

DisCoDisCo at the DISRPT2021 Shared Task: A System for Discourse Segmentation, Classification, and Connective Detection

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, Amir Zeldes
Corpling Lab

Georgetown University

{lg876, sb1796, yl879, sp1184, yz565, az364}@georgetown.edu

Abstract

This paper describes our submission to the DISRPT2021 Shared Task on Discourse Unit Segmentation, Connective Detection, and Relation Classification. Our system, called DisCoDisCo, is a Transformer-based neural classifier which enhances contextualized word embeddings (CWEs) with hand-crafted features, relying on tokenwise sequence tagging for discourse segmentation and connective detection, and a feature-rich, encoder-less sentence pair classifier for relation classification. Our results for the first two tasks outperform SOTA scores from the previous 2019 shared task, and results on relation classification suggest strong performance on the new 2021 benchmark. Ablation tests show that including features beyond CWEs are helpful for both tasks, and a partial evaluation of multiple pre-trained Transformer-based language models indicates that models pre-trained on the Next Sentence Prediction (NSP) task are optimal for relation classification.*

1 Introduction

Recent years have seen tremendous advances in NLP systems’ ability to handle discourse level phenomena, including discourse unit segmentation and connective detection (Zeldes et al., 2019) as well as discourse relation classification (e.g. Lin et al. 2014; Braud et al. 2017a; Kobayashi et al. 2021). For segmentation and connective detection, the current state of the art (SOTA) is provided by models using Transformer-based, pretrained contextualized word embeddings (Muller et al., 2019), focusing on large context windows without implementation of hand-crafted features. For relation classification, SOTA performance on the English RST-DT benchmark (Carlson et al., 2003) has been achieved by neural approaches (Guz and Carenini,

2020; Kobayashi et al., 2020; Nguyen et al., 2021; Kobayashi et al., 2021). For PDTB-style data, the 2015 and 2016 CoNLL shared tasks on shallow discourse parsing (Xue et al., 2015, 2016) have motivated work on both explicit (e.g. Kido and Aizawa 2016) and implicit (e.g. Liu et al. 2016; Wang and Lan 2016; Rutherford et al. 2017; Kim et al. 2020; Liang et al. 2020; Zhang et al. 2021) discourse relation classification in English PDTB-2 (Prasad et al., 2008) and PDTB-3 (Prasad et al., 2019) as well as on the PDTB-style Chinese newswire corpus (CDTB, Zhou and Xue 2012; Zhou et al. 2014) such as Schenk et al. (2016) and Weiss and Bajec (2016).

Our system for DISRPT 2021, called DisCoDisCo, (**D**istrict of **C**olumbia **D**iscourse **C**ognoscente) extends the current SOTA architecture by introducing hand-crafted categorical and numerical features that represent salient aspects of documents’ structural and linguistic properties. While Transformer-based contextualized word embeddings (CWEs) have proven to be rich in linguistic features, they are not perfect (Rogers et al., 2020), and there are some textual features—such as position of a sentence within a document, or the number of identical words occurring in two discourse units—which are difficult or impossible for a typical Transformer-based CWE model to know. We therefore supplement CWEs with hand-crafted features in our model, with special attention paid to features we expect CWEs to have a poor grasp of.

We implement our system with a pretrained Transformer-based contextualized word embedding model at its core, and dense embeddings of our hand-crafted features incorporated into it. Our exact approach varies by task: we use a tokenwise classification approach for EDU segmentation, a CRF-based sequence tagger for connective detection, and a BERT pooling classifier for relation classification. Our system is implemented in Py-

*Full disclosure: our team includes both organizers and dataset annotators from the shared task. All code is available at <https://github.com/gucorpling/DisCoDisCo>

Torch (Paszke et al., 2019) using the framework AllenNLP (Gardner et al., 2018). Our results show SOTA scores exceeding comparable numbers from the 2019 shared task, and ablation studies indicate the contribution of features beyond CWEs.

2 Previous Work

Segmentation and Connective Detection Following the era of rule-based segmenters (e.g. Marcu 2000, Thanh et al. 2004), Soricut and Marcu (2003) used probabilistic models over constituent trees for token-wise binary classification (i.e. boundary/no-boundary). Sporleder and Lapata (2005) used a two-level stacked boosting classifier on chunks, POS tags, tokens and sentence lengths, among other features. Hernault et al. (2010) used an SVM over token and POS trigrams as well as phrase structure trees.

More recently, Braud et al. (2017b) used bi-LSTM-CRF sequence labeling on dependency parses, with words, POS tags, dependency relations, parent, grandparent, and dependency direction, achieving an F_1 of 89.5 on the English RST-DT benchmark (Carlson et al., 2003) with parser-predicted syntax. Approaches using CWEs as the only input feature (Muller et al., 2019) have achieved an F_1 of 96.04 on the same dataset with gold sentence splits and 93.43 without, while for some smaller English and non-English datasets, approaches incorporating features and word embeddings remain superior (e.g. for English STAC and GUM, as well as Dutch RST data, Yu et al. 2019; and for Chinese, Bourgonje and Schäfer 2019; for more on these datasets see below).

For connective detection, Pitler and Nenkova (2009) used a MaxEnt classifier with syntactic features extracted from gold Penn Treebank (Marcus et al., 1993) parses of PDTB (Prasad et al., 2008) articles. Patterson and Kehler (2013) presented a logistic regression model trained on eight relation types from PDTB, with features in three categories: *Relation-level* such as the connective signaling the relation; *Argument-level* such as size or complexity of argument spans; and *Discourse-level* features, targeting dependencies between the relation and its neighboring relations in the text (cf. our approach to featurizing overall utilization of argument spans in the data below). Polepalli Ramesh et al. (2012) used SVM and CRF to identify connectives in biomedical texts (Prasad et al., 2011), with features such as POS tags, dependencies and

domain-specific semantic features included several biomedical gene/species taggers, in addition to predicted biomedical NER features.

