DisCut and DiscReT: MELODI at DISRPT 2023

[1]**Eleni Metheniti** and [1,2,3]**Chloé Braud** and [1,3]**Philippe Muller** and [1]**Laura Rivière**
[1]UT3 - IRIT ; [2]CNRS ; [3]ANITI
firstname.lastname@irit.fr

Abstract

This paper presents the results obtained by the MELODI team for the three tasks proposed within the DISRPT 2023 shared task on discourse: segmentation, connective identification, and relation classification. The competition involves corpora in various languages in several underlying frameworks, and proposes two tracks depending on the presence or not of annotations of sentence boundaries and syntactic information. For these three tasks, we rely on a transformer-based architecture, and investigate several optimizations of the models, including hyper-parameter search and layer freezing. For discourse relations, we also explore the use of adapters—a lightweight solution for model fine-tuning—and introduce relation mappings to partially deal with the label set explosion we are facing within the setting of the shared task in a multi-corpus perspective. In the end, we propose one single architecture for segmentation and connectives, based on XLM-RoBERTa large, freezed at lower layers, with new state-of-the-art results for segmentation, and we propose 3 different models for relations, since the task makes it harder to generalize across all corpora.

1 Introduction

Discourse analysis consists in building a discourse structure representing the organization of a document – a monologue or dialogue –, as the discourse tree in Figure 1. First, the document is split into minimal sub-units, called Elementary Discourse Units (EDU): the text in the example, consisting of two sentences, is divided into 5 EDUs (from 2 to 6). The EDUs are then attached together, forming larger discourse units – such as the pair (EDU2, EDU3) – that are recursively linked to form a tree or a graph, depending on the underlying framework. The links between the discourse units are semantic-pragmatic relations, such as CONCESSION, EVIDENCE, SEQUENCE etc. These relations can be triggered by an explicit lexical item, a *connective* such as BECAUSE, WHILE, or WHEN for CONDITION in the example. Relations can also be "implicit", when no such marker is present, such as the CONCESSION between EDU2 and EDU3.

There are mainly three frameworks for discourse: Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) – from which the example in Figure 1 is derived –, Segmented Discourse Theory (SDRT) (Asher and Lascarides, 2003) – where structures are graphs –, and the Penn Discourse Treebank (PDTB) (Prasad et al., 2005), where discourse relations are sparsely annotated without constraints on the overall structure. Alternatively, there have been proposals to transform discourse structure into simpler dependency structures (*dep*), e.g. in RST (Hirao et al., 2013; Hayashi et al., 2016) or SDRT (Muller et al., 2012). Recently, this view has been taken to annotate directly new data in the SciDTB corpus (Yang and Li, 2018), proposing a set of relations and segmentation rules inspired by RST but producing trees of dependency relations between EDUs.

Several corpora have been annotated under each framework for different languages: however, even within the same framework, annotation guidelines and relation sets might be different for each corpus. The DISRPT shared task intends to provide

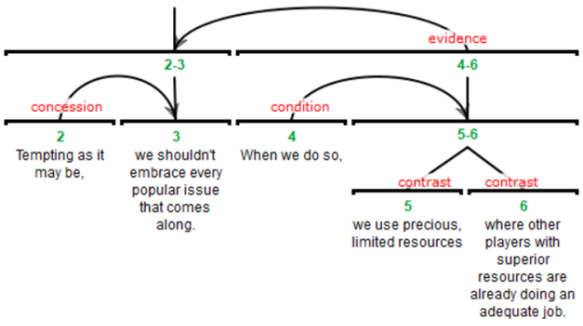

Figure 1: Example of an RST tree (Source: RST website - Common Case Analysis)

Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023), pages 29–42

a unified format for researchers to evaluate their systems against varied languages, domains, and frameworks. Three tasks were proposed: (1) discourse segmentation into EDUs, (2) identification of discourse connectives, and (3) classification of discourse relations based on attached units. The first two tasks are encoded with a BIO scheme over tokens, the latter corresponds to a multi-class classification between pairs of textual segments. The benchmark provided within DISRPT allows us to verify the robustness of our approach through 13 languages, 4 frameworks, and varied domains, including multi-party dialogues and speech transcriptions.

In this paper, we address the three tasks through two systems: DisCut[1] for tasks (1) and (2) and DiscReT[2] for task (3). These systems both rely on Transformer architectures and we thoroughly investigate different variations of the pre-trained model and the hyper-parameters values, while also varying the level of frozen layers. This latter parameter allows for lighter models, and also improvements in most cases. For task (3), we also investigate adapters (Houlsby et al., 2019) that provide a lightweight solution for transferring to new tasks. For all tasks, we favor multilingual pretrained models, in order to better generalize and experiment with corpus merging for relations, with the aim of providing a generic model that can be used for any corpus.

In the end, we ranked first on discourse segmentation on the treebanked track ($+0.87$ on the average, compared to the other system) but second for connectives (-0.47), and we are the only system with results on the plain track, with higher performance than the winner of DISRPT 2021. For relations, our system is the only one trying to mix all corpora, thus even if the performance are lower than other proposed approaches, it is possibly better at generalization.

2 Related work

Discourse parsing is the task of building the full trees/graphs. Most work focuses on attachment or discourse relation identification, and on English. Recently, a multilingual RST discourse parser has been proposed (Liu et al., 2021), building on previous work (Braud et al., 2017a; Liu et al., 2020) but proposing to jointly learn attachment and EDU segmentation and adding a cross-lingual strategy, rather than English only. It shows that multilingualism is a key component to improve performance, since data scarcity affects even English, and that good segmentation is crucial, with a loss of up to 8% with predicted EDUs for full parsing.

Discourse segmentation was considered a solved task, with scores as high as 94% (Xuan Bach et al., 2012), but it was later shown that performance drops for languages other than English, – linked to smaller corpora and lesser resources –, and when gold sentences are not given, due to sentence segmenters far from being perfect (Braud et al., 2017b). The first edition of the DISRPT shared task (Zeldes et al., 2019) also revealed the same trend with performance above 95% for some corpora, but also issues with others such as the Spanish SCDT (82.5% at best) or the Russian RRT (86.2%). The best-performing system in 2019 (Muller et al., 2019) was using a single model based on multilingual BERT for every corpus (Devlin et al., 2019), while in the second edition (Zeldes et al., 2021), the best system (Gessler et al., 2021) relied on varied language models, either mono- or multilingual, associated to hand-crafted features: best overall performance was around 91.5% on average, with a loss of about 2% when the sentences are not given.

