Improving Context Modeling in Neural Topic Segmentation

Anonymous ACL submission

Abstract

Topic segmentation is critical in key NLP tasks and recent works favor highly effective neural supervised approaches. However, current neural solutions are arguably limited in how they model context. In this paper, we enhance a segmenter based on a hierarchical attention Bi-LSTM network to better model context, by adding a coherence-related auxiliary task and restricted self-attention. Our optimized segmenter outperforms SOTA approaches on four challenging real-world datasets.

1 Introduction

Topic Segmentation is the fundamental NLP task of splitting a document into topic-coherent pieces. It can reveal important aspects of a document semantic structure that can benefit other downstream tasks such as document summarization (Mitra et al., 1997; Riedl and Biemann, 2012; Xiao and Carenini, 2019), question answering (Oh et al., 2007; Diefenbach et al., 2018) and machine reading (van Dijk, 1981; Saha et al., 2019).

Recently, several works have framed topic segmentation as neural supervised learning, because of the remarkable success achieved by such models in most NLP tasks. While one line of research builds neural models to predict segment boundaries directly (Wang et al., 2016; Koshorek et al., 2018; Badjatiya et al., 2018); another, first trains neural models for other tasks (e.g., sentence-pair coherence prediction), and then uses these models' outputs to predict boundaries (Wang et al., 2017; Arnold et al., 2019). Despite architectural differences, all these neural solutions are limited in how they model context. In essence, local contextual information is critical in predicting topical boundaries, but simple Recurrent Neural Network (RNN) and its variants are arguably not sufficiently powerful to represent the necessary information.

In this paper, we propose to enhance a SOTA segmenter based on a hierarchical attention Bi-LSTM network (Yang et al., 2016) to better model context in two complementary ways. On the one hand, we add a coherence-related auxiliary task to make our model learn more informative hidden states for all the sentences in a document. More specifically, we refine the loss function of our model to ensure that the coherence of the sentences from different segments is smaller than the coherence of the sentences from the same segment. On the other hand, we enhance context modeling by restricted self-attention (Wang et al., 2018), which enables our model to make better use of the information from the closer neighborhood of each sentence. An evaluation on four real-world test sets shows that our context modeling strategy significantly improves the performance of the SOTA neural segmenter.

2 Neural Topic Segmentation Model

In this paper, we frame topic segmentation as a sequence labeling task. More precisely, given a document represented as a sequence of sentences, our model will predict a binary label for each sentence to indicate if the sentence is the **end** of a topical coherent segment or not. Our neural topic segmentation model is shown in Figure 1.

2.1 Hierarchical Attention Bi-LSTM Network (HAN)

Our basic segmenter is inspired by Koshorek et al. (2018), which used a hierarchical Bi-LSTM network to capture the document semantic features on different levels. Formally, a sentence encoding network returns sentence embeddings from pre-trained word embeddings of sentences. Then a label prediction network processes the resulting sentence embeddings and outputs the probabilities to indicate if sentences are the segment boundaries or not. While Koshorek et al. (2018) applied max-pooling

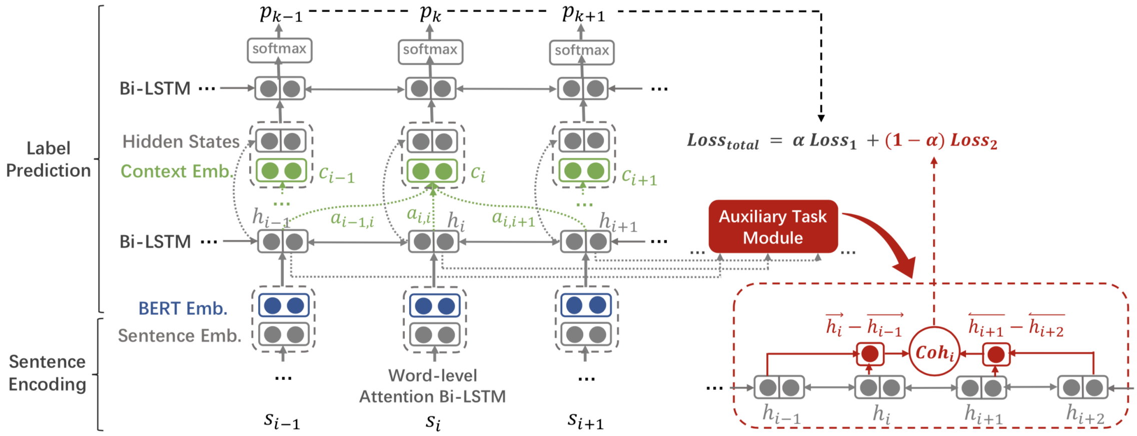

Figure 1: Our topic segmentation model with added components: restricted attention (green), coherence task (red)

