Multi-lingual Discourse Segmentation and Connective Identification: MELODI at DISRPT2021

Morteza Ezzabady and **Philippe Muller** and **Chloé Braud**
IRIT, University of Toulouse, CNRS, ANITI

Abstract

We present an approach for discourse segmentation and discourse connective identification, both at the sentence and document level, within the DISRPT 2021 shared task, a multi-lingual and multi-formalism evaluation campaign.[1] Building on the most successful architecture from the 2019 similar shared task, we leverage datasets in the same or similar languages to augment training data and improve on the best systems from the previous campaign on 3 out of 4 subtasks, with a mean improvement on all 16 datasets of 0.85%. Within the Disrpt 21 campaign the system ranks 3rd overall, very close to the 2nd system, but with a significant gap with respect to the best system, which uses a rich set of additional features. The system is nonetheless the best on languages that benefited from cross-lingual training on sentence internal segmentation (German and Spanish).

1 Introduction

Discourse segmentation, the separation of a text or conversation in elementary units that make up the arguments of the rhetorical structure of a text, has long been a neglected step in discourse analysis, considered easy and generally assumed as given in discourse parsing studies, where the focus is to predict the rhetorical structure of a document, a labelled relational structure, with properties dependent on the theorerical framework considered, Rhetorical Structure Theory (RST, Mann and Thompson, 1988), Segmented Discourse Representation Theory (SDRT, Asher and Lascarides, 2003), or the Penn Discourse Treebank (PDTB, Prasad et al., 2008).

However this important step has generated more interest recently, as illustrated by the 2019 shared task at the Discourse Relation Parsing and Treebanking (DISRPT) workshop (Zeldes et al., 2019). This campaign made available several existing corpora in different language in a common format, expressing the task as a sequence tagging problem, where tokens are to be classified as beginning a segment or not, or, in the case of PDTB corpora, being part of a discourse connective signalling a relation between textual arguments. Segmentation in itself has also shown a lot of potential as an auxiliary task in machine translation (Chen et al., 2020) and summarization (Xu et al., 2020), independently of discourse parsing.

DISRPT 2021 shared task reproduces the same setting as DISRPT 2019, with some additional data and minor modifications of the original datasets, segmentation as task 1, connective identification as task 2, and adds the prediction of relations between segments, assuming those are known, as task 3.

The following shows examples from the English datasets, illustrating respectively task 1 and task 2, with intended units to recover marked between brackets:

- *[Three seats currently are vacant] [and three others are likely to be filled within a few years] (...)*
- *[But] [in the end] his resignation as Chancellor of the Exchequer may be a good thing (...)*

In the first case, tokens "Three" and "and" would be marked as segment beginnings. In the second case, where the target connectives are "but" and "in the end", beginning tokens "but" and "in" would be marked "B" (begin) and "the", "end", would be marked as "I" (inside). In both examples, other tokens would be marked "out".

Since sentences are almost always discourse units in existing frameworks, segmentation can be seen as two sub-problems: detecting sentences, and detecting intra-sentence segment boundaries. To reflect this, the 2019 shared task introduced two

[1] Note that authors Philippe Muller and Chloé Braud were part of the organization of the shared task.

Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021), pages 22–32
November 11, 2021. ©2021 Association for Computational Linguistics

sub-tasks for segmentation and connective identification: either sentence-level segmentation, with sentence boundaries given (when annotated, or provided with a sentence splitter otherwise), also informed by syntactic parsing of sentences, or without any of that information.

With the exception of systems presented at DISRPT 2019 (Bourgonje and Schäfer, 2019; Yu et al., 2019; Muller et al., 2019), existing work on segmentation always assumed gold sentences, e.g. (Wang et al., 2018; Lukasik et al., 2020).

One interesting aspect of such a task is the availability of comparable data in different languages. This has been leveraged in the past for segmentation (Braud et al., 2017b), but not in the past 2019 campaign, where the best system relied on fine-tuning a contextual language model for each language separately, albeit using the same multilingual embedding model (Muller et al., 2019).

Here we propose to build on the previous DISRPT best system and exploit the availability of multiple corpora for the same language, or the same family of languages (romance, germanic) to augment training of dedicated models.

