Advancing Topic Segmentation and Outline Generation in Chinese Texts: The Paragraph-level Topic Representation, Corpus, and Benchmark

Feng Jiang[1,2], Weihao Liu[3], Xiaomin Chu[3], Peifeng Li[3], Qiaoming Zhu[3], Haizhou Li[1]*

[1]Shenzhen Research Institute of Big Data, School of Data Science,
The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), Guangdong, China
[2]School of Information Science and Technology,
University of Science and Technology of China, Hefei, China
[3]School of Computer Science and Technology, Soochow University, Suzhou, China
{jeffreyjiang,haizhouli}@cuhk.edu.cn,
whliu@stu.suda.edu.cn, {xmchu, pfli, qmzhu}@suda.edu.cn

Abstract

Topic segmentation and outline generation strive to divide a document into coherent topic sections and generate corresponding subheadings, unveiling the discourse topic structure of a document. Compared with sentence-level topic structure, the paragraph-level topic structure can quickly grasp and understand the overall context of the document from a higher level, benefitting many downstream tasks such as summarization, discourse parsing, and information retrieval. However, the lack of large-scale, high-quality Chinese paragraph-level topic structure corpora restrained relative research and applications. To fill this gap, we build the Chinese paragraph-level topic representation, corpus, and benchmark in this paper. Firstly, we propose a hierarchical paragraph-level topic structure representation with three layers to guide the corpus construction. Then, we employ a two-stage man-machine collaborative annotation method to construct the largest Chinese Paragraph-level Topic Structure corpus (CPTS), achieving high quality. We also build several strong baselines, including ChatGPT, to validate the computability of CPTS on two fundamental tasks (topic segmentation and outline generation) and preliminarily verified its usefulness for the downstream task (discourse parsing).

Keywords: Topic Segmentation, Outline Generation, Paragraph-level Topic Representation

1. Introduction

A well-written document usually consists of several semantically coherent text segments, each of which revolves around a specific topic. Such topic structure can be discovered by topic segmentation and outline generation, which gives an overall grasp of the document. Topic segmentation aims to detect the segments (i.e., sentence or paragraph groups) in documents, and the subsequent task outline generation is to generate the corresponding subheading of each segment. Figure 1 shows an example of two tasks at the paragraph level where the basic units are paragraphs.

Compared with sentence-level topic structure, the paragraph-level topic structure pays more attention to the document's higher-level topic structure between paragraphs, which can benefit quickly grasping and understanding the overall context of the document. It not only benefits traditional downstream NLP tasks, such as document summarization (Xiao and Carenini, 2019) and discourse parsing (Jiang et al., 2021; Huber et al., 2022a), but also play an important role in Large Language Model (LLM) era. For example, during Retrieval Augment Generation (RAG) for large language models, obtaining the required information from long documents is necessary. The paragraph-level topic structure of a document can help quickly locate the approximate location of the desired content in long documents, reducing search space.

Thanks to the development of topic structure theory (Van Dijk and Kintsch, 1983; Chafe, 1994; Watson Todd, 2003; Stede, 2011) in English, more and more work in English has focused on topic segmentation on realistic datasets since the first attempt on synthetic datasets (Choi, 2000). They are not only limited to high-quality manually annotated datasets (Eisenstein and Barzilay, 2008; Chen et al., 2009) but also large-scale datasets (Koshorek et al., 2018; Arnold et al., 2019) automatically constructed from structured source data such as WIKI. The annotation content of topic structure has gradually enriched, from only topic boundaries to using words or phrases to annotate the topics of each text segment (Liu et al., 2022).

However, there are fewer studies on Chinese topic structure compared to English. Most previous work focused on sentence-level topic segmentation

* Corresponding Author.

LREC-COLING 2024, pages 495–506
20-25 May, 2024.

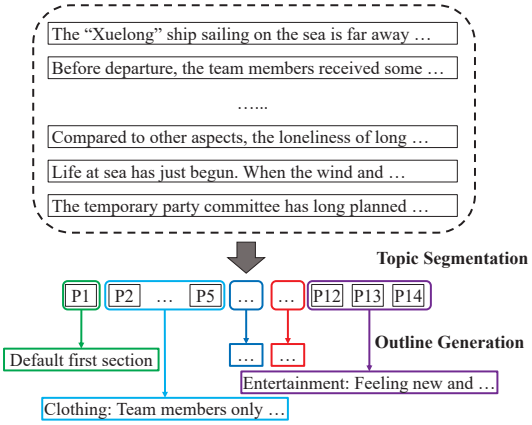


Figure 1: A document consisting of 14 basic units (paragraphs) is divided into five sections (different colors) according to the topic, and each of them has a subheading as its topic.

in dialogues (Xu et al., 2021; Zhang et al., 2023) or WIKI text (Xing et al., 2020) following English works. In the paragraph-level topic structure, Wang et al. (2016) annotated 2951 web documents with paragraph-level topic structures, but the corpus is unfortunately inaccessible, and specific annotation details are not disclosed.

Therefore, it is necessary to construct a large-scale and high-quality paragraph-level topic structure corpus to serve the Chinese research community. Considering the success of English and the current shortage of resources in Chinese, there are two challenges in filling the gap in Chinese paragraph-level topic structure research.

The first challenge is how to represent paragraph-level topic structures more richly. Most paragraph-level corpora (e.g., Cites and Elements corpora (Chen et al., 2009)) following sentence-level topic structure representation use keywords or phrases as topic contents (Koshorek et al., 2018; Arnold et al., 2019). Since basic units (paragraphs) are longer than the sentence level, using keywords or phrases cannot express richer topic information they contain (Todd, 2016). It led subsequent studies to focus more on topic segmentation (Choi, 2000; Riedl and Biemann, 2012; Badjatiya et al., 2018; Glavas and Somasundaran, 2020) and ignore outline generation. While some studies (Zhang et al., 2019; Barrow et al., 2020; Lo et al., 2021) have attempted to generate outlines, they only classify topics into limited types instead of generating real subheadings.

