and the clustering analysis to hold some degree of generalizability. This has been become increasingly important as text genres differ, among other criteria, on the relative frequencies of relation types and to what percentage relations are implicitly/explicitly signaled (Liu et al., 2023b).

Table 1 shows that version 10 of GUM has sixteen genres from which four (court, essay, letter, podcast) are currently being built up and therefore have a limited amount of data.

Genre	Train: #Doc	Train: #Rels	Dev: #Docs	Dev: #Rels	Test: #Doc	Test: #Rels
academic	14	1514	2	202	2	247
bio	16	1697	2	174	2	180
conversation	10	2075	2	430	2	340
court	4	656	1	128	1	94
essay	3	370	1	171	1	146
fiction	15	1969	2	224	2	262
interview	15	1998	2	186	2	207
letter	4	562	1	109	1	103
news	19	1356	2	202	2	198
podcast	3	517	1	186	1	109
reddit	14	1729	2	261	3	388
speech	11	1463	2	254	2	182
textbook	11	1569	2	190	2	253
vlog	11	1922	2	277	2	222
voyage	14	1470	2	155	2	149
whow	15	1944	2	214	2	221

Table 1: Comparison of the count of documents and count of relations across training, development, and test sets.

Apart from these genres being built up the remaining twelve genres all have more than ten documents which provide us with a good amount of input-label pairs. In the given data format there is one relation to the 'ROOT' of the text per document which we did not further process as it consists of text for only one EDU and therefore does not fit into the discourse relation classification setup we aim to model. The count of relations is in Table 1 is therefore disregarding the 'ROOT' case.

Input-label pairs consist of the combined text of the EDU pair with special tokens and a discourse relation label. This label can be either coarse or fine-grained. For example, the coarse label `adversative` is broken down into following fine-grained labels: `adversative-antithesis`, `adversative-concession` and `adversative-contrast`. In our experiments we use both label sets for the discourse relations.

3 Genre Clustering

As the primary data source for many discourse relation classification developments the WSJ corpus has proven immensely valuable. However, the diverse genres present in the GUM corpus do not always follow the average distribution of discourse related information of the news genre, as has been shown in Liu et al. (2023b). Therefore the generalizability of discourse models built on specific genres is not a given, as the discourse relations may be marked in various ways and certain types of discourse relations may be more dominant in one genre in comparison to another.

This begs the question whether there are discourse related characteristics that are common within a set group of genres as data sparsity prevents complex systems to be trained on the limited data of just one genre.

The current version of GUM incorporates sixteen different genres which cover a wide range of various formats and settings in which speech or text can be produced. While to some extent it is possible to intuitively find semantic similarities and differences between genres, the manual grouping of several genres into clusters of similar semantic context is at risk of being strongly opinionated by their authors.

3.1 A data-driven clustering approach

This paper mainly concerns itself with the task of discourse relation classification. Therefore we chose a data-driven approach of clustering genres into groups, focusing on descriptive statistics of the genres. Specifically, we focus on discourse related information and general text composition.

[1] https://github.com/ydarja/disco-project

[2] https://github.com/amir-zeldes/gum/tree/master/rst/dependencies

Since the signals for discourse relations form an open class, we are more concerned with distributional properties of certain types of discourse signals within a genre rather than concrete occurrences of specific signals. Apart from information on discourse signals we obtained distributional properties for coarse discourse relation labels, sentence types of the EDUs, tense of the EDUs, stop words per EDU, count of tokens per EDU and the type-token-ratio per document. We initially collected this information on a document level which in the next step was averaged over all documents within one genre.

The resulting features have been standardized to a mean of zero and variance of one. This step aligns with our data-driven approach, as it prevents bias towards features with higher variance (partially the result of different scales), allowing more subtle differences in other features to equally impact the genre differentiation.

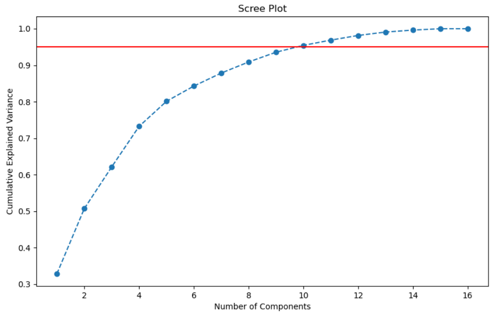

Figure 1: Scree Plot for identifying the number of necessary components. The horizontal line marks the 90 percent of cumulative explained variance.

