SegFormer: A Topic Segmentation Model with Controllable Range of Attention

Haitao Bai, Pinghui Wang*, Ruofei Zhang, Zhou Su

MOE Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University
haitao.bai@stu.xjtu.edu.cn, phwang@mail.xjtu.edu.cn, rfzhang@gmail.com, zhousu@ieee.org

Abstract

Topic segmentation aims to reveal the latent structure of a document and divide it into multiple parts. However, current neural solutions are limited in the context modeling of sentences and feature representation of candidate boundaries. This causes the model to suffer from inefficient sentence context encoding and noise information interference. In this paper, we design a new text segmentation model *SegFormer* with unidirectional attention blocks to better model sentence representations. To alleviate the problem of noise information interference, SegFormer uses a novel additional context aggregator and a topic classification loss to guide the model to aggregate the information within the appropriate range. In addition, SegFormer applies an iterative prediction algorithm to search for optimal boundaries progressively. We evaluate SegFormer's generalization ability, multilingual ability, and application ability on multiple challenging real-world datasets. Experiments show that our model significantly improves the performance by 7.5% on the benchmark WIKI-SECTION compared to several strong baselines. The application of SegFormer to a real-world dataset to separate normal and advertisement segments in product marketing essays also achieves superior performance in the evaluation with other cutting-edge models.

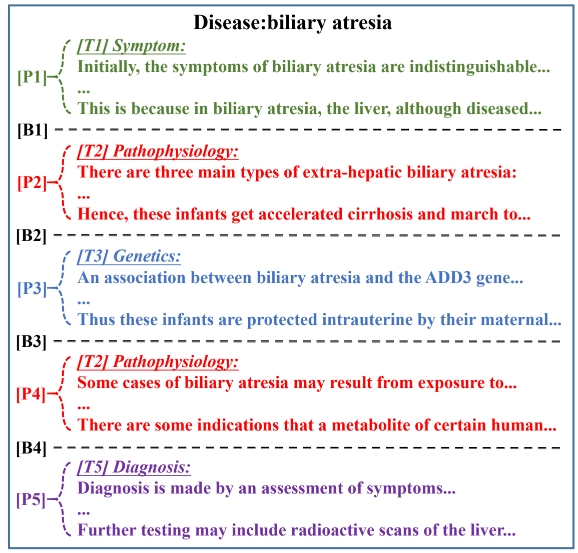

Figure 1: An essay on Wikipedia with five topic segments. Two of the five have the same topic "Pathophysiology".

Introduction

Topic segmentation aims to reveal the semantic structure of a document by dividing a document into multiple segments, such that divided segments are topically coherent inside, and the boundaries indicate changes in topic (Hearst 1994; Moens and De Busser 2001). A topic segmenter should find the correct boundaries within the essay according to topic changes and divide it into multiple parts. Figure 1 shows a real essay from Wikipedia including five parts: P1, P2, P3, P4, and P5, of which the topics are from T1 to T5, respectively. Many downstream tasks can benefit from these structured documents, including text summarization (Xiao and Carenini 2019), dialogue analysis (Xu, Zhao, and Zhang 2021), and information retrieval (Shtekh et al. 2018).

Multiple supervised and unsupervised models have been proposed for topic segmentation based on the following assumption: if a sentence is at the end of a topic segment, there must be a significant semantic difference between the context above and below this sentence. As illustrated in Figure 1, the semantic difference above and below the last sentence in P1 is significant, which helps the prediction of the segment boundary between P1 and P2. Unsupervised models such as Bayesian models (Malmasi et al. 2017) and graph-based models (Glavaš, Nanni, and Ponzetto 2016) have been proposed to predict segment boundaries by measuring semantic coherence between sentences. Supervised models (Koshorek et al. 2018; Xing et al. 2020; Lukasik et al. 2020) aim to predict labeled segment boundaries through training neural networks. These models adopt a similar hierarchical architecture and use Recurrent Neural Network (Schuster and Paliwal 1997) or Transformer (Vaswani et al. 2017) as their basic framework.

There are two major challenges in the text segmentation task: (1) First, the topic segmentation model needs to get contextual sentence embeddings because we always need to understand the meanings of a sentence with the context. On the other hand, the topic segmentation model also needs to

*Corresponding Author

get position-aware sentence embeddings because the model needs to know the relative positions among all the sentences. Otherwise, the model cannot predict the boundaries accurately. However, RNN-based encoders are difficult to extract contextual information from multiple aspects to enhance sentence representations. Moreover, the encoders based on bidirectional Transformers are insensitive to position to generate highly homogeneous sentence representation embeddings. (2) Second, the noise information on both sides of the candidate boundaries makes boundary recognition difficult because irrelevant information can distract attention and may cause adverse interference. For example, the segments P2 and P4 in Figure 1 have the same topic T2. This will cause adverse interference when predicting boundaries B2 and B3 because the similar information existing on both sides will reduce the semantic difference. On the contrary, if the model only focuses on the local part, there is no such problem. This requires the segmentation model to be able to control the context aggregation range. However, there are few models to explore what is the appropriate context range to distinguish the potential semantic difference.

