Coreference for Discourse Parsing: A Neural Approach

Grigorii Guz and Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
{gguz, carenini}@cs.ubc.ca

Abstract

We present preliminary results on investigating the benefits of coreference resolution features for neural RST discourse parsing by considering different levels of coupling of the discourse parser with the coreference resolver. In particular, starting with a strong baseline neural parser unaware of any coreference information, we compare a parser which utilizes only the output of a neural coreference resolver, with a more sophisticated model, where discourse parsing and coreference resolution are jointly learned in a neural multitask fashion. Results indicate that these initial attempts to incorporate coreference information do not boost the performance of discourse parsing in a statistically significant way.

1 Introduction and Task Description

Discourse parsing is a very useful Natural Language Processing (NLP) task involving predicting and analyzing discourse structures, which represent the coherence properties and relations among constituents of multi-sentential documents. In this work, we investigate discourse parsing in the context of Rhetorical Structure Theory (RST) Mann and Thompson (1988), which encodes documents into complete constituency discourse trees. An RST tree is defined on the sequence of a document's EDUs (Elementary Discourse Units), which are clause-like sentences or sentence fragments (propositions), acting as the leaves of the tree. Adjacent EDUs and constituents are hierarchically aggregated to form (possibly non-binary) constituents, with internal nodes containing (1) a nuclearity label, defining the importance of that subtree (rooted at the internal node) in the local context and (2) a relation label, defining the type of semantic connection between the two subtrees (e.g., Elaboration, Background).

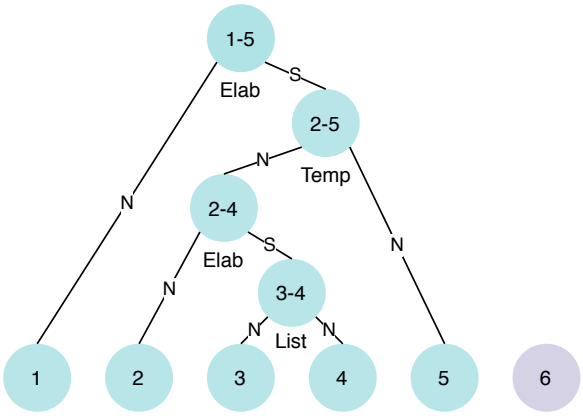

Figure 1: An example (Asher and Lascarides, 2003) of a discourse being ill-formed due to the invalid anaphoric link. The leaf EDUs are as follows: [Max had a great evening last night.]$_1$ [He had a great meal.]$_2$ [He ate salmon.]$_3$ [He devoured lots of cheese.]$_4$ [He then won a dancing competition.]$_5$ [It was a beautiful pink]$_6$

Previous research has shown that the use of RST-style discourse parsing as a system component can enhance important tasks, such as sentiment analysis, summarization and text categorization (Bhatia et al., 2015; Nejat et al., 2017; Hogenboom et al., 2015; Gerani et al., 2014; Ji and Smith, 2017). And more recently, it has been found that RST discourse structures can complement learned contextual embeddings (e.g., BERT (Devlin et al., 2018)), in tasks where linguistic information on complete documents is critical, such as argumentation analysis (Chakrabarty et al., 2019).

In this work, we present preliminary results of investigating the benefits of coreference resolution features for RST discourse parsing. From the theoretical perspective, it has long been established (Asher and Lascarides, 2003) that discourse structure can impose constraints on mention antecedent distributions, with these constraints being derived from the role of each discourse unit (sen-

