Exploring Soft-Label Training for Implicit Discourse Relation Recognition

Nelson Filipe Costa and **Leila Kosseim**
Computational Linguistics at Concordia (CLaC) Laboratory
Department of Computer Science and Software Engineering
Concordia University, Montréal, Québec, Canada
nelsonfilipe.costa@mail.concordia.ca
leila.kosseim@concordia.ca

Abstract

This paper proposes a classification model for single label implicit discourse relation recognition trained on soft-label distributions. It follows the PDTB 3.0 framework and it was trained and tested on the DiscoGeM corpus, where it achieves an F1-score of 51.38 on third-level sense classification of implicit discourse relations. We argue that training on soft-label distributions allows the model to better discern between more ambiguous discourse relations.

1 Introduction

The Penn Discourse Treebank (PDTB) framework (Miltsakaki et al., 2004; Prasad et al., 2008) defines 36 discourse relation senses organized hierarchically according to three levels of sense granularity (Prasad et al., 2019). Being able to correctly recognize these discourse relations in a text is of great importance for many downstream NLP tasks.

While current explicit discourse relation recognition (EDRR) models can already obtain F1-scores of 90.22 (Xue et al., 2016) when considering the second-level sense, the task of implicit discourse relation recognition (IDRR) remains arguably the hardest task in discourse analysis with state-of-the-art models reaching F1-scores of 55.26 (Liu and Strube, 2023) at the second-level sense. The gap in performance between the two tasks stems from the inherently subjective nature of IDRR, where even trained expert human annotators find it difficult to agree on the sense annotation of implicit discourse relations (Rohde et al., 2016; Hoek et al., 2021).

The difficulty in IDRR is evidenced by the inter-annotator agreement on different corpora. While we do not have access to the inter-annotator agreement of the last version of the PDTB 3.0 corpus (Prasad et al., 2019), the agreement at the third-level sense of PDTB 2.0 (Prasad et al., 2008) was of 80% - which also includes the easier to annotate explicit relations (45.6% of the entire corpus). Moreover, 1,075 (4.93%) of the 21,827 implicit discourse relations on the PDTB 3.0 corpus were annotated with two senses since the annotators could not agree on a single sense. This difficulty is also highlighted in the DiscoGeM corpus (Scholman et al., 2022a), where the inter-annotator agreement at the implicit third-level sense was 60%. However, if we allow implicit relations to convey multiple senses depending on the interpretation of the reader, disagreements do not necessarily indicate inaccuracies in labeling (Aroyo and Welty, 2013; Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022). In fact, it might be helpful in downstream NLP applications to have a distribution of multiple interpretations for ambiguous texts (Basile et al., 2021; Pyatkin et al., 2023).

In this paper, we propose a single label implicit discourse relation recognition model trained on soft-label distributions. The model follows the annotation guidelines of PDTB 3.0 (Prasad et al., 2019) and was trained and tested on the DiscoGeM corpus (Scholman et al., 2022a). We argue that training on soft-label distributions allows the IDRR model to better generalize and discern between the possible multiple interpretations of more ambiguous texts. Our model reaches an F1-score of 51.38 on third-level sense classification of implicit discourse relations in the DiscoGeM corpus (Scholman et al., 2022a) while state-of-the-art IDRR models (Liu and Strube, 2023) achieve an F1-score of 55.26 on second-level sense classification in the PDTB 3.0 corpus (Prasad et al., 2019).

2 Previous Work

In recent years, different models have tried to leverage the power of language models either through fine-tuning (Long and Webber, 2022; Liu and Strube, 2023) or prompt-tuning (Zhao et al., 2023; Chan et al., 2023) to face the challenging task of IDRR. So far, these efforts have relied on the prin-

Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), pages 120–126

ciple that there should be a single sense in the interpretation of implicit discourse relations. However, IDRR is an inherently ambiguous task even for expert human annotators (Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022).

Acknowledging the importance of including sources of ambiguity in human inference in the evaluation of natural language processing tasks led to a recent paradigm shift in discourse annotation. Rather than relying on expert annotators to find a single label for each implicit relation, recent annotation efforts (Yung et al., 2019; Pyatkin et al., 2020; Scholman et al., 2022a,b; Pyatkin et al., 2023) have crowdsourced this task to multiple workers in order to capture the possible multiple interpretations of more ambiguous relations.

