DiscoFlan: Instruction Fine-tuning and Refined Text Generation for Discourse Relation Label Classification

Kaveri Anuranjana
Language Science and Technology
Saarland University
kaveri@coli.uni-saarland.de

Abstract

This paper introduces DiscoFlan, our system for the DISRPT 2023 shared task on discourse relation classification. We leverage recent advances in NLP finetuning and use Flan-T5 as a multilingual discourse relation classifier. Our model uses multilingual instructional prompts to finetune on datasets from different languages and generate relation labels as classification outputs. The model's hyperparameters are tuned to enable efficient label generation by finetuning on low-resource datasets. Moreover, we introduce a post-processing step to tackle the problem of label mismatches caused by the generative nature of a seq2seq model by using the label distribution. In contrast to the previous state-of-the-art model, our approach eliminates the need for hand-crafted features in computing the discourse relation classes. Overall, DiscoFlan showcases how instruction finetuning can perform multilingual discourse relation classification for the DISRPT 2023 discourse relation classification shared task.

1 Introduction

Discourse Relation Classification (DRC) is a discourse-level task that requires the identification of discourse relations between text segments in a document. This low-resourced task contains multiple subtasks with different languages and formalisms. The numbers of unique labels vary from 9 in zho.pdtb.cdtb to 33 in nld.rst.nldt.

We train DiscoFlan and compare it with the current state-of-the-art model as well as a multilingual classification baseline on the DISRPT datasets and present our results for the 2023 DISRPT sharedtask for Discourse Relation Classification.

Supervised Large language models (LLMs) trained with human labels are truly a paradigm shift due to their zero-shot and low-resource capabilities. Improved language representation mechanisms and utilization of large pre-training corpora of LLMs have led to significant advancements in two key areas: zero-shot capabilities and low-resource prompt learning. In this paper, we focus on low-resource DRC. Such breakthroughs are a testament to the power of large-scale pretraining. With enhanced representations, language models can generalize and transfer knowledge across different tasks and domains, enabling impressive zero-shot capabilities where models can perform well on tasks they were not explicitly trained on. These improved zero-shot learners have the capability to learn efficiently on low-resource complex tasks like Relation Classification. The availability of even limited amounts of training data allows for effective low-resource prompt learning. These advancements highlight the immense potential of LLMs and their ability to tackle the real-world problem of DRC.

Our main contributions are: **1.** We perform instruction finetuning of multilingual prompts with DiscoFlan for DRC tasks of different formalisms and languages to develop a seq2seq generative label classification system. **2.** We use a simple post-processing stage harnessing the label distributions using majority label distributions for low resource dataset.[1]

2 Related Work

2.1 Instruction Finetuning

LLMs, such as InstructGPT (Ouyang et al., 2022), ChatGPT, FLAN-T5-XXL(13B) (Chung et al., 2022), LLaMA (Touvron et al., 2023) have revolutionized natural language processing. Fine-tuning, a process of training these models on specific tasks enhances their performance and is typically used for low-resource classification where creating large annotated datasets can be difficult resulting in small dataset sizes. Instruction fine-tuning leverages the models' powerful representations and contextual

[1]We release our code here: https://github.com/erzaliator/DiscoFlan

Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023), pages 22–28

understanding to achieve superior accuracy and efficiency in a wide range of NLP tasks. This approach enables adaptation of large language models to suit specific applications, making them valuable tools for natural language understanding and generation.

FlanT5 (Chung et al., 2022) is a generative LLM that has gained significant attention in the field of natural language processing (NLP). It is based on the T5 (Text-To-Text Transfer Transformer) architecture (Kale and Rastogi, 2020) and is pre-trained on a massive corpus of text data. FlanT5, demonstrates strong multi task generalization capabilities through the training paradigm of instruction finetuning, a process that involves further training the base model on specific NLP tasks with task-specific data and instructions. By providing explicit instructions during the finetuning phase, the model's underlying representations and contextual understanding can be harnessed to achieve superior performance in various NLP applications.

Motivated by the strong NLU capabilities of FlanT5 (Chung et al., 2022), similarly, we finetune FlanT5 to learn discourse relation classification by posing it as a seq2seq generative task. For the respective DISRPT discourse relation datasets, the model is tasked with generating sequences that correspond to the discourse relations from that dataset. We perform instruction finetuning on each individual dataset by using a suitable prompt template to the sentence pairs to harness the structured prompt input format that FlanT5 is pre-trained on for multi-task reasoning.