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 160–167

https://doi.org/10.18653/v1/P17

tence or EDU) with respect to the global discourse. The Veins theory (Cristea et al., 1998) is the most known formalization of anaphoric constraints with respect to RST tree structures, involving assigning to each EDU a subset of preceding EDUs defined by the nuclearity attributes of the EDU's parent nodes in the document's discourse tree (see Appendix A for the exact definition). These constrains act as a domain of referential accessibility where the antecedents must reside, for otherwise the discourse would be considered incoherent. As an example of this phenomenon, consider the discourse structure in Figure 1. In principle, a reader could apply commonsense knowledge to resolve the pronoun *it* in the last sentence to *salmon* in the third sentence, any proficient English speaker would call such a discourse ill-formed and incoherent, due to the fact that it breaks the discourse-imposed antecedent scope. In general, anaphora can only be resolved with respect to the most salient (sentence 1 in Figure 1) units of the preceding discourse (Asher and Lascarides, 2003). For our purposes, this means that having access to a document's coreference structure might be beneficial to the task of predicting the discourse structure, since the coreference structure can constrain the discourse parser's solution space. However, as shown in a corpus study by Zeldes (2017), the antecedent boundaries defined by Veins Theory are often too restrictive, suggesting that while discourse structures can be useful for predicting coreference structures and vice versa, these mutual constrains must be defined softly, at least in the context of RST theory.

To explore these ideas computationally with respect to modern neural models, we investigate the utility of automatically extracted coreference features and discourse-coreference shared representations in the context and for the benefit of neural RST discourse parsing. Our strong baseline SpanBERT-NoCoref utilizes SpanBERT (Joshi et al., 2020) as in the current SOTA coreference resolver, without utilizing any direct coreference information. Next, our SpanBERT-CorefFeats considers the output of coreference resolver as per Dai and Huang (2019), letting us test the benefit of predicted and so possibly noisy coreference features. Finally, our more sophisticated SpanBERT-Multitask model learns discourse parsing together with coreference resolution in the neural multitask learning fashion, sharing the SpanBERT contextual word encoder for both models.

2 Related Work

Dai and Huang (2019) have already explored the benefit of using coreference information for neural PDTB implicit discourse relation classification, in a way similar to our SpanBERT-CorefFeats model. In our study, we also explore the use of shared encoder architecture for both tasks to detect the additional possible synergy.

Modelwise, the most common approach to infer discourse trees is the linear bottom-up shift-reduce method, adopted from syntactic parsing. Wang et al. (2017) uses hand-crafted features and the shift-reduce method predicted by two separate Support-Vector-Machines (SVMs) for structure- and nuclearity-prediction and relation-estimation. The neural model by Yu et al. (2018) uses a similar topology, but instead relies entirely on LSTMs for automatic feature extraction and on a single multi-layer-perceptron (MLP) for classifying all possible actions. Top-down approaches to discourse parsing are also quite promising, with recent work of Kobayashi et al. (2020) applying ELMO (Peters et al., 2018) for computing span representations and achieving the new absolute SOTA performance, reporting however the scores of an ensemble of five independent runs of their proposed model instead of single-model results. In this work we follow the shift-reduce strategy and apply SpanBERT-Base (Joshi et al., 2020; Wolf et al., 2020), which we introduce below, for encoding the document contents.

The field of coreference resolution has recently been dominated by deep learning models. The current SOTA model by Joshi et al. (2020) is built upon the neural coreference resolver of (Lee et al., 2018) by incorporating SpanBERT language model, which modifies the commonly used BERT (Devlin et al., 2019) architecture with a novel span masking pretraining objective. In our work, we re-implemented their coreference resolver in PyTorch (Paszke et al., 2019). Our code for both models is available[1].

3 Shift-Reduce Architecture

All our proposed parsers share the same basic shift-reduce architecture, consisting of a Queue, which is initially filled with documents EDUs in order

[1] http://www.cs.ubc.ca/cs-research/lci/research-groups/natural-language-processing/index.html