Current SOTA approaches rely on sequence labeling in a BIO scheme with CWEs from either plain text (Muller et al., 2019) or integrating word embeddings and dependency tree features (POS, dependencies, phrase spans, Yu et al. 2019), depending on the dataset and availability of gold standard features.

Discourse Relation Classification Generally speaking, discourse relation classification assigns a relation label to two pieces of texts from a set of predefined coherence or rhetorical relation labels (Stede, 2011), which varies across different discourse frameworks, corpora, and languages. Given different perspectives and theoretical frameworks, the implementation and evaluation of the relation classification task varies considerably.

In Rhetorical Structure Theory (RST, Mann and Thompson 1988), discourse relations hold between spans of text and are hierarchically represented in a tree structure (Zeldes, 2018). Performance is evaluated and reported using the micro-averaged, standard Parseval scores for a binary tree representation, following Morey et al. (2017). Current SOTA performance (Kobayashi et al., 2021) on the English RST-DT benchmark (Carlson et al., 2003) with gold segmentation achieved a micro-averaged original Parseval score of 54.1 by utilizing both a span-based neural parser (Kobayashi et al., 2020) and a two-staged transition-based SVM parser (Wang et al., 2017) as well as leveraging silver data.

Since PDTB is a lexically grounded framework, discourse relation classification is also called sense classification in PDTB-style discourse parsing: a sense label is assigned to the discourse connective between two text spans when a discourse connective is present (i.e. explicit relation classification) or a label is assigned to an adjacent pair of sentences when no discourse connective is present (i.e. implicit relation classification) (Jurafsky and Martin, 2020). Explicit relation classification is easier as the presence of the connective itself is considered the best signal of the relation label. Most systems from the 2016 CoNLL shared task on shallow discourse parsing adopted machine learning techniques such as SVM and MaxEnt with hand-crafted features (Xue et al., 2016). For instance, for the English PDTB-2 (Prasad et al., 2008), Kido and Aizawa (2016) achieved the best performance

(an $F_1 = 90.22$) on the explicit relation classification task by implementing a majority classifier and a MaxEnt classifier while Wang and Lan (2016) achieved the best performance ($F_1 = 40.91$) on implicit relation classification using a convolutional neural network. Wang and Lan (2016) also achieved the best performance on the Chinese CDTB dataset (Zhou and Xue, 2012) in the implicit relation classification task.

More recent work on implicit relation classification has adopted a graph-based context tracking network to model the necessary context for interpreting the discourse and has gained better performance on PDTB-2 (Zhang et al., 2021). In addition, the increase in the number of implicit relation instances in PDTB-3 (Prasad et al., 2019) has sparked more interest in exploring their recognition, such as Kim et al. (2020) and Liang et al. (2020). Kim et al. (2020) presented the first set of results on implicit relation classification for both top-level senses (four labels) and more fined-grained level-2 senses (amounting to 11 labels) in PDTB-3 from two strong sentence encoder models using BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019). Due to the novelty of the DISRPT 2021 relation classification task, which combines implicit and explicit relation classification across frameworks for an unlabeled graph structure, comparable scores do not yet exist at the time of writing.

3 Approach

Our system comprises two main components: one targeting segmentation and connective detection using neural sequence tagging (as binary classification and BIO tagging respectively), and one targeting relation classification using BERT (Devlin et al., 2019) fine-tuning. We further enhance both components with the use of hand-crafted categorical and numeric features by encoding them in dense embeddings and introducing them into our neural models.

3.1 Segmentation and Connective Detection

Our model for segmentation and connective detection is structured as a sequence tagging model, as might be used for a task like POS tagging or entity recognition: the text is embedded, encoded with a single bi-LSTM, and decoded.

In the embedding layer, we rely on three kinds of embeddings: bi-LSTM encoded character embeddings ($d = 64$); language-specific fastText

(Bojanowski et al., 2017) static word embeddings ($d = 300$); and language-specific contextualized word embeddings from pretrained models posted publicly on HuggingFace’s model registry at huggingface.co and used via HuggingFace’s transformers library (Wolf et al., 2020) ($d = 768/1024$). The fastText embeddings are kept frozen during training, but the pretrained Transformer model’s parameters are trainable, at a lower learning rate. Average pooling is used to obtain word-level representations from CWE sub-word representations. Multiple CWEs were evaluated for each language, and the one that yielded the best performance on the validation splits of the corpora for that language was selected, shown in Table 1. These three representations are concatenated, yielding a vector of size d_{emb} for each word.

In the next layer, we encode the embeddings along with a variety of features and a representation of the preceding and following sentence (see below). The features we compute are tokenwise, and cover a variety of grammatical and textual information that we expected would be useful for the task. Some of the features are described in Table 2. In order to convert these features into tensors, every categorical feature is embedded in a space as big as the square root of the total number of labels for the feature, and every numerical feature is log scaled. This yields an additional d_{feat} dimensions for each word.

In addition to the features, we also compute a representation of the current sentence’s two neighboring sentences by embedding them and using a bi-LSTM to summarize them into a relatively low-dimensional ($d_{neighbors} = 400$) vector, which is concatenated onto every word’s vector. Combining the feature dimensions and the neighboring sentences’ dimensions, our input to the encoder is of size $d_{enc} = d_{emb} + d_{feat} + d_{neighbors}$.

