ESURF: Simple and Effective EDU Segmentation

Anonymous ACL submission

Abstract

Segmenting text into Elemental Discourse Units (EDUs) is a fundamental task in discourse parsing. We present a new simple method for identifying EDU boundaries, and hence segmenting them, based on lexical and character n-gram features, using random forest classification. We show that the method, despite its simplicity, outperforms other methods both for segmentation and within a state of the art discourse parser. This indicates the importance of such features for identifying basic discourse elements, pointing towards potentially more training-efficient methods for discourse analysis.

1 Introduction

A fundamental task in natural language understanding is analyzing the overall structure of a text, so that logical and coherence relations between text units are revealed. Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is a well-acccepted theoretical framework for the task within the NLP community (Kobayashi et al., 2020). RST structures a text as a tree, where the basic building blocks (leaf node) are called Elementary Discourse Units (EDUs). Discourse parsing in RST is the task of automatically constructing this hierarchical tree by identifying the EDUs and then building a parse tree by connecting adjacent EDUs and composite discourse units. Relations between adjacent units are labeled with different rhetorical relations, which are mostly assymetrical, with one unit designated the the nucleus and the other as the subordinate satellite. Parsing is generally done using a shift-reduce parser, which builds the tree incrementally by scoring transition actions (Yu et al., 2018a; Mabona et al., 2019). Recently, neural network models have achieved state-of-the-art performance in this task by leveraging sophisticated neural modules (Zhang et al., 2020)).

Rhetorical Structure Theory (RST) offers a robust approach for discourse analysis by constructing a rhetorical structure tree that captures the relationships between text elements, enhancing performance in various tasks. Although previous research efforts have advanced machine learning methods for discourse segmentation and parsing, these often rely on lexical and syntactic clues, hand-crafted features, and syntactic parse trees, and use gold-standard segmentation for training and evaluation (Yu et al., 2022; Feng and Hirst, 2014b; Ali, 2023). This coherence structure is essential for applications such as text summarization, and sentiment analysis. RST-based analysis significantly improves discourse understanding and contributes to more effective NLP applications (Nguyen et al., 2021; Liu et al., 2021).

Despite the successes of contextualized pre-trained language models (PLMs) like XLNet (Yang et al., 2020) in RST discourse parsing, challenges remain due to data insufficiency, reliance on lexical and syntactic clues, and inconsistencies between EDU-level parsing and sentence-level contextual modeling, as well as dependence on gold-standard segmentation for training. These issues, particularly the reliance on hand-crafted features and parse trees, have made EDU segmentation a significant bottleneck. In this paper, we propose a novel method for EDU segmentation which gives state-of-the-art (SOTA) performance, showing that local lexical and morphological cues can do most of the work.

We conduct experiments using the RST Discourse Treebank (RST-DT) and CNN/Daily Mail. First, we derive EDU segmentation and evaluate it with various classifiers, including transformers. We then test our proposed EDU identification method using a transition-based neural RST parser (Yu et al., 2022). Our results demonstrate improvements in EDU identification and RST parsing, with our model outperforming others and improving automated RST parsing techniques.

2 Related work

Historically, discourse processing using RST has been approached as a parsing task, using transition-based or chart parsers (Luong et al., 2015; Dai and Huang, 2019; Li et al., 2022). In recent years, performance has been improved over earlier methods by incorporating statistical models for predicting nuclearity and relation types between discourse units (Yu et al., 2018b; Kobayashi et al., 2020; Zhang et al., 2020; Guz and Carenini, 2020; Koto et al., 2021a). Such neural approaches now dominate, but many still incorporate hand-crafted features for better performance. Seq2Seq models have also been applied to both sentence and document-level parsing (Liu et al., 2019; Luong et al., 2015; Dai and Huang, 2019).

A key component of discourse parsing is identifying the Elementary Discourse Units (EDUs) defined as smallest text spans. Early methods relied on handcrafted features and syntactic clues (Mann and Thompson, 1988; Lan et al., 2013). Recent neural models like BERT and XLNet have advanced EDU segmentation and discourse coherence (Zhang et al., 2021a; Yu et al., 2022).

Some recent work has focused on developing better parsing methods independent of EDU segmentation, by using a gold-standard segmentation for training and evaluating RST parsers, and employing top-down approaches with sequence labeling for RST parsing (Nguyen et al., 2021; Mabona et al., 2019; Koto et al., 2021a). Such work gives strong baselines for parsing using different methods.

As noted above, we focus on the core subtask of EDU segmentation, and will evaluate our method both for segmentation accuracy and for its effect on parsing accuracy.

3 EDU Segmentation Using Random Forests (ESURF)

The significance of EDUs in RST parsing is crucial due to their fundamental role in understanding discourse structures. EDUs represent the smallest coherent "thought units" within a text and are the parts of which the overall discourse structure is composed. Hence, accurate segmentation and identification of EDUs is essential for an accurate analysis of rhetorical structure (Yu et al., 2018b; Lin, 2023).