The idea that discourse annotation can often be ambiguous is not new (Stede, 2008) and had already been highlighted by Huber et al. (2021) at the nuclear level of the RST framework (Mann and Thompson, 1988). In their work, Huber et al. (2021) proposed a weighted approach to the annotation of nuclearity in discourse relations following the RST framework where, similarly to the PDTB framework, a consensual annotation is hard to obtain (Demberg et al., 2019; Costa et al., 2023).

3 Dataset

In this work we used the DiscoGeM corpus (Scholman et al., 2022a) to train and test our IDRR classification model. The corpus contains 6,505 intersentential implicit discourse relations following the PDTB 3.0 annotation guidelines distributed across three different genres: 2,800 implicit discourse relations in political texts, 3,060 in literary texts and 645 in encyclopedic texts.

Rather than relying on a few trained annotators to find a sense label for each implicit discourse relation, the DiscoGeM corpus crowdsourced the annotation of each relation to multiple participants which allowed to capture a distribution of labels for each relation. Participants were asked to insert a discourse connective between the two arguments of each relation and the authors then inferred the associated sense label from the third-level senses in the PDTB 3.0 (Prasad et al., 2019). Through this method, Scholman et al. (2022a) were able to collect 65,863 annotations from 199 participants for a total of 6,505 implicit discourse relations.

3.1 Data Preparation

We generated two datasets based on the DiscoGeM corpus (Scholman et al., 2022a): one containing the arguments and the sense distribution of each discourse relation and one containing the arguments as well as their context (the adjacent text before and after each argument). We used the arguments (with or without context) as the input of our model and the sense distribution as the target values to calculate the soft cross-entropy loss. Figure 1 shows the character length distribution of both datasets.

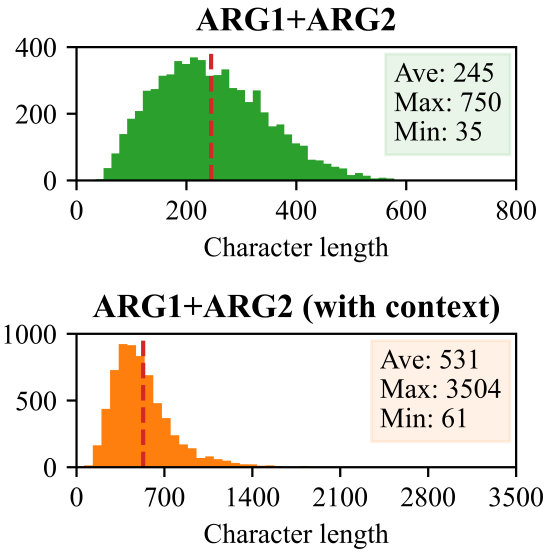

Figure 1: Distribution of character length size of the arguments of the discourse relations in the DiscoGeM corpus with and without additional textual context.

The dataset containing only the arguments (ARG1+ARG2) has an average length of 245 characters and the dataset including the context of the arguments (ARG1+ARG2 with context) has an average length of 531 characters. To ensure a balanced distribution of senses in the training and evaluation of our model, we determined the sense with the highest score for each discourse relation and then split both datasets equally while preserving the same distribution of majority-senses in training and testing. Figure 2 shows the majority-sense distribution of both datasets, after splitting 80% (5,204) of the 6,505 implicit discourse relations for training and 20% (1,301) for testing.

Note that the DiscoGeM corpus (Scholman et al., 2022a) was annotated only with 27 of the 36 third-level senses in the PDTB 3.0 (Prasad et al., 2019). The BELIEF and SPEECHACT senses were not included in the annotation process. However, as Fig-

