A Transformer Based Approach towards Identification of Discourse Unit Segments and Connectives

Sahil Bakshi and **Dipti Misra Sharma**

Language Technologies Research Center (LTRC)
International Institute of Information Technology, Hyderabad, India

sahil.bakshi@research.iiit.ac.in , dipti@iiit.ac.in

Abstract

Discourse parsing, which involves understanding the structure, information flow, and modeling the coherence of a given text, is an important task in natural language processing. It forms the basis of several natural language processing tasks such as question-answering, text summarization, and sentiment analysis. Discourse unit segmentation is one of the fundamental tasks in discourse parsing and refers to identifying the elementary units of text that combine to form a coherent text. In this paper, we present a transformer based approach towards the automated identification of discourse unit segments and connectives. Early approaches towards segmentation relied on rule-based systems using POS tags and other syntactic information to identify discourse segments. Recently, transformer based neural systems have shown promising results in this domain. Our system, SegFormers, employs this transformer based approach to perform multilingual discourse segmentation and connective identification across 16 datasets encompassing 11 languages and 3 different annotation frameworks. We evaluate the system based on F1 scores for both tasks, with the best system reporting the highest F1 score of 97.02% for the treebanked English RST-DT dataset.

1 Introduction

Discourse is defined as a coherent, structured group of sentences (Jurafsky and Martin, 2009). Discourse parsing enables the creation of models for further downstream natural language processing tasks such as question-answering, text summarization, information retrieval and extraction, sentiment analysis, and argument mining. Discourse segmentation is a fundamental task in discourse parsing, which involves identifying the minimal chunks of text that combine to form a coherent discourse. In the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) framework, these basic chunks of texts are known as Elementary Discourse Units (EDUs). These EDUs are linked together by discourse relations which may be explicit (when explicitly marked in the text by a discourse connective) or implicit. Segmentation refers to the task of identifying these EDUs. There have been several approaches towards the identification of these discourse segments. Another popular framework is the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). Both these frameworks segment the text into non-overlapping spans covering entire documents. The discourse segmentation task, in this case, corresponds to identifying the starting point of each discourse unit. In 2008, the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) was released with a corpus of over 1 million words. The unit identification task for this framework corresponds to identifying the spans of discourse connectives that explicitly identify the existence of a discourse relation.

The task of EDU segmentation has been widely researched in the past due to its importance as a building block for further downstream natural language processing tasks. Most of the previous research concerning EDU segmentation has relied on the information obtained from syntactic elements of the text such as syntactic parse trees (Tofiloski et al., 2009 ; Le Thanh et al., 2004). However, recent works have explored the task of segmentation using systems based on neural networks using the BiLSTM - CRF framework (Wang et al., 2018) or the attention mechanism (Li et al., 2016). Another important factor in all the previous works has been the presence/absence of gold sentence boundaries (Ji and Eisenstein, 2014 ; Feng and Hirst, 2014) in the data.

The DISRPT 2021 Shared Task introduces the second iteration of a cross-formalism shared task on discourse unit segmentation and connective detection, and the first iteration of a cross-formalism discourse relation classification task. The organizers provided 16 datasets in the RST (Mann and

Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021), pages 13–21

Thompson, 1988), SDRT (Asher and Lascarides, 2003) and PDTB (Prasad et al., 2008) formalisms with each dataset consisting of a training, development, and test set. The data has been provided in two formats - one is the treebanked format which provides the gold syntax and the other is the plain tokenized data without the gold syntax. This encourages the system design to be able to adapt and deal with data when the gold syntax is not provided.

In this paper, we describe our approach towards the discourse unit segmentation and connective detection tasks using a model based on the transformers architecture (Devlin et al., 2018) with a classification layer on top to provide the outputs. We report scores for discourse segmentation and connective identification on both settings - data with gold sentence boundaries as well as document level boundaries. Our best results improve the previous task iterations' (Zeldes et al., 2019) best results on 15 out of the 16 available datasets.