With 53 features of descriptive standardized statistics per genre we still have a high-dimensionality which may impede clustering algorithms to effectively find patterns to group the genres. Therefore we performed a Principal Component Analysis to reduce our dimensionality to 10 before continuing with the clustering algorithm. The scree plot in Figure 1 shows that with ten components the PCA will retain over ninety percent of the variance within the original high-dimensional representation of data.

3.2 K-Means Clustering

For the clustering algorithm we chose the K-means algorithm which iteratively assigns our genres to a fixed number of clusters which are represented by the centroids of its data points. As the number of clusters has been given to the algorithm in advance, our method of determining the best choice for this hyperparameter was informed through the elbow criteria which quantifies the within-cluster sum of squares (WCSS) for different choices of numbers of clusters.

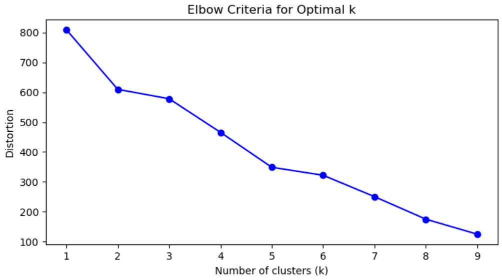

Figure 2: Elbow Plot for Optimal K

Figure 2 shows that the choice of five clusters results in a substantial drop in distortion (WCSS) while higher numbers of cluster (k) only slightly decrease this metric. Once we obtained the clusters, we assigned them intuitive names for the sake of reference and indicating some potential structural and semantic context. For visualizing our results we used the t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality to 2 for an accessible visual representation. t-SNE aims to preserve the local structure of data by calculating the likelihoods of data points being a neighbor of another thereby making it a useful tool to obtain visually interpretable representations.

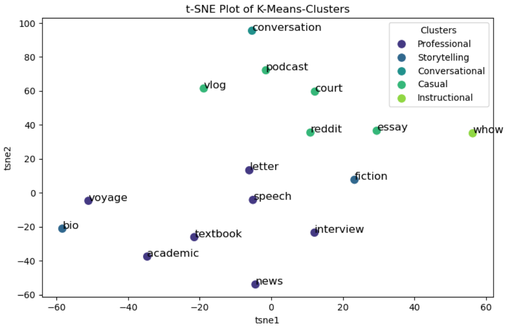

Figure 3: t-SNE plot to visually represent the clusters

The t-SNE figure 3 shows our five clusters with our naming declaration. One can see that we have two bigger clusters, one small cluster and two singular clusters. Apart from the “Storytelling” cluster encompassing the biography and fiction genres, the clusters are well locally separated and suggest that reasonable boundaries are made by our procedure. The singular clusters “Conversational” and “Instructional” indicate that the single genre within each cluster has substantial characteristics that are unique to it. At this point in time we refrain from further interpretation of the clustering results and remit to the data-driven procedures that produced them.

4 Experiments

4.1 Models

As a baseline model we decided to use a bidirectional LSTM encoder-decoder classifier, as in (Zeldes and Liu, 2020). Following their setup, the input to our model consists of EDU pairs, where each pair represents a discourse relation in the dataset. Each EDU is encoded as a sequence of word embeddings along with special tokens. These special symbols – <s> for the satellite, <n> for the nucleus, and <sep> for the separator – are essential for indicating the structural relationship between the two discourse units. An example of a single data item as a tuple of a string list and a label is illustrated in 1

1. **EDU Pair:** `[’<n>’, ’"’, ’Oh’, ’,’, ’she’, ’was’, ’great’, ’"’, ’.’, ’<sep>’, ’They’, ’all’, ’seem’, ’to’, ’know’, ’her’, ’name’, ’<n>’]`

   Label: JOINT-LIST

The EDU pairs are fed into the model using both character and word-level representations. These representations are formed by concatenating fixed 300-dimensional GloVe embeddings (Pennington et al., 2014), pretrained contextualized FLAIR word embeddings and pretrained contextualized FLAIR character embeddings (Akbik et al., 2019). Although the original paper used AllenNLP (Gardner et al., 2018) character embeddings, we opted against using them due to AllenNLP’s transition into maintenance mode. The sequences are then processed to capture contextual dependencies, and a fully connected softmax layer assigns a probability to each discourse relation. We trained and evaluated the model on both coarse and fine-grained set of labels.