Connective identification was first seen as a word disambiguation task, where the goal was, starting with a list of candidates, to decide whether each occurrence is used in a discourse reading or not (Pitler et al., 2008). It has been then recast as a sequence labeling one, where we need to decide whether a token starts, is within, or is outside a discourse connective (Stepanov and Riccardi, 2016). As for segmentation, performance drops when existing systems are trained on new domains or languages (Xue et al., 2016; Scholman et al., 2021), but fewer studies investigated this task since implicit relations are more an issue for discourse parsing. The first two editions of DISRPT demonstrated rather high performance: between $92-94$ for the English and Turkish corpora, and 87 for the Chinese one, with only a small drop when sentences are not given.

Discourse relations are the main object of study within the domain, with a specific focus on implicit ones since the connective is considered a very strong clue for guessing the relation (Pitler et al., 2008). However, again, performance drops, even for explicit relations when data are scarce (Jo-

[1]Code at https://github.com/phimit/jiant/

[2]Code at https://gitlab.irit.fr/melodi/andiamo/discret_ST3

hannsen and Søgaard, 2013). Moreover, a real-life scenario has to deal with both implicit and explicit relations, it is thus interesting to see results combining all types of relations, and for several languages. Only two systems were presented in 2021, and the winning model was based on Transformers, with a specific pretrained model depending on the target language and additional hand-crafted features: best overall performance is still low, with 61.8%.

3 Data

The 2023 DISRPT shared task, including suprise datasets, provides 26 corpora for 13 languages and 4 theoretical frameworks: 9 correspond to the PDTB framework (thus connective and relations), the others are either RST (12), dependency (3) or SDRT (2) (thus segmentation and relations). Among these, 10 new corpora are introduced in the 2023 edition: 6 are released as surprise datasets, with one new language (Thaï), and out-of-domain (OOD) data for English (COVID-DTB and TED), Portuguese (CRPC and TED) and Turkish (TED).

All statistics are given in Table 1. The largest corpora are the English PDTB (1,992 training documents), dep SciDTB (492 documents), and RST DT (309 training documents), and, for SDRT, the French Annodis (64 documents). In total, 8 corpora have less than 100 documents and are thus considered very small. The OOD corpora have no training set: the English COVDTB is rather large, with 150 in the dev set, but the other ones, based on TED talks for English, Portuguese, and Turkish are very small, their dev sets contain only 2 documents, around 100 connectives, and 200 relations to predict. For relations, label sets contain between 9 and 32 different relations, and we note that almost no corpus has the same set as another one.

We have 6 corpora for English (Prasad et al., 2019; Zeldes, 2017; Carlson et al., 2001; Asher et al., 2016; Yang and Li, 2018; Nishida and Matsumoto, 2022), 4 for Chinese (Zhou et al., 2014; Cao et al., 2018; Cheng and Li, 2019; Yi et al., 2021), 2 for Spanish (da Cunha et al., 2011; Cao et al., 2018), 2 for Portuguese (Cardoso et al., 2011; Mendes and Lejeune, 2022), 1 for German (Stede and Neumann, 2014), 1 for Basque (Iruskieta et al., 2013), 1 for Farsi (Shahmohammadi et al., 2021), 1 for French (Afantenos et al., 2012), 1 for Dutch (Redeker et al., 2012), 1 for Russian (Toldova et al., 2017), 1 for Turkish (Zeyrek and Webber, 2008; Zeyrek and Kurfalı, 2017), 1 for Italian (Tonelli et al., 2010; Riccardi et al., 2016) and 1 for Thai. In addition, OOD datasets come from the multilingual TED Discourse Bank with data for English, Portuguese and Turkish (Zeyrek et al., 2018, 2020).

4 DisCut: segmentation and connectives

4.1 DisCut: Model architecture

Identifying EDU boundaries and connectives (Tasks 1 and 2) corresponds to different corpora: PDTB-based datasets have connectives annotated, but not segmentation, while the others have no connectives. However, they can be both modeled as sequence labeling tasks (only "Beginning" labels for segmentation, "Beginning" and also "Inside" for connectives, to take into account multi-words markers). Our systems for these tasks are thus based on the same architecture with transformers pretrained models, fine-tuned on the task at hand.

The model is based on a pretrained language model (LM), with an additional linear layer for token classification. The LM is multilingual, allowing it to be used for all corpora. Contrary to systems proposed in 2019 and 2021 based on a similar architecture, we removed the CNN at the character level, and the LSTM outer layer, as additional experiments demonstrated no improvements.

The LM is based on a Transformer architecture with several layers within the encoder. It has been shown that, broadly speaking, lower layers mostly encode morpho-syntactic information, while upper contain more semantic ones (Rogers et al., 2020; Kovaleva et al., 2019; Bender and Koller, 2020). We thus experiment with freezing some lower layers while continuing the fine-tuning on higher levels, in order to have lighter models. "Freezing a layer" is the process of disallowing the update of weights for the target layer during the fine-tuning process, meaning that the layer preserves its learned information from pretraining.

Models are fed with sentences, the documents being too long for the LMs. We detail below our setting when sentences are not given ('Plain' track).

4.2 Settings

We chose to focus on multilingual LMs and experimented with mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020). We present results using XLM-RoBERTa, as preliminary experiments demonstrated improvements over mBERT. We experimented with both *base* and *large* versions, and tested the freezing of lower