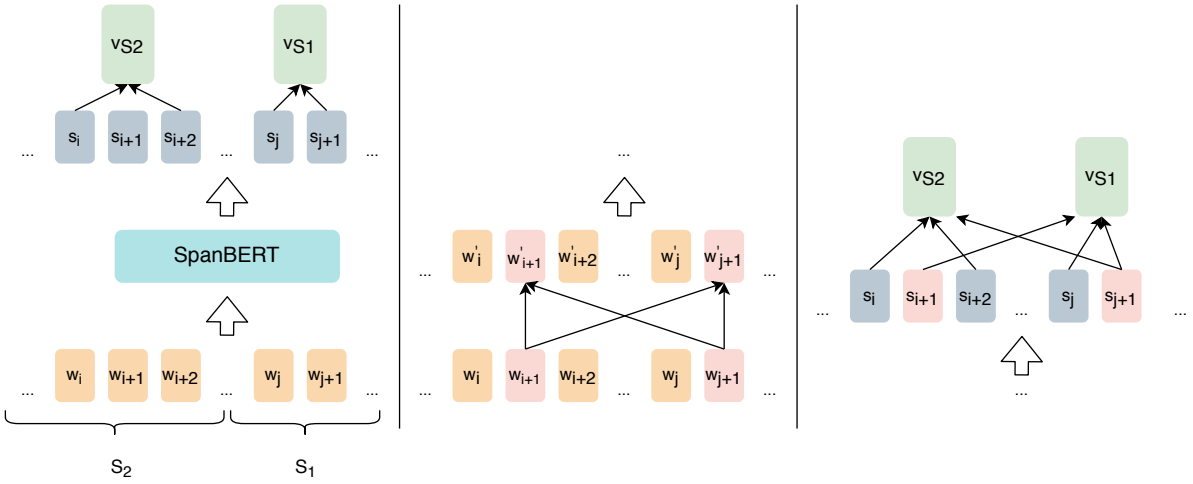

Figure 2: Overview of our models. For spans S_2 are S_1, the $w_{i:i+2}$ and $w_{j,j+1}$ respectively are the nuclear EDUs. (Left) All components of SpanBERT-NoCoref. (Middle) SpanBERT-CorefFeats modifies the initial SpanBERT embeddings according to predicted coreference clusters (in red). (Right) SpanBERT-MultiTask updates the final span representations with embeddings of mentions of shared entities.

from first to last one, and a Stack, which is initially empty, as well as the following actions on them:

The Shift delays aggregations of sub-trees at the beginning of the document by popping the top EDU Q_1 on the queue and pushing it onto the stack.

The Reduce-X aggregates the top subtrees (S_1, S_2) on the stack into a single subtree (S_{1-2}). Each reduce action further defines a nuclearity assignment $X_N \in \{\text{NN, NS, SN}\}$ to the nodes covered by S_{1-2} and a relation $X_R \in \{\text{Elaboration, Contrast, ...}\}$ holding between them.

3.1 Action Classifier Parametrization

Similarly to (Wang et al., 2017; Yu et al., 2018), all models under consideration utilize the information from top two elements S_1, S_2 of the stack, and top element Q_1 of the queue. In addition to word-/word+coference-based representations $v_{S_2}, v_{S_1}, v_{Q_1}$ for these nodes, computed differently by each model as described below, we extract textual organization features of Wang et al. (2017). In particular, for each pair $S_2 - S_1$ and $S_1 - Q_1$, we extract indicator features representing whether the pair is within the same sentence or paragraph; for each of S_2, S_1 and Q_1 we compute whether each of them are at the start/end of a sentence/paragraph/document.

In accord with Wang et al. (2017), the parsing action at each timestep is chosen by two trainable classifiers, being multi-layer perceptrons (MLPs) in our system, where each classifier takes in the concatenation of $v_{S_2}, v_{S_1}, v_{Q_1}$, together with the dense embeddings for the aforementioned organizational features. The first classifer predicts the action and nuclearity assignment among $y_{Act,Nuc} \in \{Shift, Reduce_{NN}, Reduce_{NS}, Reduce_{SN}\}$, and in case the Reduce action is chosen, the second classifier predicts the discourse relation among 18 coarse-grained RST relation classes $y_{Rel} \in \{Attribution, Elaboration, ...\}$.

3.2 Action Classifier Training and Inference

Both classifiers are trained using the Cross-Entropy loss, computed for each Stack-Queue parsing step. At test time, we apply the greedy decoding strategy to predict the discourse structure.

4 Proposed Models

We now describe the three proposed discourse parsing models which differ in the levels of coupling with the coreference model. See Figure 2 for the visual comparison.