For discourse relation classification within the RST framework, BERT-based models have demonstrated strong performance. Gessler et al. 2021 used a BERT-based architecture to classify relations between discourse units, while Kobayashi et al. 2022 highlighted the efficiency of BERT-family models in discourse parsing, benefiting from attention mechanisms to capture complex relational structures. We hypothesize that transformer-based models are well-suited for discourse relation classification because their attention mechanisms can focus on the components within each EDU that hold discourse signaling importance in the text.

In our work, we explore the transformer-based model T5 (Raffel et al., 2020), specifically the T5-small variant due to limited computational resources. Throughout this paper, we use “T5” to refer to T5-small unless otherwise specified. T5 introduces a variant of the transformer that uses relative position biases instead of traditional positional embeddings, which is particularly useful for discourse tasks. The relative positioning of discourse units is essential for understanding their relationships, especially in cases where the connected units are far apart or separated by other discourse structures.

4.2 Evaluation and Error Analysis

The study by Zeldes and Liu 2020, which serves as the foundation for our work, did not primarily focus on discourse relation classification. Consequently, the reported micro-averaged F1-score of 44.37 does not represent state-of-the-art performance at the time. A more relevant benchmark for our study is the third task of the DISRPT 2023 Shared Task (Braud et al., 2023), which involved identifying relation labels between pairs of attached discourse units. The best-performing model in this shared task, the HITS system, achieved an accuracy of 68.19 on RST relation classification for the English GUM corpus (version 9) (Liu et al., 2023a). This model was trained on a set of coarse labels largely identical to ours, except that it did not include the `same-unit` label.

Our bidirectional LSTM baseline performs slightly worse, achieving an accuracy of 64.80. However, our T5 model surpasses the HITS model with a test-set accuracy of 71.16. While these results suggest an improvement over previous work, a direct comparison is not entirely fair, as our models were trained and evaluated on version 10 of the dataset. This newer version includes four additional genres, which likely impact the distribution of discourse relations and the overall characteristics of the data. This expansion likely alters the data distribution and label frequencies, making direct comparisons with prior work less straightforward.

	baseline			T5			
relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**# train**
adversative-antithesis	0.00	0.00	0.00	0.92	0.26	0.40	380
adversative-concession	0.34	0.34	0.34	0.35	0.62	0.45	740
adversative-contrast	0.71	0.07	0.13	0.70	0.38	0.49	421
attribution-negative	0.60	0.30	0.40	0.86	0.67	0.75	82
attribution-positive	0.76	0.74	0.75	0.76	0.87	0.81	1339
causal-cause	0.25	0.33	0.29	0.52	0.44	0.48	549
causal-result	0.50	0.10	0.17	0.73	0.15	0.25	405
context-background	0.32	0.12	0.17	0.38	0.16	0.22	1042
context-circumstance	0.70	0.69	0.70	0.63	0.86	0.73	924
contingency-condition	0.86	0.81	0.83	0.81	0.89	0.84	423
elaboration-additional	0.30	0.69	0.42	0.40	0.59	0.48	2246
elaboration-attribute	0.74	0.91	0.82	0.81	0.94	0.87	2132
evaluation-comment	0.30	0.23	0.26	0.36	0.42	0.39	904
explanation-evidence	1.00	0.24	0.39	0.94	0.34	0.50	662
explanation-justify	0.25	0.04	0.07	0.00	0.00	0.00	477
explanation-motivation	0.00	0.00	0.00	0.00	0.00	0.00	222
joint-disjunction	0.92	0.48	0.63	0.75	0.60	0.67	177
joint-list	0.54	0.59	0.56	0.68	0.71	0.69	2062
joint-other	0.35	0.39	0.37	0.50	0.55	0.52	1125
joint-sequence	0.51	0.71	0.59	0.70	0.62	0.65	1169
mode-manner	0.40	0.15	0.22	0.64	0.27	0.38	255
mode-means	0.75	0.72	0.73	0.84	0.80	0.82	152
organization-heading	0.55	0.42	0.48	0.72	0.83	0.77	369
organization-phatic	0.60	0.72	0.65	0.61	0.92	0.73	688
organization-preparation	0.40	0.29	0.34	0.78	0.35	0.48	658
purpose-attribute	0.40	0.50	0.45	0.79	0.57	0.67	288
purpose-goal	0.59	0.49	0.53	0.75	0.88	0.81	499
restatement-partial	0.43	0.17	0.24	0.38	0.19	0.25	380
restatement-repetition	0.58	0.23	0.33	0.48	0.56	0.52	376
same-unit	0.79	0.71	0.75	0.80	0.92	0.86	1222
topic-question	0.60	0.51	0.55	0.64	0.90	0.74	365
topic-solutionhood	0.00	0.00	0.00	0.00	0.00	0.00	78
weighted avg	0.53	0.52	0.49	0.63	0.62	0.50	22811
macro avg	0.50	0.40	0.41	0.60	0.54	0.54	

