DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing

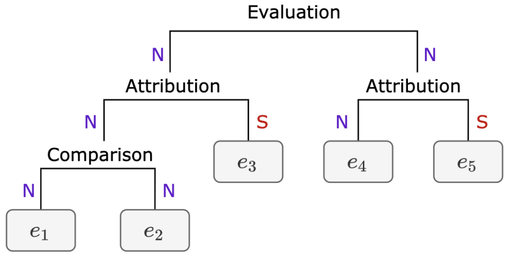
Zhengyuan Liu[†]*, **Ke Shi**[†], **Nancy F. Chen**[†]*
Institute for Infocomm Research, A*STAR, Singapore
{liu_zhengyuan,shi_ke,nfychen}@i2r.a-star.edu.sg

Abstract

Text discourse parsing weighs importantly in understanding information flow and argumentative structure in natural language, making it beneficial for downstream tasks. While previous work significantly improves the performance of RST discourse parsing, they are not readily applicable to practical use cases: (1) EDU segmentation is not integrated into most existing tree parsing frameworks, thus it is not straightforward to apply such models on newly-coming data. (2) Most parsers cannot be used in multilingual scenarios, because they are developed only in English. (3) Parsers trained from single-domain treebanks do not generalize well on out-of-domain inputs. In this work, we propose a document-level multilingual RST discourse parsing framework, which conducts EDU segmentation and discourse tree parsing jointly. Moreover, we propose a cross-translation augmentation strategy to enable the framework to support multilingual parsing and improve its domain generality. Experimental results show that our model achieves state-of-the-art performance on document-level multilingual RST parsing in all sub-tasks.

1 Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is one of the predominant theories for discourse analysis, where a document is represented by a constituency tree with discourse-related annotation. As illustrated in Figure 1, the paragraph is split to segments named Elementary Discourse Units (EDUs), as the leaf nodes of the tree, and they are further connected by rhetorical relations (*e.g., Elaboration*, *Attribution*) to form larger text spans until the entire document is included. The spans are further categorized to *Nucleus* (the core part) or *Satellite* (the subordinate part) based on their relative importance in the rhetorical relations. Thus, document-level RST discourse parsing consists of four sub-tasks: EDU segmentation, tree structure construction, nuclearity determination, and relation classification. Since discourse parsing provides structural information of the narrative flow, downstream natural language processing applications, such as reading comprehension (Gao et al., 2020), sentiment analysis (Bhatia et al., 2015), and text summarization (Liu and Chen, 2019), can benefit from incorporating semantic-related information.



e_1[The European Community's consumer price index rose a provisional 0.6% in September from August] e_2[and was up 5.3% from September 1988,] e_3[according to Eurostat, the EC's statistical agency.] e_4[The month-to-month rise in the index was the largest since April,] e_5[Eurostat said.]

Figure 1: One constituency tree with RST discourse annotation. e_i, N and S denote elementary discourse units, nucleus, and satellite, respectively. Nuclearity and discourse relations are labeled on each span pair.

RST discourse parsing has been an active research area, especially since neural approaches and large-scale pre-trained language models were introduced. On the test set of the English RST benchmark (Carlson et al., 2002), the performance of automatic parsing is approaching that of human annotators. However, compared with other off-the-shelf text processing applications like machine translation, RST parsers are still not readily applicable to massive and diverse samples due to the following challenges: (1) Most parsers take EDU segmentation as a pre-requisite data preparation step, and only conduct evaluations on samples with

[†]Equal Contribution. *Corresponding Author.

Proceedings of the 2nd Workshop on Computational Approaches to Discourse, pages 154–164
November 10–11, 2021. ©2021 Association for Computational Linguistics

gold EDU segmentation. Thus it is not straightforward to utilize them to parse raw documents. (2) Parsers are primarily optimized and evaluated in English, and are not applicable on multilingual scenarios/tasks. Human annotation under the RST scheme is labor-intensive and requires specialized linguistic knowledge, resulting in a shortage of training data especially in low resource languages. (3) Data sparsity also leads to limited generalization capabilities in terms of topic domain and language variety, as the monolingual discourse treebanks usually concentrate on a specific domain. For instance, the English RST corpus is comprised of Wall Street Journal news articles, thus its parser might not perform well on scientific articles.

In this paper, to tackle the aforementioned challenges, we propose a joint framework for document-level multilingual RST discourse analysis. To achieve parsing from scratch, we enhance a top-down discourse parsing model with joint learning of EDU segmentation. Since the well-annotated RST treebanks in different languages share the same underlying linguistic theory, data-driven approaches can benefit from joint learning on multilingual RST resources (Braud et al., 2017a). Inspired by the success of mixed multilingual training (Liu et al., 2020), we further propose a cross-translation data augmentation strategy to improve RST parsing in both language and domain coverage.

We conduct extensive experiments on RST treebanks from six languages: English, Spanish, Basque, German, Dutch, and Portuguese. Experimental results show that our framework achieves state-of-the-art performance in different languages and on all sub-tasks. We further investigate the model's zero-shot generalization capability, by assessing its performance via language-level cross validation. Additionally, the proposed framework can be readily extended to other languages with existing treebanks. The pre-trained model is built as an off-the-shelf application, and can be applied in an end-to-end manner.

