Improving Context Modeling in Neural Topic Segmentation

Anonymous AACL-IJCNLP 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 BiLSTM network to better model context, by adding a coherence-related auxiliary task and restricted self-attention. Our optimized segmenter outperforms SOTA approaches when trained and tested on three datasets. We also demonstrate our proposal's robustness in domain transfer setting by training a model on a large-scale dataset and testing it on four challenging real-world benchmarks. Furthermore, we apply our proposed strategy to two other languages (German and Chinese) and show its effectiveness in multiligual scenario.

1 Introduction

Topic segmentation is a fundamental NLP task that has received considerable attention in recent years. It can reveal important aspects of a document semantic structure by splitting the document into topical-coherent textual units. Taking the *Wikipedia* example in Table 1, without the section marks, a reliable topic segmenter should be able to detect the correct boundaries within the text and chunk this article into the topical-coherent units `T1`, `T2` and `T3`. The results of topic segmentation can further benefit other key downstream NLP tasks such as document summarization (Mitra et al., 1997; Riedl and Biemann, 2012a; Xiao and Carenini, 2019), question answering (Oh et al., 2007; Diefenbach et al., 2018) and machine reading (van Dijk, 1981; Saha et al., 2019).

A wide variety of techniques have been proposed for topic segmentation. Early unsupervised models exploit word statistic overlaps (Hearst, 1997; Galley et al., 2003), Bayesian contexts (Eisenstein and Barzilay, 2008) or semantic relatedness graphs (Glavaš et al., 2016) to measure the lexical or semantic cohesion between the sentences or paragraphs and infer the segment boundaries from them. More recently, several works have framed topic segmentation as neural supervised learning, because of the remarkable success achieved by such models in most NLP tasks (Wang et al., 2016, 2017; Sehikh et al., 2017; Koshorek et al., 2018; Arnold et al., 2019). Despite minor architectural differences, most of these neural solutions adopt Recurrent Neural Network (Schuster and Paliwal, 1997) and its variants (RNNs) as their main framework. On the one hand, RNNs are appropriate because topic segmentation can be modelled as a sequence labeling task where each sentence is either the end of a segment or not. On the other hand, this choice makes these neural models limited in how to model the context, because RNNs are designed to capture long-distance information (Lipton et al., 2015; Sehikh et al., 2017; Wang et al., 2018), while for topic segmentation, it is also critical to supervise the model to focus more on the local context.

As illustrated in Table 1, the prediction of the boundary between `T1` and `T2` hardly depends on

Preface:
Marcus is a city in Cherokee County, Iowa, United States.
[T1] ***History:***
S1: The first building in Marcus was erected in 1871.
S2: Marcus was incorporated on May 15, 1882.
[T2] ***Geography:***
S3: Marcus is located at (42.822892, -95.804894).
S4: According to the United States Census Bureau, the city has a total area of 1.54 square miles, all land.
[T3] ***Demographics:***
S5: As of the census of 2010, there were 1,117 people, 494 households, and 310 families residing in the city.
... ...

Table 1: A Wikipedia sample article about *City Marcus* covering three topics: `T1`, `T2` and `T3`

the content in T3. Bringing in excessive long-distance signals may cause unnecessary noises and hurt performance. Moreover, text coherence has strong relation with topic segmentation (Wang et al., 2017; Glavas and Somasundaran, 2020). For instance, in Table 1, sentence pairs from the same segment (like <S1, S2> or <S3, S4>) are more coherent than sentence pairs across segments (like S2 and S3). Arguably, with a proper way of modeling the coherence between adjacent sentences, topic segmenter can be further enhanced.

In this paper, we propose to enhance a SOTA topic segmenter (Koshorek et al., 2018) based on a hierarchical attention BiLSTM network to better model the local context of a sentence in two complementary ways. First, 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 objective of our model to encourage that the coherence of the sentences from different segments is smaller than the coherence of the sentences from the same segment. Secondly, we enhance context modeling by utilizing restricted self-attention (Wang et al., 2018), which enables our model to pay attention to the local context and make better use of the information from the closer neighborhood of each sentence (i.e., with respect to a window of explicitly fixed size k). Our empirical experimental results show (1) that our proposed context modeling strategy significantly improves the performance of the SOTA neural segmenter on three datasets, (2) that the enhanced segmenter is more robust in domain transfer when applied to four challenging real-world test sets sampled differently from the training data, (3) that our context modeling strategy is also effective for the segmenters trained on other challenging languages (eg., German and Chinese) rather than just English.