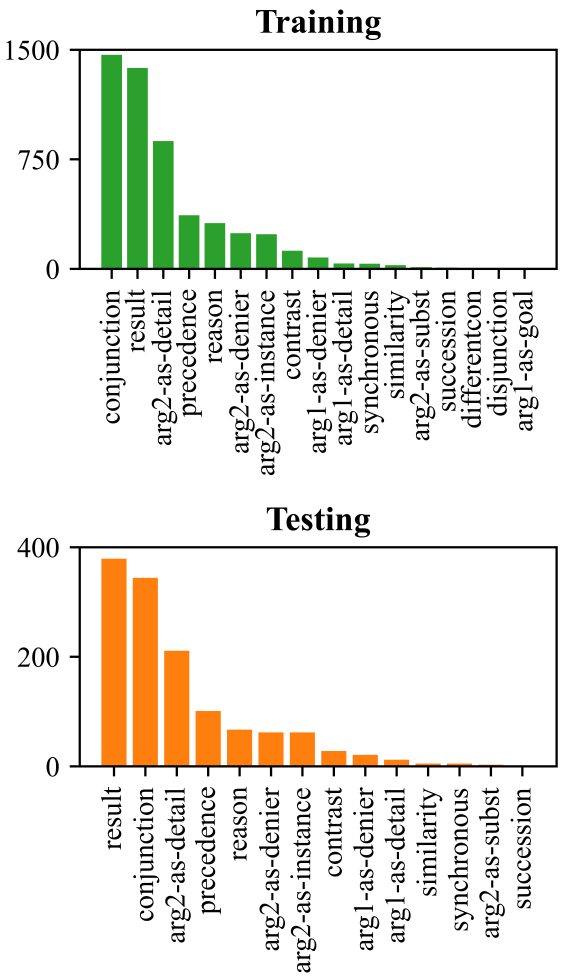

Figure 2: Distribution of the majority-sense labels in the training and testing splits of both our datasets.

ure 2 shows, not all of the 27 senses occurred in the annotated texts.

4 Classification Model

Similarly to the current state-of-the-art model in IDRR (Liu and Strube, 2023), we based our classification model on the bidirectional RoBERTa-base (Liu et al., 2019) language model. We fine-tuned the sequence classification model from Hugging Face[1] with a single classification layer using a soft cross-entropy loss with a mean reduction over batches to allow training with soft-label distributions and we optimized our model using the Adam method (Kingma and Ba, 2015). We then inferred the single label sense of each discourse relation at the evaluation stage from the element with the highest score at the output of the model. All of the code used in this paper can be found on GitHub[2].

4.1 Fine-Tuning

To optimize our model for the present task, we conducted a series of experiments with different hyperparameters to determine the configuration which yielded better results. We did not, however, experiment with different values for the beta terms in the Adam optimizer. Instead, we used the recommended values for fine-tuning RoBERTa (Liu et al., 2019): $\beta_1 = 0.9$ and $\beta_2 = 0.98$. Table 1 shows the impact of training the model with different epochs (EP) and batch sizes (BS), while keeping a constant learning rate ($\gamma = 1e^{-5}$) and no decay ($\lambda = 0$). In these experiments we considered only the dataset made of the arguments of the discourse relations (see ARG1+ARG2 in Figure 1).

Hyperparameters	F1	Precision	Recall
EP: 10 / BS: 16	49.98	49.55	51.35
EP: 10 / BS: 32	50.91	50.62	51.58
EP: 10 / BS: 64	**51.38**	**51.54**	**52.19**
EP: 20 / BS: 64	50.59	50.67	51.04

Table 1: Evaluation of our model with different epochs (EP) and batch sizes (BS), while keeping a constant learning rate ($\gamma = 1e^{-5}$) and no decay ($\lambda = 0$).

The values highlighted in bold in Table 1 show the best configuration on the test split: $EP = 10$ and $BS = 64$. For smaller batch sizes and higher epochs, the model performed better in training but worst in testing. Given the relatively small dataset, these configurations might have been more prone to over-fitting. Keeping the optimal number of epochs and batch size, in Table 2 we studied the influence of different learning rates (γ) and the impact of introducing a linear decay (λ) in the performance of the model.

The values highlighted in bold in Table 2 show the best configuration on the test split: $\gamma = 1e^{-5}$ and $\lambda = 0$. Similarly to the number of epochs and batch sizes, higher learning rates led to better results in training but worst in testing. The same phenomenon occurred with the introduction of the linear decay rate. This hints at the susceptibility of the model to over-fitting and emphasizes the importance of carefully selecting hyperparameters to ensure better generalization.

[1] https://huggingface.co/docs/transformers/model_doc/roberta

[2] https://github.com/CLaC-Lab/Implicit-Discourse-Relation-Recognition

Hyperparameters	F1	Precision	Recall
γ: 5e^{-5} / λ: 0.0	48.77	49.91	49.42
γ: 2e^{-5} / λ: 0.0	49.43	49.64	50.88
γ: 1e^{-5} / λ: 0.0	**51.38**	**51.54**	**52.19**
γ: 1e^{-5} / λ: 0.1	49.67	50.78	50.73

Table 2: Evaluation of our model with different learning rates (γ) and with decay (λ), for 10 epochs and a batch size of 64.