The sequence is fed through a bi-LSTM, and the label for each token is then predicted either by a linear projection layer or conditional random fields: CRF is used for connective detection datasets, and the linear projection layer is used for segmentation datasets.^b

For the plain text segmentation scenario, we generate automatic sentence splits and Universal Dependencies parses using the Transformer-based sentence splitter used in the AMALGUM

^bWe initially used a CRF on all datasets, but our experiments showed a small degradation on segmentation datasets when using CRF.

Lng.	Segmentation/Connective Detection	Relation Classification
deu	xlm-roberta-large	bert-base-german-cased
eng	google/electra-large-discriminator	bert-base-cased
eus	ixa-ehu/berteus-base-cased	ixa-ehu/berteus-base-cased
fas	HooshvareLab/bert-fa-base-uncased	HooshvareLab/bert-fa-base-uncased
fra	xlm-roberta-large	dbmdz/bert-base-french-europeana-cased
nld	pdelobelle/robbert-v2-dutch-base	GroNLP/bert-base-dutch-cased
por	neuralmind/bert-base-portuguese-cased	neuralmind/bert-base-portuguese-cased
rus	DeepPavlov/rubert-base-cased	DeepPavlov/rubert-base-cased-sentence
spa	dccuchile/bert-base-spanish-wwm-cased	dccuchile/bert-base-spanish-wwm-cased
tur	dbmdz/bert-base-turkish-cased	dbmdz/bert-base-turkish-cased
zho	bert-base-chinese	hfl/chinese-bert-wwm-ext

Table 1: CWE Models used, by language. All models were obtained from huggingface.co’s registry. Note that there is one exception for relation classification: on `eng.sdrt.stac`, `bert-base-uncased` is used.

Feature	Type	Example	Description
UPOS tag	Cat.	PROPN	UD POS tag
XPOS tag	Cat.	NNP	Language-specific POS tag
UD deprel	Cat.	advmod	UD dependency relation
Head distance	Num.	5	Distance from a word to its head in its UD tree
Sentence type	Cat.	subjunctive	Captures mood and other high-level sentential features
Genre	Cat.	reddit	Genre of a document (where available, as in <code>eng.rst.gum</code>)
Sentence length	Num.	23	Length, in tokens, of a sentence.

Table 2: Summary of 7 of the 12 features used for the segmentation and connective detection module. Every categorical feature is embedded in a space whose size is the square root of the total number of labels for the feature, and numerical features are log scaled.

corpus (Gessler et al., 2020) trained on the tree-banked shared task training data, tagged using Stanza (Qi et al., 2020) and parsed using DiaParser^c (Attardi et al., 2021). For `fas.rst.prstc` and `zho.rst.sctb`, we split the text based on punctuation (on ‘.’, ‘!’, ‘?’ and Chinese equivalents) since experiments revealed that this approach yields better sentence boundaries.

3.2 Relation Classification

Our relation classification module has a simple architecture: a pretrained BERT model is used (again varying by language—cf. Table 1), and a linear projection and softmax layer is used on the output of the pooling layer to predict the label of the relation. The two units involved in every relation are prepared just as if they were being prepared for BERT’s Next Sentence Prediction (NSP) task: a [CLS] token begins the sequence, a [SEP] token separates the two units in the sequence, and another [SEP] token appears at the end of the sequence. As an example, consider this instance from

`eng.sdrt.stac`:

```
[CLS] do we start ? [SEP] no [SEP]
```

Though this model was originally intended as a baseline, further experiments with e.g. a separate encoder proved to be much less competitive.

Our exact choice of pretrained model differs in most cases from the one used in the segmentation and connective detection task, primarily due to superior performance by models that were pretrained using the NSP task and had a pretrained pooler layer. This restricts the LM choice: for example, most models that are styled after RoBERTa (Liu et al., 2019) are not pretrained using an NSP task. We select models using the same process as before, based on optimal performance on the validation (dev) sets of the corpora.

The system is further enhanced with features. First, the direction feature on each relation is encoded using pseudo-tokens: if the direction of the relation is left to right (1>2), we insert the tokens } and > around the first unit. In the example above, the direction of the relation is left to right (1>2),

^c<https://github.com/Unipisa/diaparser>

Feature	Type	Example	Description
Genre	Cat.	reddit	Genre of a document (where available, as in eng.rst.gum)
Children*	Num.	2	No. of child discourse units each unit in the pair has
Discontinuous*	Cat.	false	Whether the unit’s tokens are not all contiguous in the text
Is Sentence*	Cat.	true	Whether the unit is a whole sentence
Length Ratio	Num.	0.3	Ratio of unit 1 and unit 2’s token lengths
Same Speaker	Cat.	true	Whether the same speaker produced unit 1 and unit 2
Doc. Length	Num.	214	Length of the document, in tokens
Position*	Num.	0.4	Position of the unit in the document, between 0.0 and 1.0
Distance	Num.	7	No. of other discourse units between unit 1 and unit 2
Lexical Overlap	Num.	3	No. of overlapping non-stoplist words in unit 1 and unit 2

Table 3: Sample of features used for the relation classification module. Asterisked features apply twice for each instance, once for each unit. Combination of features varies per corpus—see code for full details.

and the resulting sequence with pseudo-tokens is:

```
[CLS] } do we start ? > [SEP] no [SEP]
```

The same is done for right-to-left units, where the characters { and < are used instead, but surrounding the second unit:

```
[CLS] thanks [SEP] < im ok { [SEP]
```

Our motivation in doing this is to represent directionality for the BERT encoder in its native feature space, and experimental data show that it is helpful.