2 Related Work

Topic Segmentation Early unsupervised models exploit the lexical overlaps of sentences to measure the lexical cohesion between the sentences or paragraphs (Hearst, 1997; Galley et al., 2003; Eisenstein and Barzilay, 2008; Riedl and Biemann, 2012b). Then, by moving two sliding windows over text, the cohesion between successive text units could be plotted and a cohesion drop would signal a segment boundary. Even if these models do not require any training data, they only show limited performance in practice.

More recently, neural-based supervised methods have been devised for topic segmentation because of their more accurate predictions and greater efficiency. One line of research frames topic segmentation as a sequence labeling problem and builds neural models to predict segment boundaries directly. Wang et al. (2016) proposed a simple BiLSTM model to label if a sentence is a segment boundary or not. They demonstrated that along with engineered features based on cue phrases (eg., 'first of all', 'second'), their model can achieve marginally better performance than early unsupervised methods. Later, Koshorek et al. (2018) proposed a hierarchical neural sequence labeling model for topic segmentation and showed its superiority compared with their selected supervised and unsupervised baselines. Around the same time, Badjatiya et al. (2018) proposed an attention-based BiLSTM model to classify whether a sentence was a segment boundary or not, by considering the context around it. The work we present in this paper can be seen as pushing this line of research even further by encouraging the model to more explicitly consider contextual coherence, as well as to absorb more information from the neighbor context through restricted self-attention.

Another rather different line of works first trains neural models for other tasks, and then uses these models' outputs to predict boundaries. Wang et al. (2017) trained a CNN network to predict the coherence scores for text pairs. Sentences in a pair with large cohesion are supposed to belong to the same segment. However, their "learning to rank" framework asks for the pre-defined number of segments, which limits their model's applicability in practice. Our selected framework overcomes this constraint by tuning a confidence threshold during training stage. A sentence with the output probability over this threshold will be predicted as the end of a segment. Following a very different approach, Arnold et al. (2019) introduced a topic embedding layer into a BiLSTM model. After training their model to predict the sentence topics, the learned topic embeddings can be utilized for topic segmentation. However, one critical flaw of their method is that it requires a complicated pre-processing pipeline, which includes heading extraction and synset clustering, whose errors can propagate to the main topic segmentation task. In contrast, our proposal only requires the plain content of the training data and its

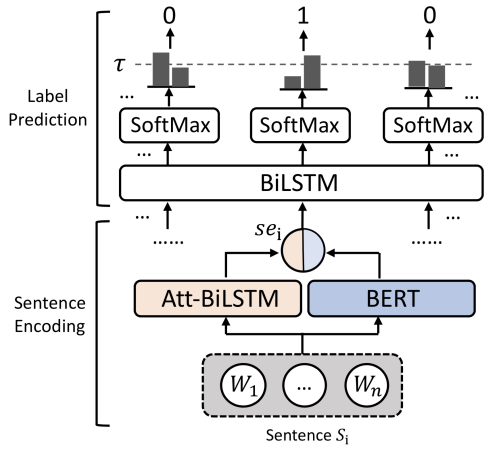

Figure 1: The architecture of our basic model. se_i is the produced sentence embedding for sentence S_i.

fully integrated neural architecture does not involve any complex pre-processing.