SpanBERT-NoCoref: in addition to the organizational features, our baseline system utilizes only the output SpanBERT-contextualized word embeddings. To predict each Stack-Queue action, a full document is passed through SpanBERT in a non-overlapping sliding window fashion, as per Joshi et al. (2020), so that the context of full document can be considered for each parsing action to account for possible context-sensitivity of discourse structures (Dai and Huang, 2018). The node representation v_{Q_1} for the first Queue element is computed as the mean of the first and the last word

Model	Structure	Nuclearity	Relation
HILDA(2010)	82.6	66.6	54.6
DPLP(2014)	82.0	68.2	57.8
CODRA(2015)	82.6	68.3	55.8
Two-Stage(2017)	86.0	72.4	59.7
Transition-Syntax(2018)	85.5	73.1	60.2
D2P2S2E (Ensemble)(2020)	87.0	74.6	60.0
SpanBERT-NoCoref	87.8 ± 0.2	75.8 ± 0.2	63.4 ± 0.3
SpanBERT-CorefFeats	88.1 ± 0.3	76.1 ± 0.6	63.6 ± 0.3
SpanBERT-MultiTask	87.9 ± 0.2	75.9 ± 0.6	63.3 ± 0.7
Human (2017)	88.3	77.3	65.4

Table 1: RST-Parseval micro precision for structure, nuclearity and relation prediction on RST-DT corpus. Scores for previous approaches are from either Morey et al. (2017) or the original papers.

embedding of the EDU that this Queue element represents. v_{S_1} and v_{S_2} are computed as the means of the first and the last word embeddings of the nuclear EDU of S_1 and S_2, as each non-leaf node in an RST structure encodes a relation between nuclear EDUs of its children (Morey et al., 2018).

SpanBERT-CorefFeats: with this architecture variant, we attempt to assess the benefit of coreference features generated by the coreference resolver for RST parsing. Given a document with n words, the coreference features will be used to update the initial (not contextualized) SpanBERT word embeddings $w_{1:n}$, which will later be passed to SpanBERT.

Specifically, for a given document we apply the pre-trained coreference parser of Joshi et al. (2020) to extract the document's coreference clusters $C_1, C_2, ...,$ each of which are equivalence classes representing different mentions of the same entity. Afterwards, we compute the vector representation c_i for each cluster C_i by performing attention-based averaging over word-vectors corresponding to mentions in that cluster:

$$c_i = \sum_{k \in C_i} a_k w_k$$

where $w_k \in \mathbb{R}^d$ is the initial SpanBERT word embedding for word k and $a_k \in [0, 1]$ are attention scores. These cluster representations are then used for updating the document's word representations using the gating mechanism Lee et al. (2018): for each word $w_k \in D$,

$$f_k = \sigma(W[c_i; w_k])$$

$$w'_k = \begin{cases} f_k \circ w_k + (1 - f_k) \circ c_i & \text{if } w_k \in C_i \\ w_k & \text{otherwise} \end{cases}$$

Finally, the embeddings w'_k are passed to SpanBERT for contextualization, and the node representations $v_{S_1}, v_{S_2}, v_{Q_1}$ are computed as in SpanBERT-NoCoref.

SpanBERT-MultiTask: learns discourse parsing and coreference resolution in a multitask learning regime, weight-sharing the SpanBERT encoder module. The coreference resolver training step proceeds in the same fashion as in (Joshi et al., 2020). For updating the discourse parsing model, we use the pre-computed coreference clusters C_i obtained from the pretrained coreference model, as running it at every training step was prohibitively time-consuming. Using the contextualized SpanBERT word embeddings $s_{1:n}$ for all words in the document, we check these coreference clusters for overlaps: considering a pair of spans S_1, S_2, if a cluster C_i has entity mentions in the spans of both stack elements S_1 and S_2, so that if there are mentions $m_j, m_k \in C_i$ such that $m_j \in S_1$ and $m_k \in S_2$, we update the span representation v_{S_1} (computed as in SpanBERT-NoCoref) with the attention weighted sum of mentions $m_k \in C_i \cap S_2$ by applying the gating mechanism as in SpanBERT-CorefFeats, so that the span representation for S_1 can incorporate more relevant context from S_2. The representation for v_{S_2} is computed similarly using mentions $m_k \in C_i \cap S_1$, and the analogous computation is performed for $S_1 - Q_1$ pair.