2 Related Work

Early work in the domain for discourse parsing involved rule-based models which used syntactic information for the prediction of discourse segments and connectives (Tofiloski et al., 2009). Le Thanh et al. (2004) used syntactic information and cue phrases to segment sentences into EDUs and integrated constraints about textual adjacency and textual organization in a beam search for text level segmentation. Soricut and Marcu (2003) used probabilistic models for EDU identification on the RST-DT corpus (Carlson et al., 2003). Building on this, Subba and Di Eugenio (2007) developed a neural network model for discourse segmentation using the same RST-DT dataset (Carlson et al., 2003). A token based classifier was introduced by Sporleder and Lapata (2005) which used POS tags, syntactic chunks, and clause information as features for the segmentation task.

Fisher and Roark (2007) investigated the approach towards segmentation using sentence level syntactic parse trees on the RST-DT corpus. On a similar note, Jain and Sharma (2016) introduced a hybrid pipeline for discourse connective and argument identification in Hindi using sub-tree extraction and linear tagging approaches. The data for this system was obtained from the Hindi Discourse Relation Bank (HDRB) (Oza et al., 2009). The importance of dependency information was investigated by Braud et al. (2017b) by introducing a segmentation model that only relies on part-of-speech information. Pitler and Nenkova (2009) proposed a method of using syntactic features for the identification of discourse connectives and the disambiguation of the connective in terms of its usage and the relation they mark.

Recent studies have been moving towards machine learning based approaches with most involving the use of sequential neural network models. Li et al. (2018) used a bidirectional recurrent neural network along with a pointer network to select text boundaries in the input sequence. An end-to-end neural segmenter was proposed by Wang et al. (2018) based on the BiLSTM-CRF framework. Their system reported new state-of-the-art performance on the RST-DT corpus (Carlson et al., 2003) with an F1 score of 94.3%. Lukasik et al. (2020) investigate a transformers based approach towards document and discourse level segmentation with the RST-DT corpus being used for discourse segmentation and the Wiki-727K dataset (Koshorek et al., 2018) for document segmentation experiments. Braud et al. (2017a) introduced the first multilingual segmenter across 5 languages and 3 non-newswire English domains using language independent tools.

Segmenters have also been developed for other languages with Lüngen et al. (2006) proposing a discourse segmenter for German, Iruskieta and Zapirain (2015) for Basque, Afantenos et al. (2010) for French, van der Vliet (2010) for Dutch, Pardo and Nunes (2008) for Brazilian Portuguese, Da Cunha et al. (2012) for Spanish and Yang and Li (2018) for Chinese.

3 Datasets

In this section, we describe the datasets provided by the organizers of the CODI-DISRPT2021: Discourse Relation Parsing and Treebanking Shared Task at EMNLP 2021[1]. The data provided consists of 16 datasets comprising of 11 languages (German, English, Basque, Persian, French, Dutch, Portuguese, Russian, Spanish, Turkish, and Mandarin Chinese). This is the first iteration of the Persian RST corpus (Shahmohammadi et al., 2021) being included for the task of discourse segmentation. The Chinese PDTB dataset (Zhou and Xue, 2015) is not available freely. Hence, the organizers provided the scores on this dataset after running the

[1] https://sites.google.com/georgetown.edu/disrpt2021/

model on the CDTB dataset during the evaluation phase.

3.1 Annotation frameworks

The data for discourse unit segmentation and connective detection tasks has been provided in the RST, SDRT, and PDTB formalisms. The tasks differ across different formalisms. In the RST and SDRT framework, the text has been segmented into non-overlapping spans covering each entire documents. The corresponding task, in this case, is finding the starting point of each discourse unit. In the PDTB framework, the segmentation task corresponds to identifying the spans of discourse connectives that explicitly identify the presence of a discourse relation.

Out of the 16 datasets, 3 corpora follow the PDTB framework (English, Turkish, and Mandarin Chinese), 2 are represented by the SDRT framework (English and French), and 11 datasets follow the RST framework. The diversity across the datasets with respect to the frameworks encourages the design of flexible systems capable of dealing with multiple formalisms.

3.2 Languages

The organizers have provided data for 11 languages. There are 4 datasets for the English language (Prasad et al., 2008 ; Zeldes, 2017 ; Carlson et al., 2003 ; Asher et al., 2016), 2 for Spanish (Da Cunha et al., 2011 ; Cao et al., 2018), 2 for Mandarin Chinese (Zhou and Xue, 2015 ; Cao et al., 2018) and 1 each for German (Stede and Neumann, 2014), Basque (Iruskieta et al., 2013 ; Aranzabe et al., 2015), Persian (Shahmohammadi et al., 2021), French (Péry-Woodley et al., 2011), Dutch (Redeker et al., 2012), Portuguese (Cardoso et al., 2011), Russian (Toldova et al., 2017), and Turkish (Zeyrek and Kurfalı, 2017).