Coherence Modeling Early works on coherence modeling merely predict the coherence score for documents by tracking the patterns of entities' grammatical role transition (Barzilay and Lapata, 2005, 2008). More recently, researchers started modeling the coherence for sentence pairs by their semantic similarities and used them for higher level coherence prediction or even other tasks, including topic segmentation. Wang et al. (2017) demonstrated the strong relation between text-pair coherence modeling and topic segmentation. They assumed that (1) a pair of texts from the same document should be ranked more coherent than a pair of texts randomly picked from different documents; (2) a pair of texts from the same segment should be ranked more coherent than a pair of texts picked from different segments of a document. With these assumptions, they created a "quasi" training corpus for text-pair coherence prediction by assigning different coherence scores to the texts from the same segment, different segments but the same document, and different documents. Then they proposed the corresponding model, and further use this model to directly conduct topic segmentation. Following their second assumption, in this paper we propose a neural solution in which by injecting a coherence-related auxiliary task, topic segmentation and sentence level coherence modeling can mutually benefit each other.

3 Neural Topic Segmentation Model

Since RNN-based topic segmenters have shown success with high-quality training data, we adopt a SOTA RNN-based topic segmenter enhanced with attention and BERT embeddings as our basic model. Then, we extend such model to make better use of the local context, something that cannot be done effectively within the RNN framework (Wang et al., 2018). In particular, we add a coherence-related auxiliary task and a restricted self-attention mechanisms to the basic model, so that predictions are more strongly influenced by the coherence between nearby sentences. As a preview of this section, we first define the problem of topic segmentation and introduce the basic model. Next, we motivate and describe our proposed extensions.

3.1 Problem Definition

Topic segmentation is naturally framed as a sequence labeling task. More precisely, given a document represented as a sequence of sentences, our model will predict the binary label for each sentence to indicate if the sentence is the end of a topical coherent segment or not. Formally,

Given: A document d in the form of a sequence of sentences $\{s_1, s_2, s_3, ..., s_k\}$.

Predict: A sequence of labels assigned to all the sentences $\{l_1, l_2, l_3, ..., l_{k-1}\}$, where l is a binary label, 1 means the corresponding sentence is the end of a segment, 0 means the corresponding sentence is not the end of a segment. We do not predict label for the last sentence s_k since it is always the end of the last segment.

3.2 Basic Model: Enhanced Hierarchical Attention Bi-LSTM Network (HAN)

Figure 1 illustrates the detailed architecture of our basic model comprising the two standard steps of sentence encoding and label prediction. Formally, a sentence encoding network returns sentence embeddings from pre-trained word embeddings of sentences. Then a label prediction network processes the sentence embeddings generated earlier and outputs the probabilities to indicate if sentences are the segment boundaries or not. Finally, to convert the numerical probabilities into binary labels, we follow the greedy decoding strategy in Koshorek et al. (2018) by setting a threshold τ. All the sentences with probabilities over τ will be labeled 1, and 0 otherwise. This parameter τ is set in the validation stage. For training, 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