For learning both tasks at the same time, we utilize the approach similar to (Sanh et al., 2018), where gradient updates are performed separately for each task and the probability of sampling a task is proportional to the relative size of each task's dataset. The initial shared SpanBERT encoder weights are set from the pretrained coreference resolver checkpoint.

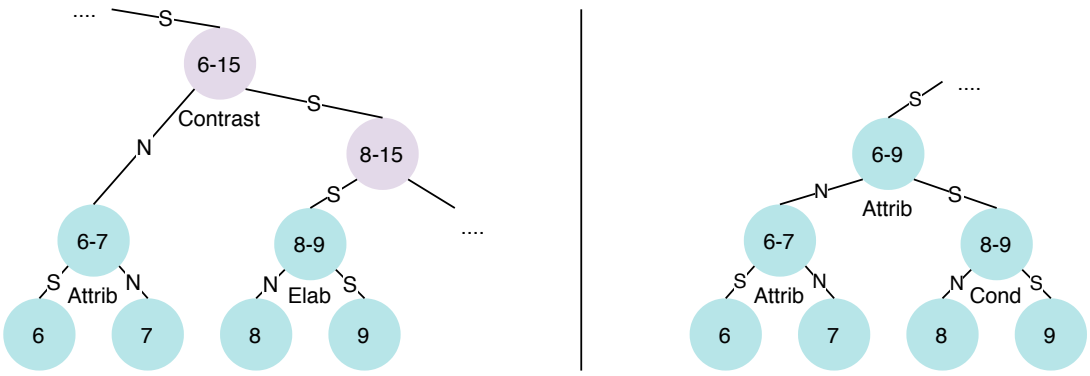

Figure 3: A subtree from SpanBERT-NoCoref prediction for *wsj_0631* (left) and gold-standard (right). Incorrect nodes are colored in purple. The EDUs are: [Finnair and SAS said]$_6$ [they plan to swap stakes in each other.]$_7$ [Neither discussed details]$_8$ [pending board meetings next month.]$_9$.

5 Experimental Settings and Results

All models were trained on the RST-DT (Carlson et al., 2002) and evaluated with RST-Parseval procedure (Marcu, 2000), with the coreference component of SpanBERT-MultiTask being trained on full OntoNotes 5.0 corpus (Weischedel et al., 2013). The details of the training procedure such as hyperparameter assignment are outlined in the Appendix B. The test results are presented in Table 1 and are the average and standard deviation single-model scores of five independent runs.

Firstly, we observe that our models strongly outperform all previous approaches, indicating huge benefit of pretrained language models for RST discourse parsing, with results approaching human performance. Then, with respect to coreference features, we notice that the models utilizing coreference information are statistically equivalent in performance to the SpanBERT-NoCoref baseline, while displaying higher variance of the test scores for Nuclearity and Relation prediction. This suggests four plausible (and not mutually exclusive) explanations: (1) the coreference information relevant to discourse parsing is already captured by SpanBERT, (2) or that coreference information is not a strong signal for discourse structure (Zeldes, 2017), or that (3) the coreference information extracted automatically is too noisy, or that (4) our specific ways of combining coreference with discourse parsing are not adequate and more work is needed to develop better solutions. It should also be noted that we only experimented on a single discourse parsing dataset, so the conclusions or generalizations should be considered preliminary.

In an attempt to shed some light on the results, we compare the predicted and gold subtree from one of the documents in our development set on Figure 3. The trees were analyzed using the RST tree visualization tool by Huber (2019). According to Veins theory, the pronoun [*neither*] in EDU 8 is a mention that should have access to its mentions ([*Finnair and SAS*] or [*they*]) in preceding EDUs. However, accoding to the discourse structure predicted by SpanBERT-NoCoref, the vein for node (8) does not contain the EDUs (6) and (7) (and in fact any of its preceding EDUs), so that [*neither*] cannot be linked to any of its preceding mentions. On the other hand, according to the gold discourse structure, EDU 8 has EDU 7 on its vein, meaning that this anaphora can be resolved. This means that if one had access to gold coreference structure and applied Veins Theory strictly, the substructure produced by SpanBERT-NoCoref would not be permitted.