Another challenge is how to build a paragraph-level topic structure corpus that is both large-scale and high-quality. On the one hand, the existing high-quality manual corpora (Chen et al., 2009; Wang et al., 2016) are relatively small since the annotation requires assigning the topic attribution of paragraphs instead of sentences, which is laborious and time-consuming (Seale and Silverman, 1997; Todd, 2011). Besides, the manual annotation for topic content may be subjective and different from the author's intention due to topic ambiguity. On the other hand, although the automatic extraction method can build large-scale corpora (Koshorek et al., 2018; Arnold et al., 2019), it only ensures the correctness of the topic structure and content in terms of form, but not in terms of semantics that is crucial for outline generation.

To address the above issues, we first propose a hierarchical paragraph-level topic structure representation for modeling the topic structure of documents more comprehensively inspired by English topic theories. It has a three-layered topic structure, not only including paragraph and topic boundaries but also subheadings and the title. Especially, it takes the real subheading (clause or sentence) rather than keywords or phrases to represent topic content, ensuring the richness of the topic information longer basic units contain.

Then, we propose a two-stage man-machine collaborative annotation method to construct the Chinese Paragraph-level Topic Structure corpus (CPTS) with about 14393 documents with high quality based on our representation. Specifically, in the first stage, we first use a heuristics automatic extraction method for the topic boundary and content (subheading) from more common unstructured new documents, keeping the subject of the topic and the large scale of the corpus. In the second stage, to ensure the high quality of the corpus, we ask the human verifiers to verify the extracted topic structure instead of manually annotating them, which can greatly reduce the workload. Using this two-stage construction process of first extracting and then verifying, we build the largest Chinese paragraph-level topic structure corpus with a high quality (94.79% Inter-Annotator Agreement and 0.849 Kappa value).

Finally, to verify the computability of the CPTS, we construct several strong baselines, including ChatGPT, on two basic tasks: topic segmentation and outline generation. Also, preliminary experiments in the downstream task discourse parsing have verified the usefulness of its topic structure. [1]

2. Related Work

2.1. Topic Structure Theory

Unlike the intra-sentence topic structure, which is often a keyword, we focus on the discourse topic structure above sentences and paragraphs. Van Dijk and Kintsch (1983) considered discourse

[1]We release the corpus and baselines at https://github.com/fjiangAI/CPTS.

topics as a series of super propositions consisting of sentence propositions. Chafe (1994) stated that a discourse topic is a collection of relevant events, states, and references that agree in some form with the speaker's semi-active consciousness. Watson Todd (2003) argued that discourse topics are clusters of similar or related concepts to create connectivity and relevance.

Although different discourse topic theories have different views on the form of topics, researchers have roughly the same definition of discourse topic boundaries. (Stede, 2011) pointed out that a document would consist of several topics, each containing one or more basic units describing the same topic. The length of a topic will vary depending on the length of the document or the purpose of the research (Moens and De Busser, 2001; Ponte and Croft, 1997). At the paragraph level, (Hearst, 1997) regarded the paragraph boundaries as the potential topic boundary.

2.2. Topic Structure Corpus

There are many corpora at the sentence level, including constructed by synthetic (Choi, 2000), manually annotated (Eisenstein and Barzilay, 2008), and automatic extracted (Koshorek et al., 2018; Arnold et al., 2019) method. In Chinese, following the schema in English, the Wiki section zh (Xing et al., 2020) is a sentence-level topic structure corpus containing 10K documents randomly selected from the Chinese Wikipedia. XZZ (Xu et al., 2021) is manually annotated 505 recorded conversations, but it only annotated sentence-level topic boundaries and did not annotate topic contents. The subsequent MUG (Zhang et al., 2023) supplemented it by annotating 654 conversations with sentence-level topic boundaries as well as the topic content.

However, due to the large basic units (paragraphs), the paragraph-level topic annotation is laborious and time-consuming (Seale and Silverman, 1997; Todd, 2011). The manually annotated corpora are relatively small, such as Cities and Elements (Chen et al., 2009), which only have about 100 documents. Therefore, recent research has shifted towards automatic extraction (Liu et al., 2022). There is relatively little research in Chinese, and the manually annotated WLX (Wang et al., 2016) is a paragraph-level topic structure corpus containing 2951 documents, but unfortunately, their dataset is not publicly accessible.

2.3. Topic Segmentation and Outline Generation Method

In English, early work mainly used unsupervised methods (Hearst, 1997; Choi, 2000; Riedl and Biemann, 2012; Glavas et al., 2016) for topic segmentation. Owing to having constructed the large-scale topic structure corpora, supervised methods have gradually become mainstream, such as the sequential labeling models (Badjatiya et al., 2018; Koshorek et al., 2018; Glavas and Somasundaran, 2020; Lukasik et al., 2020), and the pointer networks (Li et al., 2018). Only a few studies focused on the Chinese topic segmentation task by following English methods using the sequential labeling models (Wang et al., 2016; Xing et al., 2020) or local classification model (Jiang et al., 2021) to predict the topic boundary.

Most of the previous corpora in English annotated topic contents as keywords or phrases, which are very short. It caused the outline generation works to be easily formulated as a classification problem in a joint learning framework with topic segmentation (Zhang et al., 2019; Barrow et al., 2020; Lo et al., 2021). There are few research in Chinese due to a lack of suitable corpora. MUG (Zhang et al., 2023) views topic segmentation and outline generation as two separate tasks for benchmarks.