to build sentence embeddings from sentence encoding network, in this paper, we applied an attention mechanism (Bahdanau et al., 2015; Yang et al., 2016) to make the model better capture task-wise sentence semantics. After applying a softmax function, for each sentence, we will obtain the probability that it is a segment boundary. To convert the numerical probabilities into binary labels, we need to set a threshold τ. All the sentences with probabilities over τ will be labeled 1. τ is set in the validation stage.

We compute the cross-entropy loss between the ground truth labels $Y = \{y_1, ..., y_{k-1}\}$ and our predicted probabilities $P = \{p_1, ..., p_{k-1}\}$ for a document with k sentences:

$$L_1 = \sum_{i=1}^{k-1} [-y_i \log p_i - (1 - y_i) \log(1 - p_i)] \quad (1)$$

Enhancing Generality with BERT Embeddings To better deal with unseen text in test data, we utilize a pre-trained BERT sentence encoder[1] (Devlin et al., 2019) which complements our sentence encoding network. The transformer-based BERT model was trained on multi-billion sentences publicly available on the web, and therefore captures generic semantic signals. To combine BERT with task-specific information, we simply concatenate the BERT sentence embeddings with the sentence embeddings derived from our model. Such concatenation is then the input of the next level network.

2.2 Auxiliary Task Learning

In a well-structured document, the semantic coherence of a pair of sentences from the same segment should be greater than the coherence of a pair of sentences from different segments. This observation provides us with an alternative way to enable better context modeling by formulating a coherence-related auxiliary task whose objective can be jointly optimized with our original objective (Equation 1). This task is to predict the sentence-pair coherence by using the sentence hidden states generated from the Bi-LSTM network. Concurrently minimizing the loss of this task can regulate our model to reduce the semantic coherence *between segments* and increase the semantic coherence *within a segment*.

To obtain the ground truth of our introduced auxiliary task (sentence-pair coherence prediction), we leverage our segmented training set rather than requesting external annotations. For a document which contains m sentences, there are $m - 1$ consecutive sentence pairs. If this document has n segment boundaries, then among those $m - 1$ sentence pairs, n sentence pairs are from different segments, while the remaining $m - n - 1$ sentence pairs are from the same segment. In order to minimize the coherence of the sentences from different segments and maximize the coherence of the sentences in the same segment, we follow the similar strategy in Wang et al. (2017) by giving a sentence pair $sp_i =< s_i, s_{i+1} >$ a coherence label $l_i = 1$ if sp_i from the same segment, and $l_i = 0$ otherwise.

The embeddings e_i and e_{i+1} of adjacent sentences pairs $< s_i, s_{i+1} >$ used for coherence computing are calculated from Bi-LSTM forward and backward hidden states $\overrightarrow{h}$ and $\overleftarrow{h}$.

$$e_i = tanh(W_e(\overrightarrow{h_i} - \overrightarrow{h_{i-1}}) + b_e) \qquad (2)$$

$$e_{i+1} = tanh(W_e(\overleftarrow{h_{i+1}} - \overleftarrow{h_{i+2}}) + b_e) \qquad (3)$$

However, notice that instead of using the conven-

[1] https://github.com/hanxiao/bert-as-service

tional $[\overrightarrow{h_i}; \overleftarrow{h_i}]$ as the embedding of sentence i, here, similarly to Wang and Chang (2016), we subtract forward/backward states to focus on the semantics of the current sentence pair. The semantic coherence between two sentence embeddings is then computed as their cosine similarity:

$$Coh_i = cos(e_i, e_{i+1}) \tag{4}$$

We use binary cross-entropy loss to formulate the objective of our auxiliary task. For a document with k sentences, the loss can be calculated as:

$$L_2 = -\sum_{i=1,l_i=1}^{k-1} \log Coh_i - \sum_{i=1,l_i=0}^{k-1} \log(1 - Coh_i) \tag{5}$$

which penalizes high Coh across segments and low Coh within segments.

Based on Equation 1 and 5, we form the loss function of our model as:

$$L_{total} = \alpha L_1 + (1 - \alpha) L_2 \tag{6}$$

with the well-tuned trade-off parameter α, topic segmentation and the coherence-related auxiliary task are jointly optimized.