Combining this approach with a few adjustments to the base model, we manage to improve on many datasets compare to the previous best DISRPT systems, with a mean difference in F1 score of 0.46% and 1.24% on segment boundary detection for sentence and document level respectively, and 3.07% on connective detection for sentences (we didn't improve results at document level for discourse connectives).

Within the Disrpt 21 campaign the system ranks 3rd overall, very close to the 2nd system, but with a significant gap with respect to the best system, which uses a rich set of additional features.

Our code and instructions to reproduce the experiments are available online.[2]

2 Related work

Discourse segmentation appeared as an NLP task with the creation of the first annotated RST documents in English, and was primarily rule-based (Marcu, 2000). Since then the literature on discourse parsing generally assumed that elementary discourse units (discourse segments) were given, with only a handful of exceptions (Soricut and Marcu, 2003; Fisher and Roark, 2007; Tofiloski et al., 2009; Hernault et al., 2010; Joty et al., 2015), until more recent neural-based work (Wang et al., 2018; Lukasik et al., 2020), still at the sentence level, and always on English or Mandarin. Only the work of (Braud et al., 2017b,c) have considered more varied languages, after the creation of a few different datasets in the past ten years (see the Data section below).

The DISRPT 2019 workshop introduced a more general evaluation framework for discourse segmentation with a shared task considering multilingual data, and segmentation both at the sentence and document level (Zeldes et al., 2019). The best system (Muller et al., 2019) at both granularities (sentence and document) used a sequential tagging model fine-tuned on contextual embeddings.

Multi-lingual discourse parsing is also becoming more popular, see for instance (Braud et al., 2017a; Chen et al., 2020), in which it is seen as a form of multi-task learning problem, but this was not applied to discourse segmentation. In other NLP subfields, leveraging availability of corpora in different languages for the same tasks is an active area of research, with different strategies for combining tasks and languages, using meta-learning and complex sampling strategies (Nooralahzadeh et al., 2020; Tarunesh et al., 2021). A simpler approach that inspired us here, due to (Dehouck and Denis, 2019), is to use the relations between close languages to guide the training process on a task: a generic model is trained on groups of languages, further refined with models by subgroups and finally fine-tuned on individual languages.

3 Data

The 2021 shared task provides 16 corpora annotated either with discourse boundaries (13) or discourse connectives in the case of PDTB corpora (3), with the RST Farsi corpus as a surprise dataset. This covers 11 different languages, mostly indo-european languages, and with a majority of european languages: 3 romance (Spanish, French, Portuguese), 3 germanic (English, German, Dutch), Russian, the only non indo-european being Turkish, Basque, and Mandarin. Some of the datasets depend on licences for the underlying text corpus, and are not freely available. We had licences for all of them except the Mandarin corpus (zho.pdtb.cdtb), provided by the organizers for the evaluation of the task. Except Farsi, all the datasets were present in the DISRPT 2019 shared task, but the russian

[2] https://gitlab.irit.fr/melodi/andiamo/discoursesegmentation/discut

Corpus	Lang	# Doc.			Sent seg	# Sents.	# Units
		Train	Dev	Test		Train	Train
Connectives — PDTB							
eng.pdtb.pdtb	en	1,992	79	91	manual	44,563	23,850
tur.pdtb.tdb	tr	159	19	19	manual	25,080	6,841
zho.pdtb.cdtb	zh	125	21	18	manual	2,049	1,034
EDUs — RST							
eng.rst.rstdt	en	309	38	38	manual	6,672	17,646
eng.rst.gum	en	78	18	18	manual	3,600	5,012
deu.rst.pcc	de	142	17	17	manual	1,773	2,449
eus.rst.ert	eu	84	28	28	manual	991	1,713
far.rst.prstc	far	120	15	15	stanza	1713	4607
nld.rst.nldt	nl	56	12	12	manual	1,202	1,679
por.rst.cstn	pt	110	14	12	manual	1,595	3,916
rus.rst.rrt	ru	272	30	30	stanza	18932	34682
spa.rst.stb	es	203	32	32	manual	1,577	2,474
spa.rst.sctb	es	32	9	9	manual	304	473
zho.rst.sctb	zh	32	9	9	manual	344	473
EDUs — SDRT							
eng.sdrt.stac	en	29	6	6	manual	7,689	8,843
fra.sdrt.annodis	fr	64	11	11	manual	880	2,411