Table 2: Relation Classification Results on Fine-grained Labels for the Baseline and T5 Models

In the Table 2 we present the relation classification results with the set of fine-grained labels. We can observe clear differences in performance across discourse relations, largely influenced by their T5 in the training data and the extent to which they are explicitly signaled in text. The T5 model outperforms the baseline across almost all categories, scoring particularly high on relations with clear lexical markers, such as `attribution-positive` (F1 = 0.81), `contingency-condition` (F1 = 0.84), and `same-unit` (F1 = 0.86). Notably, the transformer-based d model achieves an F1-score improvement of over 20% in several categories, including adversative-antithesis, adversative-contrast, attribution-negative, organization-heading, purpose-attribute, and purpose-goal.

The results are strongly influenced by both the amount of training data available for each relation and qualitative factors, such as the presence of discourse markers or lexical cues characteristic of a given relation. For the fine-grained labels, the baseline model exhibits a moderate correlation between the number of training instances and F1-score (p=0.4138), whereas the T5-small model shows a weaker correlation of 0.2650). This suggests that the fine-tuned model is less reliant on the volume of data and may generalize better to relations with limited training examples.

Nonetheless, as shown in Table 2, the T5 model still struggles with certain relations and frequently misclassifies following labels: `explanation-justify` as `elaboration-additional` and `evaluation-comment`; `explanation-motivation` as `elaboration-additional`; `topic-solutionhood` as `topic-question`. These errors likely stem from overlapping lexical and contextual cues, particularly between additive and explanatory functions, as well as the structural similarities (e.g. inquiry-response pattern in both `topic-solutionhood` and `topic-question`). Future work could explore whether similar error patterns occur in human annotation, providing further insight into the inherent ambiguity of these discourse relations. More detailed confusion matrices can be found in Appendix ??.

	baseline			T5			
relation	**P**	**R**	**F1**	**P**	**R**	**F1**	**# train**
adversative	0.44	0.30	0.35	0.61	0.54	0.57	1541
attribution	0.76	0.76	0.76	0.80	0.85	0.82	1421
causal	0.32	0.27	0.29	0.46	0.54	0.50	954
context	0.54	0.46	0.50	0.57	0.53	0.55	1966
contingency	0.82	0.82	0.82	0.85	0.90	0.87	423
elaboration	0.59	0.79	0.68	0.68	0.80	0.74	4378
evaluation	0.27	0.20	0.23	0.35	0.34	0.34	904
explanation	0.48	0.27	0.35	0.46	0.21	0.29	1361
joint	0.83	0.95	0.89	0.90	0.94	0.92	4533
mode	0.55	0.33	0.41	0.75	0.52	0.61	407
organization	0.60	0.66	0.63	0.67	0.69	0.68	1715
purpose	0.90	0.88	0.89	0.92	0.91	0.92	787
restatement	0.32	0.09	0.14	0.57	0.30	0.40	756
same-unit	0.80	0.72	0.76	0.87	0.90	0.89	1222
topic	0.69	0.59	0.64	0.58	0.80	0.67	443
weighted avg	0.63	0.65	0.63	0.70	0.71	0.70	22811
macro avg	0.59	0.54	0.56	0.67	0.65	0.65	

Table 3: Relation Classification Results on Coarse-grained Labels for the Baseline and T5 Models

As we can see in Table 3, performance varies significantly across coarse discourse relations, with T5 generally outperforming the baseline. The model achieves high F1-scores on well-defined categories such as `joint` (F1=0.92), `purpose` (F1=0.92), and `contingency` (F1=0.87). T5 shows the largest performance gains for `restatement`, `mode` and `adversative`, while `explanation` sees a slight drop in F1-score. As with fine-grained labels, the correlation between training data size and F1-score is moderate for the baseline model (p=0.3718) and weaker for T5 (p=0.3197). As previously, some relations remain challenging despite sufficient training instances, likely due to their semantic ambiguity or overlap with other labels. More detailed confusion matrices, provided in Appendix **??**, reveal a few systematic misclassifications.