Corpus	Train				Dev				Test			
	#Doc	#Tok	#EDU/Conn	#Rel	#Doc	#Tok	#EDU/Conn	#Rel	#Doc	#Tok	#EDU/Conn	#Rel
RST												
eng.rst.rstdt	309	166854	17646	17/16002	38	17309	1797	17/1621	38	21666	2346	17/2155
rus.rst.rrt	272	390375	34682	22/28868	30	40779	3352	19/2855	30	41851	3508	20/2843
spa.rst.rststb	203	43055	2472	28/2240	32	7551	419	23/383	32	8111	460	25/426
eng.rst.gum	165	160700	20722	14/19496	24	21409	2790	14/2617	24	21770	2740	14/2575
deu.rst.pcc	142	26831	2449	26/2164	17	3152	275	24/241	17	3239	294	24/260
fas.rst.prstc	120	52309	4607	17/4100	15	7016	576	15/499	15	7369	670	16/592
eus.rst.ert	116	30690	2785	29/2533	24	7219	677	26/614	24	7871	740	26/678
por.rst.cstn	114	48469	4601	32/4148	14	6509	630	22/573	12	3815	306	21/272
nld.rst.nldt	56	17562	1662	32/1608	12	3783	343	27/331	12	3553	338	28/325
zho.rst.gcdt	40	47639	7470	31/6454	5	7619	1144	30/1006	5	7647	1092	30/953
spa.rst.sctb	32	10253	473	24/439	9	2448	103	17/94	9	3814	168	19/159
zho.rst.sctb	32	9655	473	26/439	9	2264	103	19/94	9	3577	168	20/159
SDRT												
fra.sdrt.annodis	64	22515	2255	18/2185	11	5013	556	18/528	11	5171	618	18/625
eng.sdrt.stac	33	41060	9887	16/9580	6	4747	1154	16/1145	6	6547	1547	16/1510
DEP												
eng.dep.scidtb	492	62461	6740	24/6060	154	20288	2130	24/1933	152	19744	2116	24/1911
∗**eng.dep.covdtb**	-	-	-	-	150	29369	2754	12/2399	150	31480	2951	12/2586
zho.dep.scidtb	69	11288	898	23/802	20	3852	309	18/281	20	3621	235	17/215
PDTB												
eng.pdtb.pdtb	1992	1061229	23850	23/43920	79	39768	953	20/1674	91	55660	1245	23/2257
por.pdtb.crpc	243	147594	3994	22/8797	28	20102	621	20/1285	31	19153	544	19/1248
tur.pdtb.tdb	159	391304	7063	23/2451	19	49097	831	22/312	19	46988	854	22/422
∗**tha.pdtb.tdtb**	139	199135	8277	20/8278	19	27326	1243	18/1243	22	30062	1344	18/1344
zho.pdtb.cdtb	125	52061	1034	9/3657	21	11178	314	9/855	18	10075	312	9/758
ita.pdtb.luna	42	16776	671	15/956	6	3081	139	14/210	12	6257	261	14/381
∗**eng.pdtb.tedm**	-	-	-	-	2	2574	110	20/178	4	5474	231	18/351
∗**por.pdtb.tedm**	-	-	-	-	2	2785	102	20/190	4	5405	203	18/364
∗**tur.pdtb.tedm**	-	-	-	-	2	2113	135	21/213	4	4030	247	22/364

Table 1: Statistics on the datasets: **bold** indicates a new corpus compared to DISRPT 2021, ∗ indicates a surprise corpus, '-' is for OOD corpora, without training sets. #EDU/CONN is the number of EDUs for RST, SDRT, and DEP corpora, the number of connectives for PDTB corpora; #REL corresponds to the size of label sets / total number of pairs annotated.

layers, aiming at possibly improved performance, with a lighter training.

With XLM-RoBERTa base, we tested no freezing, or freezing of either the first 3 or 8 layers (out of 12); for the large version, we increased to 6 and 12 layers (out of 24). We tested several values for the learning rate $\in [10^{-5}, 2{\cdot}10^{-5}, 10^{-4}]$ and chose 10^{-5}. We tested different batch sizes $\in [1, 4, 8, 16]$ – only the value 1 fitted our GPU for the large version –, with a gradient accumulation of 4 and a maximum of 30 epochs with patience of 10 over the performance on the development set. The input size is limited to 180. Our implementation relies on and extends the Jiant library[3] (Phang et al., 2020).

After evaluation on the dev set, we found that most models perform better with RoBERTa-large and with freezing the first 6 layers. Small improvements could be observed for some corpora with either the base version or other freezing values, but the increase was limited to less than 1.2%, and in general less than .5%, and we thus decided to favor one single model in order to make it easier to use, and better at generalizing to new data.

Dealing with raw data: The DISRPT shared task proposes two tracks for tasks 1 and 2: you can either use data segmented into sentences and syntactically parsed (*Treebanked*) – either gold or obtained with Stanza –, or raw tokenized documents (*Plain*). As the LMs have limitations on the size of their input, we can not give directly the documents as input: we thus decided to split the raw documents into sentences.

However, having observed issues with Stanza segmentation, we tried alternatives: Ersatz (Wicks and Post, 2021) and Trankit (Nguyen et al., 2021), and chose the latter based on better performance. Note that, with the evaluation being based on tokens, we had to realign tokens when the tool was modifying the tokenization. We were unable to obtain a correct sentence segmentation for the Italian ita.pdtb.luna, composed of speech transcripts, and

[3] https://jiant.info/

thus cut every 120 tokens for this corpus.

Dealing with surprise and OOD data: Dealing with the surprise Thai (tha.pdtb.tdtb) and English (eng.dep.covdtb) datasets were straightforward: since our model configuration is the same across all corpora, we retrain new models using the training data made available. This year, the organizers also include out-of-domain (OOD) data as surprise datasets, for which data are only available for evaluation (dev and test sets only). The corpora have, however, corresponding datasets within the same framework and language: we use our model trained on these available data to make predictions on the OOD ones (e.g. training on eng.pdtb.pdtb to test on eng.pdtb.tedm).

4.3 Experiments and results

We present our results in Table 2 for segmentation and connective identification. Current comparison with 2021, considering only the corpora available in 2021, demonstrate general improvements for all tasks except connective for the Plain track were results are on par: for segmentation, the average on test sets for Treebanked is 91.77% (vs 91.48 for DiscoDisco 2021) and for Plain 91.22% (vs 89.79); for connective: 91.81% (vs 91.22) for Treebanked and 91.05% (vs 91.49) for Plain. Note that our approach uses a similar architecture with much simpler inputs (only tokens), and different optimizations. When comparing the reproduced results with the ones we produced, we observed a large variance between the scores, especially for small corpora, with for example a difference of about 2 to 5 points for the TEDm corpora, and about 2 to 3 points also for other small datasets such as the spa.rst.sctb, the zho.rst.sctb, demonstrating the importance for future work to make multiple runs and indicate variance. Interested readers can find our own results on the test sets in Appendix A.

As shown in Table 2, compared to 2021, we observe a large drop in mean performance of about 10% for connective detection, for which many new corpora were added, including several OOD datasets making the task more challenging.