Second, we introduce hand-crafted features in a step between the BERT model’s embedding and encoder layers. Recall that BERT has a static embedding layer which projects each word-piece into its initial vector representation. Just before this input is sent to the Transformer encoder blocks, we expand the sequence by inserting a new vector in between the [CLS] token and the first token of unit 1. This feature vector bears sequence-level information, where categorical and numerical features have been encoded into a vector just as for the segmentation and connective detection module: numerical features are optionally log scaled or binned and embedded, and categorical features are embedded. The remaining dimensions after all features have been added to the vector are padded with 0.

Unlike our approach for segmentation and connective detection, we change which features we use on a per-corpus basis, as preliminary experiments showed that using all features for all corpora can produce significant degradations, which we hypothesize are caused by feature sparseness in the training split leading to overfitting. A sample of the features we used is in Table 3, and the list of which

features were used for which corpus is available in our code.

Specifically, for the LEXICAL OVERLAP feature in the table, we used the freely available stoplists used by the Python library spaCy (Honnibal et al., 2020). The SAME SPEAKER feature has proven very useful in the STAC dataset, which is a corpus of strategic chat conversations (Asher et al., 2016). The DISTANCE feature is used in half of the datasets and has shown effectiveness regardless of annotation framework. Similarly, the POSITION feature has been shown to be beneficial for half of the corpora. The LENGTH RATIO feature proved to be effective for the three PDTB-style datasets. For RST-style corpora, the number of CHILDREN of a nucleus or satellite unit is more effective. Moreover, the DISCONTINUOUS feature has also contributed to performance gain in several RST-style corpora such as eng.rst.gum, eng.rst.rstdt, por.rst.cstn, spa.rst.rststb, zho.rst.sctb, and fas.rst.prstc. The GENRE feature is beneficial in corpora that have a wide range of text types such as eng.rst.gum. The direction feature was also included in the feature vector, as experiments showed that including it was helpful, despite the fact that the pseudo-tokens were already expressing it to the BERT encoder.

4 Results

4.1 Segmentation and Connectives

Table 4 gives scores on EDU segmentation and connective detection in the two shared task scenarios: treebanked and plain text, as well as the best previously reported score and system for datasets which are unchanged from 2019 (see Zeldes et al. 2019 for details). We find strong performance in both the

Corpus	P	R	F ₁	2019 Best F ₁	vs. 2019
deu.rst.pcc	97.07	94.15	95.58	94.99 (ToNy)	0.59
eng.rst.gum	93.90	94.43	94.15	–	–
eng.rst.rstdt	96.39	96.89	96.64	96.04 (ToNy)	0.60
eng.sdrst.stac	96.25	93.63	94.91	95.32 (GumDrop)	−0.41
eus.rst.ert	93.42	87.73	90.46	–	–
fas.rst.prstc	92.79	93.10	92.94	–	–
fra.sdrst.annodis	89.43	90.65	90.02	–	–
nld.rst.nldt	97.50	94.50	95.97	95.45 (GumDrop)	0.52
por.rst.cstn	93.18	95.56	94.35	92.92 (ToNy)	1.43
rus.rst.rst	85.57	86.89	86.21	–	–
spa.rst.rststb	92.53	91.96	92.22	90.74 (ToNy)	1.48
spa.rst.sctb	83.44	81.55	82.48	83.12 (ToNy)	−0.64
zho.rst.sctb	90.30	77.38	83.34	81.67 (DFKI)	1.67
eng.pdtb.pdtb	92.93	91.15	92.02	–	–
tur.pdtb.tdb	93.71	94.53	94.11	–	–
zho.pdtb.cdtb	89.19	85.95	87.52	–	–
mean	92.35	90.63	91.43	–	–

(a) Results for Gold Treebanked Data.

Corpus	P	R	F ₁	2019 Best F ₁	vs. 2019	vs. Gold
deu.rst.pcc	95.15	92.86	93.94	94.68 (ToNy)	−0.74	−1.64
eng.rst.gum	92.65	92.59	92.61	–	–	−1.54
eng.rst.rstdt	96.80	95.92	96.35	93.43 (ToNy)	2.92	−0.28
eng.sdrst.stac	91.77	92.06	91.91	83.99 (ToNy)	7.92	−3.00
eus.rst.ert	92.70	88.38	90.47	–	–	0.01
fas.rst.prstc	92.95	92.78	92.86	–	–	−0.08
fra.sdrst.annodis	87.95	83.79	85.78	–	–	−4.24
nld.rst.nldt	96.97	92.54	94.69	92.32 (ToNy)	2.37	−1.29
por.rst.cstn	93.21	95.03	94.11	91.86 (ToNy)	2.25	−0.25
rus.rst.rst	87.31	84.24	85.74	–	–	−0.47
spa.rst.rststb	93.30	90.30	91.76	89.60 (ToNy)	2.16	−0.46
spa.rst.sctb	83.97	77.98	80.86	81.65 (ToNy)	−0.79	−1.62
zho.rst.sctb	84.04	70.00	76.21	73.10 (GumDrop)	3.08	−7.13
eng.pdtb.pdtb	94.29	90.92	92.56	–	–	0.54
tur.pdtb.tdb	91.98	95.22	93.56	–	–	−0.55
zho.pdtb.cdtb	90.27	86.54	88.35	–	–	0.83
mean	91.58	88.82	90.11	–	–	−1.32

(b) Results for Plain Tokenized Data.