To address the first challenge, we propose to use two unidirectional Transformer blocks to construct every sentence encoding layer. To solve the second challenge, i.e., to avoid noise information interference, we propose a novel aggregation module with a topic classification loss to learn the context aggregation range explicitly, that is, only aggregate the important information of two topics around the candidate segmentation boundary. We also propose a new training and iterative prediction strategy based on the observation that the prior discovered boundaries can be used to reduce the noise interference for subsequent boundaries' recognition.

In this work we bring the following contributions:

1. We propose a novel text segmentation model SegFormer. Specifically, we propose a new sentence contextualization encoder for text segmentation that is position-sensitive and has better sentence context modeling ability. We also propose a context aggregator using the topic classification loss and new training and inference strategy to solve the problem of noise information interference.
2. We designed multiple sets of experiments to demonstrate the generalization and multilingual abilities of the proposed model. Empirical results show that our proposed model SegFormer significantly improves the performance by 7.5% on the benchmark WIKI-SECTION dataset and achieves state-of-the-art performance.

Related Work

Unsupervised Segmentation. Early unsupervised models detect segment boundaries by quantifying lexical cohesion within text segments and low cohesion indicates a segment boundary. Lexical cohesion can be approximated easily by counting word repetitions. Some unsupervised segmentation models also use different text similarity measures to measure coherence between text segments. C99 (Choi 2000) uses inter-sentence similarity matrices to discover boundaries from divisive clustering. TextTiling (Hearst 1997) identifies major subtopic shifts and TopicTiling (Riedl and Biemann 2012) uses topic information from latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003) to split the text. Although unsupervised models do not require labeled training data, they are difficult to specialize for a specific domain and use semantic information to provide good segmentation.

Supervised Segmentation. Neural-based supervised models learn how to make accurate segmentations from large amounts of training data. Most of these models treat the text segmentation task as a sequence labeling problem and use hierarchical neural models to predict segment boundaries. (Li, Sun, and Joty 2018) propose a pointer network to point out segmentation boundaries. (Koshorek et al. 2018) train a hierarchical LSTM network TextSeg and achieve large improvements over unsupervised segmentation models. (Lukasik et al. 2020) propose three different hierarchical models based on LSTM-LSTM, Transformer-LSTM, and Transformer-Transformer architecture, respectively. (Arnold et al. 2019) propose a model which can predict segment boundaries and segment labels simultaneously. (Barrow et al. 2020) also argues that segment boundaries and segment labels contain complementary supervised signals. And they propose to jointly learn the two tasks of segment boundary prediction and segment label prediction. (Xing et al. 2020) improve sentence contextualization encoding by introducing a local attention mechanism in the LSTM encoder. (Lo et al. 2021) uses the hierarchical Transformer architecture and jointly learns the two tasks mentioned above to achieve state-of-the-art performance. However, existing models ignored the potential noise interference described in the introduction section and lack of controlling the attention range to get better boundary representations.

Our Model SegFormer

We view Topic segmentation as a sequence labeling task. Specifically, given a document containing n sentences $\{s_1, s_2, ..., s_n\}$, the segmentation model predicts the binary labels $\{l_1, l_2, ..., l_n\}$ of all these sentences to indicate whether a sentence is the end of a topic segment. When s_i is the end of a topic segment, l_i equals 1 and 0 otherwise. It should be noted that we do not need to predict l_n as s_n is always the end of the last segment, i.e., $l_n = 1$.

Overview

Figure 2 shows the architecture of our model Segformer. We propose a new text segmentation model which consists of a sentence encoder, a sentence contextualization encoder, and a context aggregator. The lower-level pre-trained sentence encoder is BERT (Devlin et al. 2018) which generates representations for each sentence respectively. The sentence contextualization encoder is responsible for generating context-based sentence representations. The context aggregator is responsible for explicitly aggregating local contexts above and below respectively from two directions to construct representations of candidate boundaries to classify.

Sentence Encoder

The lower-level sentence encoder (SE) is a bidirectional Transformer model BERT that generates sentence representations. We use the '[CLS]' token embedding $\mathbf{e}_i$ as the final

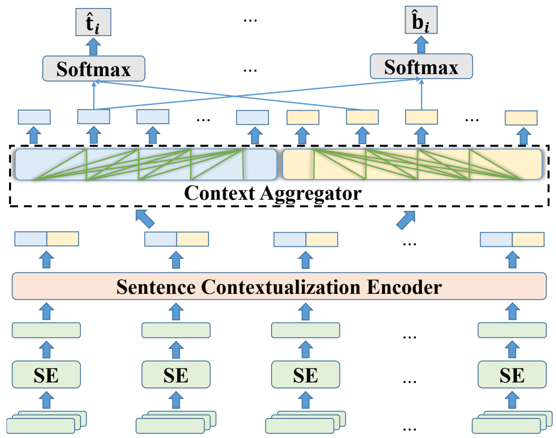

Figure 2: The architecture of SegFormer. It consists of sentence encoders (SE) with shared parameters, a sentence contextualization encoder, and a context aggregator.

sentence representation of the sentence $s_i = w_1^i, ..., w_k^i$ after passing it into the sentence encoder.