Language	Prompt
Chinese	"sent1 和 sent2 之间的话语关系是什么： sent1: 该公司报告了 2023 年第三季度的最高利润 sent2: 近期公司市值的增加对市场情绪产生了积极影响。"
English	"what discourse relation holds between sent1 and sent2: sent1: The company is reporting the highest profits for Q3 2023. sent2: The recent increase in the company's market cap has impacted market sentiments positively."
Italian	"Quale relazione discorsiva c'è tra sent1 e sent2: sent1: La società sta riportando i profitti più alti per il terzo trimestre del 2023. sent2: Il recente aumento della capitalizzazione di mercato della società ha avuto un impatto positivo sui sentimenti di mercato."

Figure 1: Prompt template for DRC in different languages for instruction finetuning. We translate the prompt across datasets.

2.2 Multilingual Discourse Classification

Discourse relations refer to the connections and dependencies between different parts of a text that contribute to its overall coherence and meaning. Various annotation frameworks have been proposed for the task of DRC such as RST(Carlson et al., 2002), PDTB(Prasad et al., 2008) and SDRT(Lascarides and Asher, 2007) among others. The Discourse Relation Parsing and Treebanking (DISRPT) provides DRC datasets across various languages and formalisms in the form of a sentence pair classification task (Zeldes et al., 2021).

Kurfalı and Östling (2019) applied cross-lingual transfer learning on the DRC task but only evaluated in a zero-shot setting. Their results were considerably below the state-of-the-art system. DiscoDisco (Gessler et al., 2021) obtains the state-of-the-art performance by using hand-crafted features to descibe the discourse segments and training with individual checkpoints for each language.

3 Methodology

3.1 Modelling Classification as a Refined Label Generation task

With instruction finetuning, the model learns to generate discourse labels given a prompt encoding the input sentence pair. The decoded output is passed through a refinement stage which ensures that mismatches are removed based on the dataset label distribution.

3.1.1 Generating labels using seq2seq model

Classification is typically performed using AutoEncoder models which are pre-trained on data-denoising objectives such as Mask Language Modelling. These models are finetuned for classification with a final prediction layer (Jin et al., 2020) to learn label representations from the model's hidden representations. FlanT5 (Chung et al., 2022) is an EncoderDecoder model which was trained with Instruction Finetuning objective. On the other hand, current-state-of-the-art Discourse Relation Classifiers are AutoEncoder based architectures (Gessler et al., 2021; Jiang et al., 2022) which use a prediction layer to encode the labels as discrete categories.

Instruction-based prompts: The input is formulated as an instruction-based prompt to utilize the Instruction Finetuning capabilities of FlanT5. It is modified as shown in Figure 1.

DiscoFlan: The FlanT5 decoder is adopted to generate discourse relation labels. The model is called DiscoFlan as the decoder hyperparameters are adjusted to generate short sequenced labels and the model is finetuned on DRC datasets.

Finetuning: The model is trained with standard loss for training EncoderDecoder models. During training, through a conditional generation cross entropy loss meant for sequence generation, the model learns to generate a sequence corresponding to the relation label rather than a typical classification cross entropy loss used to learn categorical representations.

Generating Discourse Relation labels: During inference, the decoder's output tokens are used for generating discourse relation labels. This has the added benefit of utilizing the task representations of the FlanT5 to generate sequences grounded in real-world knowledge to incorporate label meaning. This grounds the meaning of the relation labels to the model's generative space which is significantly richer than using one-hot representations. (Yung et al., 2022) also note that using one-hot encoding for label representation ignores the inherent ambiguity of discourse relation labels.

We can investigate the effect of using special numeric tokens for classification, however, we leave that to future work.

3.1.2 Refinement logic

While working with discourse relation classifiers it is empirically observed that relation classification models are prone to a large number of false predictions of the majority label. Additionally, due to a lack of training data which is generally the case for DRC datasets, DiscoFlan generates substantial mismatches. These mismatches can be partial or complete. In order to alleviate the issue of mismatches and to construct a system submission for the shared task we propose a processing step after the label generation stage. This allows the model's generated outputs to be refined to suit the shared task's analysis criterion i.e. the outputs always belong to the set of dataset labels.