Table 1: Descriptions of all corpora, according to the underlying theoretical framework. The tasks consist in finding connectives in the PDTB datasets, or the Elementary Discourse Units (segments) in RST and SDRT datasets. For each corpora are listed the number of documents in each split, the number of sentences and annotated units (connective tokens or segment boundaries in the training set, and whether the gold sentences were manually annotated or given by a parser.

dataset has been extended since, and some corpora without gold syntax annotations have been reparsed with Stanza or Spacy to provide morpho-syntaxic information for the sentence-internal subtasks.

All tasks are considered as sequence tagging, and annotated as such: for segmentation, each token is marked as being a segment boudary or not, and for connectives, which can span multiple tokens, the annotation follows the BIO convention with three labels Begin/Inside/Out for each token.

Datasets are not homogenous, as they were annotated along different principles based on three competing theoretical frameworks:

- Rhetorical Structure Theory (Mann and Thompson, 1988), which assumes a linear segmentation of documents in discourse units (no overlaps), which are then related in constituant tree structures. This is followed in the majority of corpora (11).

- Segmented Discrouse Representation Theory (Asher and Lascarides, 2003), which allows for embedded segments, and were linearized here for homogeneity of the task: a segment embedded in another one was re-annotated as forming 3 three segments. This is the case of two corpora.

- Penn Discourse Treebank (Prasad et al., 2008), which annotates discourse connectives and their arguments in a discourse relation. This gives rise to a different annotation scheme as noted above, as the task is only to locate the connective.

Table 1 presents the size of all corpora, separated by theoretical framework, and expressed in number of documents, number of sentences, and number of discourse units (segments or connectives). Note that the different corpora greatly vary in sizes and annotations. One dataset is annotated on chat conversations (STAC), while all the others are on written text, mostly news or encyclopedic.

Experimental results cannot be compared to previous multi-lingual segmentation efforts (Braud et al., 2017b), because some of the corpora have been revised (Gum, RRT) or not taken entirely (CSTN), and some have been added (TDB, PRSTC), but should be quite close to the DISRPT 2019 evaluation, as only the Russian and the GUM corpora have been extended (and there is an additional dataset).

More details can be found about all datasets in the following publications: English RSTDT (Carlson et al., 2001), PDTB (Prasad et al., 2008), SDRT-STAC (Asher et al., 2016) and GUM (Zeldes, 2016), Spanish RST (2)(da Cunha et al., 2011; Cao et al., 2018) Mandarin Chinese (Zhou et al., 2014; Cao et al., 2018), German RST (Stede and Neumann, 2014), French SDRT-Annodis (Afantenos et al., 2012), Basque RST (Iruskieta et al., 2013), Portuguese RST (Cardoso et al., 2011), Russian RST (Pisarevskaya et al., 2017), Turkish PDTB (Zeyrek et al., 2013) Dutch RST (Redeker et al., 2012) and Persian RST (Shahmohammadi et al., 2021).

4 Approach

In this paper we want to leverage combinations of multiple datasets for training, not only with corpora for the same language and task, but also with languages from the same families.

4.1 Base architecture

We started from the architecture that showed the best results on almost all languages and configurations at DISRPT 2019, namely (Muller et al., 2019), which is built around BERT (Devlin et al., 2019), a contextual language model that is easy to fine-tune on sequence tagging problems. The original architecture combined BERT contextual embeddings to the output of CNN filters over characters of each word piece, that were then fed to single-layer BiLSTM layer for the final prediction. The model is initialized with the multilingual BERT model, then fine-tune on all corpora separately as sequence tagging tasks. The original implementation used the AllenNLP library (Gardner et al., 2017), and so does our implementation.

Since BERT has a limitation on the number of word pieces it can take as input, a preprocessing step must be taken for document-level segmentation. In (Muller et al., 2019), the core-nlp library was used to predict sentence boundaries, and use this information, while we used the more recent Stanza library by the same team for that purpose (Qi et al., 2020).