We present here a comparatively simple yet highly effective method for EDU segmentation, which we call *EDU Segmentation Using Random Forests (ESURF)*, as illustrated in Fig.1. ESURF formulates the problem of EDU segmentation as a classification problem. The system considers every nine-token[1] subsequence of the text $(t_{i-3}, t_{i-2}, t_{i-1}, t_i, ..., t_{i+5})$,as a possible context for an EDU boundary, giving three tokens before and six tokens after the candidate boundary (immediately preceding t_i). The input features for classification are the individual tokens t_k given per their position in the context window, marked as ***Before*** $(t_{i-3}, t_{i-2}, t_{i-1})$, ***Leading*** (t_i, t_{i+1}, t_{i+2}), or ***Continuing*** $(t_{i+3}, t_{i+4}, t_{i+5})$ the candidate EDU. To account for morphology in a simple way, we also incorporated character subsequences of the tokens as potential features, along with their positional indices within the 9-gram sequence similarly marked as B, L, or C. (Fig. 2).

These were filtered to keep only the character subsequences that appeared in multiple corpus texts, but not in most of them, as a simple measure of informativeness.

We train a Random Forest classifier on these examples using these features. The classifier is then used to classify all candidate EDU boundaries in new text, processing each 9-token window as above for classification. Each sequence of tokens between boundaries classified as positive (i.e., as an EDU boundary) is identified as an EDU. Figure 1 gives a schematic diagram of the overall system.

4 Experiments

We perform two sets of evaluations. First, we compare the performance of ESURF against other classification methods and other EDU segmentation methods from the literature on the task of EDU segmentation. Next, we evaluate the effect on RST parsing performance of using ESURF for segmentation, as compared with the methods used in various recent RST parsing systems. The experiments were conducted on an Ubuntu 20.04.4 server with 256 CPUs (2000 MHz each) and 512 GB of RAM.

4.1 Datasets

In all of our experiments, we follow the practice of the baselines and use the RST Discourse Treebank (RST-DT) dataset, (Carlson et al., 2002), which is

[1]Some subsequences are shorter, as we do not consider sequences that cross sentence boundaries.

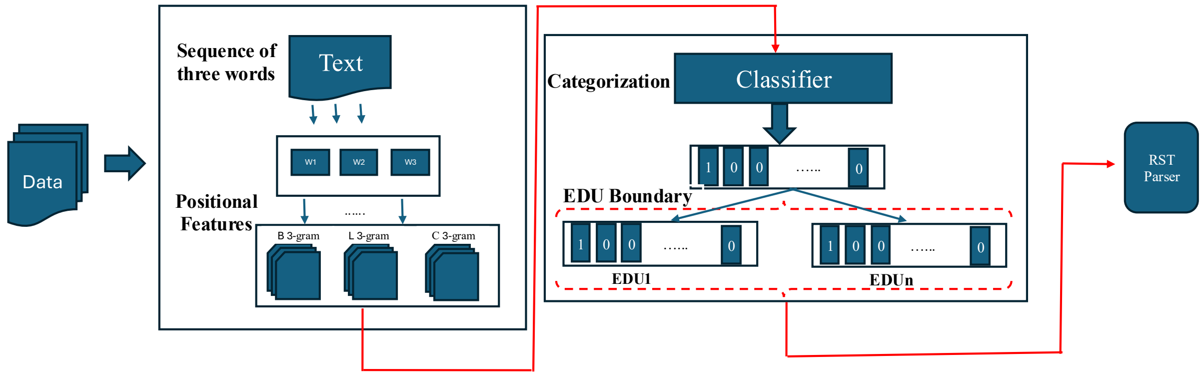

Figure 1: Illustration of ESURF framework

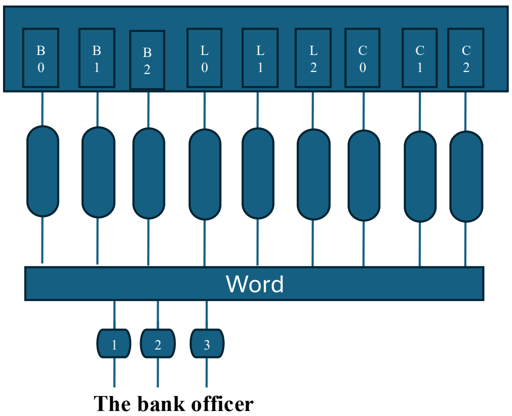

Figure 2: Illustration of a toy example for positional features

the most common and widely used in RST parsing, EDU segmentation studies, and text-based research (Lin, 2023; Wang et al., 2018; Joty et al., 2012; Pastor and Oostdijk, 2024). RST-DT consists of 347 training articles and 38 test articles annotated with full RST discourse structures and is widely used in text-based research. Additionally, we also evaluate ESURF on the CNN/Daily Mail dataset (Nallapati et al., 2016; Kobayashi et al., 2021), which includes over 300,000 articles.