Sentence Contextualization Encoder

We use this module to make sentences acquire contextual semantics. The sentence contextualization encoder consists of two bidirectional encoding layers. The proposed encoding layer is shown in Figure 3, which consists of two unidirectional attention blocks. Each sentence aggregates information from two directions respectively to enhance the heterogeneity of sentence representations. In this way, half of the output sentence representation is the forward representation embedding of the sentence, and the other half is the backward representation embedding, which together constitutes the position-aware contextualization representation.

Context Aggregator

The context aggregator is responsible for explicitly constructing local context forward and backward. We use forward embedding $\mathbf{e}_i^{\mathrm{F}}$ and backward embedding $\mathbf{e}_i^{\mathrm{B}}$ to denote the context representations above and below sentence s_i. To facilitate the aggregator only aggregating information of a single topic in one direction, we introduce the topic classification loss to guide the model to learn the aggregation range. We feed $\mathbf{c}_i^{\mathrm{topic}}$=[$\mathbf{e}_i^{\mathrm{F}}$;$\mathbf{e}_i^{\mathrm{B}}$] into the topic classifier, which is a feed-forward net with Softmax function, i.e.,

$$\hat{\mathbf{t}}_i = \mathrm{Softmax}(\mathbf{c}_i^{\mathrm{topic}} \mathbf{W}^{\mathrm{topic}} + \mathbf{b}^{\mathrm{topic}}), \tag{1}$$

where m is the number of topic categories, and $\mathbf{W}^{\mathrm{topic}} \in \mathbb{R}^{d \times m}$ and $\mathbf{b}^{\mathrm{topic}} \in \mathbb{R}^m$ are classifier's parameters. $\hat{\mathbf{t}}_i$ is the predicted probability distribution vector of sentence s_i. d is the dimension of the representation $\mathbf{c}_i^{\mathrm{topic}}$. The topic classification loss of one essay is:

$$\mathrm{L}_{\mathrm{topic}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{m} \hat{t}_{ic} \log t_{ic}, \tag{2}$$

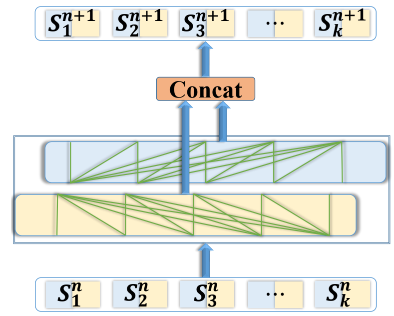

Figure 3: The bidirectional encoding layer architecture of our proposed sentence contextualization encoder.

where N is the number of sentences in this essay. $\hat{t}_{ic}$ refers to the predicted probability that the i-th sentence belongs to class c and t_{ic} is the true label. By using topic classification loss, $\mathbf{e}_i^{\mathrm{F}}$ and $\mathbf{e}_i^{\mathrm{B}}$ only have information of one topic which make the model learn the aggregation range.

Predicting Segment Boundaries

Let $\mathbf{e}_i^{\mathrm{F}}$ denote context representation before sentence s_i and $\mathbf{e}_{i+1}^{\mathrm{B}}$ denote context representation after sentence s_{i+1}. We see that $\mathbf{e}_i^{\mathrm{F}}$ and $\mathbf{e}_{i+1}^{\mathrm{B}}$ aggregate information from different topic segments when the candidate boundary between sentences s_i and s_{i+1} is the true boundary. Otherwise, they have the same topic segment information because of the introduction of the topic classification loss. We concate them as $\mathbf{c}_i^{\mathrm{boundary}}$=[$\mathbf{e}_i^{\mathrm{F}}$;$\mathbf{e}_{i+1}^{\mathrm{B}}$] to represent the candidate boundary representation between sentences s_i and s_{i+1}. We feed $\mathbf{c}_i^{\mathrm{boundary}}$ into the boundary classifier, which is a feed-forward net with the softmax function:

$$\hat{\mathbf{b}}_i = \mathrm{Softmax}(\mathbf{c}_i^{\mathrm{boundary}} \mathbf{W}^{\mathrm{boundary}} + \mathbf{b}^{\mathrm{boundary}}), \tag{3}$$

where $\mathbf{W}^{\mathrm{boundary}} \in \mathbb{R}^{d \times 2}$ and $\mathbf{b}^{\mathrm{boundary}} \in \mathbb{R}^2$ are classifier parameters. $\hat{\mathbf{b}}_i$ is the predicted probability distribution vector. d is the dimension of the representation $\mathbf{c}_i^{\mathrm{boundary}}$. The boundary classification loss of one essay is:

$$\mathrm{L}_{\mathrm{boundary}} = -\frac{1}{N-1} \sum_{i=1}^{N-1} \sum_{c=1}^{2} \hat{b}_{ic} \log b_{ic}, \tag{4}$$

where $\hat{b}_{i1}$ denotes the predicted probability that the candidate boundary between sentences s_i and s_{i+1} is not a true boundary, and $\hat{b}_{i2}$ denotes the predicted probability that the candidate boundary is a true boundary. By using the topic classification loss and the boundary classification loss jointly, our model can learn to find the true boundary using the semantic difference in the local context. We use a tunable scalar α to calculate the total loss:

$$\mathrm{L}_{\mathrm{total}} = \mathrm{L}_{\mathrm{boundary}} + \alpha \mathrm{L}_{\mathrm{topic}}. \tag{5}$$