2.3 Sentence-Level Restricted Self-Attention

The self-attention mechanism (Vaswani et al., 2017) has been widely applied to many sequence labeling tasks due to its superiority in modeling long-distance dependencies in text. However, in some cases, long-distance dependencies will instead cause noises. Wang et al. (2018) noticed this problem in discourse segmentation, where the crucial information for EDU boundary prediction comes usually only from the adjacent EDUs. Thus, they proposed a word-level restricted self-attention mechanism by adding a fixed size window constraint on the standard self-attention. In essence, this mechanism encourages the model to absorb more information directly from adjacent contexts within a fixed neighborhood. We hypothesize that similar restricted dependencies could also characterize topic segmentation, hence, instead of at word-level, we add the sentence-level restricted self-attention on top of label prediction network.

In particular, once hidden states are obtained for all the sentences of document d, we compute the similarities between the current sentence i and its nearby sentences within a window of size S. For example, the similarity between sentence s_i and s_j which is within the window size is computed as:

$$sim_{i,j} = W_a[h_i; h_j; (h_i \odot h_j)] + b_a \tag{7}$$

Dataset	Section	Wiki-50	Cities	Elements	Clinical
docs	21,376	50	100	118	227
# sent/seg	7.2	13.6	5.2	3.3	28.0
# seg/doc	7.9	3.5	12.2	6.8	5.0

Table 1: Statistics of all the topic segmentation datasets used in our experiments.

where h_i, h_j are the hidden state of s_i and s_j. W_a and b_a are attention parameters. The attention weights for all the sentences in the window are:

$$a_{i,j} = \frac{e^{sim_{i,j}}}{\sum_{s=-S}^{S} e^{sim_{i,i+s}}} \tag{8}$$

The output for sentence i after the restricted self-attention mechanism is the weighted sum of all the sentence hidden states within the window:

$$c_i = \sum_{s=-S}^{S} a_{i,i+s} h_{i+s} \tag{9}$$

where c_i denotes the *local context embedding* of sentence i generated by restricted self-attention. After getting the context embeddings for all the sentences, we concatenate them with the original sentence hidden states and input them to another Bi-LSTM layer.

3 Experiments

3.1 Datasets

Training and Validation Data Ideally, training dataset for topic segmentation should satisfy the following requirements: (1) large size; (2) covering as many topics as possible; (3) real documents with reliable segmentation either from human annotations or already specified in the documents e.g., sections. The ***WIKI-SECTION*** (Arnold et al., 2019) is a newly released dataset which satisfies those requirements. This dataset was originally generated from the most recent English and German Wikipedia dumps. To better align with our task, we only select the English samples for training. The English WIKI-SECTION consists of 3.6k wikipedia articles from domian *diseases* and 19.5k articles from domain *cities*. We deem this dataset as a reliable training source because the two domains *cities* and *diseases* cover news-based samples and scientific-based samples respectively. These two subsets complement each other and make the overall dataset contain both common expressions and precise language. We split the dataset into 80% for training and 20% for validation.

Dataset	Wiki-50	Cities	Elements	Clinical
Random	52.7	47.1	50.1	44.1
BayesSeg	49.2	36.2	**35.6**	57.2
GraphSeg	63.6	40.0	49.1	-
TextSeg	28.5	19.8	43.9	36.6
Sector	28.6	33.4	42.8	36.9
Basic Model	28.1	18.7	42.8	32.7
+AUX	27.8	17.2	41.2	31.8
+RSA	27.4	16.7	42.0	32.5
+AUX+RSA	**26.5**	**16.3**	39.2	**30.7**

Table 2: P_k error score[2](Beeferman et al., 1999) on four test sets. Results in **bold** indicate the best performance across all comparisons. Underlined results indicate the best performance in the bottom section.

Test Data We evaluate our model on four datasets that originate from different source distributions: ***WIKI-50*** (Koshorek et al., 2018) which consists of 50 samples randomly generated from the latest English Wikipedia dump, with no overlap with training and validation data. ***Cities*** (Chen et al., 2009) which consists of 100 samples generated from Wikipedia about cities. We also ensure that this dataset has no overlap with training and validation data. ***Elements*** (Chen et al., 2009) which consists of 118 samples generated from Wikipedia about chemical elements. ***Clinical Books*** (Malioutov and Barzilay, 2006) which consists of 227 chapters from a medical textbook. Table 1 gives more detailed statistics for all the datasets we use.