Table 4: Segmentation and Connective Detection Results. All numbers are averaged over five runs in order to accommodate instability in the training process which leads to varying performance. If a corpus was included in the 2019 shared task and has not been significantly modified since then, we also include the best result on the corpus in 2019 for comparison.

treebanked and plain tokenized data scenarios: our system nearly always outperforms the best score from 2019, and we observe especially large gains for connective detection.

On treebanked data, the results show that performance has improved since 2019 on nearly all unchanged datasets, with degradations of only around 0.5% for `eng.sdrst.stac` and `spa.rst.sctb` compared to the previous best systems, GumDrop and ToNy respectively. For some datasets, gains are dramatic, most notably for Turkish (14.5% gain) and Chinese connective detection (8.4%), which is perhaps due to the availability of better language models and our use of conditional random fields. On average the improvement on treebanked data is close to 3% for datasets represented in 2019.

On plain tokenized data, the improvement from 2019 is even more pronounced, with an average gain of 3.7% compared to 2.8% for treebanked data. While performance on some corpora was roughly constant regardless of whether data was treebanked or plain tokenized (e.g. `eng.rst.rstdt`, `por.rst.cstn`), it dropped considerably for some corpora on plain tokenized data. This effect is most dramatic for `zho.rst.sctb`, where we see a degradation of 7.1%. This effect cannot be explained just by the amount of training data available: correlation between training token count and degradation is low (Pearson’s $r = 0.092$, $p = 0.74$).

We speculate that these causes of the degradations are primarily due to idiosyncrasies of the corpora. `eng.sdrst.stac`, for instance, has a mild degradation (−3%), which we believe to be primarily due to the lack of punctuation and capitalization compared to e.g. a newswire corpus like

`eng.rst.rstdt`, which exhibited very little degradation. From this we might infer that degradation in the absence of treebanked data will be correlated with the degree to which predicting sentence splits from plain text is difficult. We additionally hypothesize that a lack of gold sentence breaks affects RST datasets more than PDTB datasets, since the beginning of a sentence is almost always the beginning of a new elementary discourse unit, while connectives are mainly identified lexically, and need to be identified regardless of the relative position of sentence splits.

Ablation Study In order to assess the importance of the various modules of our segmentation and connective detection system, we conduct an ablation study. Due to the large computational expense of conducting full runs over all datasets, we choose only two ablation conditions. In the first, we remove all handcrafted features described in Table 2. In the second, we remove character embeddings and fastText static word embeddings, leaving only contextualized word embeddings. The results of this study is given in Table 5.

The general trend in the results of the ablation study seems to be that both handcrafted features and supplementary word embeddings are helpful on average, though they may sometimes lead to minor degradations, and have a dramatically pronounced effect on a few corpora in particular. Handcrafted features have a mild effect on most corpora but lead to large gains for GUM, RST-DT, and PDTB. It is not immediately clear why this might be: performance on GUM, which is diverse with respect to genre, probably benefits from having a

Corpus	F ₁ (all)	F ₁ (no feats.)		F ₁ (CWE only)	
		<i>abs.</i>	<i>gain</i>	<i>abs.</i>	<i>gain</i>
deu.rst.pcc	95.58	96.28	-0.70	95.81	-0.23
eng.rst.gum	94.15	89.74	4.41	92.23	1.92
eng.rst.rstdt	96.64	91.41	5.23	94.59	2.05
eng.sdrst.stac	94.91	94.54	0.37	94.54	0.37
eus.rst.ert	90.46	91.01	-0.54	90.76	-0.30
fas.rst.prstc	92.94	92.99	-0.05	93.24	-0.30
fra.sdrst.annodis	90.02	88.80	1.22	88.48	1.54
nld.rst.nldt	95.97	95.48	0.50	95.11	0.86
por.rst.cstn	94.35	92.65	1.71	93.93	0.42
rus.rst.rst	86.21	86.18	0.03	86.01	0.20
spa.rst.rststb	92.22	92.04	0.18	92.39	-0.17
spa.rst.sctb	82.48	84.21	-1.73	83.32	-0.84
zho.rst.sctb	83.34	84.87	-1.54	82.60	0.74
eng.pdtb.pdtb	92.02	87.72	4.30	82.42	9.60
tur.pdtb.tdb	94.11	94.02	0.10	93.54	0.57
zho.pdtb.cdtb	87.52	88.53	-1.01	88.14	-0.62
mean	91.43	90.65	0.78	90.45	0.99

Table 5: F₁ scores for ablations on gold treebanked data: next to normal scores from Table 4a, we report scores without handcrafted features, and without character embeddings and fastText static word embeddings, as well as the “gain” for each (non-ablated score – ablated score). Due to time constraints, ablations are based on *three* runs instead of the standard five.

genre feature, but RST-DT and PDTB are homogeneous with respect to genre. We also note that GUM, RST-DT, and especially PDTB are large corpora, so perhaps the explanation lies in their size, but RRT is also very large and has multiple genres, yet handcrafted features led to nearly no gain on this dataset.