2 Related Work

RST Discourse Parsing Discourse structures describe the organization of documents/sentences in terms of rhetorical/discourse relations. The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) are the two most prominent theories of discourse analysis, where they are at document level and sentence level respectively. The structure-aware document analysis has shown to be useful for downstream natural language processing tasks, such as sentiment analysis (Bhatia et al., 2015) and reading comprehension (Gao et al., 2020). Many studies focused on developing automatic computational solutions for discourse parsing. Statistical approaches utilized various linguistic characteristics such as N-gram and lexical features, syntactic and organizational features (Sagae, 2009; Hernault et al., 2010; Li et al., 2014; Heilman and Sagae, 2015), and had obtained substantial improvement on the English RST-DT benchmark (Carlson et al., 2002). Neural networks have been making inroads into discourse analysis frameworks, such as attention-based hierarchical encoding (Li et al., 2016) and integrating neural-based syntactic features into a transition-based parser (Yu et al., 2018). Lin et al. (2019) explored encoder-decoder neural architectures on sentence-level discourse analysis, with a top-down parsing procedure. Recently, pre-trained language models were introduced to document-level discourse parsing, and boosted the overall performance (Shi et al., 2020).

Multilingual Parsing Aside from the English treebank, datasets in other languages have also been introduced and studied, such as German (Stede and Neumann, 2014), Dutch (Redeker et al., 2012), and Basque (Iruskieta et al., 2013). The main challenge of multilingual discourse parsing is the sparsity of annotated data. Braud et al. (2017a) conducted a harmonization of discourse treebanks across annotations in different languages, and Iruskieta and Braud (2019) used multilingual word embeddings to train systems on under-resourced languages. Recently, Liu et al. (2020) proposed a multilingual RST parser by utilizing cross-lingual language model and EDU segment-level translation, obtaining substantial performance gains.

EDU Segmentation EDU segmentation identifies the minimal text spans to be linked by discourse relations. It is the first step in building discourse parsers, and often studied as a separated task in discourse analysis. Existing segmenters on the English discourse corpus achieve sentence-level results with 95% F1 scores (Li et al., 2018), while document-level segmentation is more challenging. Muller et al. (2019) proposed a discourse segmenter that supports multiple languages and schemes. Recently, taking segmentation as a se-

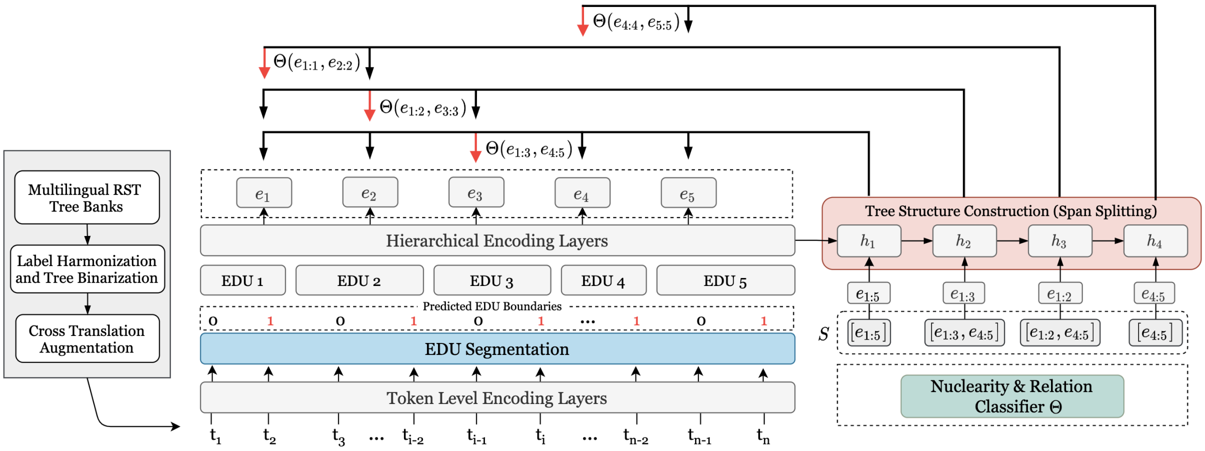

Θ(e_{4:4}, e_{5:5})
Θ(e_{1:1}, e_{2:2})
Θ(e_{1:2}, e_{3:3})
Θ(e_{1:3}, e_{4:5})
Multilingual RST Tree Banks
Label Harmonization and Tree Binarization
Cross Translation Augmentation
e1
e2
e3
e4
e5
Hierarchical Encoding Layers
EDU 1
EDU 2
EDU 3
EDU 4
EDU 5
Predicted EDU Boundaries
0 1 0 1 0 1 ... 1 0 1
EDU Segmentation
Token Level Encoding Layers
t1 t2 t3 ... ti-2 ti-1 ti ... tn-2 tn-1 tn
Tree Structure Construction (Span Splitting)
h1
h2
h3
h4
e1:5
e1:3
e1:2
e4:5
S
[e1:5]
[e1:3, e4:5]
[e1:2, e4:5]
[e4:5]
Nuclearity & Relation Classifier Θ