The Persian RST corpus was added as a surprise language dataset at the release of the test data.

3.3 Data statistics

We present a comprehensive overview of the datasets provided by the organizers in Table 1. For the tasks of discourse unit segmentation and connective identification, each dataset has a training, development, and test set available in 2 file formats, as treebanked .conllu files which have a gold dependency parse and plain .tok files which contain plain tokens. The labels in the RST and SDRT framework denote the beginning of a discourse segment. The PDTB framework consists of labels that mark the entire span of discourse connectives that explicitly identify the existence of a discourse relation.

The English PDTB corpus is the largest in the given datasets with 1,061,229 training tokens and 1,156,657 total tokens in the dataset. The Chinese section of the RST Spanish - Chinese treebank is the smallest corpus with 9,655 training tokens and a total of 15,496 tokens. These variations in the dataset size, frameworks, and languages render the task challenging. You can find the datasets at the official DISRPT 2021 GitHub repository here[2].

4 System Overview

In this section, we present our approach towards the tasks of discourse unit segmentation and connective identification. We establish a baseline using a bidirectional LSTM classifier and propose a final system using the transformers architecture (Devlin et al., 2018) with a classification layer on top.

4.1 Bidirectional LSTM

Bidirectional LSTMs are basically an extension to the LSTM model (Hochreiter and Schmidhuber, 1997) and can be thought of as joining two LSTMs together. This architecture allows the model to read input sequences in both forward and backward directions, which results in an effective capture of context, which helps in improving model performance for classification tasks.

As a baseline, we propose a PyTorch[3] implementation of a simple bidirectional LSTM which encodes input tokens using word embeddings and has a single linear layer for the segmentation task. The model takes in an input sequence of tokens along with the corresponding sequence of labels, converts them into word vectors, and passes them to the bidirectional LSTM. The final linear layer takes the hidden states as input from the bidirectional LSTM and outputs the final labels.

4.2 SegFormers

Recent studies have shown promising results with neural approaches towards NLP tasks. Transformer (Vaswani et al., 2017) based architectures like BERT (Devlin et al., 2018) achieved state-of-the-art scores on several natural language processing tasks. Following this, Devlin et al. (2018) proposed

[2] https://github.com/disrpt/sharedtask2021

[3] https://pytorch.org/

Corpus	Language	Framework	Train Tokens	Train Sent.	Train Docs.
deu.rst.pcc	German	RST	26,831	1,773	142
eng.rst.gum	English	RST	116,557	6,346	128
eng.rst.rstdt	English	RST	166,854	6,672	309
eus.rst.ert	Basque	RST	30,690	1,599	116
fas.rst.prstc	Persian	RST	52,497	1,713	120
nld.rst.nldt	Dutch	RST	17,562	1,156	56
por.rst.cstn	Portuguese	RST	52,177	1,825	114
rus.rst.rrt	Russian	RST	390,375	18,932	272
spa.rst.rststb	Spanish	RST	43,055	1,548	203
spa.rst.sctb	Spanish	RST	10,253	326	32
zho.rst.sctb	Chinese	RST	9,655	361	32
eng.sdrt.stac	English	SDRT	41,060	8,754	33
fra.sdrt.annodis	French	SDRT	22,515	1,020	64
eng.pdtb.pdtb	English	PDTB	1,061,229	44,563	1,992
tur.pdtb.tdb	Turkish	PDTB	398,515	24,960	159
zho.pdtb.cdtb	Chinese	PDTB	52,061	2,049	125

Table 1: Datasets for the segmentation tasks

a multilingual BERT model[4] pretrained on 104 languages in a self supervised fashion. We adopt the aforementioned model and modify it for the segmentation and connective detection tasks.