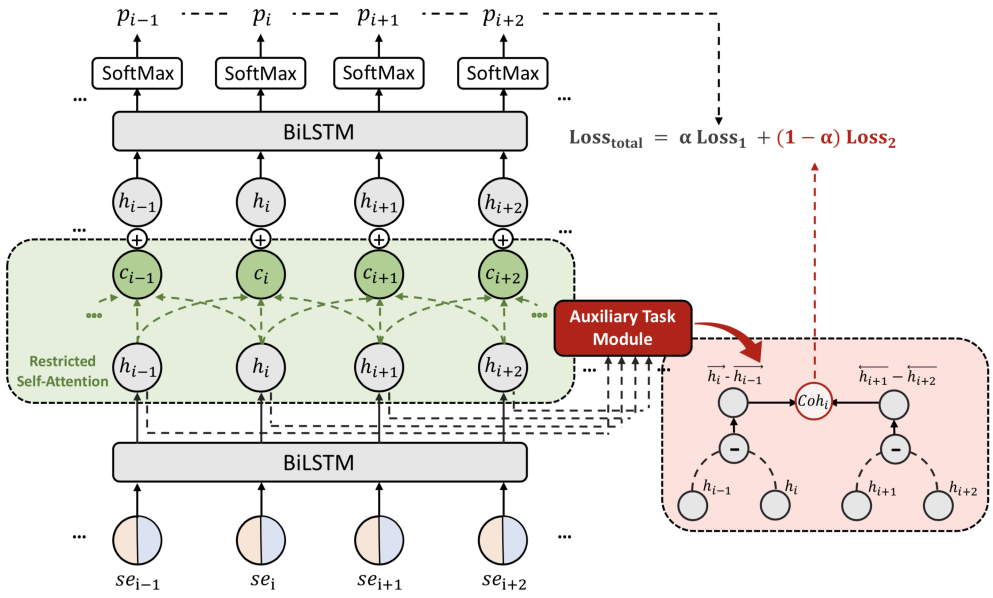

Figure 2: Our topic segmentation model with context modeling components: restricted self-attention (green), auxiliary task module (red).

sentences:

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

Looking at the details of the architecture in Figure 1, our basic model constitutes a strong baseline by extending the segmenter presented in Koshorek et al. (2018) in two ways (colored parts); namely, by improving the sentence encoder with an attention mechanism (orange) and with BERT embeddings (blue).

Enhancing Task-Specific Sentence Representations - While Koshorek et al. (2018) applied max-pooling to build sentence embeddings from sentence encoding network, we applied an attention mechanism (Bahdanau et al., 2015; Yang et al., 2016) to make the model better capture task-wise sentence semantics. The benefit of this enhancement were verified empirically (see Appendix A).

Enhancing Generality with BERT Embeddings In order to better deal with unseen text in test data and hence improve model's generality, we utilize a pre-trained BERT sentence encoder[1] which complements our sentence encoding network. The transformer-based BERT model (Devlin et al., 2019) was trained on massive data on several generic sentence-level semantic tasks, such as Natural Language Inference and Question Answering, which implies that it can arguably capture more general aspects of sentence semantics in a reliable way. To combine task-specific information with generic semantic signals from BERT, we simply concatenate the BERT sentence embeddings with the sentence embeddings derived from our encoder. Such concatenation is then the input of the next level network (see Figure 1). The benefit of this integration were also verified empirically (see Appendix A).

[1] github.com/hanxiao/bert-as-service. For languages other than English, we use their corresponding pre-trained BERT models.

3.3 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 consecutive sentence-pair coherence by using the sentence hidden states generated from the BiLSTM 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 the ground truth of our segmented training set rather than requiring 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 give a sentence pair $sp_i =< s_i, s_{i+1} >$ a coherence label $l_i = 1$ if sentences in sp_i are 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 BiLSTM forward and backward hidden states $\overrightarrow{h}$ and $\overleftarrow{h}$.

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

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

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

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

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:

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

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

Combining Equation 1 and 5, we form the loss function of our new segmenter as:

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

with the well-tuned trade-off parameter α, topic segmentation and the coherence-related auxiliary task are jointly optimized. The architecture of the auxiliary task module and its integration in our segmenter is shown in red in Figure 2.

3.4 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, when the task mainly requires modelling local context, long-distance dependencies will instead cause noise. Wang et al. (2018) noticed this problem in discourse segmentation, where the crucial information for a clause-like Elementary Discourse Unit (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 context words within a fixed range of neighborhood. We hypothesize that similar restricted dependencies also play a dominant role in topic segmentation. Hence, we add a *sentence-level* restricted self-attention on top of the label prediction network of the basic model, as shown in green in Figure 2.

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:

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

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

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

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

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

where c_i denotes the *local context embedding* of sentence i generated by restricted self-attention. After getting the local context embeddings for all the sentences, we concatenate them with the original sentence hidden states and input them to another BiLSTM layer (top of Figure 2).

4 Experimental Setup

In order to comprehensively evaluate the effectiveness of our context modeling strategy of adding a coherence-related auxiliary task and a restricted

Dataset	CHOI	RULES	SECTION	WIKI-50	CITIES	ELEMENTS	CLINICAL
documents	920	4,461	21,376	50	100	118	227
# sent/seg	7.4	7.4	7.2	13.6	5.2	3.3	28.0
# seg/doc	10.0	16.6	7.9	3.5	12.2	6.8	5.0
real world	✗	✓	✓	✓	✓	✓	✓

Table 2: Statistics of all the **English** topic segmentation datasets used in our experiments.