We explored potential improvements for that architecture, swapping the multi-lingual pretrained language model XLM (Conneau and Lample, 2019), or adding another layer to the BiLSTM stage. The final configuration was chosen based on preliminary experiments on some of the datasets, evaluated on their respective development sets. They showed that XLM didn't help, but the extra layer of LSTM could. Changes to other hyperparameters didn't improve these preliminary results so we kept them as in the original model. Details of the parameters can be found in the declarative config file of the model, also added as supplementary material.

Input	Corpus	P	R	F1
conll	deu.rst.pcc	92.58	95.27	93.91
	eng.rst.gum	94.06	90.19	92.08
	eng.rst.rstdt	96.35	95.38	95.86
	eng.sdrt.stac	94.19	95.49	94.84
	eus.rst.ert	87.23	83.75	85.46
	fas.rst.prstc	91.18	91.49	91.33
	fra.sdrt.annodis	87.13	90.11	88.59
	nld.rst.nldt	95.52	93.29	94.40
	por.rst.cstn	90.63	92.06	91.34
	rus.rst.rrt	86.04	83.83	84.92
	spa.rst.rststb	91.80	93.56	92.67
	spa.rst.sctb	85.57	80.58	83.00
	zho.rst.sctb	93.02	77.67	84.66
	mean	91.18	89.44	90.24
doc	deu.rst.pcc	94.51	93.82	94.16
	eng.rst.gum	91.80	91.68	91.74
	eng.rst.rstdt	93.48	94.94	94.20
	eng.sdrt.stac	86.10	87.52	86.81
	eus.rst.ert	87.56	85.23	86.38
	fas.rst.prstc	91.21	91.84	91.52
	fra.sdrt.annodis	87.52	90.83	89.14
	nld.rst.nldt	93.91	94.46	94.19
	por.rst.cstn	92.43	91.11	91.77
	rus.rst.rrt	83.09	81.65	82.37
	spa.rst.rststb	92.54	94.75	93.63
	spa.rst.sctb	73.23	90.29	80.87
	zho.rst.sctb	64.18	83.50	72.57
	mean	87.04	90.12	88.41

Table 2: Segmentation results on the development sets, for both sentence (conll) and document (doc) levels.

4.2 Dataset grouping

Since the shared task involves multiple datasets with the same language (2 for Spanish RST, 2 for English RST), we assumed it would be beneficial to combine them for training. Datasets in the same language are not necessarily consistent in their annotation, but we hypothetize that they have enough commonalities to help training. We also took inspiration from work on multi-lingual syntactic parsing where a lot of corpora follow the same formalism, and where past work has tried to use commonalities between different languages, particularly the approach of (Dehouck and Denis, 2019) in which the phylogenic tree of languages guides the training process: a generic model is trained on groups of languages, further refined in models by subgroups and finally fine-tuned on individual languages. DISRPT datasets are not numerous enough to provide a complex tree of languages, but we can still take advantage of the presence of languages that are relatively close: romance languages (3 languages and 4 datasets for segmentation), germanic languages (3 languages and 5 datasets for segmentation).

Input	Corpus	P	R	F1
conll	eng.pdtb.pdtb	92.27	88.55	90.37
	tur.pdtb.tdb	80.54	84.50	82.47
	zho.pdtb.cdtb	77.54	72.94	75.17
	mean	83.45	82.00	82.67
doc	eng.pdtb.pdtb	93.00	89.86	91.40
	tur.pdtb.tdb	80.66	85.66	83.08
	zho.pdtb.cdtb	71.27	64.69	67.82
	mean	81.64	80.07	80.77

Table 3: Connective identification results on the development sets, for both sentence (conll) and document (doc).

5 Results

5.1 Base model

The modification of ToNy, the best system from DISRPT 2019, gives us our base system on which we will build with multi-corpora training in a second stage.

