LREC-COLING 2024

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

Feng Jiang,

Weihao Liu, Xiaomin Chu,

Peifeng Li, Qiaoming Zhu, Haizhou Li

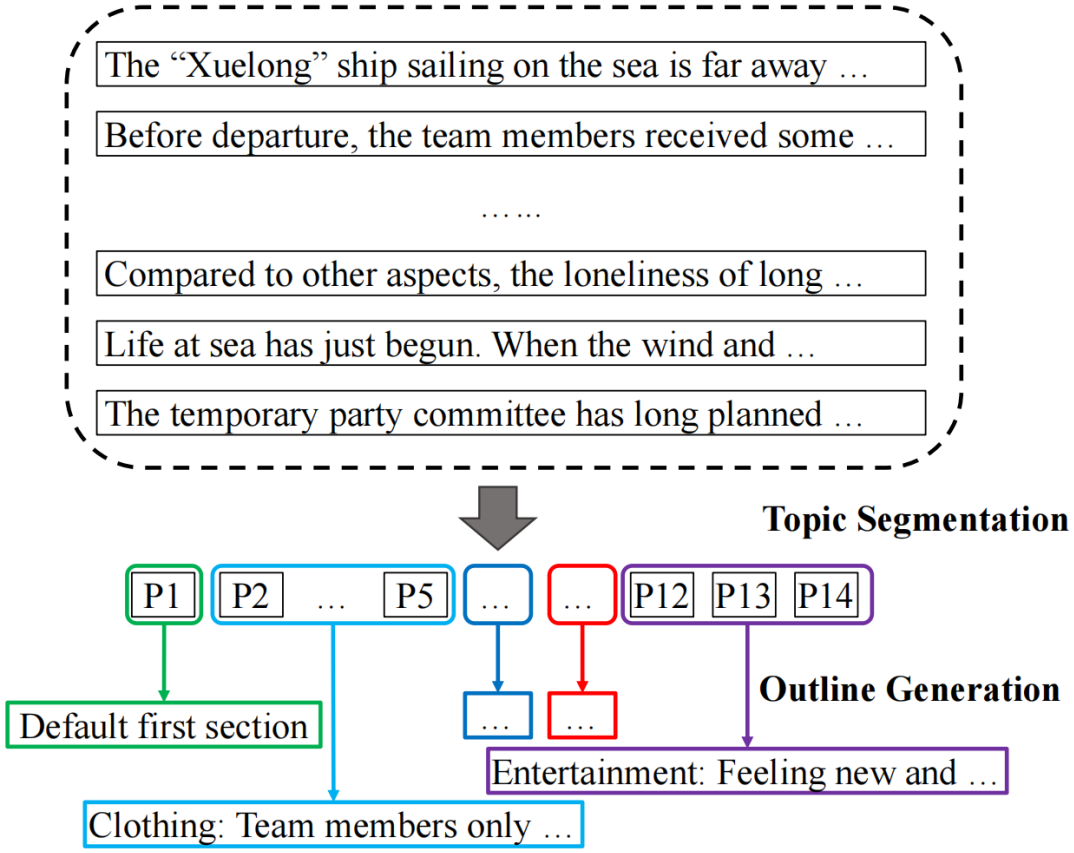
香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

Introduction

A well-written document usually consists of several semantically coherent text segments, each of which revolves around a specific topic.

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.