Turning now to the CWE-only ablation, we see a similar pattern: most corpora are only minorly affected by the inclusion of non-CWE embeddings, with a couple (GUM, RST-DT) showing a moderate gain of 2%, and one corpus (PDTB) showing an anomalous gain of 10%. Just as with the handcrafted feature ablation, it is difficult to know what could explain these corpora’s divergent behavior. Ordinarily, static word embeddings might benefit small corpora with OOV items in the test set, since the embedding space will be stable in the unseen data – however PDTB is very large and homogeneous (newswire), making this explanation unlikely. Since no other corpus showed such a dramatic drop with non-CWE embeddings ablated, and since other CWE-based systems at DISRPT 2021 score around what our system would have scored if the drop had been more in line with what was observed for other corpora (2% drop, for a score in the low 90s, as was achieved by disCut and SegFormers), we speculate that the 10% drop

Corpus	# Relations	Accuracy (w/ feats.)	Accuracy (w/o feats.)	Feature Gain
deu.rst.pcc	26	39.23	33.85	5.38
eng.pdtb.pdtb	23	74.44	75.63	-1.19
eng.rst.gum	23	66.76	62.65	5.55
eng.rst.rstdt	17	67.10	66.45	0.65
eng.sdrst.stac	16	65.03	59.67	5.36
eus.rst.ert	29	60.62	59.59	1.03
fra.sdrst.annodis	18	46.40	48.32	-1.92
nld.rst.nldt	32	55.21	52.15	3.06
por.rst.cstn	32	64.34	67.28	-2.94
rus.rst.rst	22	66.44	65.46	0.98
spa.rst.rststb	29	54.23	54.23	0.00
spa.rst.sctb	25	66.04	61.01	5.03
tur.pdtb.tdb	23	60.09	57.58	2.51
zho.pdtb.cdtb	9	86.49	87.34	-0.85
zho.rst.sctb	26	64.15	64.15	0.00
fas.rst.prstc	17	52.53	51.18	1.35
mean		61.82	60.41	1.41

Table 6: Relation Classification Results. The score for each corpus is averaged over 5 runs. Also included is score *without* any hand-crafted features.

observed here is due to some kind of implementation error or statistical fluke due to the nondeterminism of training models on GPUs, though the effect survives in the system reproduction on the Shared Task evaluators’ machine.

In sum, our ablation study for segmentation and connective detection suggests that both handcrafted features and non-CWE embeddings are not silver bullets, though they are often helpful. Degrada-tions were seen more often on smaller datasets, which perhaps indicates that in low-data situations these additional resources can serve more than anything as a source of overfitting. But both were on average responsible for a 1% gain,^d which shows they are both useful, and invites the question of whether there might be even better handcrafted features, which could be tailored more accurately to properties of specific target languages and genres.

4.2 Relation Classification

Table 6 gives scores (averaged over 5 runs for each corpus) on relation classification on all 16 corpora. We include performance on all corpora without any hand-crafted features added in order to assess their utility, and we find that they appreciably boost performance, granting on average a 1.5% accuracy gain, with some of the biggest gains on small corpora with many labels like `deu.rst.pcc` (+5.38%) and `spa.rst.sctb` (+5.03%). Since difficulty of classification increases with number of labels, we also include the number of relation types for each

^dAssuming that they can be treated independently, which is an idealization.

corpus in order to contextualize the scores. The `zho.pdtb.cdtb` corpus achieved the highest accuracy score, as there are only 9 relation types to classify,^c and scores tend to be lower for corpora with many relations like `nld.rst.nldt`.

There is much variance in how much performance on each corpus was able to benefit from having additional features. Many of the corpora that had the largest gains are small, but this is not always the case: `tur.pdtb.tdb`, one of the larger corpora, has its score improved by 2.5%. On the other hand, while small corpora generally seem to benefit more from features, not all are able to: `fra.sdrdt.annodis`, a small corpus, sees a degradation of 2% with features. We expect that much of the differential benefit of features is to be explained by the nature of the label-sets used in different corpora, and the available features. No two of these corpora use the exactly same label set, and label sets vary quite a bit in the linguistic phenomena that they encode and are sensitive to.

Additionally, different corpora have different features available, such as genre (GUM, RRT, RST-STB, PRSTC, SCTB – yes; PCC, PDTB, RST-DT, TDB – no), gold speaker information (STAC), discontinuity (Annodis, ERT, GUM, PDTB, RST-DT – yes; PCC, NLDT, STAC – no) etc., meaning that looking at gain across datasets is not comparing like with like. While some features are available for all corpora, such as distance, unit lengths, or position in document, others are restricted by framework, such as number of children, which is not relevant for PDTB-style data. A set of formalism-agnostic features (e.g. `length_ratio`, `is_sentence`, and the direction of the dependency head of the unit) were used for PDTB-style data across the board and were only effective for TDB: we hypothesized that the English PDTB dataset is so big that generic features do not add much value; for CDTB, as we found in our error analysis, the 9 relations sometimes are not very distinct from each other, and these generic features do not help with disambiguation in those cases.

Overall, a picture emerges with relations that is similar to the one that arose with segmentation and connective detection: features are helpful on the

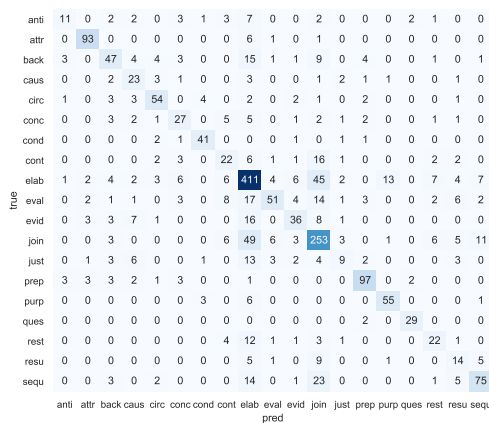
^cUnlike the other two PDTB-style corpora (i.e. `eng.pdtb.pdtb` and `tur.pdtb.tdb`), where the predicted labels are truncated at Level-2 (e.g. `TEMPORAL.ASYNCHRONOUS`), the relation labels in `zho.pdtb.cdtb` only contain one level (e.g. `TEMPORAL`).

whole, slightly harmful in some cases, and especially helpful on some corpora. More work remains to be done in understanding the contribution of individual features and how these relate to the frameworks and data types available in each language.