For segmentation, the results for the two settings, Treebanked and Plain are in general very similar, except for the Chinese zho.rst.sctb and English eng.sdrt.stac for which the Treebanked setting is clearly better (+3 to 5%). On the other hand, we have an important improvement for the French corpus fra.sdrt.annodis (almost +3%) using our new segmented files (Plain): these results are in line with the bad performance observed for Stanza. For the Russian corpus, we found that the segmentation of some parts of the documents was strange: bibliography entries were merged into very large EDUs that were split by all sentence segmenters, thus modifying the tool did not bring any improvement.

For connective detection, results are rather high for large corpora already present in the previous campaigns, even if the Chinese corpus is still challenging. As expected, the Italian Luna is associated with low performance, because it is composed of speech transcriptions of dialogues. Note that the performance for the new Thai corpus is on par, but they drop on the out-of-domain TEDm corpora for which we used the model trained on a corpus with the same language and framework, but that corresponds to a domain shift. Interestingly, the

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	**96.01**	95.58	96.60	**94.24**	93.94
eng.rst.gum	95.54	**95.50**	94.15	95.78	**94.46**	92.61
eng.rst.rstdt	97.33	**97.62**	96.64	97.60	**97.74**	96.35
eus.rst.ert	91.69	89.93	**90.46**	91.83	**91.09**	90.47
fas.rst.prstc	93.79	**93.40**	92.94	94.05	**93.36**	92.86
nld.rst.nldt	97.51	**96.54**	95.97	97.09	**97.19**	94.69
por.rst.cstn	94.06	93.98	**94.35**	93.50	**94.36**	94.11
rus.rst.rrt	86.80	85.58	**86.21**	84.75	85.41	**85.74**
spa.rst.rststb	96.19	**93.53**	92.22	96.32	**93.70**	91.76
spa.rst.sctb	86.88	**85.63**	82.48	85.44	**84.21**	80.86
zho.rst.gcdt	92.69	92.55	-	92.20	91.74	-
zho.rst.sctb	79.05	81.84	**83.34**	77.53	**78.55**	76.21
eng.sdrt.stac	94.77	**95.22**	94.91	91.57	90.67	**91.91**
fra.sdrt.annodis	90.27	88.21	**90.02**	90.17	**90.89**	85.78
∗**eng.dep.covdtb**	91.32	92.13	-	91.65	92.13	-
eng.dep.scidtb	96.18	95.07	-	95.63	94.49	-
zho.dep.scidtb	93.33	89.07	-	93.01	90.04	-
Mean	92.60	91.87	-	92.04	91.43	-
Mean corpora 2021	-	**91.77**	91.48	-	**91.22**	89.79
Connective identification						
eng.pdtb.pdtb	94.41	**93.66**	92.02	93.94	91.64	**92.56**
∗**eng.pdtb.tedm**	75.86	78.36	-	80.00	75.83	-
ita.pdtb.luna	79.72	65.85	-	74.19	71.60	-
por.pdtb.crpc	85.16	80.66	-	84.65	79.49	-
∗**por.pdtb.tedm**	73.08	80.29	-	71.22	79.45	-
∗**tha.pdtb.tdtb**	87.43	85.66	-	74.32	69.92	-
tur.pdtb.tdb	89.73	92.77	**94.11**	89.69	91.12	**93.56**
∗**tur.pdtb.tedm**	65.42	64.10	-	64.15	64.78	-
zho.pdtb.cdtb	87.66	**89.00**	87.52	87.77	**90.38**	88.35
Mean	82.05	81.15	-	79.99	79.36	-
Mean corpora 2021	-	**91.81**	91.22	-	91.05	**91.49**

Table 2: DisCut: Results (F1) on the dev and test sets for segmentation and discourse connective identification. Models with XLM-RoBERTa-large, freezing layers 0-5. Test scores come from the reproduction done by the organizers. 'DD21' stands for DiscoDisco 2021, the system ranked first in DISRPT 2021. 'Mean corpora 2021' is the mean F1 without considering the corpora added in DISRPT 2023.

use of Trankit for sentence segmentation (Plain track) leads to large improvements for Luna (almost +6%) and also allows a small increase for the Chinese zho.pdtb.cdtb (+1.4), with, on the other hand, a loss of about 2% for the English PDTB, and an impressive drop of about 16% for Thai for which the model of sentence segmentation is probably faulty. Overall, the Plain setting would lead to average results on par with the treebanked ones for connective identification, without the Thai dataset (80.54 on average for Plain against 80.59 for Treebanked, without Thai). These results indicate that the good performance of the sentence segmenter is a key component of a well performing discourse segmenter or connective identifier.

5 DiscReT: Discourse Relation Tagging

5.1 Introduction

For the third proposed task, Discourse Relation Classification across Formalisms, we submit a multilingual approach to discourse relation tagging that spans across frameworks, powered by transformer-based architectures. Our goal is to test the capacities and weaknesses of these models, given the large variety of languages and relation labels, without sacrificing the multilingual setting or the unique information captured in coarse-/fine-grained labels. Our results vary vastly between languages and frameworks but present interesting pointers for future work and model improvements.

5.2 Dataset

In order to stay faithful to the multilingual nature of the task, we decided to use all the datasets in parallel for training. Extensive earlier experiments with translations of the datasets to English, training with groups of corpora per language family, or training per annotation framework were not as successful or did not significantly outperform the accumulative approach.

We aimed to reduce label space and maximize label coverage, i.e. not having a label that only exists in one corpus if it can be rewritten as a more general one. First, we lower-cased all labels in all datasets (but preserved our modifications, in order to reverse them for the final results in accordance with the Shared Task data). Second, we manually merged labels that were either spelling variants or simplified versions of existing labels. For example, the label "qap" means "question-answer pair", which already exists as the label "question_answer_pair". Meanwhile, the label "conjunction" is a simplified version of the label "expansion.conjunction" found in RST corpora in both forms, therefore by changing the label to its more verbose form, we are preserving its information and making the labels more uniform. However, we decided against the large-scale conversion of labels based on their meaning, e.g. merging the "conjunction" and "joint" labels.These conversions reduced the number of unique labels from 163 to 135; while the number was not significantly reduced, we wanted to make the results more interpretable without sacrificing important information. We present the implemented conversions in Table 3.