3. Chinese Paragraph-level Topic Structure Representation

A proper topic structure representation is a prerequisite and necessary condition for guiding the construction of a topic corpus. It determines the form and content of corpus annotation. Most of the existing corpora only annotate basic units and topics they subordinate following Goutsos (1997)'s theory. Recent work has gradually enriched annotations, such as using keywords and phrases to annotate topic content. However, at the paragraph level, as the granularity becomes larger, richer annotations are needed to comprehensively express the high-level structure of the document, such as subheadings and titles.

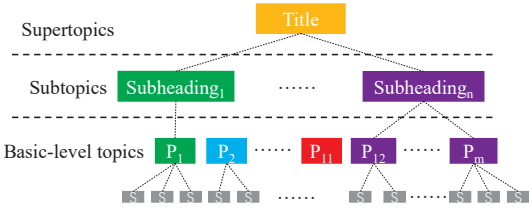


Figure 2: The Chinese paragraph-level topic structure representation, which contains m paragraphs and n subheadings. We assume a paragraph only has one topic, and a topic contains one or more paragraphs. S represents a sentence.

Recognizing this, we propose a three-layer hierarchical representation of the Chinese paragraph-level topic structure for guiding corpus construction according to discourse topic theories (Bruning et al., 1999; Van Dijk, 2014): *supertopics*, *subtopics*, and *basic level topics*, as illustrated in Figure 2. It

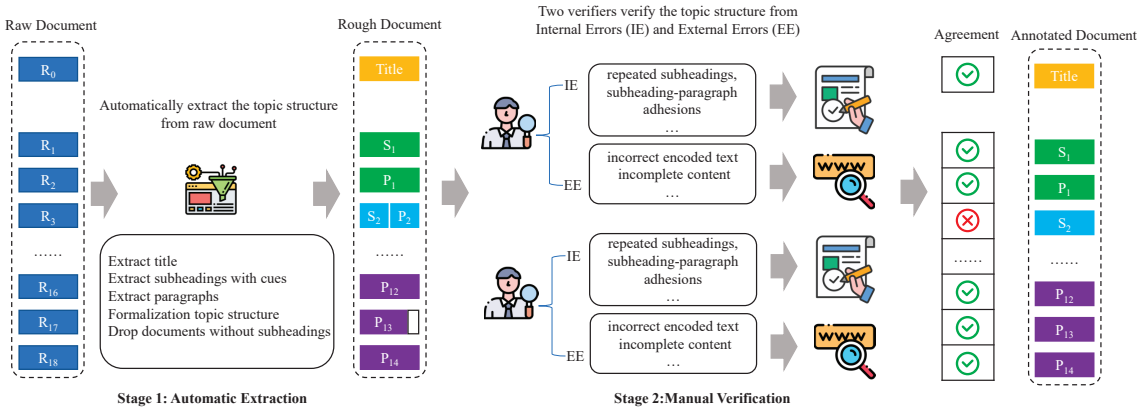


Figure 3: The two-stage man-machine collaborative annotation process. R_n means raw paragraphs, S_n means subheadings and P_n means annotated paragraphs.

not only includes paragraph boundaries and topic boundaries but also includes topic content and the higher-level title of a document. In particular, we regard the subheading and title as a clause or sentence instead of keywords to represent richer information on the paragraph level.

Specifically, we take the document's title as a supertopic, subheadings as subtopics, and paragraphs as basic-level topics, which implies a single paragraph will belong to one topic, and consecutive paragraphs describing the same topic fall under the same subheading. All subheadings, in turn, are subordinate to the title as subtopics. This hierarchical topic structure can capture relationships not only between paragraphs and subheadings but also between subheadings and the title.

In addition, we use subheadings (clauses or sentences) instead of noun phrases to represent subtopics of longer basic units (paragraphs), overcoming the limitation of keywords or phrases as subtopics that limit the amount of information conveyed by the topic at the paragraph level (the subtopic only has 4.7 tokens in latest work (Liu et al., 2022)). By integrating the semantic richness of subheadings, we can better capture the nuances of the document's content and structure.

4. Chinese Paragraph-level Topic Structure Corpus Construction

4.1. Data Source

Although our proposed representation model can be applied to various genres of documents, we still want to construct the corresponding corpus from more general text to assist downstream tasks better. Therefore, we select the news documents issued by Xinhua News Agency from Chinese Gigaword Fourth Edition (Robert Parker, David Graff, Ke Chen, Junbo Kong, Kazuaki Maeda, 2009) (Gigaword corpus) as the data source for generalization. It contains 1373448 news documents of four types (i.e., advis, multi, other and story) from January 1991 to December 2008. We chose story news as candidate documents because they account for most of them (1314198/1373448) and are more standardized than the other three types of news, which is conducive to building a more generalized topic structure.

4.2. Man-machine Collaborative Annotation

Manually constructing topic structure corpora is time-consuming and limited in scale due to topic ambiguities (Seale and Silverman, 1997; Todd, 2011), while large-scale corpora constructed by automatic extraction without manual verification do not guarantee the correctness of topic boundaries and content in semantics that are essential for finer tasks such as outline generation. Thus, inspired by previous work on automatic (Koshorek et al., 2018) and manual construction (Eisenstein and Barzilay, 2008), we design a two-stage man-machine collaborative annotation strategy involving both automatic extraction and manual verification to build a large-scale and high-quality corpus.

The details of our two-stage man-machine collaborative annotation process are shown in Figure 3. In the first stage, we take several steps to automatically extract candidate documents with topic structures, ensuring the correctness of topic boundaries and contents in terms of form. In the second stage, each document is double-checked for internal and external errors by two human validators to ensure the correctness of topic boundaries and contents in terms of semantics.