We report the results of these systems on segmentation in Table 2, and on discourse connectives identification in Table 3, with precision, recall and F1 score on the detection of segment boundary tokens, and discourse connectives. A first comparison with respect to each best subsystem from 2019 for the 4 subtasks is given in Table 6. That means ToNy for segmentation (both intra-sentential and plain), discourse connectives (plain), and Gumdrop for discourse connectives (Conll input). We can see that on average on all datasets, the base system gains +0.5, mostly due to its improvements on the plain document segmentation (connective detection only involves 3 datasets). We left out the surprise dataset for 2021 from that evaluation, since we do not have a comparison point. Note that results on this new dataset are good and consistent with the other corpora. Lower results are obtained on smaller datasets for obvious reasons (Spanish sctb, Chinese Mandarin sctb, and to a lesser extent Basque and French). Models trained without sentence information perform a little worse, as expected, with -1.6% on average, again with wider gaps for small corpora.

We do not show the breakup by dataset for the comparison with DISRPT 2019, but there is a lot of variances in results, with differences ranging from -4.5 (French Annodis) to +6.85 (Mandarin SCTB), and not necessarily only on small corpora.

Corpus	Group	P	R	F1
spa/rststb	self	91.80	93.56	92.67
	SPA	94.24	93.79	94.02
	SPO	93.81	94.03	93.92
	ROM	93.66	91.65	92.64
	FT	94.30	94.75	**94.52**
spa/sctb	self	85.57	80.58	83.00
	SPA	83.64	89.32	86.38
	SPO	86.24	91.26	**88.68**
	ROM	81.98	88.35	85.05
	FT	88.89	85.44	87.13
por/cstn	self	90.63	92.06	91.34
	SPO	92.88	89.05	90.92
	ROM	90.14	90.00	90.07
	FT	90.98	92.86	**91.91**
fra/annodis	self	87.13	90.11	**88.59**
	ROM	91.12	83.09	86.92

Table 4: Intra sentential results on romance development datasets, with different training setups: SPA means the grouping of both spanish datasets for training, SPO the grouping of spanish and portuguese data, ROM the addition of French to the group. FT means a model fine-tuning the SPO model. Lines with "self" are just copied from the basic evaluation (training on the dataset only) for comparison. In bold are indicated the best F1 results per dataset on their development set, and the corresponding model was thus kept for the final evaluation.

5.2 Multi-dataset training

As shown above, a lot of smaller datasets have lower results than larger ones, which is to be expected. We present here the result when applying the strategy described in Section 4.2. We tried it on two groups of languages: romance languages, and germanic languages.

For romance languages, since there are two spanish RST corpora, we grouped them to trained a "spanish" more generic model, then we trained a model with all Spanish and Portuguese data, then a generic romance model by including also French. We then used those models for predictions on the datasets respective development part. We did something similar with germanic languages, grouping all English datasets into one, joining the Dutch and German datasets into another one, and finally training a generic germanic model on all of them. Following a procedure similar to what was done in (Dehouck and Denis, 2019), we also fine-tuned some of these models on the individual datasets before using them for prediction. Lack of time during the campaign prevented us from trying all combinations and all datasets like this, but we tested this on all Spanish and English datasets, respectively fine-tuning the global Spanish-Portuguese on Spanish and Portuguese datasets (since it showed a good compromise on the dev sets) and the global English model on all English datasets. Due to time constraints, we tested this only on one type of input, the sentence level (conll files).

Results are presented in Tables 4 and 5. For the final evaluation, we kept for each dataset the model that performed better on the dataset development test.

Corpus	Group	P	R	F1
deu.rst.pcc	self	92.58	95.27	93.91
	GD	95.51	92.73	94.10
	GER	96.64	94.18	**95.40**
nld.rst.nldt	self	95.52	93.29	94.40
	GD	96.95	92.71	**94.78**
	GER	94.44	94.17	94.31
eng.rst.gum	self	94.06	90.19	92.08
	ENG	94.08	92.64	**93.36**
	GER	94.11	92.42	93.26
	FT	93.63	92.12	92.87
eng.rst.rstdt	self	96.35	95.38	**95.86**
	ENG	96.27	94.77	95.51
	GER	95.47	95.10	95.29
	FT	94.96	96.55	95.75
eng.sdrt.stac	self	94.19	95.49	94.84
	ENG	95.80	92.81	94.28
	GER	95.98	93.07	94.50
	FT	95.55	94.97	**95.26**

Table 5: Intra sentential results on germanic development datasets, with different training setups: GD means the grouping of German and Dutch for training, GER the grouping of English, German and Dutch, and ENG the grouping of all 3 English datasets. FT means a model fine-tuning the global ENG model on the specific corpus training set. Lines with "self" are just copied from the basic evaluation (training on the dataset only) for comparison. In bold are indicated the best F1 results per dataset on their development set.