Mismatches are strings that do not belong to the label space. For example - The RST label *elaboration* being incorrectly generated as *elab* by the model. While developing DiscoFlan it was observed that a significant portion of the mismatches were of the form - *elaboration of, elaborated, elaborating*. Hence, a simple post-processing stage is used to refine the decoder's generated sequences during evaluation. After removing the noisy affixes (such as *"-er", "-ed"*, etc.), the remaining lemma is matched against the training set's labels (for out-of-domain datasets, the validation set labels are used). The label matching with the lemma is used as the final output.

When the prediction lemma does not belong to the label space, it is replaced with the majority label of the training dataset. Figure 2 provides such an example.

This modification is denoted as **DiscoFlan+Ref**.

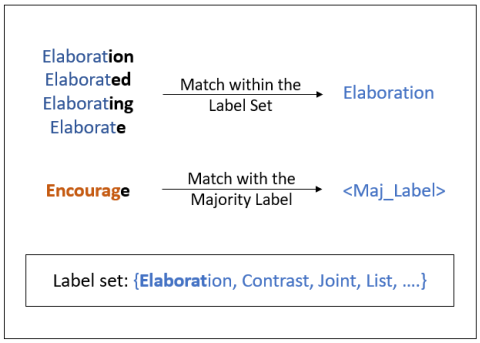

Figure 2: Refinement stage for the decoded output implemented for DiscoFlan+Ref.

3.2 Baseline (Xlm-R classifier)

DiscoFlan is a multilingual model whereas previous models for DRC have been monolingual with separate language models for each dataset. In order to assess the impact of using multilingual representations, we also compare our results with **Xlm-R** (Conneau et al., 2020) model for multilingual representations which has been shown to perform well on NLI tasks across a diverse set of languages.

3.2.1 Training Setup

The performances of DiscoFlan and other variants are assessed for Discourse Relation Classification using the weights provided by HuggingFace library [2]. Due to practical considerations, FlanT5-small is used (specifically, google/flan-t5-small is used as the model type). The FlanT5-small model consists of significantly lesser parameters as compared to FlanT5-base.

For comparability, the same hyperparameters are kept for all models across all language pairs. A batch size of 16 is used for all runs. The models are trained for 50 epochs with an Early Stopping patience of 12 calls. 5 or 10 epochs are used to

[2] https://huggingface.co/docs/transformers/model_doc/flan-t5

train the larger datasets. Details of the epoch hyperparameter can be found in our code. The smaller datasets are trained with a high learning rate, $1e-3$ while the larger datasets use a smaller learning rate of $1e-5$.

The huggingface Transformer and Pytorch library are used. Each instance of a model is run on a 32 GB Nvidia Tesla V100 GPU card.

3.2.2 Model Setup

It is noted that raw generations are sensitive to the model parameters - max generation length, min generation length. Figure 4 shows the average label length for the datasets. The average length varies from 8 to 22 characters within the shared task.

The minimum generation length and maximum generation length are set on a per-dataset basis. Readers are suggested to refer to our code to obtain these values for each dataset.

Reducing the beam width improves the quality of generations. A smaller beam width means that the model only considers a limited number of candidates at each decoding step. When generating small text, such as labels for classification tasks, small beam width is suitable. Additionally, smaller beam width leads to faster model convergence as the generation will favour a specific set of candidates early on. A beam width of 4 is used.

4 Results

4.1 Learning seq2seq representations

We train DiscoFlan on the DISRPT datasets and present our results for the 2023 DISRPT sharedtask for Discourse Relation Classification. Firstly, we make predictions using raw generated tokens. The results are presented in Table 1 (column DiscoFlan). Secondly, we apply simple refinement logic to exploit the distribution of discourse labels. The results are also presented in Table 1 (column DiscoFlan+Ref).

5 Analysis

5.1 Refinement improves low resource relation classification

Table 1 shows how the refinement logic helps the model to infer better labels. We find that supervision alone is not enough to produce good labels. Many of the labels that the model generates are not in the label space. We fix this by replacing them with the most common label prediction. This improves the model performance for all the datasets. The 2023 DISRPT sharedtask adds 11 more datasets to the 15 datasets that the previous best models used. DiscoFlan+Ref does not use hand crafted features, but it is close to the best model for fas.rst.rpstc. We note that the model often overfits on one label. This means that we need to improve the instruction fine-tuning, because just using the text and a suitable loss function is not sufficient.