Additionally, in GUM, certain discourse relations are predominantly implicit (Liu et al., 2023b). For instance, more than 95% of `evaluation` and `restatement` relations lack explicit discourse markers, which may explain their lower classification performance in our experiments. This suggests that models relying heavily on lexical cues may struggle with implicitly signaled relations, highlighting the need for improved contextual modeling. Furthermore, we conducted descriptive statistical analyses of both fine-grained and coarse labels, which are available on GitHub. A potential future direction is to examine correlations between these statistics—such as the percentage of explicit discourse markers per relation—and model performance, which could offer deeper insights into what factors contribute to classification difficulty.

4.3 Evaluation across Genre Clusters

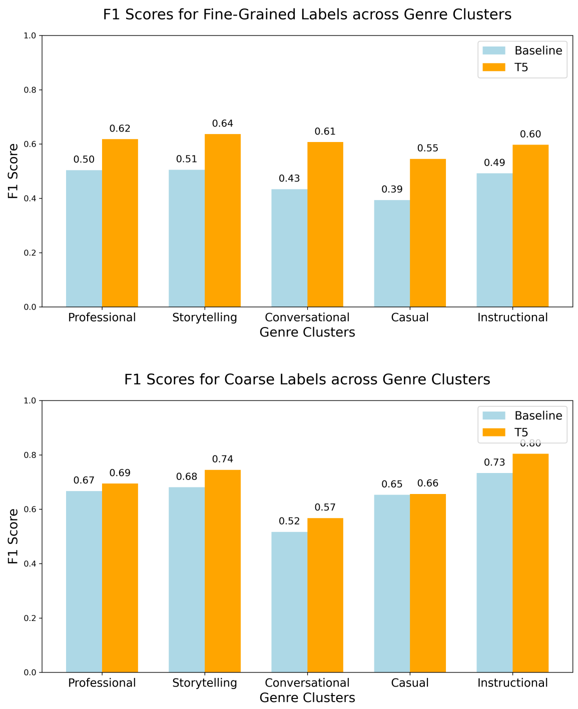

Figure 4: Comparison of Baseline and T5-small Model performance across Genre Clusters

In Figure 4, we compare the performance of the T5 and baseline models across five genre clusters, obtained as described in Section 3.

The T5 model consistently achieves higher scores than the baseline, particularly when fine-grained labels are employed. This performance gap underscores T5's ability to capture nuanced discourse relations, likely due to its transformer architecture that deals better with long-range dependencies and rich contextual information.

A detailed analysis across the clusters reveals that while overall performance differences among genres are moderate, distinct patterns emerge. Notably, the instructional cluster – comprising primarily WikiHow-style content with its clear, step-by-step structure –yields the highest performance for both models, despite having the least amount of data. This result suggests that the inherent regularity and explicit discourse markers in instructional texts provide strong cues for relation classification, enabling robust model performance even under data scarcity.

Conversely, both models exhibit lower performance in conversational and casual clusters, where informal language and less predictable structures introduce greater ambiguity. T5's consistent superiority across all clusters indicates its enhanced capacity to manage such variability, whereas the baseline model appears more sensitive to the structural clarity of the input data.

5 Discussion

In this paper, we presented a comprehensive study of discourse relation classification on a newly expanded version of the GUM corpus (version 10.2.0). We introduced a data-driven approach to cluster genres based on descriptive statistics, which may also contribute to improved genre definition and classification. To the best of our knowledge, this is the first work on discourse relation classification using this newest dataset version.

Our experiments compared a bidirectional LSTM baseline, following the setup by (Zeldes and Liu, 2020) with a transformer-based T5-small model, revealing that while both models are influenced by the amount of training data and lexical cues, T5 consistently achieves higher overall performance and demonstrates better generalization to underrepresented relations. Notably, T5 improved F1-scores by at least 20% for several relations, although categories such as explanation and evaluation remain challenging.

While our findings highlight the potential of transformer-based models for discourse relation classification, several aspects remain open for further exploration. First, we did not conduct extensive hyperparameter tuning for either model; more systematic optimization and the incorporation of additional training data from other datasets could likely enhance performance.