Figure 2: The architecture of the proposed joint document-level neural parser. A segmenter is first utilized to predict the EDU breaks, and a hierarchical encoder is used to generate the EDU representations. Then, the pointer-network-based decoder and the relation classifier predict the tree structure, nuclearity, and rhetorical relations. t, e and h denote input tokens, encoded EDU representations, and decoded hidden states. The stack S is maintained by the decoder to track top-down depth-first span splitting. With each splitting pointer k, sub-spans $e_{i:k}$ and $e_{k+1:j}$ are fed to a classifier Φ for nuclearity and relation determination.

quence labeling task was shown to be effective in reaching strong segmentation results. Fusing syntactic features to language models was also introduced (Desai et al., 2020). In this work, to the best of our knowledge, we are the first to build a joint framework for document-level multilingual RST discourse analysis that supports parsing from scratch, and can be potentially extended to any language by text-level transformation.

3 Methodology

In this section, we elaborate on the proposed joint multilingual RST discourse parsing framework. We first integrate EDU segmentation into a top-down Transformer-based neural parser, and show how to leverage dynamic loss weights to control the balance of each sub-task. We then propose cross-translation augmentation to improve the multilingual and domain generalization capability.

3.1 Transformer-based Neural Parser

The neural model consists of an EDU segmenter, a hierarchical encoder, a span splitting decoder for tree construction, and a classifier for nuclearity/relation determination.

3.1.1 EDU Segmentation

The EDU segmentation aims to split a document into continuous units and is usually formulated to detect the span breaks. In this work, we conduct it as a sequence labeling task (Muller et al., 2019; Devlin et al., 2019). Given a document containing n tokens, an embedding layer is employed to generate the token-level representations $T = \{t_1, ..., t_n\}$, in particular, a pre-trained language backbone is used to leverage the resourceful prior knowledge. Instead of detecting the beginning of each EDU as in previous work (Muller et al., 2019), here we propose to predict both EDU boundaries via token-level classification. In detail, a linear layer is used to predict the type of each token in one EDU span, i.e., at the begin/intermediate/end position.[1] For extensive comparison, we also implement another segmenter by using a pointer mechanism (Vinyals et al., 2015). Results in Table 3 show that the token-level classification approach consistently produces better performance.

3.1.2 Hierarchical Encoding

To obtain EDU representations with both local and global views, spans are hierarchically modeled from token and EDU-level to document-level. For the document containing n tokens, the initial EDU-level representations are calculated by averaging the token embeddings $t_{i:j}$ of each EDU, where i, j are its boundary indices. Then they are fed into a Bidirectional-GRU (Cho et al., 2014) to capture context-aware representations at the document level. Boundary information has been shown to be effective in previous discourse parsing studies (Shi et al., 2020), thus we also incorporate boundary embeddings from both ends of each EDU to

[1] For the EDU that only contains one token, its begin and end position are the same.

implicitly exploit the syntactic features such as part-of-speech (POS) and sentential information. Then, the ensemble representations are fed to a linear layer, and we obtain the final contextualized EDU representations $E = \{e_1, ..., e_m\}$, where m is the total number of EDUs.

3.1.3 Tree Structure Construction

The constituency parsing process is to analyze the input by breaking down it into sub-spans also known as constituents. In previous studies (Lin et al., 2019; Shi et al., 2020), with a generic constituency-based decoding framework, the discourse parsing results of depth-first and breadth-first manner are similar. Here the decoder builds the tree structure in a top-down depth-first manner. Starting from splitting a span with the entire document, a pointer network iteratively decides the delimitation point to divide a span into two sub-spans, until it reaches the leaf nodes with only one EDU. As the parsing example illustrated in Figure 2, a stack S is maintained to ensure the parsing is conducted under the top-down depth-first manner, and it is initialized with the span containing all EDUs $e_{1:m}$. At each decoding step, the span $e_{i:j}$ at the head of S is popped to the pointer network to decide the split point k based on the attention mechanism (Bahdanau et al., 2015).

$$s_{t,u} = \sigma(h_t, e_u) \ \ \textbf{for} \ \ u = i...j \qquad (1)$$

$$a_t = \text{softmax}(s_t) = \frac{\exp(s_{t,u})}{\sum_{u=i}^{j} \exp(s_{t,u})} \qquad (2)$$

where $\sigma(x, y)$ is the dot product used as the attention scoring function. The span $e_{i:j}$ is split into two sub-spans $e_{i:k}$ and $e_{k+1:j}$. The sub-spans that need further processing are pushed to the top of the stack S to maintain depth-first manner. The decoder iteratively parses the spans until S is empty.

3.1.4 Nuclearity and Relation Classification

At each decoding step, a bi-affine classifier is employed to predict the nuclearity and rhetorical relations of two sub-spans $e_{i:k}$ and $e_{k+1:j}$ split by the pointer network. More specifically, the nuclearity labels *Nucleus* (N) and *Satellite* (S) are attached together with rhetorical relation labels (e.g., *NS-Evaluation*, *NN-Background*). In particular, the EDU representations are first fed to a dense layer with Exponential Linear Unit (ELU) activation for latent feature transformation, and then a bi-affine layer (Dozat and Manning, 2017) with softmax activation is adopted to predict the nuclearity and rhetorical relations.