4.2.1. Automatic Extraction

Following previous work (Koshorek et al., 2018), we automatically extracted the topic structure of

the document in the first stage. Different from easily extracting topic structure from structured text in WIKI, news text is harder because it is typically composed of unstructured paragraphs with natural texts. Therefore, we designed a heuristic automatic extraction method to extract topic structures from raw documents automatically. Firstly, we extract the first paragraph that only includes one sentence and without ending punctuation as the title of a document. Then, we traverse the following paragraphs in the documents. If a paragraph has only one sentence and has a special token "(subheading)", we regard it as a subheading. Otherwise, it will be added to the paragraph list. After traversing all paragraphs, we formalize the topic structure representation based on the position of the paragraphs and subheadings and drop the document that does not contain any subheadings. This automatic extraction method can quickly extract topic structures from a large number of documents, but due to its simple heuristic rules, the accuracy cannot be guaranteed. Therefore, we take the second stage of manual verification to make it up.

4.2.2. Manual Verification

Hearst (1997) pointed out that most documents do not contain explicit subheadings that indicate the topic structure. Therefore, after the first stage, we obtain 14393 (about 1% of raw documents) rough documents containing subheadings extracted automatically since few documents explicitly have two or more subheadings with special tokens in the Gigaword corpus. To ensure the correctness of topic boundaries and content in terms of semantics, we ask verifiers to verify the topic structure of each rough document. Our verification team consists of one Ph.D. student, six master's students, and one senior undergraduate student, all of whom are engaged in natural language processing. They are divided into four groups, and each document will be verified by one group (two verifiers) to ensure the objectivity and accuracy of verification. It is worth mentioning that since the sub-topics are automatically extracted in the first stage, the verifier simply needs to check the correctness of paragraphs, subheadings, and title rather than label boundaries and write topic contents, which significantly reduces the annotation effort.

As shown in Figure 3, manual verification mainly verifies the correctness of automatic extraction from a semantic perspective at this stage and also quickly re-verifies the form correctness of topic boundaries that have been automatically extracted. For semantic issues, they are mainly checked from both internal and external errors. Internal errors refer to errors that verifiers can correct through the document itself, including repeated subheadings, title-paragraph adhesions, etc. For external errors such as incorrect encoded text or incomplete content in some subheadings or paragraphs that cannot be fixed from the document itself, verifiers use the help of search engines to retrieve the source news and make corresponding modifications.

During the second verification stage, two verifiers in each group verify the documents separately. When two verifiers have a conflict in their annotations, they discuss resolving. Finally, 36% of documents containing errors are revised by our verifiers. Thanks to the automatic extraction of most of the correct topic structures in the first stage, the average Inter-Annotator Agreement (IAA) between two verifiers of the same group is 94.79%, and the Kappa value (Cohen, 1960) between them is 0.849. It demonstrates that our two-stage man-machine collaborative annotation method only requires the verifier to validate the topic structure rather than directly generate it, significantly reducing the workload while maintaining high quality.

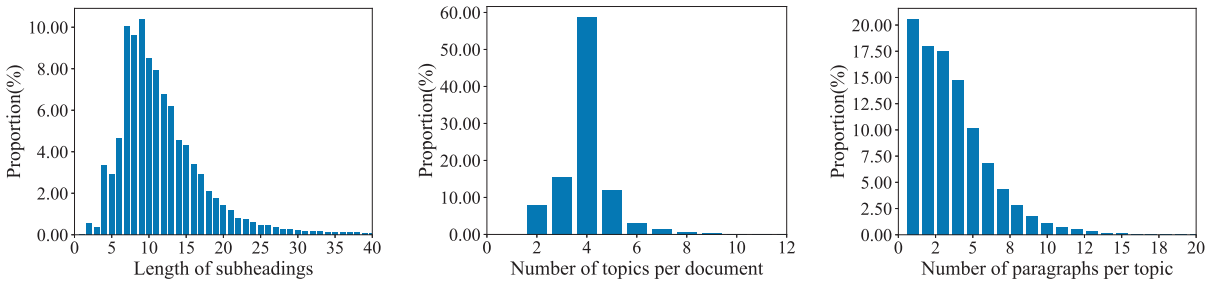
Item	Max	Min	Avg.
# words/document	5791	180	1727.96
# paragraphs/document	40	2	14.76
# words/subheading	147	1	12.33
# paragraphs/subheading	33	1	3.70
# subheadings/document	20	2	4.00

Table 1: The statistical details of CPTS.

5. Statistics and Analysis on CPTS

5.1. Details of CPTS corpus

The details of CPTS are shown in Table 1, the range and average figures for various aspects like the number of words per document (ranging from 180 to 5791, with an average of 1727.96), paragraphs per document (ranging from 2 to 40, averaging at 14.76), words per subheading (averaging at 3.70) and subheadings per document (ranging from 2 to 20, with an average of 4.00). Furthermore, Figure 4 depicts the main distributions of the length of subheadings, topics per document, and paragraphs per topic. Figure 4a shows that about 90% subheadings have more than seven words, and only a few subheadings have less than four words. It shows that subheadings in CPTS are usually clauses or sentences rather than words or phrases (Arnold et al., 2019), which could fully express the information of a paragraph-level topic. Figure 4b shows that about 60% of the documents have four topics, demonstrating the topic granularity will change with the document length (Todd, 2016). We also notice that over 70% of topics contain less than four paragraphs in Figure 4c. They indicate the usefulness of the paragraph-level topic: A document can be divided into two more simple



(a) Distribution of subheadings length.

(b) Distribution of topics per document.

(c) Distribution of paragraphs per topic.

Figure 4: The main distribution of the length of subheadings, topics per document, and paragraphs per topic.

structures through paragraph-level topics (the discourse structure among paragraph-level topics and that in one topic).