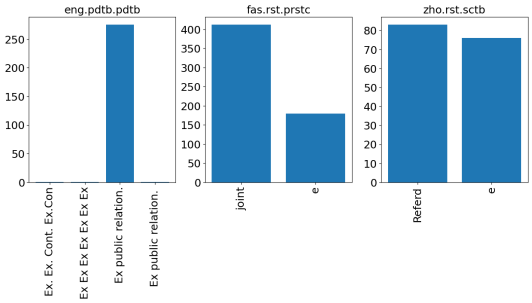

Figure 3: Labels predicted by DiscoFlan for datasets eng.pdtb.pdtb, fas.rst.prstc, zho.rst.sctb

Additionally, **DiscoFlan+Ref** outperforms the **DiscoFlan** model for low-resource datasets as it resolves mismatches. This is due to the fact that language generation requires a large amount of data for finetuning. Mismatches are also subsequently higher for generative models. Refinement addresses the issue of complete and partial mismatching caused to generation issues.

This highlights that weak label learners can be augmented with simple distributional logic to improve model classification. The column "Probs" denotes the accuracy achieved by always predicting the majority label. This chance probability bounds the gains that can achieved by **DiscoFlan+Ref**.

Figure 3 shows the labels produced for three sample dataset for **DiscoFlan**. In the case for complex labels like eng.pdtb.pdtb, the model is prone to generating out-of-vocabulary labels. Where the labels are not significantly complex, the model learns to overfit on a single label.

Note that Xlm-R and DiscoDisco [3] are also prone to majority label generation.

6 Conclusion and Future Work

In conclusion, our paper introduces DiscoFlan, a multilingual discourse relation classifier submitted for the DISRPT 2023 shared task. We addressed

[3]The numbers for DiscoDisco reported in Table 1 are taken from the paper

Corpus	DD w/ feats.	DD w/o feats.	DiscoFlan	DiscoFlan+Ref	Baseline	Probs
deu.rst.pcc	39.23	33.85	0.00	13.08	15.51	9.70
eng.dep.covdtb	na	na	0.00	50.15	na	50.25*
eng.dep.scidtb	na	na	0.00	34.12	na	34.59
eng.pdtb.pdtb	74.44	75.63	0.00	24.41	66.95	27.92
eng.pdtb.tedm	na	na	0.00	33.05	na	29.6*
eng.rst.gum	66.76	62.65	0.00	25.39	53.07	21.86
eng.rst.rstdt	67.1	66.45	0.00	36.94	62.47	40.33
eng.sdrt.stac	65.03	59.67	0.00	22.65	43.4	23.74
eus.rst.ert	60.62	59.59	7.96	28.61	22.74	21.4
fas.rst.prstc	52.53	51.18	19.59	45.44	34.01	23.78
fra.sdrt.annodis	46.4	48.32	19.36	19.36	33.12	20.50
ita.pdtb.luna	na	na	0.00	22.37	na	22.3
nld.rst.nldt	55.21	52.15	0.00	35.08	33.84	26.43
por.pdtb.crpc	na	na	7.93	43.83	na	32.1
por.pdtb.tedm	na	na	0.00	29.95	na	25.1*
por.rst.cstn	64.34	67.28	0.36	35.29	58.7	27.74
rus.rst.rrt	66.44	65.46	0.00	23.60	58.05	23.53
spa.rst.rststb	54.23	54.23	5.86	26.76	31.53	20.17
spa.rst.sctb	66.04	61.01	0.00	44.65	46.12	34.16
tha.pdtb.tdtb	na	na	0.00	19.35	na	23.03
tur.pdtb.tdb	60.09	57.58	36.49	36.49	35.23	25.05
tur.pdtb.tedm	na	na	35.71	35.71	na	27.10*
zho.dep.scidtb	na	na	29.00	33.49	48.72	30.92
zho.rst.gcdt	na	na	59.36	59.37	na	18.93
zho.rst.sctb	64.15	64.15	0.00	20.46	47.92	33.25
zho.pdtb.cdtb	86.49	87.34	0.00	43.40	na	66.01

Table 1: Comparing results of Relation Classification results against Xlm-R baseline and state-of-the-art DisCoDisCo (DD) Gessler et al. (2021) in terms of accuracy. We report the accuracy from the DISRPT 2023 sharedtask for DiscoFlan+Ref. Using the released test set and metric, we also report the accuracy for DiscoFlan. Improved numbers are denoted in bold. Accuracy of the current year's new shared task datasets are underlined where model outperforms chance probability.