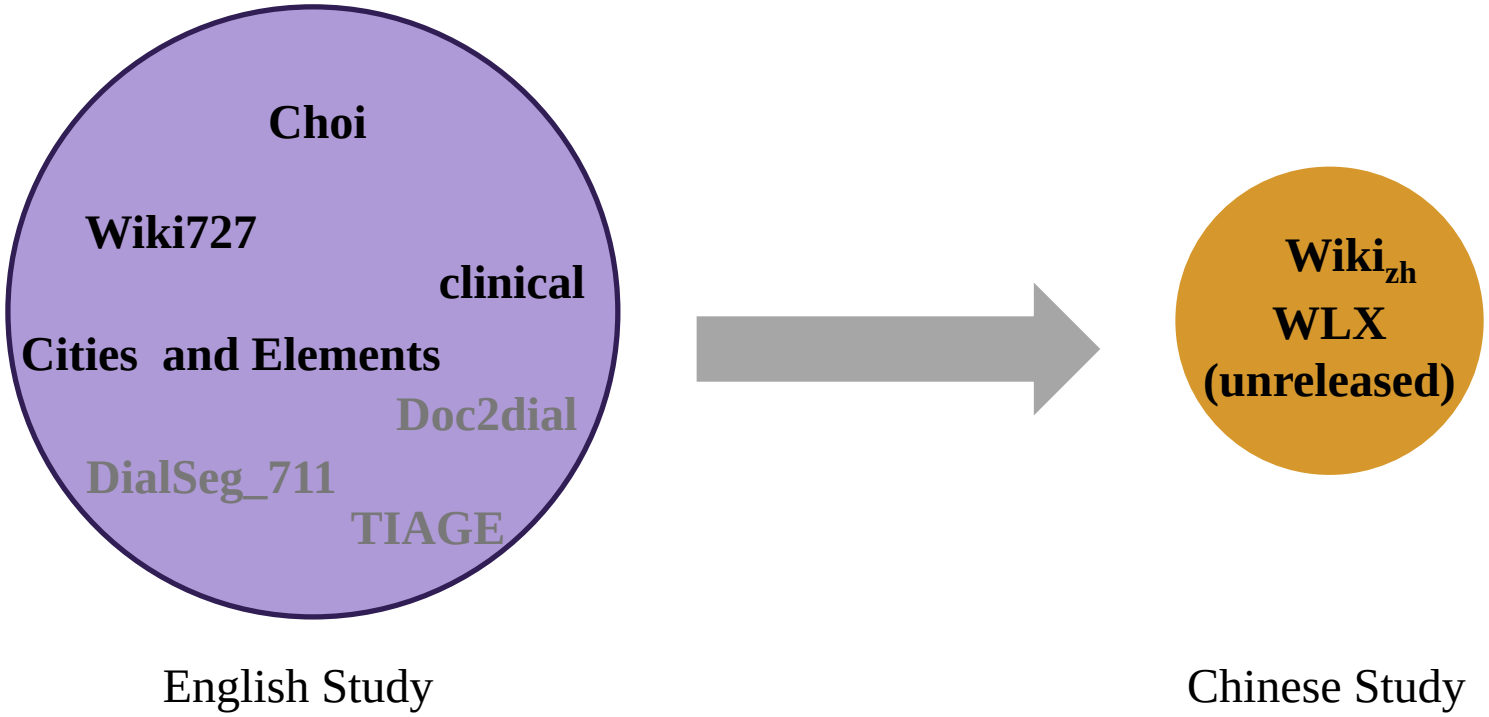
Compared with *sentence-level* topic structure, the ***paragraph-level* topic structure** pays more attention to the document's higher-level topic structure between paragraphs.



Paragraph-level topic segmentation and outline generation

Challenges

There are fewer studies on **Chinese topic structure** compared to English, especially in paragraph-level



English Study

Chinese Study

Challenges

There are fewer studies on **Chinese topic structure** compared to English, especially in paragraph-level

- **Lack of Representation**:

Existing work more focus on modeling design and ignore the corresponding topic theory in Chinese