Dataset	EN	DE	ZH
documents	21,376	12,993	10,000
# sent/seg	7.2	6.3	5.1
# seg/doc	7.9	7.0	6.4
real world	✓	✓	✓

Table 3: Statistics of the the WIKI-SECTION data in English(EN), German(DE) and Chinese(ZH).

self-attention mechanisms to the basic model, we conduct three sets of experiments: **`(i) Intra Domain`** : we train and test the models in the same domain, repeating this evaluation for three different domains (datasets). **`(ii) Domain Transfer`** : we train the models on a large dataset which covers a variety of topics and test them on four challenging real-world datasets. **`(iii) Multilingual`** : we train and test our model on three datasets in different languages (English, German and Chinese), to assess our proposed strategy's generality in the multilingual scenario.

4.1 Datasets

Data for Intra Domain Evaluation High quality training dataset for topic segmentation usually satisfies the following criteria: (1) large size; (2) cover many topics; (3) contains real documents with reliable segmentation either from human annotations or already specified in the documents e.g., sections. In order to comprehensively evaluate the effectiveness of our context modeling strategy when dealing with data of different quality, we train and test models on the following three datasets:

CHOI (Choi, 2000) whose articles are synthesized artificially by stitching together different sources (i.e., they were not written as one document by one author). Hence, it does not reflect naturally occurring topic drifts. While the quality of this dataset is low, it is a popular benchmark for topic segmentation evaluation. We include this dataset to allow comparison with the previous work.

RULES is a new dataset we collected from the U.S. Federal Register issues[2]. When U.S. federal agencies make changes to regulations or other policies, they must publish a document called a "Rule" in the Federal Register. The Rule describes what is being changed and discusses the motivation and legal justification for the action. Since each paragraph in a document discusses one topic, we consider the last sentence of each paragraph as a ground truth topic boundary. The discussion paragraphs cover diverse topics in complex formal, technical language that can be hard to find online, so we deem it as an additional well-labelled dataset for testing topic segmentation to complement our other datasets which contain more informal language.

WIKI-SECTION (Arnold et al., 2019) is a newly released dataset which was originally generated from the most recent English and German Wikipedia dumps. To better align with the purpose of intra domain experiment, we only select the English samples for training and the German samples will be used in multilingual evaluation. The English *WIKI-SECTION* (labeled *SECTION* in the tables) consists of wikipedia articles from domain *diseases* and *cities*. We deem this dataset as the most reliable training source among the three datasets. It has the largest size and the two domains (*cities* and *diseases*) cover news-based samples and scientific-based samples respectively.

We split *CHOI* and *RULES* into 80% for training, 10% for validation and 10% for testing. For *SECTION*, we follow Arnold et al. (2019) and split it into 70% training, 10% validation, 20% test sets. Table 2(left) contains the statistical details for these three sets.

Data for Domain Transfer Evaluation We pick *WIKI-SECTION* as our training set in this line of experiments, due to its largest size and variety of covered topics. Following previous work, we evaluate our model and baselines 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

[2]https://www.govinfo.gov/

Dataset	CHOI	RULES	SECTION	MEAN
Random	49.4	50.6	51.3	50.4
BayesSeg	20.8	41.5	39.5	33.9
GraphSeg	6.6	39.3	44.9	30.3
TextSeg	1.0	7.7	12.6	7.1
Sector	-	-	12.7	-
Transformer	4.8	9.6	13.6	9.3
Basic Model	0.81	7.0	11.3	6.4
+AUX	0.64†	6.1†	10.4†	5.7
+RSA	0.72†	6.3†	10.0†	5.7
+AUX+RSA	**0.54**†	**5.8**†	**9.7**†	**5.3**

Table 4: P_k error score on four test sets. Results in **bold** indicate the best performance across all comparisons. Underlined results indicate the best performance in the bottom section. † indicates the result is significantly different ($p < 0.05$) from basic model.

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 2(right) gives more detailed statistics for these datasets.