3.2 Dynamic Weighted Loss

The training objective of our framework is to minimize the sum of the loss $\mathcal{L}_e$ of document-level EDU segmentation, the loss $\mathcal{L}_s$ of parsing the correct tree structure, and the loss $\mathcal{L}_l$ of predicting the corresponding nuclearity and relation labels:

$$\mathcal{L}_e(\theta_e) = -\sum_{n=1}^{N} \text{log} P_{\theta_e}(y_n|X) \qquad (3)$$

$$\mathcal{L}_s(\theta_s) = -\sum_{t=1}^{T} \text{log} P_{\theta_s}(y_t|y_1, ..., y_{t-1}, X) \qquad (4)$$

$$\mathcal{L}_l(\theta_l) = -\sum_{m=1}^{M}\sum_{r=1}^{R} \text{log} P_{\theta_l}(y_m = r|X) \qquad (5)$$

$$\mathcal{L}_{total}(\theta) = \lambda_1 \mathcal{L}_e(\theta_e) + \lambda_2 \mathcal{L}_s(\theta_s) + \lambda_3 \mathcal{L}_l(\theta_l) \qquad (6)$$

where X is the given document, θ_e, θ_s and θ_l are the parameters of the EDU segmenter, the tree structure decoder, and the nuclearity-relation classifier, respectively. N and T are the total token number and span number. $y_1, ..., y_{t-1}$ denote the sub-trees that have been generated in the previous steps. M is the number of spans with at least two EDUs, and R is the total number of pre-defined nuclearity-relation labels.

To find the balance of training multiple objectives, we adopt the adaptive weighting (Liu et al., 2019) to dynamically control the weights of multiple tasks. Specifically, each task k is weighted by λ_k, where λ_k is calculated as:

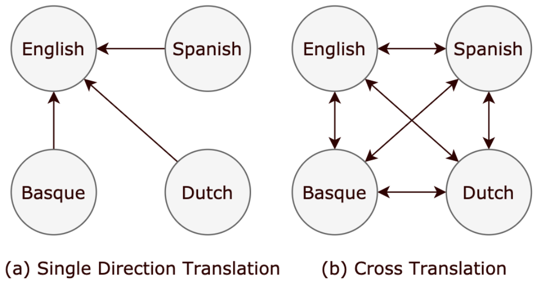
$$w_k(i-1) = \frac{\mathcal{L}_k(i-1)}{\mathcal{L}_k(i-2)} \qquad (7)$$

$$\lambda_k(i) = \frac{K \cdot \exp(w_k(i-1)/Temp)}{\sum_j \exp(w_j(i-1)/Temp)} \qquad (8)$$

where i is the training iterations, K is the task number, and $Temp$ represents the temperature value that smooths the loss from re-weighting. In our experimental settings, adopting dynamic weighted loss brought about relative 2.5% improvement on all sub-tasks.

3.3 Cross Translation Augmentation

Data augmentation is an effective approach to tackle the drawbacks of low resource training by creating additional data from existing samples. For



(a) Single Direction Translation

(b) Cross Translation

Figure 3: Overview of single direction translation (a) and cross-translation strategy (b). Here we take 4 languages as an example. Arrows denote the translate directions.

Treebank Lang.	Train No.	Dev No.	Test No.
English (En)			
- English RST-DT	309	38	38
- English GUM-DT	78	18	18
Portuguese (Pt)	256	38	38
Spanish (Es)	203	32	32
German (De)	142	17	17
Dutch (Nl)	56	12	12
Basque (Eu)	84	28	28

Table 2: The collected RST discourse treebanks from 6 languages. We use the split of train, developmental and test set, as well as the data pre-processing following (Braud et al., 2017a).

English (Source Text)
e_1[The European Community's consumer price index rose a provisional 0.6% in September from August] e_2[and was up 5.3% from September 1988,] e_3[according to Eurostat, the EC's statistical agency.] e_4[The month-to-month rise in the index was the largest since April,] e_5[Eurostat said.]
Dutch (Translated Text)
e_1[De consumentenprijsindex van de Europese Gemeenschap is in september met een voorlopige 0,6% gestegen ten opzichte van augustus] e_2[en steeg met 5,3% ten opzichte van september 1988,] e_3[volgens Eurostat, het statistiekbureau van de EC.] e_4[De maand-op-maand stijging van de index was de grootste sinds april,] e_5[aldus Eurostat.]
Spanish (Translated Text)
e_1[O índice de preços ao consumidor da Comunidade Europeia subiu 0.6 % provisório em setembro ante agosto] e_2[e aumentou 5.3 % em relação a setembro de 1988,] e_3[de acordo com o Eurostat, a agência de estatísticas da CE.] e_4[A alta mensal do índice foi a maior desde abril,] e_5[Eurostat disse.]

Table 1: One example of EDU segment-level translation. The three text samples share the same discourse tree structure, nuclearity, and relation annotation.

instance, back translation, a popular data augmentation method, is widely applied to tasks like machine translation (Edunov et al., 2018). Since the well-annotated RST treebanks in different languages share the same underlying linguistic theory, data-driven approaches can benefit from joint learning on multilingual RST resources. In previous work, Liu et al. (2020) uniformed the multilingual task to a monolingual one by translating all discourse tree samples at the EDU level to English.