5.3 Final evaluation

For the final evaluation of the campaign, every team provided their code and instructions for reproducing the experiments, and one member of the organization team reproduced entirely the experiments of one model they were not involved with (two of the four teams involved organization members, including the present system).

We reported the official scores on the test sets in Tables 7 for segmentation and 8 for connective detection.

Overall our system is ranked 3rd out of 4, with results very close to the 2nd-ranked system (Segformers), with only 0.15% difference on average for treebanked data segmentation, and 0.45% on plain data segmentation. The gap with the first ranked system (DiscoDisco) is 0.7% on average on treebanked segmentation and 1.28% on plain text segmentation. Our system performed less well on connective detection, especially with respect to models taking dependency between labels into account (such as a CRF in the case of DiscoDisco): about 4.5% less than Segformers and 6% less than DiscoDisco, mostly due to lower results on the Mandarin dataset.[3]

It is to be noted that we achieved our best results in relation to the other systems with datasets used in cross-training with similar languages (see Sections 4.2 and 5.2), and for which we observed on the development data that it had an impact: German and Spanish corpora, in the case of treebanked data (conll), since we didn't have time to try this strategy

[3] Full results for all systems are not shown for space constraint reasons, but are displayed on the Shared task website at https://sites.google.com/georgetown.edu/disrpt2021/results and are summarized in the introductory paper to the Shared Task proceedings.

task	input	Δbase/19	Δ grouped/19
pdtb	conll	3.07	3.07
pdtb	tok	-1.31	-1.31
seg	connl	-0.41	0.46
seg	tok	1.24	1.24
all	all	0.51	0.85

Table 6: Mean comparison of our base system and our improved system with respect to the best Disrpt 2019 system for each sub task, and the mean on all 16 datasets. This is the only evaluation we made on the test set prior to the official evaluation. Note that grouped training was tested only on conll segmentation, so the other scores are just copied from the base system.

on plain documents. We have the best scores on these languages, but note that cross-training cannot really explain the good results on German, since results are surprisingly similar between treebanked and plain data.

6 Conclusion

We presented an approach for discourse segmentation and discourse connective identification, both at the sentence and document level, in a multi-lingual and multi-formalism context. Building on the successful architecture from the 2019 DISRPT shared task, we leverage datasets in the same or similar languages to augment training data and improve on the best systems from the previous campaign, on 3 of the 4 sub-tasks. While below the best system which uses a rich set of features over a similar architecture, we still manage to have the best scores on some of the languages where we experimented with cross-lingual training: German and Spanish for sentence-internal segmentation.

Due to time constraints, we could not fully explore all potentially useful language combinations and fine-tuning on specific datasets, that could help improve on the tasks, and give insights on how different languages help each other addressing the discourse segmentation problem. Further progress on multi-lingual embeddings or alignments of different embeddings could be a source of future investigations, as well as more elaborate multi-lingual training procedures.

Acknowledgments

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, and through the project SLANT (ANR-19-CE23-0022) and Quantum (ANR-19-CE23-0025).