Data For Multilingual Evaluation In order to test the effectiveness of our context modeling strategy across languages, besides the English *WIKI-SECTION*, we train and test our model on two other Wikipedia datasets in German and Chinese:

SECTION-DE which was released together with English *WIKI-SECTION* in Arnold et al. (2019). It also contains articles about cities and diseases. The section marks are used as the ground truth labels.

SECTION-ZH which was randomly generated from the Chinese Wikipedia dump[3]. As before, section marks are also used here as ground truth boundaries. The statistical details of these two datasets can be found in Table 3.

4.2 Baselines

These include two popular unsupervised topic segmentation methods, *BayesSeg* (Eisenstein and Barzilay, 2008) and *GraphSeg* (Glavaš et al., 2016), as well as the three recently proposed supervised neural models, *TextSeg* (Koshorek et al., 2018), *Sector* (Arnold et al., 2019) and *Hierarchical Transformer* (labeled *Transformer* in the tables) (Glavas and Somasundaran, 2020). We use the original implementation code of *BayesSeg*, *GraphSeg* and *TextSeg*. We reimplement the *Hierarchical Transformer* by ourselves. In Table 5, we adopt the results of *BayesSeg*, *GraphSeg* and *Sector* on four test sets from Arnold et al. (2019)[4].

4.3 Evaluation Metric

We use the standard P_k error score (Beeferman et al., 1999) as our evaluation metric, since it has become the standard for comparing topic segmenters. P_k is calculated as:

$$P_k(ref, hyp) = \sum_{i=0}^{n-k} \delta_{ref}(i, i+k) \neq \delta_{hyp}(i, i+k)$$

where δ is an indicator function which is 1 if sentence i and $i+k$ are in the same segment, 0 otherwise. It measures the probability of mismatch between the ground truth segments (*ref*) and model predictions (*hyp*) within a sliding window k. Window size k is the average segment length of *ref*. Since P_k is a penalty metric, lower score indicates better performance.

4.4 Neural Model Setup

Following Koshorek et al. (2018), our initial word embeddings are GoogleNews word2vec ($d = 300$). We also use word2vec embeddings ($d = 300$) and Fasttext embeddings ($d = 300$), which are both derived from Wikipedia corpora for German and Chinese respectively. We use the Adam optimizer, setting the learning rate to 0.001 and batch size to 8. The BiLSTM hidden state size is 256 following Koshorek et al. (2018). Model training is done for 10 epochs and performance is monitored over the validation set. We generate BERT sentence embeddings with the pre-trained 12-layer model released by Google AI (embedding size 768). The window size of restricted self-attention is 3 and α is 0.8. These were tuned on the validation set.

5 Results and Discussion

5.1 Intra Domain Evaluation

Table 4 shows the models' performance on the three datasets, when all supervised models are trained and evaluated on the training and test set from the same domain. To investigate the effectiveness of auxiliary task (AUX) and restricted self-attention (RSA), Table 4 also shows the results of individually adding each component to our basic segmenter.

[3]https://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[4]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.

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	64.6
TextSeg	28.5	19.8	43.9	36.6
Sector	28.6	33.4	42.8	36.9
Transformer	29.3	20.2	45.2	35.6
Basic Model	28.7	17.9	43.5	33.8
+AUX	27.9	17.0†	41.8†	31.5†
+RSA	27.8†	16.8†	42.7	31.9†
+AUX+RSA	**26.8**†	**16.1**†	39.4†	**30.5**†

Table 5: P_k error score on four test sets. Results in **bold** indicate the best performance across all comparisons. Underlined results indicate the best performance in the bottom section. † indicates the result is significantly different ($p < 0.05$) from basic model.

The most important observation from the table is that our model enhanced by context modeling outperforms all the supervised and unsupervised baselines with a substantial performance gain. With our context modeling strategy, the average P_k scores of our model over the three datasets improves on the best model (TextSeg) among the baselines by 25%. Compared with the basic model, adding AUX or RSA equally gives significant and consistent improvement across all three sets. Adding both AUX and RSA results in the biggest improvement by up to 17% on the mean across the three datasets.