5.2. Compared with Other Chinese Topic Structure Corpora

The comparison of CPTS and the other Chinese topic structure corpora (Wiki$_{zh}$ (Xing et al., 2020), XZZ (Xu et al., 2021), MUG (Zhang et al., 2023) and WLX (Wang et al., 2016)) are shown in Table 2. Firstly, thanks to our two-stage human-machine collaborative annotation method, we have constructed the largest high-quality Chinese topic structure corpus. It is about four times larger than the existing largest paragraph-level one (WLX), and even larger than the largest sentence-level corpus (Wiki$_{zh}$) by automatic extraction. At the same time, incorporating manual verification after automatic extraction makes our corpus maintain the same high quality as the manually annotated corpus (94.79% IAA and 0.849 Kappa value).

Secondly, compared with other written corpus, CPTS annotated more comprehensive topic structures based on our Chinese paragraph-level topic structure representation, including Paragraph Boundaries (PB), Topic Boundaries (TB), subheadings, and titles, which benefit more tasks like Outline Generation (OG) and Title Generation (TG).

Finally, CPTS will be open access to the community to fill the gap in the Chinese paragraph-level topic segmentation resources. It complements the sentence-level dialogue topic structure corpus mutually and fully supports the relevant tasks, including paragraph-level topic segmentation and outline generation, and even speeds up the retrieval of needed information from documents in the RAG process of LLM.

We also noticed the similarities between our CPTS and MUG, but there are two aspects of differences. Firstly, the annotated genre and topic level are different. We focus on paragraph-level topic structures in written texts, while MUG focuses on sentence-level topic structures in spoken dialogues. Secondly, the source of annotating content is different. In our CPTS, paragraph boundaries, topic boundaries, subheadings, and the title are manually verified after being automatically extracted from the original document, which is closer to the author's intention. However, in the MUG, annotators manually write these from the reader aspect.

6. Experiments on Corpus Evaluation

To verify the computability of our annotated CPTS, we select several strong baselines to experiment on two basic tasks (i.e., topic segmentation and outline generation) of CPST as benchmarks for further research. Following previous work (Zhang et al., 2023), we build the baselines on these two tasks separately to obtain the objective and absolute performance of them. We randomly divided the dataset into training (90%) and testing sets (10%) according to the paragraph length distribution of the document for topic segmentation and outline generation. Specifically, there are 12953 documents in the training set and 1440 documents in the test set.

6.1. Topic Segmentation

6.1.1. Task Definition and Baselines

Following Koshorek et al. (2018), we view topic segmentation as a supervised learning task, as shown in Eq. 1. The input P is a document, represented as a sequence of n paragraphs $(p_1, ..., p_n)$, and the label $Y = (y_1, ..., y_{n-1})$ is a segmentation of the document, represented by $n-1$ binary values, where y_i denotes whether p_i ends a segment. $Model_{ts}$ is a topic segmentation model.

$$(y_1, ..., y_{n-1}) = Model_{ts}(p_1, ..., p_n) \quad (1)$$

For the topic segmentation task, we select the following three representative kinds of models as baselines. **Segbot** (Li et al., 2018) and **PN-XLNet**

Dataset	Scale	Genre	Topic level	Topic Form	Annotation Method	Annotation content	Support Tasks	Accessible
XZZ	505	Dialogue	sentence	-	manual	TB	TS	√
MUG	654	Dialogue	sentence	clause or sentence	manual	PB, TB, Subheadings, Title	TS, OG, TG	√
Wiki$_{zh}$	10000	Wikipedia	sentence	phrase	automatic	TB	TS	×*
WLX	2951	Web doc	paragraph	unknown	manual	Unknown	TS	×
CPTS(Ours)	14393	News text	paragraph	clause or sentence	man-machine collaborative	PB, TB, Subheadings, Title	TS, OG, TG	√

Table 2: The comparison of CPTS and the other Chinese corpora. The asterisk* means that Wiki section zh (Wiki$_{zh}$) contains 10000 documents randomly selected from ZhWiki and is not directly available.TB means Topic Boundary, PB means Paragraph Boundary, TS means Topic Segmentation, OG means Outline Generation, and TG means Title Generation.

(a variant of Segbot where using XLNet replaces the GRU encoder) are two pointer network models that first encode input text by GRU or XLNet and then use a pointer network to select topic boundaries in the input sequence. **TM-BERT** (Jiang et al., 2021) is local classification model that identifies the topic boundary by a triple semantic BERT-based matching mechanism. **BERT+Bi-LSTM** and **Hier. BERT** (Lukasik et al., 2020) are sequential labeling models, using LSTM and Transformer as base-architecture, separately. Following the published papers, we reproduce them, and the experimental settings of each model in topic segmentation are shown in Table 3. We also take **ChatGPT** as a baseline and adopt the prompt and settings from the probing ChatGPT in conversation topic segmentation (Fan et al., 2023).

Model	BS	LR	Epoch	PLM
Segbot	20	1E-03	10	Word2Vec
PN-XLNet	20	1E-03	10	XLNet-base
TM-BERT	2	1E-05	10	Bert-base
Bert+Bi-LSTM	2	1E-05	10	Bert-base
Hier. BERT	2	1E-05	10	Bert-base

Table 3: The main hyper-parameters of baselines on topic segmentation. BS is the batch size, LR is learning rate and PLM is the pre-trained language model.

6.1.2. Evaluation and Results

For topic segmentation evaluation, we use the following commonly used metrics[2]: P_k, WindowDiff, Segmentation Similarity and Boundary Similarity. We also report the macro-F1 of each model for a comprehensive evaluation.

The experimental results are shown in Table 4. Although ChatGPT is a powerful LLM, its performance in topic segmentation on text still lags far behind other fine-tuned pre-trained models due to 0-shot setting. Compared to Segbot without a pre-trained model, PN-XLNet improves the performance in all metrics. As a local classification model, TM-BERT outperforms PN-XLNet by 2.92 in F1 value, however, it performs worse in the popular topic segmentation evaluation matrix (Pk, WD, S, and B). Besides, with the two-layer architecture, BERT+Bi-LSTM and Hier. BERT achieve the best performance. Both models outperform other models in terms of F1 values and other metrics, as BERT better captures the semantic representation of paragraphs and hierarchical modeling better captures global information.

