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 WIKISECTION 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.

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

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

Disease:biliary atresia

[T1] Symptom:
Initially, the symptoms of biliary atresia are indistinguishable...

[T2] Pathophysiology:
There are three main types of extra-hepatic biliary atresia:

[T3] Genetics:

[T4] Pathophysiology:

[T5] Diagnosis:
Diagnosis is made by an assessment of symptoms...

Further testing may include radioactive scans of the liver...

*Corresponding Author

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
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. 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.

Our Model SegFormer

We view Topic segmentation as a sequence labeling task. Specifically, given a document containing \( n \) sentences \( \{s_1, s_2, \ldots, s_n\} \), the segmentation model predicts the binary labels \( \{l_1, l_2, \ldots, l_n\} \) 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 \( e_i \) as the final
sentence representation of the sentence $s_i = w_i^1, \ldots, w_i^k$ 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 $e_i^F$ and backward embedding $e_i^B$ 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 $c_i^{\text{topic}} = [e_i^F, e_i^B]$ into the topic classifier, which is a feed-forward net with Softmax function, i.e.,

$$t_i = \text{Softmax}(c_i^{\text{topic}} W^{\text{topic}} + b^{\text{topic}}),$$

where $m$ is the number of topic categories, and $W^{\text{topic}} \in \mathbb{R}^{d \times m}$ and $b^{\text{topic}} \in \mathbb{R}^m$ are classifier’s parameters. $t_i$ is the predicted probability distribution vector of sentence $s_i$. $d$ is the dimension of the representation $c_i^{\text{topic}}$. The topic classification loss of one essay is:

$$L_{\text{topic}} = - \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{m} t_{ic} \log \hat{t}_{ic},$$

where $N$ is the number of sentences in this essay. $t_{ic}$ 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, $e_i^F$ and $e_i^B$ only have information of one topic which make the model learn the aggregation range.

**Predicting Segment Boundaries**

Let $e_i^F$ denote context representation before sentence $s_i$ and $e_i^B$ denote context representation after sentence $s_{i+1}$. We see that $e_i^F$ and $e_i^B$ 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 $c_i^{\text{boundary}} = [e_i^F; e_i^B]$ to represent the candidate boundary representation between sentences $s_i$ and $s_{i+1}$. We feed $c_i^{\text{boundary}}$ into the boundary classifier, which is a feed-forward net with the softmax function:

$$\hat{b}_i = \text{Softmax}(c_i^{\text{boundary}} W^{\text{boundary}} + b^{\text{boundary}}),$$

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

$$L_{\text{boundary}} = - \frac{1}{N-1} \sum_{i=1}^{N-1} \sum_{c=1}^{2} \hat{b}_{ic} \log \hat{b}_{ic},$$

where $\hat{b}_{11}$ denotes the predicted probability that the candidate boundary between sentences $s_i$ and $s_{i+1}$ is a true boundary, and $\hat{b}_{12}$ 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 $\alpha$ to calculate the total loss:

$$L_{\text{total}} = L_{\text{boundary}} + \alpha L_{\text{topic}}.$$

**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
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 ∈ {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, x1, x2, x3, x4] to control the training process, where x1 is an integer which denotes the training epoch. We set the mask ratio = 25% from epoch 0 to x1, mask ratio = 50% from epoch x1 to x2, mask ratio = 75% from epoch x2 to x3 and mask ratio = 100% from epoch x3 to x4. And x4 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 work.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 (Ltopic), 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 contextualizer 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_Disease 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:

¹https://github.com/nlgandnlu/SegFormer
The main differences between our model SegFormer, S-LSTM, and Transformer\(^2\) are shown in Table 1. We follow the same metric and dataset settings with Transformer\(^2\) 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 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 German 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 \(\alpha\) 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.

\(^1\)https://github.com/zhanzecheng/SOHU_competition
\(^2\)https://github.com/kelvinlo-uni/Transformer-squared

### 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\text{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

<table>
<thead>
<tr>
<th>Differences</th>
<th>S-LSTM/Transformer(^2)</th>
<th>SegFormer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeling of (f^b)</td>
<td>Implicitly</td>
<td>Explicitly</td>
</tr>
<tr>
<td>Controllable Attention</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Iterative Inference</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

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

### 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.

### 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 Table 1. We follow the hyper-parameter settings for all the models in their official implementations.

### Datasets for Application Experiments

- **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.

### 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 \(\alpha\) 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.