5 Results and Analysis

Having selected the optimal hyperparameter configuration ($EP = 10$, $BS = 64$, $\gamma = 1e^{-5}$ and $\lambda = 0$), we applied our classification model to the task of IDRR under two different settings. In the first setting we considered only the arguments of the discourse relations as input to our model, while in the second setting we also took into consideration their adjacent textual context. In both settings, the model outputs a soft-label distribution over the possible third-level senses in the PDTB 3.0 (Prasad et al., 2019), from which the sense with the highest score is selected and evaluated against the respective majority-sense from the DiscoGeM corpus (Scholman et al., 2022a). Table 3 presents the results of our model under both settings.

Input	F1	Precision	Recall
ARG1+ARG2	**51.38**	**51.54**	**52.19**
ARG1+ARG2 (with context)	43.67	43.22	45.43

Table 3: Results of third-level sense classification of implicit discourse relations considering the arguments without and with additional textual context.

As indicated in Section 3.1, our model is based on the RoBERTa (Liu et al., 2019) language model, whose maximum input length size is 512. However, the average length of the input with context is 531 characters, while the average length of the input without context is 245 characters (see Figure 1). The results in Table 3 indicate that the extra contextual information gain does not outweigh the information lost to truncation, as we obtain higher scores on all metrics for the shorter inputs without context. We include the confusion matrix of the output of our model without context in Table 4 of Appendix A.

Although we did not test our model directly on the PDTB 3.0 corpus (Prasad et al., 2019), our results suggest the benefits of training IDRR classification models on soft-label distributions. Our model obtained an F1-score of 51.38 on a subset of the DiscoGeM corpus (Scholman et al., 2022a), while the current best model in IDRR (Liu and Strube, 2023) obtained an F1-score of 55.26 on a subset of the PDTB 3.0 corpus (Prasad et al., 2019). In their work, Pyatkin et al. (2023) obtained an accuracy of 41% on a subset of the PDTB 3.0 corpus when training their model on the union of the DiscoGeM and the QADiscourse (Pyatkin et al., 2020) corpora.

6 Conclusion

In this paper we proposed a single label implicit discourse relation recognition model trained on soft-label distributions from the DiscoGeM corpus and evaluated it on single label classification to allow an easier comparison against existing state-of-the-art IDRR models. We obtained an F1-score of 51.38 on third-level sense classification of implicit discourse relations on the DiscoGeM corpus following the PDTB 3.0 annotation guidelines. Our results hint at the possible benefits of training IDRR classification models on soft-label distributions to help generalize and discern between possible multiple interpretations of ambiguous texts.

7 Limitations and Future Work

In this work we trained and evaluated our model using only the DiscoGeM corpus. Although the training was done using soft-labels, the evaluation considered only single labels. We would now like to evaluate the performance of our model also on the soft-label prediction task itself using soft evaluation metrics. In addition, since most state-of-the-art IDRR models are trained and evaluated on the PDTB 3.0 corpus, we would also like to evaluate the performance of our model on single label classification using the PDTB 3.0 corpus. This would allow us to draw a direct comparison between our approach and other existing IDRR models.

Finally, our proposed classification model consists of a rather simple configuration of the RoBERTa-base model with a sequential classification layer on top. In future work, we would like to explore more elaborate model configurations. We would also like to train our model on the traditional single label IDRR classification task and use it as a baseline to evaluate the true potential of training our model on soft-labels.

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback on the previous version of this paper. They would also like to thank the Linguistics Data Consortium (LDC) for providing access to the PDTB 3.0 corpus and the authors of the DiscoGeM corpus for providing free access to their corpus. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Lora Aroyo and Chris Welty. 2013. Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of the 5th Annual Association for Computing Machinery Web Science Conference (WebSci'13)*, Paris, France. Association for Computing Machinery (ACM).

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We Need to Consider Disagreement in Evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics (ACL).

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023. DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 35–57, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

Nelson Filipe Costa, Nadia Sheikh, and Leila Kosseim. 2023. Mapping Explicit and Implicit Discourse Relations between the RST-DT and the PDTB 3.0. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (RANLP'23)*, pages 344–352, Varna, Bulgaria.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1):87–135.

Jet Hoek, Merel C.J. Scholman, and Ted J.M. Sanders. 2021. Is there less annotator agreement when the discourse relation is underspecified? In *Proceedings of the 1st Workshop on Integrating Perspectives on Discourse Annotation*, pages 1–6, Tübingen, Germany. Association for Computational Linguistics (ACL).