Previous approaches towards EDU segmentation adopt a sequence classification approach. Wang et al. (2018) employed a self attention mechanism (Vaswani et al., 2017) restricting the attention area for the model to a neighborhood of fixed size which would prevent the unnecessary tokens from misleading the model. Building upon this, we present the task of segmentation as a token classification problem to the multilingual BERT model. This enables the architecture to learn and gain information from each individual token and spares it from committing errors due to unnecessary tokens. The multilingual BERT model has been trained on a huge corpus of multilingual data which makes it capable of being fine tuned to handle downstream tasks such as token classification even when the available training dataset is not very large.

Our system, SegFormers, takes sequences of tokens and corresponding labels as input which are then fed into a tokenizer. The WordPiece tokenizer converts the words into tokens to be fed into the model. One issue that arises, in this case, is the problem of out of vocabulary words. In such cases, the tokenizer splits the input word into subtokens which are present in the tokenizer vocabulary. This creates a mismatch between the number of input tokens and labels. We handle this issue by modifying the label tokenization process to match the output length. The tokenizer returns an offset mapping for each split word, according to which we then modify the label encodings. This ensures efficiency in the training process. We obtain the segmentation outputs by adding a linear layer that takes the outputs from the hidden states and converts them into the segmentation boundary labels.

This system is suited for the treebanked (.conllu files) data since the lengths of the input sequences stay below the threshold of 512 tokens. However, in the case of plain tokenized documents (.tok files), the input sequences are entire documents whose length goes beyond the 512 token limit. We work around this by adopting the approach of Muller et al. (2019) and using the StanfordNLP sentence splitter[5] for splitting the input document sequences into sentences.

Figure 1 illustrates the basic architecture and data flow in our proposed system. The results obtained from our system indicate that fine-tuning the multilingual BERT model by considering discourse unit segmentation as a token classification task instead of a sequence labeling task leads to better model performance.

[4] https://github.com/google-research/bert

[5] https://stanfordnlp.github.io/CoreNLP/ssplit.html

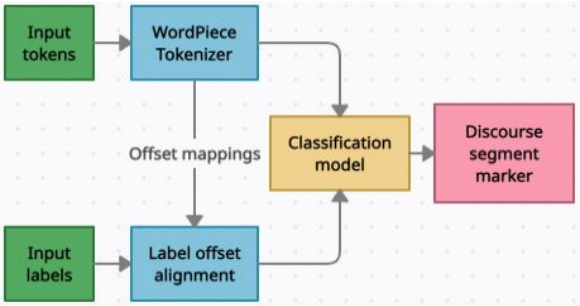

Figure 1: System architecture and data flow

5 Experimental Settings

For the bidirectional LSTM model, we used 300 dimensional randomly initialized word embeddings to encode the input tokens and pass it to a hidden layer with 100 dimensions and a dropout rate of 0.5. We used the Adam optimizer to update the model weights with a learning rate of 1e-3 and a batch size of 1 for training. We used the negative log-likelihood loss and trained the model for 5 epochs to obtain the optimal results.

For our system, we used HuggingFace's[6] PyTorch implementation of the multilingual BERT model. We fine tuned the model to work with the discourse unit segmentation task and present the optimal model settings for the same. We used a batch size of 8 for the training of the model. We experimented with various preprocessing and training parameters, changing the padding and truncation length, tweaking the weight update rate, and changing the number of training epochs. We found that the best results were obtained when the input sequences in the batch were padded to the length of the longest sequence in that given batch. We used the AdamW optimizer with a learning rate of 1e-5 and trained for 5 epochs to obtain the best F1 scores.

6 Results

Our system, SegFormers, improves the best scores of the previous iteration's results for 15 out of the 16 provided datasets for the discourse unit segmentation and the connective identification tasks. In this section, we present the detailed results of all our systems. The final precision, recall and F1 scores have been provided by the organizers by averaging the scores obtained over 5 runs.