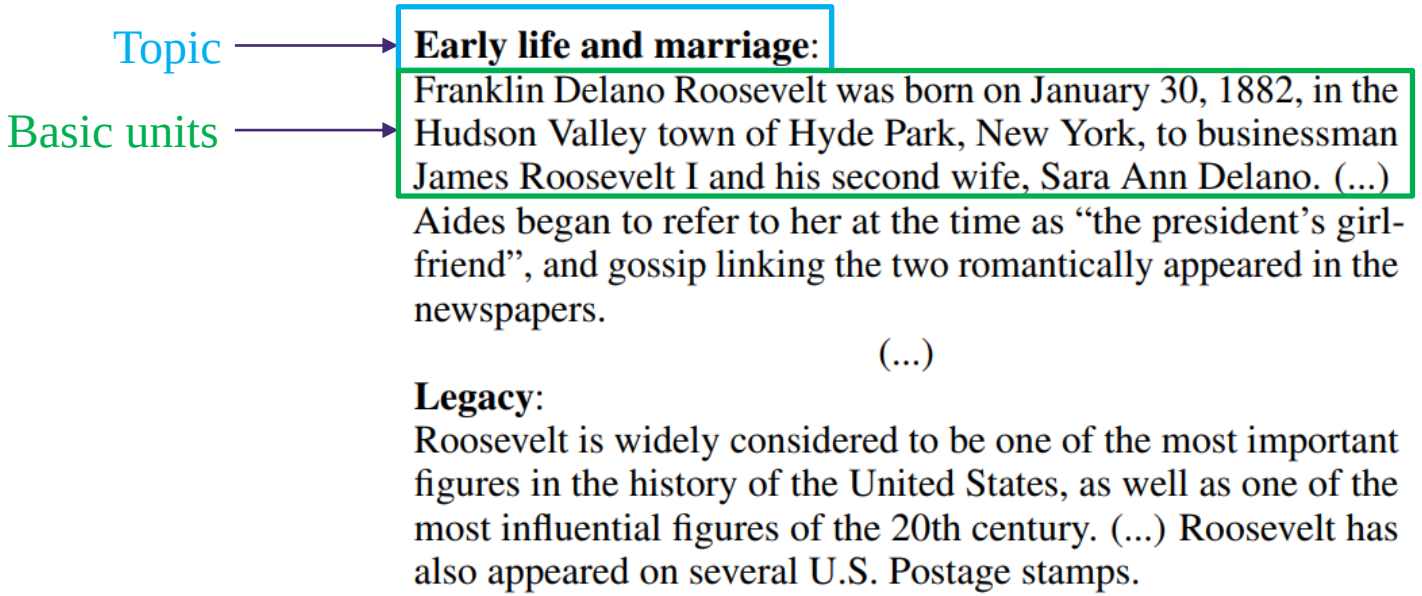
- **Lack of Corpus**:

Existing work more focus on sentence-level corpus ignore the higher level topic structure in Chinese

- **Representation**: How to represent paragraph-level topic structures more richly.
- **Corpus**: How to build a paragraph-level topic structure corpus that is both large-scale and high-quality.

Chinese Paragraph-level Topic Structure Representation

Most of the existing corpora only annotate **basic units** and **topics** they subordinate.

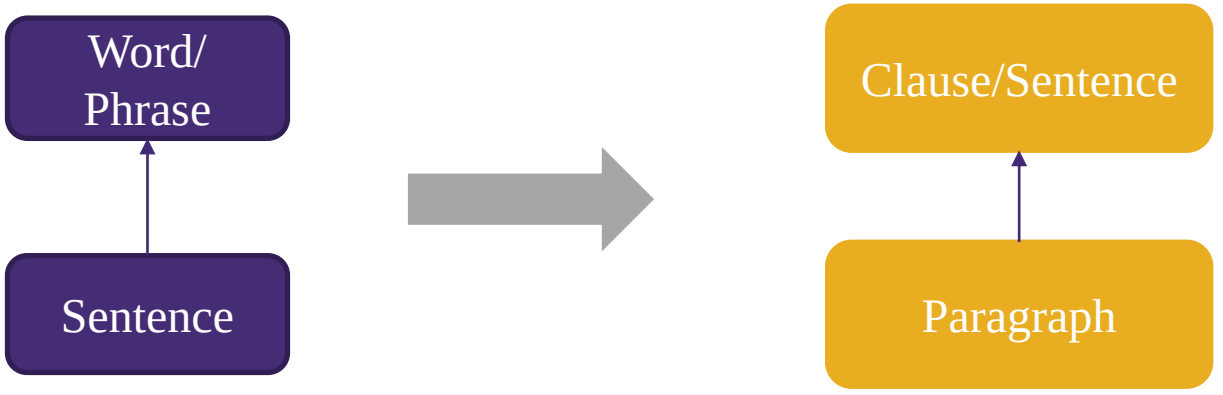


Lukasik M, Dadachev B, Papineni K, et al. Text Segmentation by Cross Segment Attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4707-4716.

