Toward Fast and Accurate Neural Discourse Segmentation

Yizhong Wang **Sujian Li** **Jingfeng Yang**
MOE Key Lab of Computational Linguistics, School of EECS, Peking University
{yizhong, lisujian, yjfllpyym}@pku.edu.cn

Abstract

Discourse segmentation, which segments texts into Elementary Discourse Units, is a fundamental step in discourse analysis. Previous discourse segmenters rely on complicated hand-crafted features and are not practical in actual use. In this paper, we propose an end-to-end neural segmenter based on BiLSTM-CRF framework. To improve its accuracy, we address the problem of data insufficiency by transferring a word representation model that is trained on a large corpus. We also propose a restricted self-attention mechanism in order to capture useful information within a neighborhood. Experiments on the RST-DT corpus show that our model is significantly faster than previous methods, while achieving new state-of-the-art performance. [1]

1 Introduction

Discourse segmentation, which divides text into proper discourse units, is one of the fundamental tasks in natural language processing. According to Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), a complex text is composed of non-overlapping Elementary Discourse Units (EDUs), as shown in Table 1. Segmenting text into such discourse units is a key step in discourse analysis (Marcu, 2000) and can benefit many downstream tasks, such as sentence compression (Sporleder and Lapata, 2005) or document summarization (Li et al., 2016).

> [Mr. Rambo says]$_{e_1}$ [that a 3.2-acre property]$_{e_2}$ [overlooking the San Fernando Valley]$_{e_3}$ [is priced at \$4 million]$_{e_4}$ [because the late actor Erroll Flynn once lived there.]$_{e_5}$

Table 1: A sentence that is segmented into five EDUs

Since EDUs are initially designed to be determined with lexical and syntactic clues (Carlson et al., 2001), existing methods for discourse segmentation usually design hand-crafted features to capture these clues (Feng and Hirst, 2014). Especially, nearly all previous methods rely on syntactic parse trees to achieve good performance. But extracting such features usually takes a long time, which contradicts the fundamental role of discourse segmentation and hinders its actual use. Considering the great success of deep learning on many NLP tasks (Lu and Li, 2016), it's a natural idea for us to design an end-to-end neural model that can segment texts fast and accurately.

The first challenge of applying neural methods to discourse segmentation is data insufficiency. Due to the limited size of labeled data in existing corpus (Carlson et al., 2001), it's quite hard to train a data-hungry neural model without any prior knowledge. In fact, some traditional features, such as the POS tags or parse trees, naturally provide strong signals for identifying EDUs. Removing them definitely increases the difficulty of learning an accurate model. Secondly, many EDU boundaries are actually not determined locally. For example, to recognize the boundary between e_3 and e_4 in Table 1, our model has to be aware that e_3 is an embedded clauses starting from "overlooking", otherwise it could regard "San Fernando Valley" as the subject of e_4. Such kind of long-distance dependency can be precisely extracted from parse trees but is difficult for neural models to capture.

To address these challenges, in this paper, we propose a neural discourse segmenter based on the BiLSTM-CRF (Huang et al., 2015) framework and further improve it from two aspects. Firstly, since the discourse segmentation corpus is too small to learn precise word representations, we transfer a word representation model trained on a large corpus into our task, and show that this trans-

[1] Our code is available at https://github.com/PKU-TANGENT/NeuralEDUSeg

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 962–967
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

ferred model can provide very useful information for our task. Secondly, in order to model long-distance dependency, we employ the self-attention mechanism (Vaswani et al., 2017) when encoding the text. Different from previous self-attention, we restrict the attention area to a neighborhood of fixed size. The motivation is that effective information for determining the boundaries is usually collected from adjacent EDUs, while the whole text may contain many disturbing words, which could mislead the model into incorrect decisions. In summary, the contributions of this work are as follows:

- Our neural discourse segmentation model doesn't rely on any syntactic features, while it can outperform other state-of-the-art systems and achieve significant speedup.
- To our knowledge, we are the first to transfer word representations learned from large corpus into discourse segmentation task and show that they can significantly alleviate the data insufficiency problem.
- Based on the nature of discourse segmentation, we propose a restricted attention mechanism , which enables the model to capture useful information within a neighborhood but ignore unnecessary faraway noises.

2 Neural Discourse Segmentation Model

We model discourse segmentation as a sequence labeling task, where the start word of each EDU (except the first EDU) is supposed to be labeled as 1 and other words are labeled as 0. Figure 1 gives an overview of our segmentation model. We will introduce the BiLSTM-CRF framework in Section 2.1, and describe the two key components of our model in Section 2.2, 2.3.