[6] https://huggingface.co/

Corpus	P	R	F1
deu.rst.pcc	93.71	96.26	94.97
eng.rst.gum	91.47	95.67	93.52
eng.rst.rstdt	96.15	98.04	**97.09**
eng.sdrt.stac	96.73	95.67	96.20
eus.rst.ert	87.23	88.65	87.94
fas.rst.prstc	90.53	94.18	92.32
fra.sdrt.annodis	86.76	88.03	87.39
nld.rst.nldt	98.16	94.67	96.39
por.rst.cstn	92.60	94.12	93.35
rus.rst.rrt	84.93	85.18	85.06
spa.rst.rststb	89.12	94.35	91.66
spa.rst.sctb	92.31	78.57	84.89
zho.rst.sctb	79.89	82.74	81.29
mean	**90.74**	**91.24**	**90.93**

Table 2: SegFormers scores for the EDU segmentation task on the treebanked (.conllu files) data. The best score has been marked in bold. The model performs the best on the English RST-DT dataset with an F1 score of 97.09%.

6.1 Bidirectional LSTM

We report the baseline accuracy, precision, recall, and F1 scores on the development and test sets for the 15 available datasets based on the outputs obtained from the bidirectional LSTM model in Table 3. The bidirectional LSTM model gives largely decent scores on most of the datasets, with the lowest being 65.979% F1 score for the segmentation task on the zho.rst.sctb dataset and 66.292% F1 score for the connective identification task on the tur.pdtb.tdb dataset. For the segmentation task, the model attains its best performance on the nld.rst.nldt dataset with an F1 score of 85.362% and similarly, 78.629% F1 score on the eng.pdtb.pdtb dataset for connective identification task. The low F1 scores on the Chinese RST dataset can be attributed to the small training dataset size with only 32 documents consisting of 9,655 tokens available for training. On many datasets, the precision scores are quite higher than the recall (Basque dataset - 95% precision and 61% recall, Russian dataset - 84% precision and 60% recall), indicating that the model is primarily aiming for the generic discourse unit boundary detection at the beginning of the discourse segments.

6.2 SegFormers

We report the discourse segmentation results from our multilingual transformers based model, Seg-

	Development Set				Test Set			
Corpus	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
deu.rst.pcc	97.239	89.166	77.818	83.106	97.190	90.763	76.870	83.241
eng.pdtb.pdtb	98.873	81.446	76.117	78.692	98.882	78.870	76.578	77.707
eng.rst.gum	95.226	80.000	83.399	81.664	95.247	78.860	84.227	81.455
eng.rst.rstdt	96.614	82.961	84.808	83.874	96.436	81.785	86.317	83.990
eng.sdrt.stac	90.878	79.049	85.008	81.920	90.835	79.246	82.934	81.048
eus.rst.ert	95.927	93.821	60.561	73.608	96.112	95.974	61.216	74.752
fas.rst.prstc	97.534	90.380	78.298	83.906	97.272	90.500	78.208	83.907
fra.sdrt.annodis	96.120	91.666	71.223	80.161	95.678	90.890	70.806	79.601
nld.rst.nldt	97.488	90.522	80.758	85.362	97.382	91.525	79.881	**85.308**
por.rst.cstn	95.974	75.698	86.031	80.534	96.749	74.863	89.542	81.547
rus.rst.rrt	95.892	84.981	60.769	70.864	95.734	84.391	60.262	70.314
spa.rst.rststb	98.119	84.367	81.145	82.725	97.756	80.616	79.565	80.087
spa.rst.sctb	97.834	79.761	65.048	71.657	97.771	81.679	63.690	71.571
tur.pdtb.tdb	98.741	63.533	70.037	66.627	98.693	62.632	70.405	66.292
zho.rst.sctb	97.040	68.000	66.019	66.995	97.232	78.048	57.142	65.979

Table 3: Results for the baseline bidirectional LSTM model on the 15 available datasets for the tasks of EDU segmentation and connective detection. The model performs the best on the Dutch RST dataset with an F1 score of 85.308%.

Formers, in Tables 2 and 4. Our system returns high F1 scores across all 16 provided datasets and manages to improve the previous best scores on all except the French ANNODIS corpus in the treebanked scenario. We also report better scores for 14 out of the 16 provided datasets in the plain document level (.tok files) setting with only the French ANNODIS corpus and the Chinese section of the RST Spanish-Chinese Treebank, yielding scores lower than the previous best results.

We also report the first ever scores for the Persian RST corpus (Shahmohammadi et al., 2021) with our system returning the best F1 score of 92.32% in the treebanked scenario with the gold syntax and the best F1 score of 91.90% on the plain tokenized documents.