Chinese Paragraph-level Topic Structure Representation

However, at the paragraph level

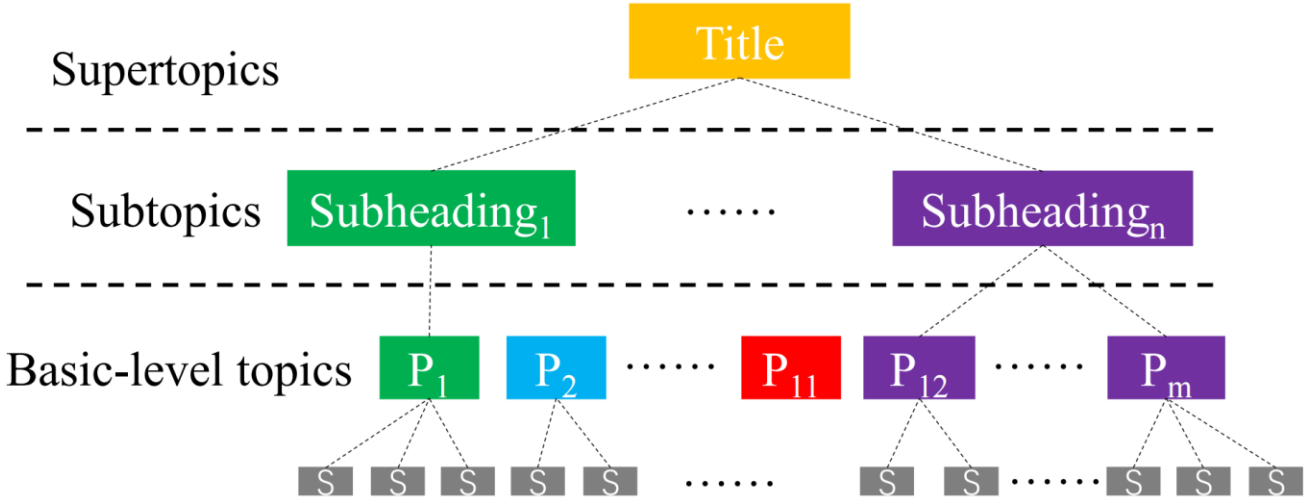
(1) It needs to contain more information due to the granularity of basic units is larger



(2) It needs to express the high-level structure of the document, such as subheadings and titles.

Chinese Paragraph-level Topic Structure Representation

Therefore, 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)



(1) Three levels: title as a **supertopic**, subheadings as **subtopics**, and paragraphs as **basic-level topics**
(2) Each level of unit belongs to only one higher one
(3) Each level is in sentence form instead of words