[2]https://github.com/cfournie/segmentation.evaluation

Model	P_k↓	WD↓	S↑	B↑	F1↑
ChatGPT (0-shot)	41.12	63.57	37.45	59.51	52.51
Segbot	24.06	25.85	89.73	58.94	75.23
PN-XLNet	22.02	23.34	91.27	65.19	77.70
TM-BERT	22.86	24.44	89.93	58.84	80.62
BERT+Bi-LSTM	**19.45**	**20.89**	91.76	65.88	**81.62**
Hier. BERT	19.76	21.00	**91.92**	**66.54**	81.40

Table 4: The performance on topic segmentation. Different from Segmentation Simlarity (S) and Boundary Similarity (B), P_k and WindowDiff (WD) are penalty measures.

6.2. Outline Generation

6.2.1. Task Definition and Baselines

Unlike the previous work (Barrow et al., 2020) on outline generation that takes only the first-level heading (usually a word or a phrase) of the document in Wikipedia as its subheading, the subheading in CPTS is usually a clause or sentence. It is more challenging to joint learning of topic segmentation and outline generation (Zhang et al., 2019). Therefore, we align with MUG (Zhang et al., 2023) and treat outline generation as a separate task like summary generation instead of text classification (Barrow et al., 2020; Lo et al., 2021) to simplify the problem and achieve a more intuitive performance. Given a section s_j that contains m consecutive paragraphs $(p_j^1, ..., p_j^m)$ in one topic, the outline generation model ($Model_{og}$) needs to generate the subheading h_j of them, as shown in Eq. 2.

$$h_j = Model_{og}(p_j^1, ..., p_j^m) \quad (2)$$

We select the following popular text generation models as our baselines. **BART** (Lewis et al., 2020) and **T5** (Raffel et al., 2020) are popular encoder-decoder-based pre-trained models. All of them have the same settings: Batch-size is 8, LR is

1E-05 and epoch is 10. We also take the **ChatGPT** as a baseline since the strong generation ability (The corresponding prompt can be seen in Appendix A.1).

6.2.2. Evaluation and Results

Since outline generation is regarded as a summarization task, we evaluate baseline models using the evaluation methods commonly used in summarization. Specifically, we use ROUGE (Lin, 2004a) and BLEU (Papineni et al., 2002) to evaluate the quality of generated subheadings from the word overlap, and use BertScore (Zhang et al., 2020) to evaluate the quality of generated subheadings from the semantics. Furthermore, we also did a manual evaluation by ranking their results on 100 randomly selected samples. The manual evaluation details can be seen in Appendix A.2.

Model	R-1	R-2	R-L	BLEU	BertScore	Rank ↓
ChatGPT (0-shot)	22.64	12.04	20.58	6.49	61.39	**2.49**
ChatGPT (3-shot)	22.25	11.87	20.22	6.47	61.45	2.61
BART	25.86	16.20	24.50	12.55	63.49	3.68
T5	27.14	16.00	25.44	12.04	63.74	3.25
T5 (24)	**28.91**	**17.88**	**27.06**	**14.46**	**64.67**	2.98

Table 5: The performance on outline generation.

Table 5 shows the performance of baselines in outline generation. Compared to the poor performance in topic segmentation, ChatGPT's 0-shot performance is close to the other popular fine-tuned models in outline generation due to its strong text generation ability. However, the 3-shot setting does not achieve further significant improvement. One reason may be that three long texts as input affect the in-context-learning (ICL) capability of ChatGPT. BART and T5 achieve similar performance with the same scale (the 6-layer encoder and the 6-layer decoder). In Table 5, the T5 model outperforms BART on R-1, R-L, and BertScore by 1.28, 0.94, and 0.25, respectively, while BART performs better on R-2 and BLEU. Not surprisingly, the larger T5 model performs best among all baselines.

It is worth noting that subheadings are more refined, shorter, and more abstract than summaries. Therefore, even the performance of the SOTA model is still not ideal in evaluation (e.g., the ROUGE in outline generation is lower than that in the summarization task.). It shows that outline generation is challenging, and there is still room for improvement in future work.

We observe two interesting trends from the last right column that is the manual evaluation. On the one hand, ChatGPT (0-shot) and ChatGPT (3-shot), which are models that are not fine-tuned, perform better (average 2.49 and 2.61 separately) than the other three fine-tuned models on manual ranking. It indicates that ChatGPT-generated subheadings are more friendly for humans, even though they may not align precisely with the original subheadings. On the other hand, the fine-tuned models (BART, T5, and T5 (24)) show the consistency of orders between manual ranking and their respective Rouge, BLEU, and BertScore metrics evaluations. It shows the reliability of our automatic evaluation and the usefulness of our corpus.

6.2.3. Title Generation Results

Since we have proposed a three-level topic structure representation, we further added title generation as a supplement to outline generation. In this task, we use all subtitles as input to generate the title.

Model	R-1	R-2	R-L	BLEU	BertScore
ChatGPT(0-shot)	16.87	7.79	15.08	3.85	59.52
ChatGPT(3-shot)	16.81	7.60	15.00	3.69	59.31
BART	25.85	**16.62**	24.67	**11.86**	63.79
T5	25.06	14.19	23.47	8.86	62.76
T5 (24)	**28.01**	16.55	**26.11**	10.96	**64.61**

Table 6: The performance on title generation.

Table 6 shows the performance of each model on the title generation task. The performance of each model is similar to that in the outline generation task. The lower performance of ChatGPT on title generation shows it is more challenging than outline generation since it needs more abstraction and reasoning at a higher level. It also demonstrates that outline and title generation are harder than traditional text summarization.