Training and Inference Strategy

Inference Strategy. By setting the mask matrix in the context aggregator, we can constrain the attention range of context aggregation. We use the prior boundaries found to form

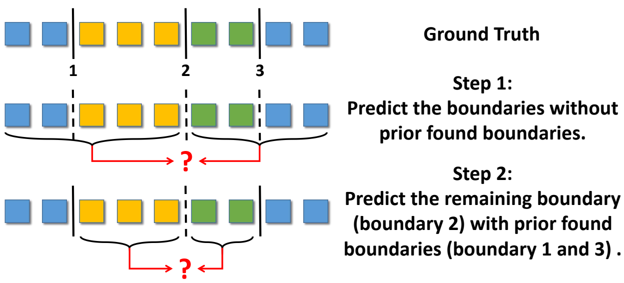

Figure 4: The procedure of our iterative algorithm. The essay has four topic segments and two of the four have the same topic. Due to the interference of homogeneous information, it is difficult to predict all boundaries at the first step. However, we can utilize the part of the boundaries found at the first step (boundary 1 and boundary 3) to predict the remaining boundary (boundary 2) more easily at the second step.

a barrier to eliminate noise information. Therefore, it is easier to find new boundaries that are difficult to find previously. Therefore, we find boundaries iteratively until no new boundaries are predicted. We show our iterative algorithm through an example in Figure 4.

Training Strategy. To let the model find the other boundaries from the observed boundaries, inspired by the curriculum learning idea, we develop a new training strategy. For one essay, we set the $mask\_ratio \in \{25\%, 50\%, 75\%, 100\%\}$, which denotes the proportion of the total ground-truth boundaries that our model needs to predict. And the other $1 - mask\_ratio$ boundaries are inputted in SegFormer as the observed boundaries. We train the model to predict the $mask\_ratio$ boundaries using the observed $1 - mask\_ratio$ boundaries. Training is from easy (low $mask\_ratio$) to hard (high $mask\_ratio$). We use the $mask\_epoch = [0, x_1, x_2, x_3, x_4]$ to control the training process, where x_i is an integer which denotes the training epoch. We set the $mask\_ratio = 25\%$ from epoch 0 to x_1, $mask\_ratio = 50\%$ from epoch x_1 to x_2, $mask\_ratio = 75\%$ from epoch x_2 to x_3 and $mask\_ratio = 100\%$ from epoch x_3 to x_4. And x_4 denotes the total training epochs.

Experiments

To comprehensively evaluate the effectiveness of our model, we conduct multiple sets of evaluation experiments. We make our source code and datasets publicly available to facilitate future study[1].

- **Intra-domain and Multilingual Experiments.** In this set of experiments, we train and test our proposed model using the same domain (dataset). We use the benchmark dataset WIKI-SECTION (Arnold et al. 2019) to evaluate.
- **Ablation Study.** To investigate the effectiveness of key components used in our model, we perform the ablation study by training multiple ablated versions of the proposed model. We study the following components: topic loss (L_{topic}), training inference strategy (T&I), sentence contextualization encoder (SCE), and context aggregator (CA).
- **Domain Transfer Experiments.** Following previous work (Xing et al. 2020), we test the models trained with WIKI-SECTION on another four real datasets to evaluate the transferability of the proposed model.
- **Application Experiments.** With the rapid development of social platforms, users tend to communicate and obtain information on social media. However, due to the high influence of these social platforms, some accounts are hired by advertisers for product marketing. They attract followers by writing popular essays and discreetly placing ads within them. Obviously, if the proposed text segmentation model can naturally divide normal content and marketing advertisements, it can assist in purifying the text content. However, the current text segmentation models mainly study the segmentation of content with different topic classes but not the segmentation of general classes. There are two classes in a product marketing essay: normal content and advertising content. Both normal content and advertising content can contain multiple narrative topics like the products' performance, appearance, etc. We expect the model to predict boundaries between normal content and advertising content and ignore the influence of different narrative topics. We evaluate our model SegFormer in this challenging real-world segmentation scenario: advertising text segmentation.
- **Motivation Experiments.** We verify whether SegFormer effectively addresses the two challenges mentioned in the introduction section. First, we replace the sentence contextualization encoder with LSTM and bidirectional Transformer respectively to compare with the proposed model SegFormer. Second, we test the ability to mitigate homogeneous information interference of the proposed model and baseline models. Specifically, we randomly construct three test datasets with different proportions of the same topic segments based on the original En_Diseaese test dataset, in which every synthetic essay has 5 topic segments. Every synthetic dataset has 500 essays. For example, the same topic segment ratio is 40% means that 2 of the 5 random topic segments have the same topic and they are not adjacent to each other. We use the models pre-trained on the original En_Disease training dataset to test on the three synthetic test datasets. We repeat the experiment three times with different random seeds and average the results.