4.2 ESURF on CNN/DailyMail and RST-DT Datasets

We evaluate ESURF against various classifiers on the task of classifying sections as EDUs or non-EDUs. This comparison includes CRF (as a classifier), a 3-layer MLP (Multi-layer Perceptron), BERT (bert-base-uncased), and XLNet, using the CNN/Daily Mail and RST-DT datasets. The evaluation is performed on a similarly sized subset of data points (50% positive / 50% negative) with preprocessing consistent with our previous approach. As shown in Table 1, ESURF outperforms the other models in accuracy, precision, recall, and F1 score.

For the CNN/Daily Mail dataset, ESURF achieves an accuracy of 91.5% and an F1-score of 91.4%, outperforming BERT and the other segmenters in this comparison.

These results show that, despite its simplicity, ESURF is highly effectiveness in EDU segmentation, indicating the centrality of lexical and morphological context as cues for discourse segmentation.

We further evaluate our model against several established discourse segmenters, which are widely recognized as baselines in EDU segmentation studies, as shown in Table 2. For a fair comparison with our model, the same dataset is used. Comparison includes JCN, which uses a Logistic Regression model with features from sentence context, combining syntactic tree structures and statistical estimates. We also assess CRF and WLY, which apply sequence labeling and a BiLSTM-CRF framework, HILDA (HIL) and SPADE (SP), which employ statistical models integrating syntactic and lexical information to identify discourse boundaries and build sentence-level discourse trees, and Joint, which employs a pointer network with a depth-first parsing strategy to construct discourse trees.

Results on the RST-DT test set show that our model, ESURF, again outperforms the other methods. Specifically, while the Joint Model achieves an F1-score of 95.5%, ESURF improves upon this with an F1-score of 96 .1%. This improvement in performance suggests a potential increase in RST parsing accuracy.

4.3 RST Parsing Using ESURF

We evaluate the impact of various parsing methods, including our EDU segmentation model, by using it in a state-of-the-art RST parser. Our analysis

Model	Acc. (CNN)	Prec. (CNN)	Rec. (CNN)	F1 (CNN)	Acc. (RST-DT)	Prec. (RST-DT)	Rec. (RST-DT)	F1 (RST-DT)
SVM	0.862	0.884	0.859	0.871	0.892	0.891	0.896	0.893
CRF	0.891	0.849	0.881	0.865	0.928	0.858	0.933	0.894
Gradient Boosted	0.859	0.851	0.888	0.868	0.873	0.850	0.899	0.874
ESURF	**0.915**	**0.912**	**0.918**	**0.914**	**0.958**	**0.944**	**0.979**	**0.961**
BERT	0.889	0.884	0.853	0.869	0.877	0.931	0.841	0.884
XLNet	0.615	0.515	0.586	0.549	0.662	0.532	0.507	0.519
MLP	0.785	0.822	0.759	0.789	0.793	0.840	0.761	0.798
XGboost	0.862	0.848	0.837	0.842	0.896	0.939	0.874	0.905

Table 1: Classifier performance for EDU identification on the CNN/Daily Mail and RST-DT datasets. ESURF achieved the highest metrics for both.

Model	Precision	Recall	F1 Score
Hill (Hernault et al., 2010)	0.779	0.706	0.741
SP(Soricut and Marcu, 2003)	0.838	0.868	0.852
CRF (Feng and Hirst, 2014a)	0.903	0.918	0.905
JCN (Joty et al., 2012)	0.880	0.923	0.901
WYL (Wang et al., 2018)	0.924	0.944	0.932
F&R (Fisher and Roark, 2007)	0.913	0.897	0.905
Joint Model (Lin, 2023)	0.933	0.978	0.955
ESURF	**0.944**	**0.979**	**0.961**

Table 2: Performance comparison of various established EDU segmentation methods for RST-DT.

Model	S	N	R	F
(Yu et al., 2022)	0.764	0.661	0.545	0.535
(Zhang et al., 2021b)	0.763	0.655	0.556	0.538
(Yu et al., 2018b)	0.714	0.603	0.492	0.481
(Zhang et al., 2020)	0.672	0.555	0.453	0.443
(Nguyen et al., 2021)	0.743	0.643	0.516	0.502
(Koto et al., 2021b)	0.731	0.623	0.515	0.503
(Yu et al., 2022)+ESURF	**0.783**	**0.673**	**0.564**	**0.562**

Table 3: Performance comparison of various RST Parser with metrics S , N , R , and F.

demonstrates that enhancing EDU segmentation significantly improves discourse parsing performance. we employ a Shift-reduce transition-based neural RST parser to assess the effectiveness of our EDU segmentation method. We employ the parser developed by(Yu et al., 2022), which leverages neural RST parsing to produce action sequences from EDU representations.