<table>
<thead>
<tr>
<th>Models</th>
<th>En_Disease</th>
<th>De_Disease</th>
<th>En_City</th>
<th>De_City</th>
</tr>
</thead>
<tbody>
<tr>
<td>C99</td>
<td>37.4</td>
<td>42.7</td>
<td>36.8</td>
<td>38.3</td>
</tr>
<tr>
<td>TextSeg</td>
<td>43.4</td>
<td>45.4</td>
<td>30.5</td>
<td>41.3</td>
</tr>
<tr>
<td>SECTOR</td>
<td>24.3</td>
<td>35.7</td>
<td>19.3</td>
<td>27.5</td>
</tr>
<tr>
<td>S-LSTM</td>
<td>26.3</td>
<td>27.5</td>
<td>15.5</td>
<td>16.2</td>
</tr>
<tr>
<td>Local-LSTM</td>
<td>21.1</td>
<td>28.0</td>
<td>9.3</td>
<td>11.3</td>
</tr>
<tr>
<td>Bert-LSTM</td>
<td>23.6</td>
<td>22.1</td>
<td>10.2</td>
<td>9.8</td>
</tr>
<tr>
<td>Hibert</td>
<td>32.2</td>
<td>29.1</td>
<td>16.5</td>
<td>17.1</td>
</tr>
<tr>
<td>Transformer(^2)</td>
<td>18.8</td>
<td>16.0</td>
<td>9.1</td>
<td>7.3</td>
</tr>
<tr>
<td>SegFormer</td>
<td>17.6</td>
<td>14.9</td>
<td>8.2</td>
<td>6.8</td>
</tr>
</tbody>
</table>

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.

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.

<table>
<thead>
<tr>
<th>Models</th>
<th>En_Disease</th>
<th>De_Disease</th>
<th>En_City</th>
<th>De_City</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegFormer</td>
<td>17.6</td>
<td>14.9</td>
<td>8.2</td>
<td>6.8</td>
</tr>
<tr>
<td>w/o L_topic</td>
<td>22.7</td>
<td>18.7</td>
<td>8.6</td>
<td>7.1</td>
</tr>
<tr>
<td>w/o SCE</td>
<td>23.3</td>
<td>18.6</td>
<td>10.6</td>
<td>11.1</td>
</tr>
<tr>
<td>w/o T&amp;I</td>
<td>18.4</td>
<td>15.6</td>
<td>8.6</td>
<td>7.2</td>
</tr>
<tr>
<td>w/o T&amp;I+CA</td>
<td>23.7</td>
<td>19.9</td>
<td>9.0</td>
<td>8.2</td>
</tr>
<tr>
<td>w/o all</td>
<td>26.6</td>
<td>22.9</td>
<td>9.7</td>
<td>8.5</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Models</th>
<th>WIKI-50 Cities</th>
<th>Elements</th>
<th>Clinical Books</th>
</tr>
</thead>
<tbody>
<tr>
<td>C99</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Topic-Tiling</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TextSeg</td>
<td>28.5</td>
<td>19.8</td>
<td>43.9</td>
</tr>
<tr>
<td>SECTOR</td>
<td>28.6</td>
<td>33.4</td>
<td>42.8</td>
</tr>
<tr>
<td>S-LSTM</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Local-LSTM</td>
<td>26.8</td>
<td>16.1</td>
<td>39.4</td>
</tr>
<tr>
<td>Bert-LSTM</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hibert</td>
<td>29.3</td>
<td>20.2</td>
<td>45.2</td>
</tr>
<tr>
<td>Transformer2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SegFormer2</td>
<td>25.3</td>
<td>15.2</td>
<td>49.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Encoders</th>
<th>En_Disease</th>
<th>De_Disease</th>
<th>En_City</th>
<th>De_City</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our encoder</td>
<td>17.6</td>
<td>14.9</td>
<td>8.2</td>
<td>6.8</td>
</tr>
<tr>
<td>LSTM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transformer</td>
<td>22.8</td>
<td>18.3</td>
<td>10.6</td>
<td>10.4</td>
</tr>
</tbody>
</table>

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 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.

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
Symptoms are very similar to those...
- Fasciculations (Primary Symptom)
- Muscle cramping (Primary Symptom)
- Muscle Stiffness
- Generalized fatigue
- Anxiety
- Exercise intolerance
- Globus sensations
- Paraesthesias,
- Hyperreflexia
The procedure of diagnosis for Cramp Fasciculation...
The differentiation between a diagnosis of BFS...
Treatment is similar to treatment for benign...
Carbamazepine therapy has been found to provide...

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 (U22B2109), MoE-CMCC "Artificial Intelligence” Project (MCM20190701).

References