In this paper, we propose a cross-translation data augmentation strategy.[2] The method with single direction translation converts all samples to one language in both the training and the inference stage (see Figure 3(a)). This approach cannot exploit the capability of multilingual language backbones. It also increases the test time due to additional computation for translation. In contrast, cross-translation will convert samples from one language to other languages, to produce multilingual training data (see Figure 3(b)). Thus the model is able to process multilingual input during inference. As shown in Table 1, adopting segment-level translation retains the original EDU segmentation as the source text, thus the converted sample in a target language will share the same discourse tree structure and nuclearity/relation labels. We postulate that this text-level transformation will bridge the gaps among different languages. Moreover, since different RST treebanks use articles from different domains (Liu et al., 2020), we speculate that adopting cross-translation can also increase domain coverage in the monolingual space, and further improve the model's overall generalization ability.

4 Experimental Results

In this section, we elaborate on experiment settings of the multilingual RST segmentation and parsing task, compare our proposed framework with previous models, and conduct result analysis.

4.1 Multilingual Dataset

We constructed a multilingual data collection by merging RST treebanks from 6 languages: English (En) (Carlson et al., 2002), Brazilian Portuguese (Pt)[3] (Cardoso et al., 2011; Pardo and Nunes, 2004; Collovini et al., 2007; Pardo and Seno, 2005), Spanish (Es) (Da Cunha et al., 2011), German (De) (Stede and Neumann, 2014), Dutch (Nl) (Redeker et al., 2012), and Basque (Eu) (Iruskieta et al., 2013), and their details are shown in Table

[2] The neural machine translation engine from Google is used: https://cloud.google.com/translate.

[3] The Portuguese RST dataset consists of 140 samples from CST-News (Cardoso et al., 2011), 100 samples from CorpusTCC (Pardo and Nunes, 2004), 50 samples from Summ-it (Collovini et al., 2007), and 40 samples from Rhetalho (Pardo and Seno, 2005).

	English (En)	Portuguese (Pt)	Spanish (Es)	German (De)	Dutch (Nl)	Basque (Eu)
Braud et al. (2017b)	89.5	82.2	79.3	85.1	82.6	-
Muller et al. (2019)	93.7	91.3	88.2	94.1	90.7	85.8
Pointer-Net Segmenter	91.8	92.5	93.6	93.4	94.9	87.3
Boundary CLS Segmenter (Ours)	**96.5**	**92.8**	**93.7**	**95.1**	**95.5**	**88.7**

Table 3: Document-level multilingual EDU Segmentation performance on 6 languages. Micro F1 scores are reported as in (Muller et al., 2019).

	English (En)			Portuguese (Pt)			Spanish (Es)		
Model	Sp.	Nu.	Rel.	Sp.	Nu.	Rel.	Sp.	Nu.	Rel.
Yu et al. (2018)	85.5	73.1	60.2	-	-	-	-	-	-
Iruskieta and Braud (2019)	80.9	65.5	52.1	79.7	62.8	47.8	85.4	65.0	45.8
Cross Rep. (Liu et al., 2020)	87.5	74.7	63.0	86.3	71.7	60.0	86.2	71.1	54.4
Segment Trans. (Liu et al., 2020)	87.8	75.4	63.5	86.5	72.0	60.3	87.9	71.4	56.1
DMRST w/o Cross Trans.	87.9	75.3	64.0	86.5	73.3	61.5	88.2	73.7	60.3
DMRST (Our Framework)	**88.2**	**76.2**	**64.7**	**87.0**	**74.3**	**62.1**	**88.7**	**75.7**	**63.4**
	German (De)			Dutch (Nl)			Basque (Eu)		
Model	Sp.	Nu.	Rel.	Sp.	Nu.	Rel.	Sp.	Nu.	Rel.
Cross Rep. (Liu et al., 2020)	83.6	62.2	45.1	85.9	64.5	49.4	85.1	65.8	47.7
Segment Trans. (Liu et al., 2020)	82.3	58.9	41.0	84.6	62.7	47.2	84.4	65.5	47.3
DMRST w/o Cross Trans.	83.1	62.2	45.9	85.5	64.4	50.6	80.2	59.8	42.1
DMRST (Our Framework)	**84.3**	**64.1**	**47.3**	**85.6**	**66.3**	**52.3**	**85.1**	**67.2**	**48.3**

Table 4: Document-level multilingual RST parsing comparison of baseline models and our framework. *Sp.*, *Nu.*, and *Rel.* denote span splitting, nuclearity determination, and relation classification, respectively. Micro F1 scores of RST Parseval (Marcu, 2000) are reported. Here gold EDU segmentation is used for baseline comparison.

2. We conducted label harmonization (Braud et al., 2017a) to uniform rhetorical definitions among different treebanks. The discourse trees were transformed into a binary format. Unlinked EUDs were removed. Following previous work, we reorganized the discourse relations to 18 categories, and attached the nuclearity labels (i.e., *Nucleus-Satellite* (NS), *Satellite-Nucleus* (SN), and *Nucleus-Nucleus* (NN)) to the relation labels (e.g., *Elaboration*, *Attribution*). For each language, we randomly extracted a set of samples for validation. The original training size was 1.1k, and became 6.7k with cross-translation augmentation. The sub-word tokenizer of the *'XLM-RoBERTa-base'* (Conneau et al., 2020) is used for input pre-processing.