5.2 Domain Transfer Evaluation

Table 5 compares the performance of the baselines and our model on four challenging real-world test datasets. All supervised models are trained on the training set of *WIKI-SECTION*. 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. Admittedly, *BayesSeg* performs better on *Elements*, possibly because that merely word embedding similarity is sufficient to indicate segment boundaries in this dataset. However, *BayesSeg* is completely dominated by our model on the other test sets. Overall, this indicates that our proposed context modeling strategy can not only enhance the model under the intra domain setting, but also produce robust models that transfer to other unseen domains. Furthermore, we observe that AUX and RSA are both necessary for our model, since they do not only improve performance individually, but they achieve the best results when synergistically combined.

Dataset	EN	DE	ZH
Random	51.3	48.7	52.2
Basic Model	11.3	18.2	20.5
+AUX	10.4†	17.7	20.5
+RSA	10.0†	16.6†	**19.8**†
+AUX+RSA	**9.7**†	**15.9**†	20.0†

Table 6: P_k error score on the datasets in three languages (English, German and Chinese).

5.3 Multilingual Evaluation

Table 6 shows results for our context modeling strategy across three different languages: English (EN), German (DE) and Chinese (ZH). Remarkably, even our basic model without any add-on component outperforms the random baseline by a wide margin. Looking at the gains from AUX and RSA, for German we observe a pattern similar to English, with our complete context modeling strategy (AUX+RSA) delivering the strongest gains. However, the performance on Chinese is not as strong as on English and German. Employing RSA still achieves a significant 0.7 P_k score drop, but introducing AUX does not help. One possible reason is that the sentences in the Chinese Wikipedia pages are relatively short and fragmented. Thus, the semantics of these sentences may be too simple to sufficiently guide the coherence auxiliary task.

6 Conclusions and Future Work

We address a serious limitation of current neural topic segmenters, namely their inability to effectively modelling context. To this end, we propose a novel neural model that adds a coherence-related auxiliary task and restricted self-attention on top of a hierarchical BiLSTM attention segmenter to to make better use of the contextual information. Experimental results on three datasets show that our strategy is effective within domains. Further, results on four challenging real-world benchmarks demonstrate its effectiveness in domain transfer settings. Finally, the application to other two languages (German and Chinese) suggests that our strategy has potential in multilingual scenarios.

As future work, we will investigate whether our proposed context modeling strategy is also effective for segmenting dialogues (Takanobu et al., 2018). Secondly, we will explore how to capture even more accurate and informative contextual information by integrating document structures generated by discourse parsers (Huber and Carenini, 2019).

References

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 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*.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

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.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 562–569.

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.

Goran Glavas and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 2306–2315.

Marti A. Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. 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 2306–2316.

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.

Zachary C. Lipton, John Berkowitz, and Charles Elkan. 2015. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019.

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.

Martin Riedl and Chris Biemann. 2012a. 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.

Martin Riedl and Chris Biemann. 2012b. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42.

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.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45:2673–2681.

Imran Sehikh, Dominique Fohr, and Irina Illina. 2017. Topic segmentation in asr transcripts using bidirectional rnns for change detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Ryuichi Takanobu, Minlie Huang, Zhongzhou Zhao, Fenglin Li, Haiqing Chen, Xiaoyan Zhu, and Liqiang Nie. 2018. A weakly supervised method for topic segmentation and labeling in goal-oriented dialogues via reinforcement learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4403–4410.

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.

Appendix A: Architecture Comparison for Baisc Model

Table 7 shows that replacing the max-pooling with the attention based BiLSTM sentence encoder yields better performance. Also, the concatenation of BERT embedding and the output of Att-BiLSTM yields the best performance compared with only one of them respectively.

Dataset	**CHOI**	**RULES**	**SECTION**	MEAN
MaxPooling	1.04	7.74	12.62	7.14
BLSTM	0.92	7.47	11.60	6.66
BERT	0.93	8.35	12.08	7.12
BLSTM+BERT	**0.81**	**6.90**	**11.30**	**6.34**

Table 7: P_k error score (lower the better, see Section 4.3 for details) of different sentence encoding strategies on three datasets (Section 4.1). Results in **bold** are the best performance across the comparisons.