treebanked	P	R	F1
deu.rst.pcc	98.91	92.52	**95.61**
eng.rst.gum	93.27	93.65	93.46
eng.rst.rstdt	96.16	95.99	96.08
eng.sdrt.stac	97.41	92.37	94.82
eus.rst.ert	90.04	83.11	86.44
fas.rst.prstc	93.54	90.75	92.12
fra.sdrt.annodis	87.26	88.67	87.96
nld.rst.nldt	94.15	95.27	94.71
por.rst.cstn	90.31	94.44	92.33
rus.rst.rrt	88.19	81.10	84.50
spa.rst.rststb	92.04	93.04	**92.54**
spa.rst.sctb	85.39	90.48	**87.86**
zho.rst.sctb	92.48	73.21	81.73
mean	92.24	89.58	90.78
plain	P	R	F1
deu.rst.pcc	94.67	96.60	**95.63**
eng.rst.gum	92.76	89.54	91.13
eng.rst.rstdt	93.39	94.50	93.94
eng.sdrt.stac	85.30	87.01	86.14
eus.rst.ert	91.45	83.78	87.45
fas.rst.prstc	93.59	89.40	91.45
fra.sdrt.annodis	89.90	86.41	88.12
nld.rst.nldt	94.35	93.79	94.07
por.rst.cstn	93.36	91.83	92.59
rus.rst.rrt	83.60	84.01	83.80
spa.rst.rststb	92.19	89.78	90.97
spa.rst.sctb	78.65	89.88	**83.89**
zho.rst.sctb	68.11	75.00	71.39
mean	88.56	88.58	88.51

Table 7: Final official segmentation results on the test set, reproduced by the organizers. In bold, F1 scores for which our system has the best performance of the shared task.

treebanked	P	R	F1
PDTB	93.32	88.67	90.94
TDB	90.55	86.93	88.70
CDTB	84.43	66.03	74.10
mean	89.43	80.54	84.58
plain			
PDTB	88.84	92.09	90.43
TDB	90.12	88.10	89.10
CDTB	77.40	72.44	74.83
mean	85.45	84.21	84.79

Table 8: Final official connective detection results on the test set, reproduced by the organizers.

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Lydia-Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Pery-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In *Proceedings of LREC*.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of LREC*.

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

Peter Bourgonje and Robin Schäfer. 2019. Multilingual and cross-genre discourse unit segmentation. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 105–114, Minneapolis, MN. 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.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017c. Does syntax help discourse segmentation? not so much. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442, Copenhagen, Denmark. Association for Computational Linguistics.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The RST spanish-chinese treebank. In *Proceedings of LAW-MWE-CxG*.

Paula C.F. Cardoso, Erick G. Maziero, María Lucía Castro Jorge, Eloize R.M. Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago A. S. Pardo. 2011. CSTNews - 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.

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

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

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*.

Mathieu Dehouck and Pascal Denis. 2019. Phylogenic multi-lingual dependency parsing. 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 192–203, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Seeger Fisher and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 488–495, Prague, Czech Republic. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue Discourse*, 1(3):1–33.

Mikel Iruskieta, María J. Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez-Dios, Mikel Lersundi, and Oier Lopez de la Calle. 2013. The RST Basque Treebank: an online search interface to check rhetorical relations. In *Proceedings of the 4th Workshop RST and Discourse Studies*.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, 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.

Daniel Marcu. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics*, 26(3):395–448.

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.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, A Nasedkin, S Nikiforova, I Pavlova, and A Shelepov. 2017. Towards building a discourse-annotated corpus of russian. In *Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies "Dialogue"*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *The Sixth International Conference on Language Resources and Evaluation*, pages 2961 – 2968, Marrakech, Morocco. ELRA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a Dutch text corpus. In *Proceedings of LREC*.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian rhetorical structure theory. *CoRR*, abs/2106.13833.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 228–235.

Manfred Stede and Arne Neumann. 2014. Potsdam Commentary Corpus 2.0: Annotation for discourse research. In *Proceedings of LREC*.

Ishan Tarunesh, Sushil Khyalia, Vishwajeet Kumar, Ganesh Ramakrishnan, and Preethi Jyothi. 2021. Meta-learning for effective multi-task and multilingual modelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3600–3612, Online. Association for Computational Linguistics.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80, Suntec, Singapore. Association for Computational Linguistics.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967, Brussels, Belgium. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. GumDrop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 133–143, Minneapolis, MN. Association for Computational Linguistics.

Amir Zeldes. 2016. The GUM corpus: Creating multilayer resources in the classroom. In *Proceedings of LREC*.

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.

Deniz Zeyrek, Demirsahin Isın, A. Sevdik-Çallı, and Ruket Çakıcı. 2013. Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue and Discourse*.

Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese discourse treebank 0.5 LDC2014T21. Web Download. Philadelphia: Linguistic Data Consortium.