6 Conclusions and Future Work

We empirically compare different levels of coupling between a shift-reduce neural discourse parser and a neural coreference resolver. Remarkably, our baseline delivers SOTA performance on RST-DT, but does not seem to benefit from coreference features.

For future work, we plan to experiment with (1) alternative discourse parsing architectures and approaches for neural multitasking, along with more powerful coreference models (2) alternative ways of augmenting a neural discourse parser with coreference information and other tasks like summarization (3) improving the coreference resolution performance by leveraging information provided by a discourse parser.

References

Ralph Weischedel et al. 2013. Ontonotes release 5.0. *Linguistic Data Consortium*.

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

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.

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

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathleen McKeown, and Alyssa Hwang. 2019. Ampersand: Argument mining for persuasive online discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2926–2936.

Dan Cristea, Nancy Ide, and Laurent Romary. 1998. Veins theory: A model of global discourse cohesion and coherence. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 281–285, Montreal, Quebec, Canada. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2976–2987, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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

Alexander Hogenboom, Flavius Frasincar, Franciska De Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Commun. ACM*, 58(7):69–77.

Patrick Huber. 2019. Discourse-sentiment alignment tool (dsat).

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)*, volume 1, pages 13–24.

Yangfeng Ji and Noah A Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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

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*, 34:8099–8106.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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*.

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.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Bita Nejat, Giuseppe Carenini, and Raymond Ng. 2017. Exploring joint neural model for sentence level discourse parsing and sentiment analysis. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 289–298.

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. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2018. A hierarchical multi-task approach for learning embeddings from semantic tasks.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

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.

Amir Zeldes. 2017. A distributional view of discourse encapsulation: Multifactorial prediction of coreference density in RST. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 20–28, Santiago de Compostela, Spain. Association for Computational Linguistics.

A Veins Theory Definitions

The following definitions are from Cristea et al. (1998). For each node in an RST tree, its head is defined as follows:

1. The head of the terminal (leaf) node is itself.
2. The head of a non-terminal node is the concatenation of the heads of its nuclear children.

Next, we define the vein expression of each node recursively top-down. When the node is a leaf, the preceding nodes on its vein correspond to its domain of referential accessibility.

1. The vein expression of the root is its head.
2. For each nuclear node, its vein expression is the union of its head with:
   - its parent's vein, if this node has no left siblings that are satellites.
   - its parent's vein and its left sibling's vein, if this sibling is a satellite.
3. For every satellite node, its vein expression is the union of its head with:
   - its parent's vein, if this node is a left child.
   - its parent's vein with heads of prior (up in the tree) satellite nodes removed.

B Hyperparameters and Training Settings

As RST-DT does not specify a standard training-validation split, we select 10% of the training documents for the validation set, stratifying the split by the number of EDUs in each document. Similarly to Joshi et al. (2020), we train all of our models with AdamW (Loshchilov and Hutter, 2019) optimizer with learning rate of $1e^{-5}$ for SpanBERT and $2e^{-4}$ for model-specific components, with the batch size of 5 and linear decay for 20 epochs. All of our MLPs consist of 2 linear layers, with a GeLU (Hendrycks and Gimpel, 2016) nonlinearity and a Dropout layer with a value of 0.3 between them. Each organizational feature of Wang et al. (2017) is represented using a learnable 10-dimentional embedding, or a vector of zeros if the feature is missing (for example, the feature specifying if the 2-top elements of the stack are in the same sentence when the stack contains only one element). With regards to multitask regime, the probability of discourse parsing task being sampled over coreference resolution was ≈ 0.72 (each Stack-Queue state was treated as a datapoint), but due to highly demanding computational requirements of the coreference resolver and time constrains, this probability was increased to 0.9. Nonetheless, the results for the correct task proportions will be provided through other sources.