4.2 Evaluation Metrics

For EDU segmentation evaluation, micro-averaged F1 score of token-level segment break classification as in (Muller et al., 2019) was used. For tree parsing evaluation, we applied the standard micro-averaged F1 scores on *Span* (**Sp.**), *Nuclearity-Satellite* (**Nu.**), and *Rhetorical Relation* (**Rel.**), where *Span* describes the accuracy of tree structure construction, *Nuclearity-Satellite* and *Rhetorical Relation* assesses the ability to categorize the nuclearity and the discourse relations, respectively. We also adopted *Full* to evaluate the overall performance considering both *Nuclearity-Satellite* and *Relation* together with *Span* as in (Morey et al., 2017). Following previous studies, we adopted the same 18 relations defined in (Carlson and Marcu, 2001). We reported the tree parsing scores in two metrics: the Original Parseval (Morey et al., 2017) and the RST Parseval (Marcu, 2000) for ease of comparison with previous studies.

4.3 Training Configuration

The proposed framework was implemented with PyTorch (Paszke et al., 2019) and Hugging Face (Wolf et al., 2019). We used *'XLM-RoBERTa-base'* (Conneau et al., 2020) as the language backbone, and fine-tuned its last 8 layers during training. Documents were processed with the sub-word tokenization scheme. The dropout rate of the language backbone was set to 0.2 and that of the rest layers was 0.5. AdamW (Kingma and Ba, 2015) optimization algorithm was used, with the initial learning rate of 2e-5 and a linear scheduler (decay ratio=0.9). Batch size was set to 12. We trained each model for 15 epochs, and selected the best checkpoints on the validation set for evaluation. For each round of evaluation, we repeated the training 5 times with different random seeds and averaged their scores. The

Model	English (En)				Portuguese (Pt)				Spanish (Es)			
	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.
Original Parseval (Morey et al., 2017)												
DMRST (Gold Seg.)	76.7	66.2	56.5	100.0	72.5	61.8	53.1	100.0	79.2	70.3	57.1	100.0
DMRST (Predicted Seg.)	**70.4**	**60.6**	**51.6**	**96.5**	**62.5**	**51.6**	**44.7**	**92.8**	**71.2**	**60.1**	**50.9**	**93.7**
w/o Cross Trans. (Predicted Seg.)	70.3	60.4	51.3	96.4	65.3	53.6	46.3	93.7	70.2	59.3	51.1	93.7
RST Parseval (Marcu, 2000)												
DMRST (Gold Seg.)	88.2	76.2	64.7	100.0	87.0	74.3	62.1	100.0	88.7	75.7	63.4	100.0
DMRST (Predicted Seg.)	**83.2**	**71.1**	**60.5**	**96.5**	**77.8**	**64.9**	**53.2**	**92.8**	**79.5**	**67.4**	**56.7**	**93.7**
w/o Cross Trans. (Predicted Seg.)	83.0	70.8	60.7	96.4	78.4	65.3	54.7	93.7	79.4	66.9	56.5	93.7

Model	German (De)				Dutch (Nl)				Basque (Eu)			
	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.
Original Parseval (Morey et al., 2017)												
DMRST (Gold Seg.)	68.6	45.9	37.1	100.0	71.2	54.1	43.1	100.0	66.6	48.3	34.7	100.0
DMRST (Predicted Seg.)	**58.1**	**40.1**	**32.3**	**95.1**	**62.3**	**46.6**	**39.4**	**95.5**	**53.3**	**39.1**	**31.2**	**88.7**
w/o Cross Trans. (Predicted Seg.)	56.3	39.6	31.2	94.6	63.1	44.9	37.8	95.5	44.4	31.1	23.3	87.8
RST Parseval (Marcu, 2000)												
DMRST (Gold Seg.)	84.3	64.1	47.3	100.0	85.6	66.3	52.3	100.0	85.1	67.2	48.3	100.0
DMRST (Predicted Seg.)	**76.4**	**57.8**	**41.8**	**95.1**	**80.2**	**62.3**	**49.4**	**95.5**	**71.2**	**52.7**	**37.2**	**88.7**
w/o Cross Trans. (Predicted Seg.)	75.4	57.0	41.1	94.6	80.1	61.9	48.3	95.5	66.0	47.5	33.0	87.8

Table 5: Multilingual parsing performance comparison of using gold and predicted EDU segmentation. *Sp.*, *Nu.*, *Rel.* and *Seg.* denote span splitting, nuclearity classification, relation determination, and segmentation, respectively. Micro F1 scores of RST Parseval (Marcu, 2000) and Original Parseval (Morey et al., 2017) are reported. Scores from the proposed framework are in bold for better readability.

total trainable parameter size was 91M, where 56M parameters were from fine-tuning *'XLM-RoBERTa-base'*. All experiments were run on a single Tesla A100 GPU with 40GB memory.