Original Label	**Conversion**
alternation	expansion.alternative
alternative	expansion.alternative
bg-general	background
causation	cause
cause-result	cause-effect
conditional	condition
conjunction	expansion.conjunction
correction	expansion.correction
disjunction	expansion.disjunction
evidence	explanation-evidence
exp-evidence	explanation-evidence
expansion.genexpansion	expansion
expansion.level	expansion.level-of-detail
findings	result
goal	purpose-goal
joint-disjunction	expansion.disjunction
justify	explanation-justify
list	joint-list
motivation	explanation-motivation
otherwise	adversative
qap	question_answer_pair
qap.hypophora	hypophora
repetition	restatement-repetition
restatement	expansion.restatement
sequence	joint-sequence
temporal.synchrony	temporal.synchronous
textual-organization	organization
unconditional	expansion.disjunction
unless	contrast

Table 3: List of label conversions that we implemented (apart from lower-casing). Underlined labels were found exclusively in the surprise datasets.

We make use of the directional information of the relations, available in the datasets in the column "dir". We do not change the input in sentences with the direction "1>2", but we switch the input position of sentences with the direction "1<2" to "2>1". An example can be found in Table 4. Even though the models we use in this task are bidirectional, we observed an increase in performance when the direction of relations was unified.

We do not further process the text input, as the

Corpus:	spa.rst.rststb
unit1_txt	La diferenciación como un modelo para el análisis de las relaciones de pareja
unit2_txt	El presente artículo hace una revisión sobre este concepto
dir	1>2
label	preparation
input	[CLS] La diferenciación como un modelo para el análisis de las relaciones de pareja [SEP] El presente artículo hace una revisión sobre este concepto

Corpus:	deu.rst.pcc
unit1_txt	Und die Zeit drängt .
unit2_txt	Der große Einbruch der Schülerzahlen an den weiterführenden Schulen beginnt bereits im Herbst 2003 .
dir	1<2
label	reason
input	[CLS] Der große Einbruch der Schülerzahlen an den weiterführenden Schulen beginnt bereits im Herbst 2003 . [SEP] Und die Zeit drängt .

Table 4: Examples of inputs with different directions. In the first example, the direction of the relation is 1>2, therefore the model input is in the same order as in the data. In the second example, the direction is 1<2, so the model input has the two sentences in reversed order.

necessary conversions (e.g. tokenization, lowercasing) are specified by each model. However, at the tokenization stage, we ensured that the input length complied with the restrictions of maximum input length that transformer-based models impose; each sentence is truncated to half of the maximum input length, if necessary.

5.3 DiscReT: Model architectures

We opted for transformer-based architectures for our experiments and tested several of them (mBERT, xml-RoBERTa, DistilBERT) in order to decide on which one to focus our research effort on. After preliminary tests, the multilingual BERT base cased model (mBERT) (Devlin et al., 2019) was the most successful overall and included all the languages of the Shared Task in its pretrained version available from Huggingface.[4]

As a "baseline" for our experiments, we trained an mBERT classifier built with PyTorch (Paszke et al., 2019), without frozen layers, and trained for a maximum of 5 epochs.

In order to inject additional information in the finetuning process of the classifier, without further changing the input data, we used *adapters* alongside our mBERT classifier. Adapters (Houlsby et al., 2019) are an alternative lightweight method to finetuning with equivalent good results on most NLP tasks. An adapter is a transformer architecture with layer-specific pretrained parameters Θ_l which are frozen and a small set of new parameters Φ_l (where l is the transformer layer). During finetuning, only the adapters' Φ_l parameters are updated from the loss function L on dataset D (see Equation 1). This enables efficient parameter sharing between tasks, languages, etc.

$$\Phi_l^* \leftarrow \arg\min_{\Phi_l} L(D; \{\Theta_l, \Phi_l\}) \quad (1)$$

We are using the tool AdapterHub (Pfeiffer et al., 2020) which allows for easier finetuning and integration of adapters to transformer-based models. After several experiments, we observed that the finetuning process of an adapter is quite different than that of a model; the adapter set of parameters learns most effectively with more finetuning epochs than a normal model and the training process per epoch is longer. Additionally, we experimented with freezing the parameters of certain layers for the models and the adapters, in order to determine the best model.

We trained multiple mBERT adapters, out of which the most successful were:

1. mBERT adapter trained on the entire dataset for 15 epochs and with frozen layer 1 (A1)
2. mBERT adapter trained on the entire dataset for 15 epochs and with frozen layers 1-3 (A1-3)

5.4 Results

5.4.1 Shared Task results

While evaluating our models, we observed that the best accuracy in each development set was not always achieved by one model. Our final submission is composed of three models:

1. the "baseline" finetuned mBERT model without adapter with multiple epochs (B)
2. the finetuned mBERT model for 3 epochs with an mBERT adapter trained for 15 epochs and layer 1 frozen (A1)
3. the finetuned mBERT-cased model for 4 epochs with an mBERT adapter trained for 15 epochs and layers 1-3 frozen (A1-3)

The results on the test set, as recreated and reported by the organizers of the Shared Task, are found in Table 5. Our poor performance is, to some

[4] https://huggingface.co/bert-base-multilingual-cased

extent, due to the problems we faced to convert the lower-cased and converted labels back to their upper-cased format, which was required for the Shared Task evaluation. This dramatically lowered the test results reproduced and published for the Shared Task. For clarity, we are reporting the Shared Task results from the organizers, but also include the results on the dev and test sets that were produced before converting the labels to their original in Tables 6 and 7 respectively. These results were calculated with scikit-learn (Pedregosa et al., 2011) and the process of calculating them is transparent in our code.

Our goal was to create a truly multilingual approach for discourse relation parsing. We did not aim to establish a new state-of-the-art, but to observe whether multilingual word embeddings can work in synergy (to learn common labels) and specialize at the same time (to learn corpus-unique labels). We also deliberately focused and submitted a combination of three models, instead of proposing the best model for each dataset, thus sacrificing performance for reproducibility. During our experiments, there were other combinations of adapters and models with frozen layers that yielded slightly better results on specific corpora, however, the training times for multiple models would be problematic for a Shared Task entry.