2.1 BiLSTM-CRF for Sequence Labeling

Conditional Random Fields (CRF) (Lafferty et al., 2001) is an effective method to sequence labeling problem and has been widely used in many NLP tasks (Sutton and McCallum, 2012). To approach our discourse segmentation task in a neural way, we adopt the BiLSTM-CRF model (Huang et al., 2015) as the framework of our system. Formally, given an input sentence $\mathbf{x} = \{x_t\}_{t=1}^{n}$, we first embed each word into a vector $\mathbf{e}_t$. Then these word embeddings are fed into a bi-directional LSTM layer to model the sequential information:

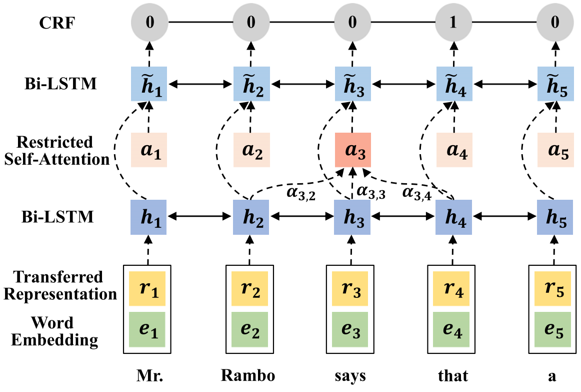

Figure 1: Overview of our model for discourse segmentation

$$\mathbf{h}_t = \text{BiLSTM}(\mathbf{h}_{t-1}, \mathbf{e}_t) \tag{1}$$

where $\mathbf{h}_t$ is the concatenation of the hidden states from both forward and backward LSTMs. After encoding this sentence, we make labeling decisions for each word. Instead of modeling the decisions independently, the CRF layer computes the conditional probability $p(\mathbf{y}|\mathbf{h}; \mathbf{W}, \mathbf{b})$ over all possible label sequences $\mathbf{y}$ given $\mathbf{h}$ as follows:

$$p(\mathbf{y}|\mathbf{h}; \mathbf{W}, \mathbf{b}) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, \mathbf{h})}{\sum\limits_{y' \in \mathcal{Y}} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, \mathbf{h})} \tag{2}$$

where $\psi_i(y_{i-1}, y_i, \mathbf{h}) = \exp(\mathbf{w}^T \mathbf{h}_i + b)$ is the potential function and $\mathcal{Y}$ is the set of possible label sequences. The training objective is to maximize the conditional likelihood of the golden label sequence. During testing, we search for the label sequence with the highest conditional probability.

2.2 Transferring Representations Learned from Large Corpus

Due to the large parameter space, neural models usually require much more training data in order to achieve good performance. However, to the best of our knowledge, nearly all existing discourse segmentation corpora are limited in size. After we remove all the syntactic features, which has been proven useful in many previous work (Bach et al., 2012; Feng and Hirst, 2014; Joty et al., 2015), it's expected that our neural model will not achieve very satisfying results.

To tackle this issue, we propose to leverage model learned from other large datasets, aiming that this transferred model has been well trained to encode text and capture useful signals. Instead of training the transferred model by ourselves, in this paper, we adopt the ELMo word representations (Peters et al., 2018), which are derived from a bidirectional language model (BiLM) trained on one billion word benchmark corpus (Chelba et al., 2014). Specifically, this BiLM has one character convolution layer and two biLSTM layers, and correspondingly there are three internal representations for each word x_t, which are denoted as $\{\mathbf{h}_{t,l}^{\text{LM}}\}_{l=1}^{3}$. Following (Peters et al., 2018), we compute the ELMo representation $\mathbf{r}_t$ for word x_t as follows:

$$\mathbf{r}_t = \gamma^{\text{LM}} \sum\nolimits_{l=0}^{3} s_l^{\text{LM}} \mathbf{h}_{t,l}^{\text{LM}} \tag{3}$$

where $\mathbf{s}^{\text{LM}}$ are normalized weights and γ^{LM} controls the scaling of the entire ELMo vector. Then we concatenate $\mathbf{r}_t$ with the word embedding $\mathbf{e}_t$, and take them as the input of Equation (1).