Chinese Paragraph-level Topic Structure Corpus Construction

Data Source

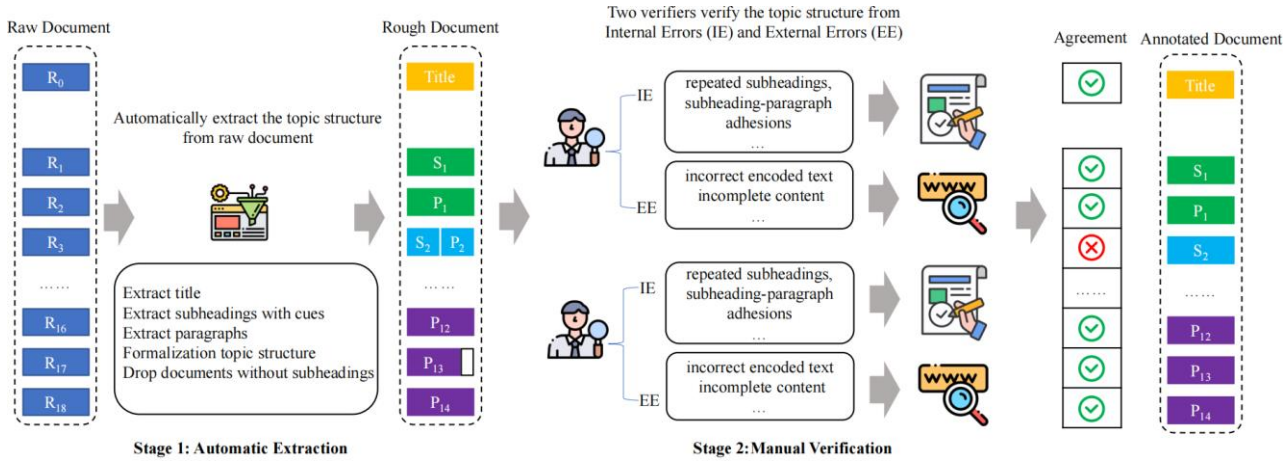
Annotators

1 PhD Student
6 Master Student
1 Undergraduate Student

News documents issued by Xinhua News Agency from Chinese **Gigaword** Fourth Edition [1]
1314198 story news documents from January 1991 to December 2008

[1] https://catalog.ldc.upenn.edu/LDC2009T27

Chinese Paragraph-level Topic Structure Corpus Construction



The Two-stage Man-machine Collaborative Annotation

Stage 1: Automatic Extraction

A heuristic automatic extraction method to extract topic structures from raw documents automatically.

Stage 2: Manual Verification

(1) Verify the correctness of automatic extraction from a semantic perspective

- **Internal Errors (IE)** including repeated subheadings, title-paragraph adhesions, etc.

It could be corrected by the document content

- **External Errors (EE)** such as incorrect encoded text or incomplete content in some subheadings or paragraphs:

It only be corrected by using search engines

(2) Quickly re-verifies the form correctness of topic boundaries that have been automatically extracted

Statistics and Analysis on CPTS

CPTS contains 14393 annotated documents

- 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).

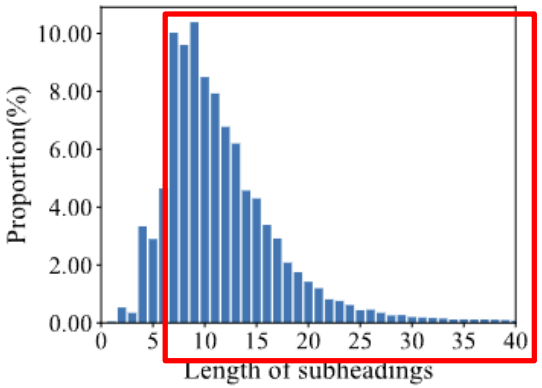
Item	Max	Min	Avg.
# words/document	5791	180	1727.96
# paragraphs/document	40	2	14.76
# words/subheading	147	1	12.33
# paragraphs/subheading	33	1	3.70
# subheadings/document	20	2	4.00

Diversity of annotation

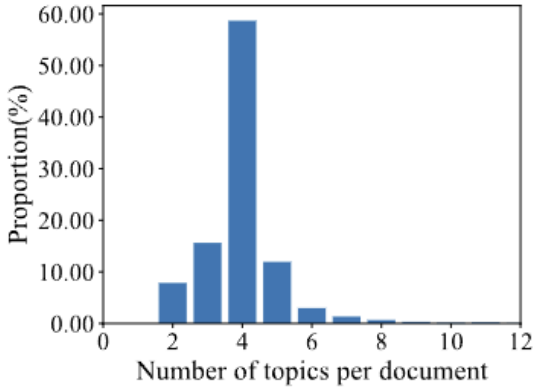
One group	Two annotators
Inter-group Annotated Agreement	94.79%
Kappa Value	0.849

High-quality of annotation

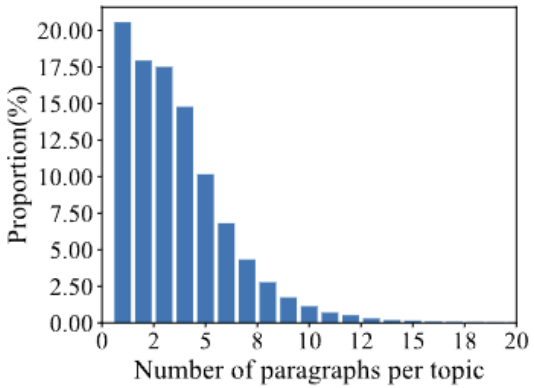
Statistics and Analysis on CPTS



(a) Distribution of subheadings length.