Datasets

Datasets for Intra-domain and Multilingual Experiments. Following previous works, we conduct experiments on the following benchmark dataset:

- **WIKI-SECTION (Arnold et al. 2019)** is generated from the Wikipedia dumps and is a large-scale multi-domain and multilingual dataset. It covers two domains (cities and diseases) and two languages (English and German). The dataset has the following four datasets: En_Disease, De_Disease, En_City, and De_City including 3590, 2323, 19539, and 12537 articles, respectively.

Datasets for Domain Transfer Experiments. Following previous works, we evaluate SegFormer trained on the WIKI-SECTION dataset on the other four datasets from different distributions to test the domain transfer ability:

[1]https://github.com/nlgandnlu/SegFormer

Differences	S-LSTM/Transformer^2	SegFormer
Modeling of f^b	Implicitly	Explicitly
Controllable Attention	No	Yes
Iterative Inference	No	Yes

Table 1: The main differences between our model SegFormer, S-LSTM, and Transformer^2. f^b means the feature of the candidate boundary.

• **WIKI-50 (Koshorek et al. 2018)** has 50 articles randomly generated from the English Wikipedia dump.

• **Cities (Chen et al. 2009)** has 100 articles about cities.

• **Elements (Chen et al. 2009)** has 118 chemical elements articles generated from Wikipedia.

• **Clinical Books (Barzilay and Malioutov 2006)** has 227 articles from a medical textbook.

Datasets for Application Experiments.

• **Advertisements**[2] is a Chinese advertising dataset. We use 3313 advertorials that label each sentence as an advertising sentence or not in the dataset. The numbers of documents for training, validation, and testing are 2319, 331, and 663, respectively. This is a real-world advertisement dataset that can be used to test the application of segmentation models on the task of general class segmentation.

Evaluation Metrics and Baselines

We evaluate the results with P_k metric which is proposed by (Beeferman, Berger, and Lafferty 1999). We follow the same metric and dataset settings with Transformer^2 (Lo et al. 2021) to get comparable results[3].

The baselines we compared are: (1) unsupervised segmentation models: C99 (Choi 2000) and Topic-Tiling (Riedl and Biemann 2012). (2) supervised segmentation models: TextSeg (Koshorek et al. 2018), SECTOR (Arnold et al. 2019), S-LSTM (Barrow et al. 2020), Local-LSTM (Xing et al. 2020), Transformer^2 (Lo et al. 2021), Bert-LSTM and Hibert (Lukasik et al. 2020). The main differences between our model SegFormer, S-LSTM, and Transformer^2 are shown in the table 1. We follow the hyper-parameter settings for all the models in their official implementations.

Experiment Settings

We use the pre-trained model *Bert-base* for English datasets and *German Bert* for German datasets. The dimension of token embedding is 768, and the size of the dictionary is 30,522. The sentence contextualization encoder has 2 layers with 12 self-attention heads. We have used the Adam optimizer with the learning rate being 0.00001 for BERT and 0.0001 for sentence contextualization encoder and context aggregator. The dropout rate is 0.1. The tunable scalar α is 1. The batch_size is 32 and we train our model for 20 epochs. The $mask\_epoch = [0, 2, 6, 10, 20]$. All the baseline models are implemented following the settings mentioned by corresponding works and the open source code.

[2]https://github.com/zhanzecheng/SOHU_competition

[3]https://github.com/kelvinlo-uni/Transformer-squared

Models	En_Disease	De_Disease	En_City	De_City
C99	37.4	42.7	36.8	38.3
Topic-Tiling	43.4	45.4	30.5	41.3
TextSeg	24.3	35.7	19.3	27.5
SECTOR	26.3	27.5	15.5	16.2
S-LSTM	20.0	18.8	9.1	9.5
Local-LSTM	21.1	28.0	9.3	11.3
Bert-LSTM	23.6	22.1	10.2	9.8
Hibert	32.2	29.1	16.5	17.1
Transformer^2	18.8	16.0	9.1	7.3
SegFormer	**17.6**	**14.9**	**8.2**	**6.8**

Table 2: Results of intra-domain and multilingual experiments. We evaluate the model performance with P_k metric. The best performance is highlighted in bold.