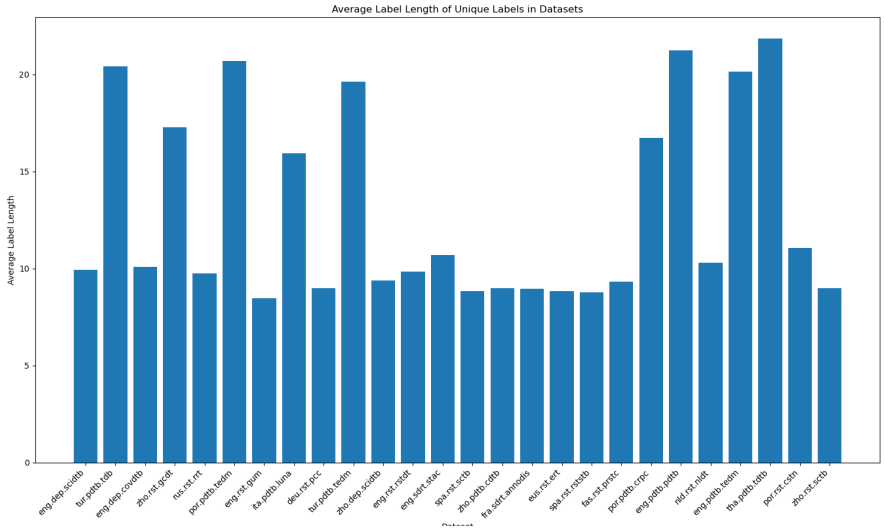

Figure 4: Average unique Label sequence length

the challenge of mismatched in seq2seq models by leveraging label distribution information for label generation.

Our approach eliminates the need for handcrafted features and introduces a novel label generation mechanism that anchors the labels to a fixed set. Emperical results demostrate promising results for DiscoFlan+Ref as well as DiscoFlan compared to the state-of-the-art model and a multilingual baseline.

We analyzed the limitations of multilingual models as weak learners and showed that larger models with richer pre-training objectives, in the form of instruction fine-tuning, yield more meaningful representations.

Post-processing refinement logic improves low-resource relation classification, as evidenced by the consistent outperformance of DiscoFlan+Ref over the baseline model. It addresses issues of mismatches caused by generation problems, leading to enhanced classification accuracy. Our findings highlight the potential of augmenting weak label learners with distributional logic to improve model classification. DiscoFlan showcases instruction finetuning for multilingual discourse relation classification for the DISRPT 2023 shared task and provides valuable insights for future research in this area.

We recognize the potential of larger models to improve prediction quality; however, due to constraints in terms of resources and time, we were unable to test the performance of Flan-T5-Large in our study. Furthermore, we acknowledge that further advancements in decoding strategies and imporved prompts have the potential to enhance label representations and generation. In our future work, we intend to explore these topics to enhance our current models.

Limitations

While using the majority label solves the problem of handing out-of-vocabulary labels during fine-tuning, we acknowledge that label refinement method relies on the majority label. This makes a strong assumption about our dataset bias, namely, that the majority label outnumbers the rest of the labels significantly to impact accuracy. Hence, this method may not be applicable to well-balanced datasets.

We also note that simply predicting the majority label is simple method of label prediction which does not generalized to new unseen datasets. Improving label prediction by enriching datasets manually or automatically might make the task more representative of natural data.

Using larger models can improve model prediction however due to time and machine constraints we leave the evaluation using FlanT5-large for future work.

Ethics Statement

We note that low resource classifiers are prone to overfitting. We encourage users to thoroughly analyse the predicted labels before using our provided

models. No other ethical considerations need to be made regarding the data and models.

Acknowledgements

This work is supported by the German Research Foundation (DFG) under Grant SFB 1102 ("Information Density and Linguistic Encoding", Project-ID 232722074). We thank Prof. Vera Demberg, Amir Zeldes, Chloe Braud and Laura Riviere for their valuable time and suggestions in improving our submission. We also thank the DISRPT 2023 organisers for their assistance and feedback on our system submission.

References

Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. 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.

Yuxin Jiang, Linhan Zhang, and Wei Wang. 2022. Global and local hierarchy-aware contrastive framework for implicit discourse relation recognition. *arXiv preprint arXiv:2211.13873*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Murathan Kurfalı and Robert Östling. 2019. Zero-shot transfer for implicit discourse relation classification. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 226–231, Stockholm, Sweden. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. *Computing meaning*, pages 87–124.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Frances Yung, Kaveri Anuranjana, Merel Scholman, and Vera Demberg. 2022. Label distributions help implicit discourse relation classification. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 48–53.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. 2021. The disrpt 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12.