7. Application in Discourse Parsing

To validate the effectiveness of our constructed representation and corpus, we also utilized it to assist in discourse parsing, a downstream task. Different from topic segmentation, which splits the documents into several segments, discourse parsing is a more complex task where the model needs to build a tree structure on basic discourse units. Previous researches (Jiang et al., 2021; Huber et al., 2022a) have shown that topic structure can imply a skeleton of the rhetorical structure of a document, and there is a consistency in local discrimination for topic segmentation and discourse parsing. However, the lack of a topic structure corpus limits the method application in Chinese. Therefore, we attempt to use CPTS for training the model to predict topic structure in the document without explicit subheadings, helping discourse parsing.

Model	Span
Dist(Paragraph Boundary)	50.23
Dist(Topic Boundary)	55.33

Table 7: The performance on MCDTB.

Specifically, we conducted our preliminary experiment on MCDTB (Jiang et al., 2018), which is a popular corpus for discourse parsing in Chinese. Then, following the previous work (Huber et al., 2022a), we employed two distant supervision methods for paragraph-level discourse parsing. The Dist(Paragraph Boundary) model utilizes paragraph boundaries from the MCDTB corpus as topic boundaries to learn the discourse structure of the document, while the Dist(topic Boundary) model leverages the real topic structure provided by CPTS as the learning goal to learn the discourse structure of the document. Table 7 shows that the real topic structure in CPTS enhances the parser's performance on paragraph-level discourse parsing in Chinese from 50.23 to 55.33, demonstrating the usefulness of the topic structure corpus.

8. Discussion and Future Work

8.1. The Applicability of Annotation Method

Our methodology is not limited to the news documents used in this paper and is equally applicable to other genres as long as they possess some markers indicating paragraph-level topic structures. This includes a wide range of textual materials such as legal documents, novels, and academic documents, where structured topics are often revealed by special tokens. Then, in our method, the annotator only needs to verify the internal and external errors after the automatic extraction with the special token, reducing the annotation workload greatly.

A particularly promising application is in the analysis of scripts or novels, where our method could be used to identify broader narrative shifts or stage changes indicated by unique transition and voiceover tokens in the script. We believe that by constructing paragraph-level topic structures for these diverse categories, our method will not only aid in a deeper understanding and grasp of the paragraph-level topic structure of documents but also could guide large language models (LLMs) in generating more controlled content at the higher level according to given structures.

8.2. The Potential Challenges

Expansion of the Joint Learning Framework of Topic Segmentation and Outline Generation. There are some works attempting to jointly learn topic segmentation and outline generation by two classification tasks due to shorter subheadings. Expanding the current joint learning framework is a viable approach to deal with longer subheading challenges. A possible solution is integrating heterogeneous text classification tasks (such as topic segmentation) into text generation tasks (like outline generation) into a single and unified generation model based on powerful LLMs.

Exploration of Hierarchical Topic Structures. Existing methods usually view the topic structure as a flat structure and ignore the longer dependency of different topics. Therefore, another promising direction is delving into the hierarchical nature of paragraph-level topic structures. This exploration can be conducted from both bottom-up and top-down perspectives. By harnessing the capabilities of large-scale language models, it's possible to model the hierarchical relationships among topics (including parent-child connections between different layers) and the interconnections between various topic segments within the same document, such as writing style and semantic relationships with more words.

9. Conclusion

To fill the gap in Chinese paragraph-level topic structure resources, we first propose a three-layer discourse topic structure representation to guide the construction of our corpus. It takes the sentence as the topic to express richer paragraph-level information and incorporates paragraph boundaries, topic boundaries, subheadings, and the title into the topic structure. Then, we designed a two-stage human-machine collaborative annotation method to construct the largest high-quality Chinese paragraph-level topic structure corpus. By combining automatic extraction and manual verification, we ensure the correctness of the topic structure not only formally but also semantically. We described the construction process of the corpus in detail and conducted an in-depth analysis and comparison of it. Finally, we verified the computability of the corpus through eight topic segmentation and outline generation baselines, including ChatGPT. In the future, we will focus on improving the performance of Chinese topic segmentation and outline generation by designing appropriate methods to assist other downstream tasks in the LLM era.

10. Acknowledgements

This research is supported by the project of Shenzhen Science and Technology Research Fund (Fundamental Research Key Project Grant No. JCYJ20220818103001002), the Internal Project Fund from Shenzhen Research Institute of Big Data under Grant No. T00120-220002, Shenzhen Key Laboratory of Cross-Modal Cognitive Computing (grant number ZDSYS20230626091302006), and the National Natural Science Foundation of China (No. 62376181).

11. Bibliographical References

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

Joe Barrow, Rajiv Jain, Vlad I. Morariu, Varun Manjunatha, Douglas W. Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *ACL*, pages 313–322.

Roger H Bruning, Gregory J Schraw, and Royce R Ronning. 1999. *Cognitive psychology and instruction*. ERIC.

Wallace Chafe. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*.

Freddy YY Choi, Peter Wiemer-Hastings, and Johanna D Moore. 2001. Latent semantic analysis for text segmentation. In *EMNLP*.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling News Discourse Structure Using Explicit Subtopic Structures Guided Critics. In *Findings of EMNLP*, pages 1594–1605.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *NAACL-HLT*, pages 353–361.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study.

Chris Fournier. 2013. Evaluating text segmentation using boundary edit distance. In *ACL*, pages 1702–1712.

Chris Fournier and Diana Inkpen. 2012. Segmentation similarity and agreement. In *NAACL-HLT*, pages 152–161.

Goran Glavas, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In **SEM@ACL*.

Goran Glavas and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *AAAI*, pages 7797–7804.

Dionysis Goutsos. 1997. *Modeling discourse topic: sequential relations and strategies in expository text*, volume 59. Greenwood Publishing Group.

Joseph E Grimes. 2015. *The thread of discourse*. De Gruyter Mouton.