Intra-domain and Multilingual Results

The intra-domain and multilingual evaluation results are shown in Table 2. We see our model achieves the best performance on all datasets compared to all baselines. In detail, our model achieves 6.4% and 9.9% relative improvement of P_k over the second best baseline model Transformer^2 on En_Disease and En_City. This indicates that our model has better intra-domain generalization ability. In addition, SegFormer outperforms 6.9% and 6.8% on the datasets of De_Disease and De_City than Transformer^2. This shows that our model performs consistently improvement across multiple language settings. Overall, SegFormer improves the average performance on WIKI-SECTION by 7.5% relative to Transformer^2 and achieves state-of-the-art performance, which shows the effectiveness of our proposed architecture.

Ablation Study

Table 3 shows the evaluation results of the ablation study. Compared with the full model, removing each of these components causes significant and consistent performance loss. We see some important conclusions from the results.

• **The contextual representation of sentences is necessary for the topic segmentation task.** We see the performance given by SegFormer is increased by 26.4% relative to the ablated version of 'without SCE' on average. Without the sentence contextualization encoder, the generated sentence representations will lose the meaningful context information and lead to poor segmentation results.

• **Our proposed modules and strategies can significantly reduce information interference and improve performance.** In general, we find that our proposed modules and strategies lead to significant improvements: compared with the ablation model 'without T&I+L_{topic}+CA', SegFormer increases the performance by 26.1% on average. Specifically, the topic loss significantly improves the average performance by 12.9%, which indicates that the topic supervision loss successfully guides the model to aggregate the context information and help alleviate the noise interference. In addition, we see the use of training and inference strategy can also improve the performance by 4.8% on average, indicating that the strategy of finding boundaries progressively from easy to hard is effective. To evaluate the

Models	En_Disease	De_Disease	En_City	De_City
SegFormer	**17.6**	**14.9**	**8.2**	**6.8**
w/o L_{topic}	22.7	18.7	8.6	7.1
w/o SCE	23.3	18.6	10.6	11.1
w/o T&I	18.4	15.6	8.6	7.2
w/o T&I+CA	23.7	19.9	9.0	8.2
w/o all	26.6	22.9	9.7	8.5

Table 3: Results of ablation experiments. The best performance is highlighted in bold. 'w/o' denotes without, 'L_{topic}' denotes the topic loss, 'SCE' denotes the sentence contextualization encoder, 'T&I' denotes the training and inference strategy, 'CA' denotes the context aggregator and 'w/o all' denotes without T&I+L_{topic}+CA.

Models	WIKI-50	Cities	Elements	Clinical Books
C99	-	-	-	-
Topic-Tiling	-	-	-	-
TextSeg	28.5	19.8	43.9	36.6
SECTOR	28.6	33.4	42.8	36.9
S-LSTM	-	-	-	-
Local-LSTM	26.8	16.1	**39.4**	30.5
Bert-LSTM	-	-	-	-
Hibert	29.3	20.2	45.2	35.6
Transformer^2	-	-	-	-
SegFormer	**25.3**	**15.2**	49.4	**28.6**

Table 4: Results of domain transfer experiments. The best performance is highlighted in bold. We use the results given by Xing et al. and '-' means the authors did not give the result of the model on the corresponding dataset.

effectiveness of our proposed context aggregator, we show the results of the ablation version 'without T&I+CA'. Experimental results show that the average performance improves by 19.2%, which demonstrates the effectiveness of the proposed context representation aggregating and comparing strategy.

Domain Transfer Results

Table 4 compares the performance of SegFormer and the baseline models on four challenging real-world test datasets. Following previous works, we train all the models on WIKI-SECTION and then test them on the four datasets. Our model outperforms all the baseline methods on three test sets, which indicates that our model generalizes better on out-of-domain datasets. SegFormer does not perform well on Elements because the topics of chemical elements articles are very different from WIKI-SECTION. In fact, the topic-aware context aggregator in SegFormer has difficulty adapting to drastic changes in topics. In future work, we need to reduce SegFormer's dependence on topics to enhance the transfer performance for out-of-domain datasets.

Application Results

Results on Advertisements are shown in Figure 5. Experiments show that SegFormer still outperforms the compared baseline methods by a large margin. This shows that SegFormer is also suitable for general text segmentation tasks. In future work, we will consider extending the segmentation model to other meaningful application scenarios. Hibert achieves the worst results because it suffers from the problem of over-smoothing, which further shows the importance of position-aware ability in text segmentation tasks.

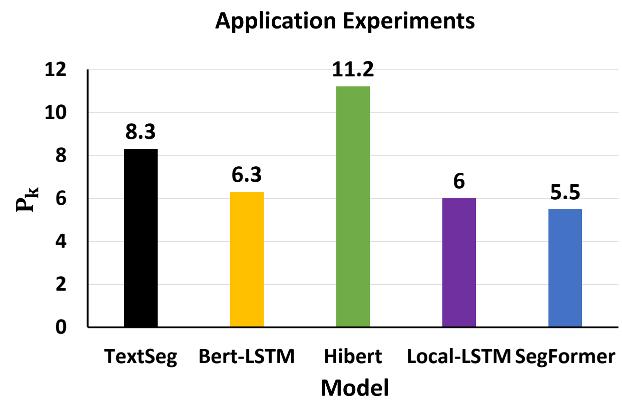

Figure 5: Results of application experiments.