Given that our results are not much worse than approaches with a combination of monolingual models and independent training, it is possible to derive benefits from joint training and evaluating multiple languages. Our multilingual models showed strengths (e.g. in the spa.rst.rstsb dataset) and weaknesses (e.g. in English, Turkish and Chinese datasets) that cannot be pinpointed directly to a specific framework, the size of the corpus, or the size of the specific language data, and will need to be further explored. Our submission was marred by implementation issues, but we are hopeful that in future work we will tackle these issues and implement improvements on our multilingual approach.

Corpus	**DiscReT**	**DiscoDisco**	**Difference**
deu.rst.pcc	26.92	39.23	-12.31
eng.rst.gum	55.34	66.76	-11.42
eng.rst.rstdt	49.98	67.1	-17.12
eus.rst.ert	51.77	60.62	-8.85
fas.rst.prstc	50.34	52.53	-2.19
nld.rst.nldt	43.69	55.21	-11.52
por.rst.cstn	62.87	64.34	-1.47
rus.rst.rrt	61.52	66.44	-4.92
spa.rst.rststb	58.22	54.23	**3.99**
spa.rst.sctb	33.33	66.04	-32.71
zho.rst.gcdt	55.72	-	-
zho.rst.sctb	49.06	64.15	-15.09
eng.sdrt.stac	56.89	65.03	-8.14
fra.sdrt.annodis	44.96	46.4	-1.44
∗**eng.dep.covdtb**	41.3	-	-
eng.dep.scidtb	67.56	-	-
zho.dep.scidtb	67.44	-	-
eng.pdtb.pdtb	69.25	74.44	-5.19
∗**eng.pdtb.tedm**	19.94	-	-
ita.pdtb.luna	58.42	-	-
∗**por.pdtb.crpc**	72.76	-	-
∗**por.pdtb.tedm**	54.95	-	-
∗**tha.pdtb.tdtb**	95.24	-	-
tur.pdtb.tdb	49.05	60.09	-11.04
∗**tur.pdtb.tedm**	49.73	-	-
zho.pdtb.cdtb	69.13	86.49	-17.36
MEAN (all)	54.44	-	
MEAN (2021)	52.02	61.82	-9.8

Table 5: The results that organizers provided for discourse relation classification (Task 3), evaluating the test sets and reporting accuracy in %. 'DiscoDisco' was the best-performing model of DISRPT 2021 (Gessler et al., 2021) and 'Diff.' is the comparison with our models. MEAN (all) provides the mean for the currently available datasets, while MEAN (2021) averages only DISRPT 2021's corpora.

6 Conclusion

In this paper, we presented our submissions for the three tasks of the DISRPT Shared Task. Our main goals were to rely on only a few architectures variants for generality, and experiment with parameter efficient methods. For Tasks 1-2, we employed multi-task, multi-corpora approaches; however, at this stage of our research our results are not optimal. In future work, we aim to further explore this strategy, as it seems promising for lower-resource languages. Additionally, we are interested in approaches beyond the scope of this campaign, such as domain transfer. Furthermore, it was possible to perform segmentation and connective detection on datasets without training data, as shown by the surprise TEDm test sets. It would be interesting to examine whether the DISRPT framework could be transferred to new languages, for which there are no training data for segmentation or connective detection, such as the rest of the TEDm corpus. As for Task 3, our focus was on a unified, purely multilingual approach with parameter optimization, as well as dataset preprocessing for unification. Even though we faced problems on the Shared Task submission results, our approach showed promising results compared to language-specific models.

Corpus	B (1)	B (2)	B (3)	B (4)	B (5)	B (6)	A1-3 (4)	A1 (3)
deu.rst.pcc	25.31	29.46	26.97	**31.54**	31.54	29.88	29.05	30.71
eng.rst.gum	48.03	51.66	52.46	53.31	53.76	53.69	55.79	**56.67**
eng.rst.rstdt	46.33	49.85	44.97	47.25	48.49	47.62	**51.02**	50.28
eus.rst.ert	41.53	43.65	43.16	44.95	42.18	44.46	45.44	**47.07**
fas.rst.prstc	**53.31**	50.1	52.1	51.3	49.5	51.9	52.91	52.71
nld.rst.nldt	43.5	40.79	**46.53**	41.09	46.53	42.9	45.92	45.32
por.rst.cstn	54.8	59.51	55.15	60.38	60.56	58.12	**62.48**	61.43
rus.rst.rrt	57.41	58.84	60.04	60.04	58.77	59.61	**61.16**	60.91
spa.rst.rststb	51.44	54.31	55.61	**62.4**	60.05	61.62	60.31	59.01
spa.rst.sctb	47.87	62.77	56.38	58.51	62.77	**67.02**	59.57	64.89
zho.rst.gcdt	53.98	55.57	56.86	57.55	57.75	58.05	58.85	**59.34**
zho.rst.sctb	40.43	**50**	46.81	46.81	42.55	**50**	47.87	46.81
eng.sdrt.stac	45.5	55.02	53.8	55.28	55.55	54.15	**57.82**	56.59
fra.sdrt.annodis	30.3	44.32	47.16	46.4	**49.81**	47.92	48.3	47.54
∗eng.dep.covdtb	40.39	42.81	35.22	36.64	36.18	42.93	**43.56**	43.1
eng.dep.scidtb	59.39	59.34	66.48	66.06	70.2	66.17	70.56	**71.03**
zho.dep.scidtb	47.33	61.57	62.99	59.79	60.85	62.63	**66.19**	65.48
eng.pdtb.pdtb	67.32	67.44	71.39	70.85	70.43	69.41	**72.4**	71.09
∗**eng.pdtb.tedm**	10.67	14.04	19.1	15.73	15.17	14.04	**20.79**	19.1
ita.pdtb.luna	45.93	53.59	51.67	50.72	54.07	54.07	54.55	**56.46**
∗**por.pdtb.crpc**	65.76	66.69	67.39	67.16	67.16	65.29	**68.25**	67.94
∗**por.pdtb.tedm**	50	45.79	47.37	49.47	48.95	46.32	**54.21**	51.05
∗**tha.pdtb.tdtb**	92.68	93.56	93.08	93.97	93.64	92.76	93.72	**93.97**
tur.pdtb.tdb	42.95	39.1	**42.95**	40.06	39.42	41.67	41.03	39.1
∗**tur.pdtb.tedm**	42.72	42.72	44.13	42.25	41.31	**46.48**	43.66	43.66
zho.pdtb.cdtb	73.8	75.09	**76.37**	74.62	73.8	74.15	75.44	73.92
MEAN	49.18	52.6	52.93	53.24	53.5	53.96	**55.42**	55.2

Table 6: Results on the dev set for discourse relation classification, before converting labels to their original form. In parenthesis is the number of epochs for which the model was trained.