3.2 Baselines

These include two popular unsupervised topic segmentation methods, *BayesSeg* (Eisenstein and Barzilay, 2008) and *GraphSeg* (Glavaš et al., 2016), as well as two SOTA supervised neural models, *TextSeg* (Koshorek et al., 2018) and *Sector* (Arnold et al., 2019). We use the original implementation code of TextSeg and train it on our training dataset. We adopt the results of *BayesSeg*, *GraphSeg* and *Sector* on four test sets from Arnold et al. (2019)[3].

3.3 Experimental Setup

Following Koshorek et al. (2018), we utilize the pre-trained GoogleNews word2vec embeddings ($d = 300$) as our initial word embeddings. We optimize our model with Adam optimizer, and set the learning rate to 0.001 and batch size to 8. The Bi-LSTM hidden state size is set to 256 following the the same setting in Koshorek et al. (2018). Model training is done for 10 epochs and performance is monitored over the validation set. To generate BERT sentence embeddings, we adopt the pre-trained 12-layer model released by Google AI (embedding size 768). We set the window size of restricted self-attention to 3 based on the average segment length of our training data and α to 0.8 which is tuned on the validation set.

3.4 Experimental Results

Table 2 compares the performance of the baselines and our model on four test datasets. To investigate the effectiveness of auxiliary task (AUX) and restricted self-attention (RSA), Table 2 also shows the results of individually adding each component to our basic segmenter. One important observation is that our model enhanced by context modeling outperforms all the baseline methods on three out of four test sets with a substantial performance gap. *BayesSeg* performs better on *Elements*, but it does not consistently perform as well as our model on other test sets. This clearly demonstrates the effectiveness of our proposed context modeling strategy. Furthermore, we observe that the auxiliary task module and restricted self-attention are both necessary for our model, since they do not only improve performance individually, but they achieve the best results when synergistically combined. Interestingly, improvements are more substantial on *Cities* and *Elements* than on the other datasets. One possible explanation is that, even though *Cities* and *Elements* has no overlap content with our training data, its topic and context relations are the closest to the dominant topic of our training data.

4 Conclusion and Future Work

The current neural topic segmenters are limited in how they model context. We propose to add a coherence-related auxiliary task and restricted self-attention on top of a hierarchical Bi-LSTM attention segmenter to make better use of the contextual information. The experimental results on four test sets demonstrate the effectiveness of our strategy. As future work, we plan to investigate how to integrate document structures or sentence dependencies obtained from other NLP tasks (e.g., discourse parsing), which could provide even more accurate and informative context modelling.

[2] P_k is a penalty metric, lower score indicates better performance. We also used *Windiff* (Pevzner and Hearst, 2002) as another evaluation metric. Since it was highly correlated with P_k, we omit it and only present performance by P_k.

[3] Arnold et al. (2019) reported *Sector*'s performance on multiple model settings. Here we pick the best performance of their model on each test set despite the setting difference.

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudr-Mauroux, Felix A. Gers, and Alexander Lser. 2019. Sector: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Pinkesh Badjatiya, Litton J. Kurisinkel, Manish Gupta, and Vasudeva Varma. 2018. Attention-based neural text segmentation. In *European Conference on Information Retrieval 2018*, pages 180–193.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations*.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379.

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.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, 55(3):529–569.

Teun van Dijk. 1981. Episodes as units of discourse analysis. *Analyzing Discourse: Text and Talk*.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130. Association for Computational Linguistics.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. 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 469–473.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1997. Automatic text summarization by paragraph extraction. In *Intelligent Scalable Text Summarization*.

HyoJung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. 2007. Semantic passage segmentation based on sentence topics for question answering. *Information Sciences*, 177(18):3696–3717.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557.

Swarnadeep Saha, Malolan Chetlur, Tejas Indulal Dhamecha, W M Gayathri K Wijayarathna, Red Mendoza, Paul Gagnon, Nabil Zary, and Shantanu Godbole. 2019. Aligning learning outcomes to learning resources: A lexico-semantic spatial approach. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5168–5174.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

Liang Wang, Sujian Li, Yajuan Lv, and Houfeng Wang. 2017. Learning to rank semantic coherence for topic segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1340–1344.

Liang Wang, Sujian Li, Xinyan Xiao, and Yajuan Lyu. 2016. Topic segmentation of web documents with automatic cue phrase identification and blstm-cnn. In *Natural Language Understanding and Intelligent Applications*, pages 177–188.

Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2315.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 962–967.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3009–3019.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.