4.4 EDU Segmentation Results

EDU segmentation is the first step of discourse analysis from scratch, and its accuracy is important for the follow-up parsing steps. Thus in this section, we evaluate the performance of our boundary detection segmenter, and compare it with state-of-the-art document-level multilingual EDU segmenters (Braud et al., 2017b; Muller et al., 2019). Additionally, we implemented our model with a pointer mechanism (Vinyals et al., 2015; Li et al., 2018) as a control study.

From the results shown in Table 3, our segmenter outperforms baselines significantly in all languages. This potentially results from adopting the stronger contextualized language backbone (Conneau et al., 2020). Moreover, conducting EDU segmentation in a sequence labeling manner is more computationally efficient, and achieves higher scores than the pointer-based approach, which is consistent with the observation from a recent sentence-level study (Desai et al., 2020).

4.5 Multilingual Parsing Results

We compare the proposed framework with several strong RST parsing baselines: Yu et al. (2018) proposed a transition-based neural parser, obtaining competitive results in English. Iruskieta and Braud (2019) introduced a multilingual parser for 3 languages (English, Portuguese, and Spanish). Liu et al. (2020) proposed a multilingual parser that utilized cross-lingual representation (**Cross Rep.**), and adopted segment-level translation (**Segment Trans.**), and produced state-of-the-art results on 6 languages. Aside from the proposed model (**DMRST**), we added an ablation study on the cross-translation strategy (**DMRST w/o Cross Trans.**). In this section, we use the gold EDU segmentation during the inference stage for a fair comparison to the baselines.

Model	Sp.	Nu.	Rel.	Full
(Zhang et al., 2020)	62.3	50.1	40.7	39.6
(Nguyen et al., 2021)	68.4	59.1	47.8	46.6
DMRST (only EN)	69.8	59.4	49.4	48.6
DMRST (Multilingual)	**70.4**	**60.6**	**51.6**	**50.1**

Table 6: Performance comparison on the English RST treebank with predicted EDU segmentation.

From the results shown in Table 4: (1) Adopting multilingual pre-trained language backbone significantly boosts the RST parsing performance. (2) The multilingual model obtains further improvement with the cross-translation augmentation in all sub-tasks and languages. (3) All sub-tasks are improved substantially compared to previous mul-

	English (En)				Portuguese (Pt)				Spanish (Es)			
Model	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.
Original Parseval (Morey et al., 2017)												
DMRST w/o Cross Trans.	36.9	26.2	17.8	78.4	39.2	29.5	23.1	80.9	40.0	33.0	26.4	76.6
DMRST (Our Framework)	**43.9**	**30.8**	**23.3**	**82.7**	**44.7**	**35.8**	**28.9**	**83.7**	**48.1**	**36.8**	**29.5**	**82.2**
RST Parseval (Marcu, 2000)												
DMRST w/o Cross Trans.	57.8	40.7	27.0	78.4	60.4	44.4	31.8	80.9	58.1	42.8	28.3	76.6
DMRST (Our Framework)	**63.4**	**46.5**	**30.2**	**82.7**	**64.5**	**50.0**	**37.7**	**83.7**	**65.2**	**49.3**	**34.3**	**82.2**
	German (De)				Dutch (Nl)				Basque (Eu)			
Model	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.	Sp.	Nu.	Rel.	Seg.
Original Parseval (Morey et al., 2017)												
DMRST w/o Cross Trans.	43.8	29.3	21.7	87.6	51.8	35.3	27.2	89.0	30.7	17.7	8.5	80.5
DMRST (Our Framework)	**49.0**	**30.7**	**22.8**	**88.2**	**56.5**	**36.0**	**27.1**	**91.0**	**41.0**	**30.1**	**21.3**	**79.1**
RST Parseval (Marcu, 2000)												
DMRST w/o Cross Trans.	66.1	45.4	30.1	87.6	70.6	50.6	36.4	89.0	55.5	32.5	16.8	80.5
DMRST (Our Framework)	**68.9**	**46.2**	**30.3**	**88.2**	**73.9**	**52.3**	**36.1**	**91.0**	**60.3**	**43.3**	**28.3**	**79.1**

Table 7: Zero-shot performance comparison of models w/ and w/o cross-translation strategy. *Sp.*, *Nu.*, *Rel.* and *Seg.* denote span splitting, nuclearity classification, relation determination, and segmentation, respectively. Micro F1 scores of RST Parseval (Marcu, 2000) and Original Parseval (Morey et al., 2017) are reported.

tilingual baselines (Braud et al., 2017a; Liu et al., 2020). Moreover, our model also outperforms the state-of-the-art English RST parsers (see Table 6), demonstrating that fusing multilingual resources is beneficial for monolingual tasks.

4.6 Parsing from Scratch

In most previous work on RST parsing, EDU segmentation is regarded as a separate data preprocessing step, and the test samples with gold segmentation are used for evaluation. However, in practical cases, gold EDU segmentation is unavailable. Thus in this section, we assess the proposed framework with the predicted segmentation, simulating the real-world scenario. We compare our model **DMRST** to the model without cross-translation augmentation (**DMRST w/o Cross Trans.**). Aside from the common metric RST Parseval (Marcu, 2000) used in many prior studies, we also report test results on the Original Parseval (Morey et al., 2017).