We observe significant improvements compared to the previous iteration results on the Basque dataset (F1 score of 88.52% compared to the previous best F1 score of 84.06%) with the plain tokenized documents. To the best of our knowledge, we also report state-of-the-art performance so far for discourse unit segmentation on the English RST-DT dataset with an F1 score of 97.09% on the test dataset with the gold syntax.

With respect to the task of connective identification on the datasets with the PDTB framework, we could only obtain results on the English and Turkish PDTB datasets since the Chinese CDTB

Corpus	**P**	**R**	**F1**
deu.rst.pcc	96.81	92.86	94.79
eng.rst.gum	90.38	94.01	92.16
eng.rst.rstdt	94.31	96.16	**95.23**
eng.sdrt.stac	88.38	87.01	87.69
eus.rst.ert	86.47	90.68	88.52
fas.rst.prstc	94.20	89.70	91.90
fra.sdrt.annodis	89.67	87.06	88.34
nld.rst.nldt	96.07	94.08	95.07
por.rst.cstn	92.60	94.12	93.35
rus.rst.rrt	84.32	84.15	84.23
spa.rst.rststb	90.95	89.57	90.25
spa.rst.sctb	79.78	86.90	83.19
zho.rst.sctb	65.52	79.17	71.70
mean	**88.42**	**89.65**	**88.96**

Table 4: SegFormers scores for the EDU segmentation task on the plain tokenized (.tok files) data. The best score has been marked in bold. The model performs the best on the English RST-DT corpus with an F1 score of 95.23%.

	Treebanked (.conllu) data			Plain tokenized (.tok) data		
Corpus	**P**	**R**	**F1**	**P**	**R**	**F1**
eng.pdtb.pdtb	89.72	92.61	91.14	90.37	91.96	**91.16**
tur.pdtb.tdb	90.42	91.16	**90.79**	89.36	91.04	90.19
zho.pdtb.cdtb	85.04	87.50	**86.25**	85.24	83.33	84.27
mean	**88.39**	**90.42**	**89.39**	**88.32**	**88.78**	**88.54**

Table 5: Final SegFormers scores for the connective detection task on the PDTB framework datasets. The best scores for each dataset have been marked in bold. The model performs the best on the English PDTB dataset with an F1 score of 91.16%.

dataset was not freely available. The organizers provided the scores on the Chinese CDTB dataset after the evaluation phase. We report the F1 scores for our model on these datasets in Table 5 for the treebanked and plain scenario.

Comparing the results of the model on the treebanked data with the results on the plain tokenized documents, we can see that the performance of the model is consistently better on the treebanked datasets, with the Chinese section of the RST Spanish-Chinese Treebank showing the most significant drop in F1 scores (81.29% to 71.70%). There are a couple of exceptions, with the model giving better scores for the Basque and Spanish RST datasets. However, in general, we can conclude that the presence of gold sentence boundaries helps the system perform better compared to the datasets with predicted sentence boundary markers.

7 Conclusion and Future Work

In this paper, we looked at the topic of discourse segmentation and its importance in the development of robust models for further downstream natural language processing tasks. We report our approach towards building a system for the EDU segmentation and connective identification tasks across 16 datasets consisting of 11 languages and 3 frameworks. We present a multilingual system leveraging the new transformer based approach and fine-tune it to obtain quality results for the two above-mentioned tasks. Our results showed that a token classification approach towards the task of segmentation helps the model perform better and results in state-of-the-art scores.

We are also currently working on developing a curated dataset and model for the connective identification task for the Hindi language. This aims to increase the diversity of the available data and add a dataset to the existing corpora.

References

Stergos Afantenos, Pascal Denis, Philippe Muller, and Laurence Danlos. 2010. Learning recursive segments for discourse parsing. *arXiv preprint arXiv:1003.5372*.

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2015. Automatic conversion of the basque dependency treebank to universal dependencies. In *Proceedings of the fourteenth international workshop on treebanks an linguistic theories (TLT14)*, pages 233–241.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Nicholas Asher and Alex Lascarides. 2003. *Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure*. Cambridge University Press.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017a. Cross-lingual and cross-domain discourse segmentation of entire documents. *arXiv preprint arXiv:1704.04100*.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017b. Does syntax help discourse segmentation? not so much. In *Conference on Empirical Methods in Natural Language Processing*, pages 2432–2442.