For evaluation, we adopt the framework proposed by (Morey et al., 2017; Shahmohammadi and Stede, 2024; Yu et al., 2022; Zhang et al., 2021b), which evaluates performance using micro-averaged F1 across four scoring metrics: **S**pan (tree structure without labels), **N**uclearity (structure with just nuclearity labels), **R**elation (structure with just relation labels), and **F**ull (structure with both nuclearity and relation labels). These metrics are widely used in RST parsing research, ensuring consistency with the existing literature. They provide a comprehensive view of the parsing performance, capturing different levels of structural information. This approach allows for a balanced evaluation of the model's effectiveness, offering insights into both the structural accuracy and the quality of label assignments. In Table 3, we compare our results with various leading RST parsers.

Our ESURF segmentation method improves the performance of the transition-based RST parser from (Yu et al., 2022) by approximately 2.48% in span, 1.66% in nuclearity, more than 3.48% in relationship, and 5% in full metrics. (Yu et al., 2022) reimplemented the EDU segmenter from Muller (Muller et al., 2019) for segmenting large-scale unlabeled texts. This improvement highlights how improved EDU segmentation can enhance RST parsing performance.

5 Conclusion and Future Work

We demonstrate that our method, ESURF, despite its simplicity, achieves SOTA performance on EDU segmentation and also improves SOTA RST parsing performance. This shows that lexical and morphological context give strong cues for identifying basic discourse structure constituents. For future work, we plan to apply ESURF to large unlabeled datasets like the GUM corpus for semi-supervised EDU segmentation, potentially improving training efficiency for lower-resource languages. In addition, we will explore advanced feature extraction and embedding techniques to enhance performance and deepen our understanding of discourse structure by identifying linguistic cues and predicting discourse markers.

6 Limitation

This study solely focuses on sentence-level discourse parsing and assumes accurate sentence segmentation. Future work should relax this assumption by using an automated sentence segmenter, explore different features and parameter settings to evaluate their impact, and evaluate/compare methods on a broader range of datasets. Furthermore, future work should explore advanced feature extraction and embedding techniques to enhance performance and improve the identification of linguistic cues and discourse markers.

References

Omar Ali. 2023. *Fuzzy text segmentation using syntactic features for rhetorical structure theory*. Ph.D. thesis, University of Portsmouth.

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

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.

Vanessa Wei Feng and Graeme Hirst. 2014a. 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.

Vanessa Wei Feng and Graeme Hirst. 2014b. Two-pass discourse segmentation with pairing and global features. *arXiv preprint arXiv:1407.8215*.

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.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167.

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

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down rst parsing utilizing granularity levels in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8099–8106.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural rst parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1600–1612.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021a. Top-down discourse parsing via sequence labelling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 715–726, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021b. Top-down discourse parsing via sequence labelling. *arXiv preprint arXiv:2102.02080*.

Man Lan, Yu Xu, and Zheng-Yu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 476–485.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):165329.

Xiang Lin. 2023. Natural language processing as autoregressive generation.

Jinxian Liu, Bingbing Ni, Caiyuan Li, Jiancheng Yang, and Qi Tian. 2019. Dynamic points agglomeration for hierarchical point sets learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7546–7555.

Zhengyuan Liu, Ke Shi, and Nancy F Chen. 2021. Dmrst: A joint framework for document-level multilingual rst discourse segmentation and parsing. *arXiv preprint arXiv:2110.04518*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Amandla Mabona, Laura Rimell, Stephen Clark, and Andreas Vlachos. 2019. Neural generative rhetorical structure parsing. *arXiv preprint arXiv:1909.11049*.

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.

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 *Conference on Empirical Methods on Natural Language Processing (EMNLP 2017)*, pages pp–1330.

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.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, and 1 others. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. Rst parsing from scratch. *arXiv preprint arXiv:2105.10861*.

Martial Pastor and Nelleke Oostdijk. 2024. Signals as features: Predicting error/success in rhetorical structure parsing. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 139–148.

Sara Shahmohammadi and Manfred Stede. 2024. Discourse parsing for german with new rst corpora. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 65–74.

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.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

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

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

Nan Yu, Meishan Zhang, Guohong Fu, and Min Zhang. 2022. Rst discourse parsing with second-stage edu-level pre-training. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4269–4280.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021a. Adversarial learning for discourse rhetorical structure parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3946–3957, Online. Association for Computational Linguistics.

Longyin Zhang, Fang Kong, and Guodong Zhou. 2021b. Adversarial learning for discourse rhetorical structure parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3946–3957.

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. *arXiv preprint arXiv:2005.02680*.