2.3 Restricted Self-Attention

As we have introduced in Section 1, some EDU boundaries rely on relatively long-distance signals to recognize, while normal LSTM model is still weak at this. Recently, self-attention mechanism, which relates different positions of a single sequence, has been successfully applied to many NLP tasks (Vaswani et al., 2017; Wang et al., 2017) and shows its superiority in capturing long dependency. However, we found that most boundaries are actually only influenced by nearby EDUs, thereby forcing the model to attend to the whole sequence will bring in unnecessary noises. Therefore, we propose a restricted self-attention mechanism, which only collects information from a fixed neighborhood. To do this, we first compute the similarity between current word x_i and each nearby word x_j within a window:

$$s_{i,j} = \mathbf{w}_{attn}^{T} [\mathbf{h}_i, \mathbf{h}_j, \mathbf{h}_i \odot \mathbf{h}_j] \tag{4}$$

Then the attention vector $\mathbf{a}_i$ is computed as a weighted sum of nearby words:

$$\alpha_{i,j} = \frac{e^{s_{i,j}}}{\sum_{k=-K}^{K} e^{s_{i,i+k}}} \tag{5}$$

$$\mathbf{a}_i = \sum\nolimits_{j=-K}^{K} \alpha_{i,i+k} \mathbf{h}_{i+k} \tag{6}$$

where hyper-parameter K is the window size. This attention vector $\mathbf{a}_i$ is then put into another BiLSTM layer together with $\mathbf{h}_i$ in order to fuse the information:

$$\tilde{\mathbf{h}}_t = \text{BiLSTM}(\tilde{\mathbf{h}}_{t-1}, [\mathbf{h}_t, \mathbf{a}_t]) \tag{7}$$

We use $\tilde{\mathbf{h}}_t$ as the new input to the CRF layer.

3 Experiments and Results

3.1 Dataset and Metrics

We conduct experiments on the RST Discourse Treebank (RST-DT) (Carlson et al., 2001). The original corpus contains 385 Wall Street Journal articles from the Penn Treebank, which are divided in to training set (347 articles, 6132 sentences) and test set (38 articles, 991 sentences). We randomly sample 34 (10%) articles from the train set as validation set in order to tune the hyper-parameters and only train our model on the remained train set. We follow mainstream studies (Soricut and Marcu, 2003; Joty et al., 2015) to measure segmentation accuracy only with respect to the intra-sentential segment boundaries, and we report Precision (P), Recall (R) and F1-score (F1) for segmentation performance.

3.2 Implementation Details

We tune all the hyper-parameters according to the model performance on the separated validation set. The 300-D Glove embeddings (Pennington et al., 2014) are employed and kept fixed during training. We use the AllenNLP toolkit (Gardner et al., 2018) to compute the ELMo word representations. The hidden size of our model is set to be 200 and the batch size is 32. L2 regularization is applied to trainable variables with its weight as 0.0001 and we use dropout between every two layers, where the dropout rate is 0.1. For model training, we employ the Adam algorithm (Kingma and Ba, 2014) with its initial learning rate as 0.0001 and we clip the gradients to a maximal norm 5.0. Exponential moving average is applied to all trainable variables with a decay rate 0.9999. The window size K for restricted attention is set to be 5.

3.3 Performance

The results of our model and other competing systems on the test set of RST-DT are shown in Table 2. We compare our results against the following systems: (1) **SPADE** (Soricut and Marcu, 2003) is an early system using simple lexical and syntactic features; (2) **NNDS** (Subba and Di Eugenio, 2007) uses a neural network classifier to do the

Model	Tree	P(%)	R(%)	F1(%)
SPADE	Gold	84.1	85.4	84.7
NNDS	Gold	85.5	86.6	86.0
CRFSeg	Gold	92.7	89.7	91.2
Reranking	Gold	**93.1**	94.2	93.7
CRFSeg	Stanford	91.0	87.2	89.0
CODRA	BLLIP	88.0	92.3	90.1
Reranking	Stanford	91.5	90.4	91.0
Two-Pass	BLLIP	92.8	92.3	92.6
Our Model	No	92.9	**95.7**	**94.3**
- Attention	No	92.4	94.8	93.6
- ELMo	No	87.9	84.5	86.2
- Both	No	87.0	82.8	84.8
Human	No	98.5	98.2	98.3

Table 2: Performance of our model and other systems on the RST-DT test set [3]

segmentation after extracting features; (3) **CRFSeg** (Hernault et al., 2010) is the first discourse segmenter using CRF model; (4) **CODRA** (Joty et al., 2015) uses fewer features and a simple logistic regression model to achieve impressive results; (5) **Reranking** (Bach et al., 2012) reranks the N-best outputs of a base CRF segmenter; (6) **Two-Pass** (Feng and Hirst, 2014) conducts a second segmentation after extracting global features from the first segmentation result. All these methods rely on tree features and we list their performance given different parse trees, where **Gold** are the trees extracted from the Penn Treebank (Prasad et al., 2005), **Stanford** represents trees from the Stanford parser (Klein and Manning, 2003) and **BLLIP** represents those from the BLLIP parser (Charniak and Johnson, 2005). It should be noted that the results of SPADE and CRFSeg are taken from Bach et al. (2012) since the original papers adopt different evaluation metrics. All the other results are taken from the corresponding original papers.