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

Patrick Huber, Linzi Xing, and Giuseppe Carenini. 2022a. Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *AAAI*.

Patrick Huber, Linzi Xing, and Giuseppe Carenini. 2022b. Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *AAAI*.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Fang Kong. 2021. Hierarchical macro discourse parsing based on topic segmentation. In *AAAI*, pages 13152–13160.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A macro-level Chinese discourse TreeBank. In *COLING*, pages 3493–3504.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Jing Li, Aixin Sun, and Shafiq Joty. 2018. SEGBOT: A generic neural text segmentation model with pointer network. In *IJCAI*, pages 4166–4172.

Chin-Yew Lin. 2004a. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3728–3738.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray L. Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of EMNLP*, pages 3334–3340.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text Segmentation by Cross Segment Attention. In *EMNLP*, pages 4707–4716.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *ACL*.

Marie-Francine Moens and Rik De Busser. 2001. Generic topic segmentation of document texts. In *SIGIR*, pages 418–419.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Jay M Ponte and W Bruce Croft. 1997. Text segmentation by topic. In *International Conference on Theory and Practice of Digital Libraries*, pages 113–125.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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

Clive Seale and David Silverman. 1997. Ensuring rigour in qualitative research. *The European journal of public health*, 7(4):379–384.

Manfred Stede. 2011. Discourse processing. *Synthesis Lectures on Human Language Technologies*, 4(3):1–165.

Richard Watson Todd. 2011. Analyzing discourse topics and topic keywords.

Richard Watson Todd. 2016. *Discourse topics*, volume 269. John Benjamins Publishing Company.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *ACL*, pages 491–498.

Teun A Van Dijk. 2014. *Discourse and knowledge: A sociocognitive approach*. Cambridge University Press.

Teun A Van Dijk and Walter Kintsch. 1983. *Strategies of discourse comprehension*. Acadamic Press.

Richard Watson Todd. 2003. *Topics in classroom discourse*. Ph.D. thesis, UK: University of Liverpool.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *EMNLP-IJCNLP*, pages 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Linzi Xing, Patrick Huber, and Giuseppe Carenini. 2022. Improving topic segmentation by injecting discourse dependencies. In *Proceedings of 3rd Workshop on Computational Approaches to Discourse (CODI 2022)*, page 7.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2019. Outline generation: Understanding the inherent content structure of documents. In *SIGIR*, pages 745–754.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*.

12. Language Resource References

Sebastian Arnold and Rudolf Schneider and Philippe Cudré-Mauroux and Felix A. Gers and Alexander Löser. 2019. *SECTOR: A Neural Model for Coherent Topic Segmentation and Classification*.

Harr Chen and S. R. K. Branavan and Regina Barzilay and David R. Karger. 2009. *Global Models of Document Structure using Latent Permutations*.

Freddy Y. Y. Choi. 2000. *Advances in domain independent linear text segmentation*.

Jacob Eisenstein and Regina Barzilay. 2008. *Bayesian Unsupervised Topic Segmentation*.

Omri Koshorek and Adir Cohen and Noam Mor and Michael Rotman and Jonathan Berant. 2018. *Text Segmentation as a Supervised Learning Task*.

Liu, Yang and Zhu, Chenguang and Zeng, Michael. 2022. *End-to-End Segmentation-based News Summarization*. Association for Computational Linguistics.

Robert Parker, David Graff, Ke Chen, Junbo Kong, Kazuaki Maeda. 2009. *Chinese Gigaword Fourth Edition*. Linguistic Data Consortium. ISLRN 261-416-300-929-8.

Liang Wang and Sujian Li and Xinyan Xiao and Yajuan Lyu. 2016. *Topic Segmentation of Web Documents with Automatic Cue Phrase Identification and BLSTM-CNN*.

Linzi Xing and Brad Hackinen and Giuseppe Carenini and Francesco Trebbi. 2020. *Improving Context Modeling in Neural Topic Segmentation*.

Yi Xu and Hai Zhao and Zhuosheng Zhang. 2021. *Topic-Aware Multi-turn Dialogue Modeling*.

Zhang, Qinglin and Deng, Chong and Liu, Jiaqing and Yu, Hai and Chen, Qian and Wang, Wen and Yan, Zhijie and Liu, Jinglin and Ren, Yi and Zhao, Zhou. 2023. *Mug: A general meeting understanding and generation benchmark*. IEEE.

A. Appendix

A.1. The prompt we designed for outline generation

The prompt we designed for the 0-shot setting in the outline generation is the following:

instruction:"一个文章包含几个段落(paragraph)和一个小标题(subheading)，请根据下面的段落生成文章的小标题，填写在subheading属性中，并以json的格式返回。" (A document contains several paragraphs and a subheading. Please generate a subheading for the document based on the following paragraphs, fill it in the subheading attribute, and return it in JSON format.)

input: <sample>

where <sample> is a dictionary containing several paragraphs on one topic: "input": "{"paragraph": [...], "subheading": ""}"

The prompt for the 3-shot setting is similar to the above.

A.2. The details of manual evaluation on outline generation task

We randomly selected 100 samples and asked three evaluators to rank them based on the following evaluation settings.

In the manual evaluation, evaluators were provided with a text fragment with several paragraphs belonging to a single topic as the prompt. Subsequently, they were presented with output results from the five models as ranking candidates. Importantly, these options were initially randomized and anonymized to ensure an unbiased evaluation process. The evaluators were asked to rank these candidate subheadings (from 1 to 5, 1 is the best and 5 is the worst) based on the following three criteria:

- Relevance: Whether the subheading accurately represents the content described in the text.

- Appropriateness: Whether the subheading conforms to typical styles and formats of subheadings.

- Fluency: Whether the subheading is correctly formulated and flows smoothly.

We then compiled the average rankings for each model from three evaluators to determine the final scores, ensuring the objectivity of our evaluation.