Shuyuan Cao, Iria da Cunha, and Mikel Iruskieta. 2018. The rst spanish-chinese treebank. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 156–166.

Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS 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. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Iria Da Cunha, Eric San Juan, Juan Manuel Torres-Moreno, Marina Lloberese, and Irene Castellóne. 2012. Diseg 1.0: The first system for spanish discourse segmentation. *Expert Systems with Applications*, 39(2):1671–1678.

Iria Da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the rst spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

M Iruskieta, MJ Aranzabe, A Diaz de Ilarraza, I Gonzalez, M Lersundi, and O Lopez de la Calle. 2013. Euskal rsttreebanka (euskal rst-ko erlazio eta zuhaitz bankua). the rst basque treebank: an online search interface to check rhetorical relations. In *4th Workshop"RST and Discourse Studies", Brasil*.

Mikel Iruskieta and Benat Zapirain. 2015. Euseduseg: A dependency-based edu segmentation for basque. *Procesamiento del Lenguaje Natural*, (55):41–48.

Rohit Jain and Dipti Misra Sharma. 2016. Explicit argument identification for discourse parsing in hindi: A hybrid pipeline. In *Proceedings of the NAACL Student Research Workshop*, pages 66–72.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 13–24.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. *arXiv preprint arXiv:1803.09337*.

Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. 2004. Generating discourse structures for written text. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 329–335.

Jing Li, Aixin Sun, and Shafiq R Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *IJCAI*, pages 4166–4172.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of EMNLP 2016*, pages 362–371.

Michal Lukasik, Boris Dadachev, Gonçalo Simoes, and Kishore Papineni. 2020. Text segmentation by cross segment attention. *arXiv preprint arXiv:2004.14535*.

Harald Lüngen, Csilla Puskás, Maja Bärenfänger, Mirco Hilbert, and Henning Lobin. 2006. Discourse segmentation of german written texts. In *International Conference on Natural Language Processing (in Finland)*, pages 245–256. Springer.

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

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. Association for Computational Linguistics.

Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. 2009. The hindi discourse relation bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161.

Thiago Alexandre Salgueiro Pardo and Maria das Graças Volpe Nunes. 2008. On the development and evaluation of a brazilian portuguese discourse parser. *Journal of Technical and Applied Informatics*, 15(2):43–64.

Marie-Paule Péry-Woodley, Stergos Afantenos, Lydia-Mai Ho-Dac, and Nicholas Asher. 2011. La ressource annodis, un corpus enrichi d'annotations discursives. *Traitement Automatique des Langues*, 52(3):71–101.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of LREC*.

Gisela Redeker, Ildikó Berzlánovich, Nynke Van Der Vliet, Gosse Bouma, and Markus Egg. 2012. Multi-layer discourse annotation of a dutch text corpus. *age*, 1:2.

Sara Shahmohammadi, Hadi Veisi, and Ali Darzi. 2021. Persian rhetorical structure theory. *arXiv preprint arXiv: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.

Caroline Sporleder and Mirella Lapata. 2005. Discourse chunking and its application to sentence compression. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 257–264.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *LREC*, pages 925–929.

Rajen Subba and Barbara Di Eugenio. 2007. Automatic discourse segmentation using neural networks. In *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue*, pages 189–190.

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.

Svetlana Toldova, Dina Pisarevskaya, Margarita Ananyeva, Maria Kobozeva, Alexander Nasedkin, Sofia Nikiforova, Irina Pavlova, and Alexey Shelepov. 2017. Rhetorical relations markers in russian rst treebank. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 29–33.

Nynke van der Vliet. 2010. Syntax-based discourse segmentation of dutch text. *Proceedings of the 15th Student Session, ESSLLI*, pages 203–210.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. *arXiv preprint arXiv:1808.09147*.

Jingfeng Yang and Sujian Li. 2018. Chinese discourse segmentation using bilingual discourse commonality. *arXiv preprint arXiv:1809.01497*.

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

Amir Zeldes, Debopam Das, Erick Galani Maziero, Juliano Desiderato 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*.

Deniz Zeyrek and Murathan Kurfalı. 2017. Tdb 1.1: Extensions on turkish discourse bank. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 76–81.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: a chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49(2):397–431.