From Table 2, we can see that our model achieves state-of-the-art performance without extra parse trees. Especially, if no gold parse trees are provided, our system outperforms other methods by more than 1.7 points in F1 score. Since the gold parse trees are not available when processing new sentences, this improvement becomes more valuable when the system is put into use.

System	Speed (Sents/s)	Speedup
Two-Pass	1.39	1.0x
SPADE	3.78	2.7x
Ours (Batch=1)	9.09	6.5x
Ours (Batch=32)	76.23	54.8x

Table 3: Speed comparison with two open-sourced discourse segmenter

To further explore the influence of different components in our model, we also report the results of ablation experiments in Table 2. We can see that the transferred ELMo representations provide the most significant improvement. This accords with our assumption that the RST-DT corpus itself is not large enough to train an expressive neural model sufficiently. With the help of the transferred representations, we are capable of capturing more semantic and syntactic signals. Also, comparing the models with and without the restricted self-attention, we find that this attention mechanism can further boost the performance. Especially, if there are no ELMo vectors, the improvement provided by the attention mechanism is more noticeable.

3.4 Speed Comparison

We also measure the speedup of our model against traditional systems in Table 3. The **Two-Pass** system has the best performance among all existing methods, while **SPADE** is much simpler with less features. We test these systems on the same machine (CPU: Intel Xeon E5-2690, GPU: NVIDIA Tesla P100). The results show that our system is 2.4-6.5 times faster than the compared systems if the batch size is 1. Moreover, if we process the test sentences in parallel, we can achieve 20.2-54.8 times speedup with the batch size as 32. This makes our system more practical in actually use.

3.5 Effect of Restricted Self-Attention

We propose to restrict the self-attention within a neighborhood instead of the whole sequence. Table 4 demonstrates the performance of our model over different window size K. We can see that all these results is better than the performance our model without attention mechanism. However, a proper restriction window is helpful for the attention mechanism to take better effect.

[3] In parallel with our work, Li et al. (2018) proposes another neural model with its performance as: P-91.6, R-92.8, F1-92.2. We didn't see their paper at the time of submission, but it's worth mentioning here for the readers' reference.

Window Size	1	5	10	∞
F1-score	94.0	94.3	94.2	93.8

Table 4: Performance of our model over different attention window size

4 Conclusion

In this paper, we propose a neural discourse segmenter that can segment text fast and accurately. Different from previous methods, our segmenter doesn't rely on any hand-crafted features, especially the syntactic parse tree. To achieve our goal, we propose to leverage the word representations learned from large corpus and we also propose a restricted self-attention mechanism. Experimental results on RST-DT show that our system can achieve state-of-the-art performance together with significant speedup.

Acknowledgments

We thank the anonymous reviewers for their insightful comments on this paper. This work was partially supported by National Natural Science Foundation of China (61572049 and 6187022165). The corresponding author of this paper is Sujian Li.

References

Ngo Xuan Bach, Minh Le Nguyen, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the SIGDIAL 2012 Conference, The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 5-6 July 2012, Seoul National University, Seoul, South Korea*, pages 160–168.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pages 173–180.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 2635–2639.

Vanessa Wei Feng and Graeme Hirst. 2014. Two-pass discourse segmentation with pairing and global features. *CoRR*, abs/1407.8215.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

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

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Shafiq R. Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center, Sapporo, Japan.*, pages 423–430.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4166–4172.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 13-15 September 2016, Los Angeles, CA, USA*, pages 137–147.

Zhengdong Lu and Hang Li. 2016. Recent progress in deep learning for NLP. In *Tutorial Abstracts, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pages 11–13.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Daniel Marcu. 2000. *The theory and practice of discourse parsing and summarization*. MIT press.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *CoRR*, abs/1802.05365.

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The penn discourse treebank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156. Association for Computational Linguistics.

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

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

Charles A. Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6000–6010.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*.