From the results shown in Table 5, we observe that: (1) EDU segmentation performance of the two models are similar. This is likely because using lexical and syntactic information is sufficient to obtain a reasonable result. (2) For both metrics, our framework achieves overall better performance in all sub-tasks and languages, especially in the lower resource languages like Basque and Dutch. (3) Since the tree structure and nuclearity/relation classification are calculated on the EDU segments, their accuracy are affected significantly by the incorrect segment predictions. For instance, when gold segmentation is provided, *DMRST* outperforms *DMRST w/o Cross Trans.* at all fronts. However, the former produces slightly lower scores than the latter in Portuguese, due to its suboptimal segmentation accuracy (92.8 vs. 93.7). This also emphasizes the importance of EDU segmentation in a successful end-to-end RST parsing system.

5 Analysis on Zero-Shot Generalization

Incorporating discourse information is beneficial to various downstream NLP tasks, but only a small number of languages possess RST treebanks. Such treebanks have limited annotated samples, and it is difficult to extend their sample size due to annotation complexity. To examine if our proposed multilingual framework can be adopted to languages without any monolingual annotated sample (e.g., Italian, Polish), we conducted a zero-shot analysis via language-level cross validation.

In each round, we select one language as the target language, and RST treebanks from the remaining 5 languages are used to train the multilingual parser. We then evaluate it on the test set from the target language. For example, we assume that a small set of Portuguese articles is to be parsed, and we only have training samples from the other 5 languages (i.e., En, Es, De, Nl, and Eu). Then zero-shot inference is conducted on Portuguese. As shown in Table 7, compared with full training (see Table 5), all the zero-shot evaluation scores drop significantly, especially on English, since the

English corpus is the most resourceful and well-annotated RST treebank. Aside from English, the other 5 languages result in acceptable performance for zero-shot inference. With the cross-translation augmentation, the proposed multilingual discourse parser achieves higher scores, this is because (1) the text transformation helps language-level generalization, and (2) the mixed data have a larger domain coverage. For example, combining samples from Basque (science articles) with English (finance news) makes model perform better on Portuguese (science and news articles). This also suggests that the multilingual parser can be extended to other languages via cross-translation augmentation from existing treebanks of 6 languages.

6 Conclusions

In this work, we proposed a joint framework for document-level multilingual RST discourse parsing, which supports EDU segmentation as well as discourse tree parsing. Experimental results showed that the proposed framework achieves state-of-the-art performance on document-level multilingual discourse parsing on six languages in all aspects. We also demonstrated its inference capability when limited training data is available, and it can be readily extended to other languages.

Acknowledgments

This research was supported by funding from the Institute for Infocomm Research (I2R) under A*STAR ARES, Singapore. We thank Ai Ti Aw for the insightful discussions and Chloé Braud for sharing linguistic resources. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. 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.

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 and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. 2007. Summ-it: Um corpus anotado com informaç oes discursivas visandoa sumarizaç ao automática. *Proceedings of TIL*.

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

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

Takshak Desai, Parag Pravin Dakle, and Dan Moldovan. 2020. Joint learning of syntactic features helps discourse segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1073–1080.

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.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven CH Hoi, Caiming Xiong, Irwin King, and Michael R Lyu. 2020. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Michael Heilman and Kenji Sagae. 2015. Fast rhetorical structure theory discourse parsing. *arXiv preprint arXiv:1505.02425*.

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Mikel Iruskieta, Marıa J Aranzabe, Arantza Diaz de Ilarraza, Itziar Gonzalez, Mikel Lersundi, and Oier Lopez de Lacalle. 2013. The RST Basque treebank: an online search interface to check rhetorical relations. In *4th workshop RST and discourse studies*, pages 40–49.

Mikel Iruskieta and Chloé Braud. 2019. EusDisParser: improving an under-resourced discourse parser with cross-lingual data. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 62–71, Minneapolis, MN. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence*, IJCAI-ECAI-2018, Stockholm, Sweden.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 362–371.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35.

Xiang Lin, Shafiq Joty, Prathyusha Jwalapuram, and M Saiful Bari. 2019. A unified linear-time framework for sentence-level discourse parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4200, Florence, Italy. Association for Computational Linguistics.

Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1871–1880.

Zhengyuan Liu and Nancy Chen. 2019. Exploiting discourse-level segmentation for extractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 116–121, Hong Kong, China. Association for Computational Linguistics.

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

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 rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics*, 26(3):395–448.

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 the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124, Minneapolis, MN. Association for Computational Linguistics.

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.

Thiago Alexandre Salgueiro Pardo and Eloize Rossi Marques Seno. 2005. Rhetalho: um corpus de referência anotado retoricamente. *Anais do V Encontro de Corpora*, pages 24–25.

Thiago AS Pardo and Maria das Graças Volpe Nunes. 2004. Relações retóricas e seus marcadores superficiais: Análise de um corpus de textos científicos em português do brasil. *Relatório Técnico NILC*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

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.

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 81–84. Association for Computational Linguistics.

Ke Shi, Zhengyuan Liu, and Nancy F Chen. 2020. An end-to-end document-level neural discourse parser exploiting multi-granularity representations. *arXiv preprint arXiv:2012.11169*.

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

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570.

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6386–6395, Online. Association for Computational Linguistics.