Encoders	En_Disease	De_Disease	En_City	De_City
Our encoder	17.6	**14.9**	**8.2**	**6.8**
LSTM	**17.4**	15.2	8.3	7.1
Transformer	22.8	18.3	10.6	10.4

Table 5: Results of sentence contextualization encoder test. The best performance is highlighted in bold.

Motivation Experiments Results

• **The results of sentence contextualization representation test.** We use the bidirectional LSTM and bidirectional Transformer respectively as our sentence contextualization encoder to compare with SegFormer. The experimental results are shown in Table 5. The results show that the proposed encoder performs best on three of four datasets. We reduce the over-smoothing problem of encoded sentence embeddings in the bidirectional Transformer by introducing directionality. The proposed encoder also has the multi-head attention mechanism which is not used in the bidirectional LSTM encoder. Therefore, the proposed encoder can model sentence representations from multiple aspects and pay attention to important information easier than a bidirectional LSTM encoder and thus achieves the best results.

• **The results of homogeneous information interference test.** Homogeneous information is an important type of noise information in topic segmentation. To study the effectiveness of our model in reducing the impact of homogeneous information interference, we use En_Disease to construct synthetic datasets with different proportions of homogeneous information. Then we observe the trends of P_k value of SegFormer and other baseline models when the proportion of homogeneous information in the dataset changes. The results are shown in Figure 7. We see the P_k value

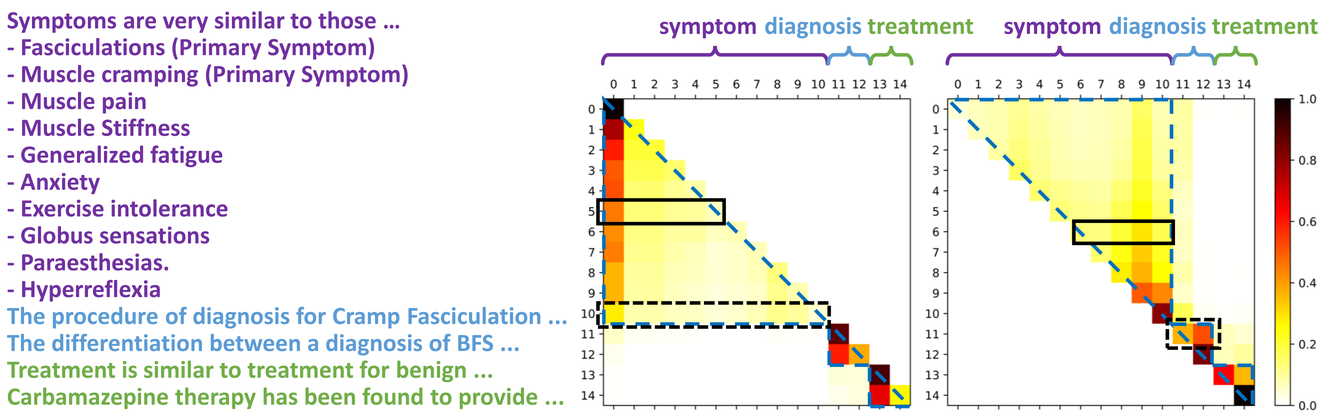

Figure 6: Results of the case study. We show the average attention weights of 12 heads of context aggregators on the right and the original essay on the left. The left attention heatmap is for forward attention and the right is for backward attention.

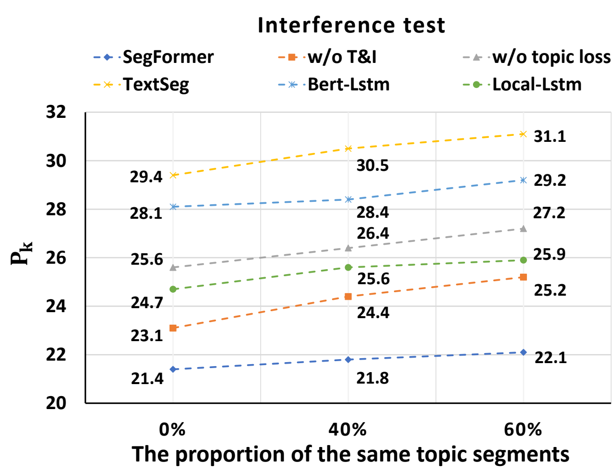

Figure 7: Results of interference test. T&I denotes the proposed training and inference strategy.

of SegFormer is stabler than all the other baseline models, which shows that our model can better alleviate the homogeneous noise information interference. We see the performance of the two ablation versions drops rapidly, which shows that the introduction of topic loss and T&I strategy alleviates the interference of homogeneous information. The topic loss reduces the interference by guiding the model to learn the aggregation range, and T&I strategy reduces the interference in the iterative prediction process.