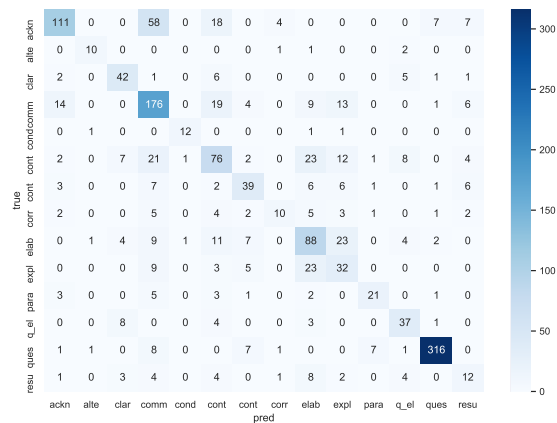
5 Discussion

Figure 1 shows confusion matrices for common relations in the highest and lowest scoring EDU datasets, `eng.rst.gum` (GUM) and `eng.sdrdt.stac` (STAC). Both panels reveal issues with over-prediction of the most common labels, which can be thought of as ‘defaults’: the most common ELABORATION in GUM and the second most common, COMMENT in STAC. The actual most common relation in STAC, QUESTION, does not suffer from false positives, likely due to a combination of the frequent and reliable question mark cue, *wh*-words or subject-verb inversion, combined with the availability of gold speaker and direction information (QUESTION only links units from different speakers, left to right). The same is true for QUESTION in GUM, which also obtained a comparatively high score. Conversely, rarer relations are hardly predicted, with ANTITHESIS in GUM being predicted for less than half of its true instances, and similarly for RESULT in STAC, suggesting a class imbalance problem, in particular given that both these relations are sometimes marked by overt discourse markers.

Although EDU datasets (RST/SDRT) do not distinguish explicit and implicit relations, analysis suggests that explicit signals are important. For GUM, the model scored high on medium-frequency relations with clear cues such as ATTRIBUTION, which are always signalled by attribution verbs such as *believe* and *say*. This also holds true for the `eng.rst.rstdt` corpus where ATTRIBUTION is the highest-scoring relation. GUM’s ELABORATION and JOINT as well as SEQUENCE and JOINT are two pairs of relation labels that are most frequently confused with each other, as the former pair contains the two labels that are both overgeneralized and the latter contains the two labels that are both MULTINUCLEAR relation types that could be confusing when no explicit connective or lexical item indicating a sequential order of actions or events is present (Liu, 2019). Relations with relatively unambiguous markers, such as CONDITION, show good results in both GUM and STAC, indicating that even relatively rare relations can be identified



(a) eng.rst.gum (GUM)



(b) eng.sdrt.stac (STAC)

Figure 1: Confusion Matrices for Common Relations in the Highest and Lowest Scoring EDU Datasets.

if they are usually explicitly signalled.

Relations such as JUSTIFY, RESULT, and CAUSE scored low in both matrices as such instances in the test data often do not have explicit discourse markers to help understand the rhetorical relation between the units in context. In the presence of an ambiguous discourse marker, predictions prefer the relation that is more prototypically associated with that marker: for instance, the gold relation label for (1) is RESULT whereas the model classified it as SEQUENCE, likely due to the fact that the discourse marker *then* tends to be a strong signal for SEQUENCE indicating something sequential is involved.

- (1) **Then** she lets go and falls . You scream .
[GUM_fiction_falling]

In fact, if it were not the case that RST mandates a single outgoing relation per discourse unit, it would be possible to claim a concurrent sequential relation (likely to a previous unit) next to the annotated RESULT relation for the pair in the example (see [Stede 2008](#) on concurrent relations in RST).

6 Conclusion

We have presented DisCoDisCo, a system for all tasks in the DISPRT 2021 Shared Task: EDU segmentation, connective detection, and relation classification. Our system relies on sequence tagging and sentence pair classification architectures powered by CWEs and supported by rich, handcrafted, instance-level features, such as position in the document, distance between units, gold speaker infor-

mation, document metadata, and more.

Our results suggest that powerful pretrained language models are the main drivers of performance, with additional features providing small to medium improvements (with some exceptions, such as the high importance of speaker information for chat data as in STAC). For relation classification, CWEs pretrained using an NSP task proved to be superior.

Our error analysis suggests unsurprisingly that class imbalances, especially in the case of relations that tend to be implicit (i.e. lack overt lexical signals), lead to over-prediction of majority classes, suggesting a need for more training data for the minority ones. However, we are encouraged by improvements on datasets that were featured in the 2019 Shared Task ([Zeldes et al., 2019](#)), and the overall high scores obtained by the system across a range of datasets, all while including some correct predictions for relatively rare relations. We hope that the growing availability of annotated data, coupled with architectures that can harness pre-trained models, will lead to further improvements in the near future.

References

- 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).