(b) Distribution of topics per document.



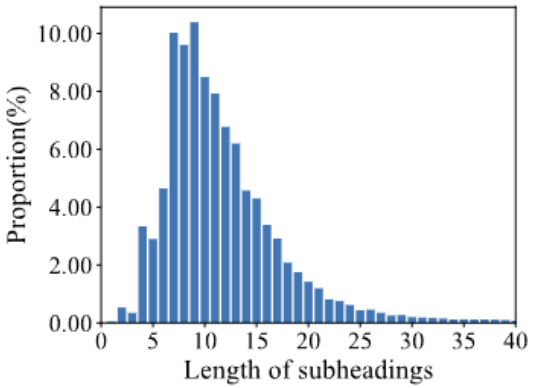
(c) Distribution of paragraphs per topic.

(a) **About 90% subheadings have more than seven words**, which could fully express the information of a paragraph-level topic by clauses or sentences

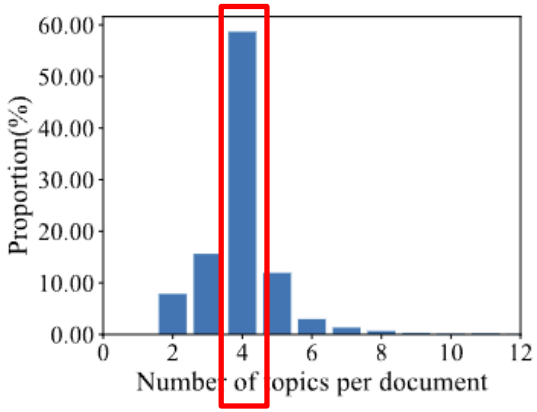
(b) About 60% of the documents have four topics, demonstrating the topic granularity will change with the document length

(c) Over 70% of topics contain less than four paragraphs, which indicate the usefulness of the paragraph-level topic: it can divide a document into two more simple structures: the discourse structure among paragraph-level topics and that in one topic

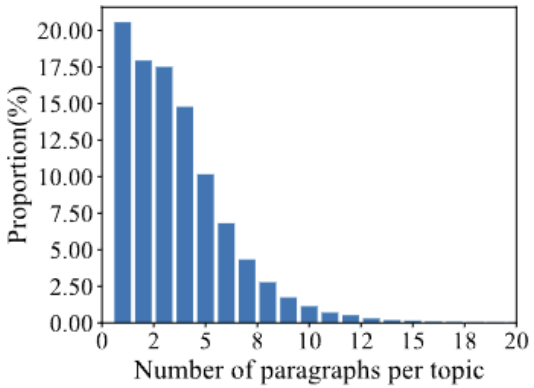
Statistics and Analysis on CPTS



(a) Distribution of subheadings length.



(b) Distribution of topics per document.



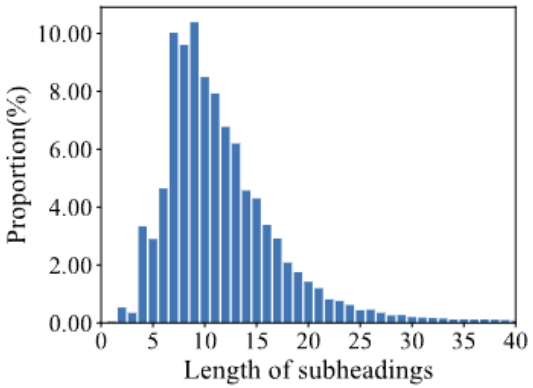
(c) Distribution of paragraphs per topic.