Moreover, the choice of the T5-small variant was driven by limited computational resources, and future studies should examine the impact of larger transformer models. Another promising direction is to analyze genre clusters in more detail, for instance by training models on individual clusters or by training on some clusters and evaluating on others to assess the robustness of genre-specific discourse cues. Additionally, our current evaluation relies on traditional accuracy metrics, which may be too strict given the complexity of discourse relation classification. Using distributions over labels or probabilistic outputs could provide a more nuanced assessment, especially since our error analysis indicates that misclassifications often involve relations with similar characteristics. Finally, it would be valuable to test whether the attention weights of transformer-based classifiers can be interpreted in relation to discourse signals, potentially shedding light on the linguistic factors driving model decisions.

Overall, our work provides new insights into discourse relation classification in a diverse corpus setting and could provide a basis for improving model robustness, investigating genre-specific discourse patterns, and exploring the interpretability of transformer-based models.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Chloé Braud, Yang Janet Liu, Eleni Metheniti, Philippe Muller, Laura Rivière, Attapol Rutherford, and Amir Zeldes. 2023. The DISRPT 2023 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 1–21, Toronto, Canada. The Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. DisCoDisCo at the DISRPT2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2022. A simple and strong baseline for end-to-end neural RST-style discourse parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6725–6737, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35.

Wei Liu, Yi Fan, and Michael Strube. 2023a. HITS at DISRPT 2023: Discourse segmentation, connective detection, and relation classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 43–49, Toronto, Canada. The Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. 2023b. What's hard in english rst parsing? predictive models for error analysis. *Preprint*, arXiv:2309.04940.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2024. eRST: A signaled graph theory of discourse relations and organization. *Preprint*, arXiv:2403.13560.

Amir Zeldes and Yang Liu. 2020. A neural approach to discourse relation signal detection. *arXiv preprint arXiv:2001.02380*.

A Appendix

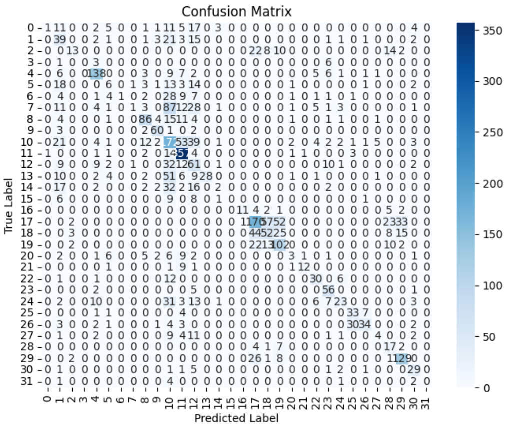

Figure 5: Confusion matrices for the baseline model using fine-grained labels

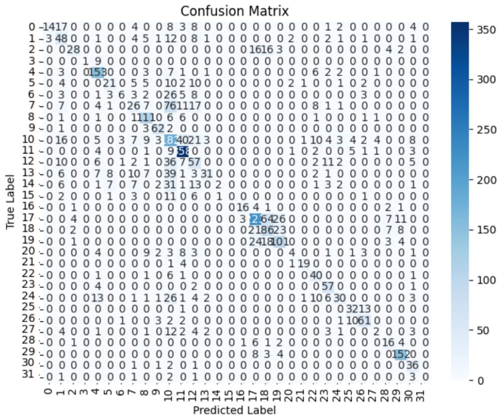

Figure 6: Confusion matrices for the T5 model using fine-grained labels

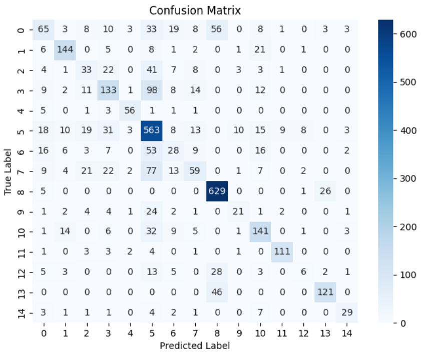

Figure 7: Confusion matrix for the baseline model using coarse labels

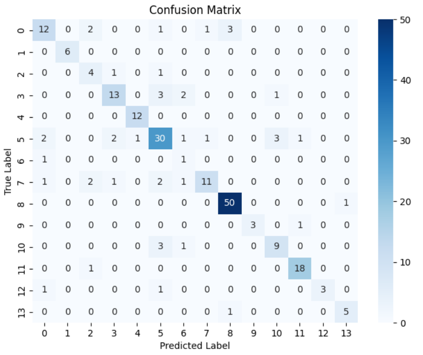

Figure 8: Confusion matrix for the T5-small model using coarse labels