Patrick Huber, Wen Xiao, and Giuseppe Carenini. 2021. W-RST: Towards a Weighted RST-style Discourse Framework. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21)*, pages 3908–3918, Online. Association for Computational Linguistics (ACL).

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics (TACL)*, 10:1357–1374.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*, pages 1–15, San Diego, California, USA.

Wei Liu and Michael Strube. 2023. Annotation-Inspired Implicit Discourse Relation Classification with Auxiliary Discourse Connective Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 15696–15712, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

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.

Wanqiu Long and Bonnie Webber. 2022. Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP'22)*, pages 10704–10716, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics (ACL).

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.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7:677–694.

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)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05. Web Download. Philadelphia: Linguistic Data Consortium.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*, pages 2804–2819, Online. Association for Computational Linguistics.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1014–1032.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task. In *Proceedings of the 10th Linguistic Annotation Workshop (LAW'16)*, pages 49–58, Berlin, Germany. Association for Computational Linguistics.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2022a. DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 3281–3290, Marseille, France. European Language Resources Association (ELRA).

Merel Scholman, Valentina Pyatkin, Frances Yung, Ido Dagan, Reut Tsarfaty, and Vera Demberg. 2022b. Design Choices in Crowdsourcing Discourse Relation Annotations: The Effect of Worker Selection and Training. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22)*, pages 2148–2156, Marseille, France. European Language Resources Association (ELRA).

Manfred Stede. 2008. Disambiguating Rhetorical Structure. *Research on Language and Computation*, 6(3):311–332.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics (ACL).

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing Discourse Relation Annotations by a Two-Step Connective Insertion Task. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 16–25, Florence, Italy. Association for Computational Linguistics (ACL).

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL'23)*, pages 6477–6492, Toronto, Ontario, Canada. Association for Computational Linguistics (ACL).

A Appendix - Confusion Matrix

Targets \ Predictions	Synchronous	Precedence	Succession	Reason	Result	ARG1-as-Cond	ARG2-as-Cond	ARG1-as-NegCond	ARG2-as-NegCond	ARG1-as-Goal	ARG2-as-Goal	ARG1-as-Denier	ARG2-as-Denier	Contrast	Similarity	Conjunction	Disjunction	ARG1-as-Instance	ARG2-as-Instance	ARG1-as-Detail	ARG2-as-Detail	Equivalence	ARG1-as-Manner	ARG2-as-Manner	ARG1-as-Exception	ARG2-as-Exception	ARG2-as-Substitution	Different-Conn	NoRel
Synchronous	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0
Precedence	2	66	0	1	20	0	0	0	0	0	0	0	0	0	0	13	0	0	0	0	8	0	0	0	0	0	0	0	0
Succession	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Reason	0	0	0	48	11	0	0	0	0	0	0	0	2	3	0	12	0	0	1	0	12	0	0	0	0	0	0	0	0
Result	0	22	0	20	245	0	0	0	0	0	0	0	0	23	2	0	63	0	0	8	0	31	0	0	0	0	0	0	0
ARG1-as-Cond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Cond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-NegCond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-NegCond	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-Goal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Goal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-Denier	0	0	0	3	2	0	0	0	0	0	0	2	3	0	0	6	0	0	0	4	0	0	0	0	0	0	0	0	0
ARG2-as-Denier	0	0	0	1	11	0	0	0	0	0	0	1	19	1	0	12	0	0	1	0	3	0	0	0	0	0	0	0	0
Contrast	0	1	0	6	4	0	0	0	0	0	0	1	2	3	0	10	0	0	0	0	1	0	0	0	0	0	0	0	0
Similarity	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	1	0	0	0	0	0	0	0	0
Conjunction	0	17	0	24	46	0	0	0	0	0	0	3	11	1	0	187	0	0	5	0	31	0	0	0	0	0	0	0	0
Disjunction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-Instance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Instance	0	2	0	6	4	0	0	0	0	0	0	0	0	0	0	11	0	0	21	0	11	0	0	0	0	0	0	0	0
ARG1-as-Detail	0	0	0	3	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
ARG2-as-Detail	0	1	0	21	16	0	0	0	0	0	0	1	1	0	0	37	0	0	5	0	88	0	0	0	0	0	0	0	0
Equivalence	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-Manner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Manner	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG1-as-Exception	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Exception	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ARG2-as-Substitution	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Different-Conn	0	1	0	2	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0
NoRel	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: Confusion matrix for the majority third-level sense classification of implicit discourse relations considering only the arguments without the context of the relation as input. Color gradients are calculated at the target level (row-wise).