Case Study

We show a random example in the test dataset of En_Disease in Figure 6. SegFormer successfully predicts all the boundaries in this example. We see the attention range of the aggregator from the attention heatmap in Figure 6. The blue dashed boxes in the two heatmaps denote the learned main distribution areas of the context aggregator's attention. As expected, the forward aggregator mainly aggregates the content of the topic segment above the candidate boundary, and the backward aggregator mainly aggregates the content of the topic segment below the candidate boundary. We see there is still some cross-segment attention in the aggregator which may introduce extra noise. In future work, we will consider how to control the attention range more strictly to further reduce the noise information. We also find that SegFormer tends to aggregate the information of the central sentences (sentences 0, 11, and 13) to represent the meaning of the segment. This shows that the attention mechanism is effective in the topic segmentation task because the model always needs to pay attention to the summary sentences to represent the meaning of the segments.

We use the candidate boundary between sentences 10 and 11 and the candidate boundary between sentences 5 and 6 as two examples to illustrate how SegFormer works. How does the proposed model determine whether there is a boundary between sentences 10 and 11? The forward aggregator aggregates the information of sentences from 0 to 10 in the black dotted box in the left attention heatmap. The backward aggregator aggregates the information of sentences from 11 to 12 in the black dotted box in the right attention heatmap. The aggregator compares the semantic difference and finds the true boundary. Similar to the above procedure, the aggregator compares the semantic difference between the context in the black solid line boxes and finds that there is no boundary between sentences 5 and 6. Because the aggregated information of them are similar as they come from the same topic segment. By aggregating and comparing the information on both sides of the candidate boundaries as stated, SegFormer can find the true boundaries accurately.

Conclusion

This paper proposes SegFormer which improves sentence contextualization encoding and significantly reduces the influence of noise information interference. Experiments show that SegFormer performs better on the topic segmentation tasks than baseline models and also has better generalization ability, multilingual ability, and application ability. In future work, We plan to explore a more efficient sentence contextualization module and better attention range to construct boundary representations for topic segmentation.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments and suggestions. This work was supported in part by the National Key R&D Program of China (2021YFB1715600), National Natural Science Foundation of China (U22B2019), MoE-CMCC "Artificial Intelligence" Project (MCM20190701).

References

Arnold, S.; Schneider, R.; Cudré-Mauroux, P.; Gers, F. A.; and Löser, A. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7: 169–184.

Barrow, J.; Jain, R.; Morariu, V.; Manjunatha, V.; Oard, D. W.; and Resnik, P. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 313–322.

Barzilay, R.; and Malioutov, I. 2006. Minimum cut model for spoken lecture segmentation. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Citeseer.

Beeferman, D.; Berger, A.; and Lafferty, J. 1999. Statistical models for text segmentation. *Machine learning*, 34(1): 177–210.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Chen, H.; Branavan, S.; Barzilay, R.; and Karger, D. R. 2009. Global models of document structure using latent permutations. Association for Computational Linguistics.

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Glavaš, G.; Nanni, F.; and Ponzetto, S. P. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 125–130. Association for Computational Linguistics.

Hearst, M. A. 1994. Multi-Paragraph Segmentation Expository Text. In *32nd Annual Meeting of the Association for Computational Linguistics*, 9–16.

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

Koshorek, O.; Cohen, A.; Mor, N.; Rotman, M.; and Berant, J. 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)*, 469–473.

Li, J.; Sun, A.; and Joty, S. 2018. SEGBOT: a generic neural text segmentation model with pointer network. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4166–4172.

Lo, K.; Jin, Y.; Tan, W.; Liu, M.; Du, L.; and Buntine, W. 2021. Transformer over Pre-trained Transformer for Neural Text Segmentation with Enhanced Topic Coherence. In *Findings of the Association for Computational Linguistics*, 3334–3340.

Lukasik, M.; Dadachev, B.; Papineni, K.; and Simões, G. 2020. Text Segmentation by Cross Segment Attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4707–4716.

Malmasi, S.; Dras, M.; Johnson, M.; Du, L.; and Wolska, M. 2017. Unsupervised text segmentation based on native language characteristics. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1457–1469.

Moens, M.-F.; and De Busser, R. 2001. Generic topic segmentation of document texts. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval*, 418–419.

Riedl, M.; and Biemann, C. 2012. TopicTiling: a text segmentation algorithm based on LDA. In *Proceedings of the Association for Computational Linguistics 2012 student research workshop*, 37–42.

Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.

Shtekh, G.; Kazakova, P.; Nikitinsky, N.; and Skachkov, N. 2018. Applying topic segmentation to document-level information retrieval. In *Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia*, 1–6.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Xiao, W.; and Carenini, G. 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 and the 9th International Joint Conference on Natural Language Processing*, 3011–3021.

Xing, L.; Hackinen, B.; Carenini, G.; and Trebbi, F. 2020. Improving Context Modeling in Neural Topic Segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 626–636.

Xu, Y.; Zhao, H.; and Zhang, Z. 2021. Topicaware multi-turn dialogue modeling. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*.