Corpus	B (1)	B (2)	B (3)	B (4)	B (5)	B (6)	A1-3 (4)	A1 (3)	DiscoDisco	Diff.
deu.rst.pcc	25.77	30.77	26.54	32.31	32.31	33.08	**33.85**	33.08	39.23	-5.38
eng.rst.gum	50.49	54.72	55.96	57.09	57.36	55.69	**58.56**	58.41	66.76	-8.2
eng.rst.rstdt	46.91	**50.16**	46.73	47.94	48.54	48.77	49.84	49.88	67.1	-16.94
eus.rst.ert	40.56	43.66	44.99	47.94	46.17	48.97	50.44	**51.33**	60.62	-9.29
fas.rst.prstc	47.47	47.13	48.31	**50.84**	47.47	49.16	50.51	49.66	52.53	-1.69
nld.rst.nldt	43.38	42.46	43.38	42.46	43.38	40.31	46.15	**47.38**	55.21	-7.83
por.rst.cstn	64.34	65.44	65.07	64.34	63.97	64.71	**65.44**	65.07	64.34	**1.1**
rus.rst.rrt	59.44	60.11	60.75	60.96	60.11	59.41	**62.29**	61.98	66.44	-4.15
spa.rst.rststb	48.83	51.17	53.76	57.04	54.69	53.76	57.75	**59.15**	54.23	**4.92**
spa.rst.sctb	58.49	64.15	64.78	**69.81**	64.15	63.52	65.41	61.64	66.04	**3.77**
zho.rst.gcdt	47.32	49.32	49.32	52.78	53.2	52.47	53.73	**54.67**	-	-
zho.rst.sctb	45.28	53.46	55.35	**57.86**	44.03	48.43	47.8	50.31	64.15	-6.29
eng.sdrt.stac	40.99	50.46	50.73	52.52	52.32	51.19	**55.76**	55.17	65.03	-9.27
fra.sdrt.annodis	31.68	42.56	45.92	44	44.8	45.12	**46.88**	45.12	46.4	**0.48**
∗**eng.dep.covdtb**	38.09	**41.38**	33.37	35.11	35.77	39.44	41.14	40.87	-	-
eng.dep.scidtb	59.45	61.38	67.29	67.09	69.65	68.18	69.81	**70.38**	-	-
zho.dep.scidtb	53.02	61.86	64.19	56.28	60.93	60.93	**64.65**	64.19	-	-
eng.pdtb.pdtb	64.91	64.2	68.41	68.01	65.62	64.82	**68.85**	68.63	74.44	-5.59
∗**eng.pdtb.tedm**	10.83	12.25	18.23	15.67	12.54	15.67	**20.8**	19.94	-	-
ita.pdtb.luna	45.53	52.11	52.11	52.37	56.32	53.68	57.63	**57.63**	-	-
∗**por.pdtb.crpc**	69.15	67.71	70.59	71.07	70.67	68.51	71.07	**72.04**	-	-
∗**por.pdtb.tedm**	58.24	54.12	58.52	56.04	55.49	56.04	56.32	**58.52**	-	-
∗**tha.pdtb.tdtb**	94.12	95.16	95.24	**95.39**	95.16	94.79	94.94	95.31	-	-
tur.pdtb.tdb	**51.9**	48.1	48.82	48.58	46.92	50.95	51.42	50.71	60.09	-8.19
∗**tur.pdtb.tedm**	45.33	45.05	44.51	45.33	44.23	46.7	49.18	**50.55**	-	-
zho.pdtb.cdtb	68.34	71.24	**73.61**	67.41	66.75	65.17	68.6	66.89	86.49	-12.88
MEAN (2021)	49.3	52.49	53.32	54.32	52.41	52.69	**54.97**	54.65	61.82	-7.17
MEAN (all)	50.38	53.08	54.1	54.47	53.56	53.83	**56.11**	56.1	-	-

Table 7: Results on the test set for discourse relation classification, before converting labels to their original form. In parenthesis is the number of epochs for which the model was trained. 'DiscoDisco' was the best-performing model of DISRPT 2021 (Gessler et al., 2021) and 'Diff.' is the comparison with our models. MEAN (all) provides the mean for the currently available datasets, while MEAN (2021) averages only DISRPT 2021's corpora.

Acknowledgements

This work is supported by the AnDiaMO project (ANR-21-CE23-0020). This work was partially supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program. This work is also partially supported by the SLANT project (ANR-19-CE23-0022). Chloé Braud and Philippe Muller are part of the programme DesCartes and are also supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The work was also supported by the ANR grant SUMM-RE (ANR-20-CE23-0017).

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anaïs Cadilhac, Cédric Degremont, Pascal Denis, Markus Guhe, Simon Keizer, Alex Lascarides, Oliver Lemon, Philippe Muller, Soumya Paul, Verena Rieser, and Laure Vieu. 2012. Developing a corpus of strategic conversation in the settlers of catan. In *Proceedings of the workshop on Games and NLP (GAMNLP)*.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Chloé Braud, Maximin Coavoux, and Anders Søgaard. 2017a. Cross-lingual RST discourse parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 292–304, Valencia, Spain. Association for Computational Linguistics.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017b. Cross-lingual and cross-domain discourse segmentation of entire documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 237–243, Vancouver, Canada. Association for Computational Linguistics.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST Spanish-Chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paula Christina Figueira Cardoso, Erick Galani Maziero, Maria Lucía del Rosario Castro Jorge, M. Eloize, R. Kibar Aji Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Graças Volpe Nunes, and Thiago Alexandre Salgueiro Pardo. 2011. CST-News - a discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, Cuiabá, Brazil.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Yi Cheng and Sujian Li. 2019. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*, pages 24–29, Tokyo, Japan. Association for Computational Linguistics.

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.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish Treebank. In *Proceedings of the Fifth Linguistic Annotation Workshop (LAW V)*, pages 1–10, Portland, OR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Katsuhiko Hayashi, Tsutomu Hirao, and Masaaki Nagata. 2016. Empirical comparison of dependency conversions for RST discourse trees. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 128–136, Los Angeles. Association for Computational Linguistics.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019.

Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Mikel Iruskieta, María Jesús Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque TreeBank: An online search interface to check rhetorical relations. In *4th Workshop on RST and Discourse Studies*, pages 40–49, Fortaleza, Brasil.

Anders Johannsen and Anders Søgaard. 2013. Disambiguating explicit discourse connectives without oracles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 997–1001, Nagoya, Japan. Asian Federation of Natural Language Processing.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2020. Multilingual neural RST discourse parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6730–6738, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. DMRST: A joint framework for document-level multilingual RST discourse segmentation and parsing. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 154–164, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Amália Mendes and Pierre Lejeune. 2022. Crpc-db a discourse bank for portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, page 79–89, Berlin, Heidelberg. Springer-Verlag.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India. The COLING 2012 Organizing Committee.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90, Manchester, UK. Coling 2008 Organizing Committee.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The penn discourse treebank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multilayer discourse annotation of a Dutch text corpus. In *Proceedings of LREC 2012*, pages 2820–2825, Istanbul, Turkey.

Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. 2016. Discourse connective detection in spoken conversations. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6095–6099.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Merel Scholman, Tianai Dong, Frances Yung, and Vera Demberg. 2021. Comparison of methods for explicit discourse connective identification across various domains. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 95–106, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian Rhetorical Structure Theory. *arXiv preprint arXiv:2106.13833*.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Evgeny Stepanov and Giuseppe Riccardi. 2016. UniTN end-to-end discourse parser for CoNLL 2016 shared task. In *Proceedings of the CoNLL-16 shared task*, pages 85–91, Berlin, Germany. Association for Computational Linguistics.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in Russian RST treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33, Santiago de Compostela, Spain. Association for Computational Linguistics.

Sara Tonelli, Giuseppe Riccardi, Rashmi Prasad, and Aravind Joshi. 2010. Annotation of discourse relations for conversational spoken dialogs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 160–168, Seoul, South Korea. Association for Computational Linguistics.

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.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Cheng Yi, Li Sujian, and Li Yueyuan. 2021. Unifying discourse resources with dependency framework. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1058–1065, Huhhot, China. Chinese Information Processing Society of China.

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

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Antonio, and Mikel Iruskieta. 2019. The DISRPT 2019 shared task on elementary discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 97–104, Minneapolis, MN. Association for Computational Linguistics.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deniz Zeyrek and Murathan Kurfalı. 2017. TDB 1.1: Extensions on Turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81, Valencia, Spain. Association for Computational Linguistics.

Deniz Zeyrek, Amália Mendes, Yulia Grishina, Murathan Kurfalı, Samuel Gibbon, and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): a parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 54:587–613.

Deniz Zeyrek, Amália Mendes, and Murathan Kurfalı. 2018. Multilingual extension of PDTB-style annotation: The case of TED multilingual discourse bank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Deniz Zeyrek and Bonnie Webber. 2008. A discourse resource for Turkish: Annotating discourse connectives in the METU corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese discourse treebank 0.5 ldc2014t21. *Web Download. Philadelphia: Linguistic Data Consortium*.

Corpus	Treebanked			Plain		
	F1 dev	F1 test	DD21	F1 dev	F1 test	DD21
Segmentation						
deu.rst.pcc	96.79	**97.30**	95.58	96.60	**96.60**	93.94
eng.rst.gum	95.54	**95.55**	94.15	95.78	**94.97**	92.61
eng.rst.rstdt	97.33	**97.11**	96.64	97.60	**97.45**	96.35
eus.rst.ert	91.69	**91.56**	90.46	91.83	**92.38**	90.47
fas.rst.prstc	93.79	**93.88**	92.94	94.05	92.56	**92.86**
nld.rst.nldt	97.51	**97.47**	95.97	97.09	**97.63**	94.69
por.rst.cstn	94.06	93.48	**94.35**	93.50	94.08	**94.11**
rus.rst.rrt	86.80	85.86	**86.21**	84.75	85.46	**85.74**
spa.rst.rststb	96.19	**92.67**	92.22	96.32	**92.31**	91.76
spa.rst.sctb	86.88	**84.40**	82.48	85.44	**87.16**	80.86
zho.rst.gcdt	92.69	92.30	-	92.20	91.78	-
zho.rst.sctb	79.05	81.18	**83.34**	77.53	75.29	**76.21**
eng.sdrt.stac	94.77	94.83	**94.91**	91.57	90.69	**91.91**
fra.sdrt.annodis	90.27	89.54	**90.02**	90.17	**91.40**	85.78
∗**eng.dep.covdtb**	91.32	91.41	-	91.65	92.26	-
eng.dep.scidtb	96.18	95.44	-	95.63	94.89	-
zho.dep.scidtb	93.33	90.40	-	93.01	89.64	-
Mean	92.60	92.02	-	92.04	91.56	-
Mean corpora 2021	-	**91.91**	91.48	-	**91.38**	89.79
Connective identification						
eng.pdtb.pdtb	94.41	**92.38**	92.02	93.94	92.25	**92.56**
∗**eng.pdtb.tedm**	75.86	77.88	-	80.00	80.63	-
ita.pdtb.luna	79.72	64.08	-	74.19	70.17	-
por.pdtb.crpc	85.16	81.74	-	84.65	80.26	-
∗**por.pdtb.tedm**	73.08	75.23	-	71.22	77.60	-
∗**tha.pdtb.tdtb**	87.43	86.42	-	74.32	69.32	-
tur.pdtb.tdb	89.73	92.48	**94.11**	89.69	**93.57**	93.56
∗**tur.pdtb.tedm**	65.42	66.33	-	64.15	64.27	-
zho.pdtb.cdtb	87.66	**89.95**	87.52	87.77	**90.43**	88.35
Average	82.05	80.72	-	79.99	79.83	-
Mean corpora 2021	-	**92.30**	91.22	-	**91.78**	91.49

Table 8: DisCut: Results (F1) on the dev and test sets for segmentation and discourse connective identification. Models with RoBERTa-large, freezing layers 0-5. 'DD21' stands for DiscoDisco 2021, the system ranked first in DISRPT 2021. 'Mean corpora 2021' is the mean F1 without considering the corpora added in DISRPT 2023.

A Additional results

The table 8 corresponds to the scores we obtain on the test sets, that can be compared to the ones obtained by the organizers when reproducing our system, as given in Table 2.