- Giuseppe Attardi, Daniele Sartiano, and Yu Zhang. 2021. DiaParser attentive dependency parser.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- 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. Does syntax help discourse segmentation? Not So Much. In *Proceedings of EMNLP 2017*, pages 2432–2442, Copenhagen.
- Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Current and New Directions in Discourse and Dialogue*, Text, Speech and Language Technology 22, pages 85–112. Kluwer, Dordrecht.
- 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.
- Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of ACL 2018*, pages 1–6, Melbourne, Australia.
- Luke Gessler, Siyao Peng, Yang Liu, Yilun Zhu, Shabnam Behzad, and Amir Zeldes. 2020. [AMALGUM - a free, balanced, multilayer English web corpus](#). In *Proceedings of LREC 2020*, pages 5267–5275, Marseille, France.
- Grigori Guz and Giuseppe Carenini. 2020. [Coreference for discourse parsing: A neural approach](#). In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.
- Hugo Hernault, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue and Discourse*, 1(3):1–33.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Dan Jurafsky and James H. Martin. 2020. *Speech and Language Processing*. <https://web.stanford.edu/jurafsky/slp3/>.
- Yusuke Kido and Akiko Aizawa. 2016. [Discourse relation sense classification with two-step classifiers](#). In *Proceedings of the CoNLL-16 shared task*, pages 129–135, Berlin, Germany. Association for Computational Linguistics.
- Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. [Implicit discourse relation classification: We need to talk about evaluation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.
- Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. [Top-down rst parsing utilizing granularity levels in documents](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8099–8106.
- Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. [Improving neural RST parsing model with silver agreement subtrees](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612, Online. Association for Computational Linguistics.
- Li Liang, Zheng Zhao, and Bonnie Webber. 2020. [Extending implicit discourse relation recognition to the PDTB-3](#). In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 135–147, Online. Association for Computational Linguistics.
- Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. [A pdtb-styled end-to-end discourse parser](#). *Natural Language Engineering*, 20(2):151–184.
- Yang Liu. 2019. [Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres](#). In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 72–81, Minneapolis, MN. Association for Computational Linguistics.
- Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI’16, page 2750–2756. AAAI Press.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#).
- William C Mann and Sandra A Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.
- Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330.
- Mathieu Morey, Philippe Muller, and Nicholas Asher. 2017. [How much progress have we made on RST discourse parsing? a replication study of recent results on the RST-DT](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1319–1324, Copenhagen, Denmark. Association for Computational Linguistics.
- Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. ToNy: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, pages 115–124, Minneapolis, MN.
- Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. [RST parsing from scratch](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625, Online. Association for Computational Linguistics.
- 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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proceedings of EMNLP 2013*, pages 914–923.
- Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore.
- Balaji Polepalli Ramesh, Rashmi Prasad, Tim Miller, Brian Harrington, and Hong Yu. 2012. Automatic discourse connective detection in biomedical text. *Journal of the American Medical Informatics Association*, 19(5):800–808.
- Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltasakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakesh, Morocco.
- Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The Biomedical Discourse Relation Bank. *BMC bioinformatics*, 12(1):188.
- Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. 2019. Penn Discourse Treebank Version 3.0. LDC2019T05.
- 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*, pages 101–108, Online. Association for Computational Linguistics.
- 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.
- Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. [A systematic study of neural discourse models for implicit discourse relation](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291, Valencia, Spain. Association for Computational Linguistics.
- Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. [Do we really need all those rich linguistic features? a neural network-based approach to implicit sense labeling](#). In *Proceedings of the CoNLL-16 shared task*, pages 41–49, Berlin, Germany. Association for Computational Linguistics.
- Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of HLT-NAACL 2003*, pages 149–156, Edmonton.
- Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of EMNLP 2005*, pages 257–264, Vancouver.

- Manfred Stede. 2008. [Disambiguating rhetorical structure](#). *Research on Language and Computation*, 6(3):311–332.
- Manfred Stede. 2011. Discourse Processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.
- Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. 2004. Generating discourse structures for written text. In *Proceedings of COLING 2004*, pages 329–335, Geneva, Switzerland.
- Jianxiang Wang and Man Lan. 2016. [Two end-to-end shallow discourse parsers for English and Chinese in CoNLL-2016 shared task](#). In *Proceedings of the CoNLL-16 shared task*, pages 33–40, Berlin, Germany. Association for Computational Linguistics.
- Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. [A two-stage parsing method for text-level discourse analysis](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.
- Gregor Weiss and Marko Bajec. 2016. [Discourse sense classification from scratch using focused RNNs](#). In *Proceedings of the CoNLL-16 shared task*, pages 50–54, Berlin, Germany. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. [The CoNLL-2015 shared task on shallow discourse parsing](#). In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. 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.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Yue Yu, Yilun Zhu, Yang Liu, Yan Liu, Siyao Peng, Mackenzie Gong, and Amir Zeldes. 2019. [Gum-Drop at the DISRPT2019 shared task: A model stacking approach to discourse unit segmentation and connective detection](#). In *Proceedings of Discourse Relation Treebanking and Parsing (DISRPT 2019)*, pages 133–143, Minneapolis, MN.
- Amir Zeldes. 2018. *Multilayer Corpus Studies*. Routledge Advances in Corpus Linguistics 22. Routledge, London.
- 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.
- Yingxue Zhang, Fandong Meng, Peng Li, Ping Jian, and Jie Zhou. 2021. [Context tracking network: Graph-based context modeling for implicit discourse relation recognition](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1592–1599, Online. Association for Computational Linguistics.
- Yuping Zhou, Jill Lu, Jennifer Zhang, and Nianwen Xue. 2014. Chinese Discourse Treebank 0.5. LDC2014T21.
- Yuping Zhou and Nianwen Xue. 2012. [PDTB-style discourse annotation of Chinese text](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.