(a) About 90% subheadings have more than seven words, which could fully express the information of a paragraph-level topic by clauses or sentences

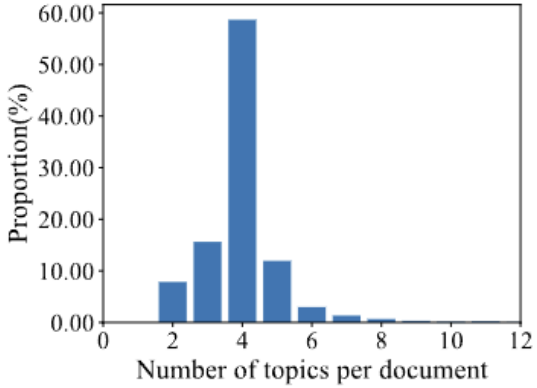
(b) **About 60% of the documents have four topics**, demonstrating the topic granularity will change with the document length

(c) Over 70% of topics contain less than four paragraphs, which indicate the usefulness of the paragraph-level topic: it can divide a document into two more simple structures: the discourse structure among paragraph-level topics and that in one topic

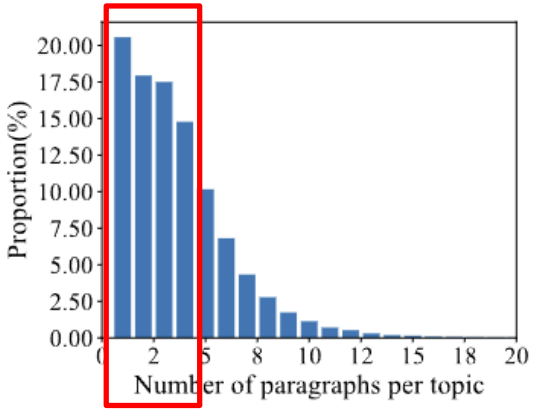
Statistics and Analysis on CPTS



(a) Distribution of subheadings length.



(b) Distribution of topics per document.



(c) Distribution of paragraphs per topic.

(a) About 90% subheadings have more than seven words, which could fully express the information of a paragraph-level topic by clauses or sentences

(b) About 60% of the documents have four topics, demonstrating the topic granularity will change with the document length

(c) **Over 70% of topics contain less than four paragraphs**, which indicate the usefulness of the paragraph-level topic: it can divide a document into two more simple structures: the discourse structure among paragraph-level topics and that in one topic

Compared CPTS with Other Chinese Topic Structure Corpora

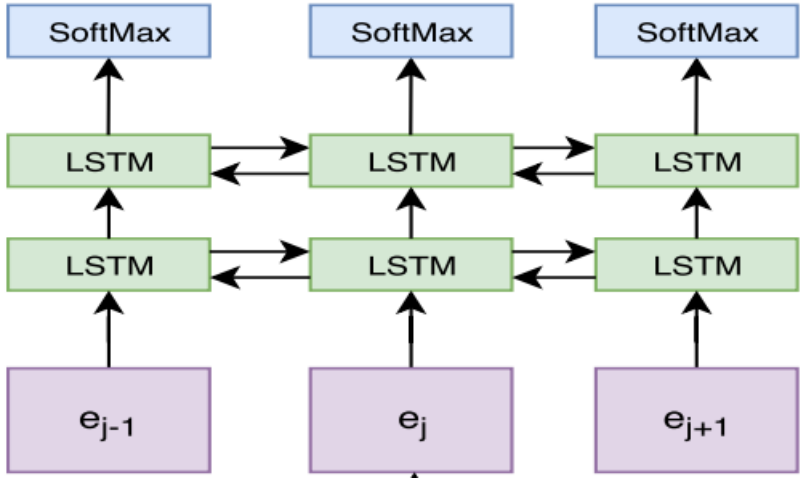
Dataset	Scale	Genre	Topic level	Topic Form	Annotation Method	Annotation content	Support Tasks	Accessible
XZZ	505	Dialogue	sentence	-	manual	TB	TS	√
MUG	654	Dialogue	sentence	clause or sentence	manual	PB, TB, Subheadings, Title	TS, OG, TG	√
Wiki_{zh}	10000	Wikipedia	sentence	phrase	automatic	TB	TS	×*
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	√

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.

(1) CPTS is **the largest high-quality** Chinese topic structure corpus (14393).
(2) CPTS annotated **more comprehensive paragraph-level topic structures**, including Paragraph Boundaries (PB), Topic Boundaries (TB), subheadings, and titles.
(3) CPTS will **be open access to the community**.

https://github.com/fjiangAI/CPTS

Experiments on Corpus Evaluation



Supervised Topic Segmentation [1]

- Topic Segmentation Task

Model	$P_k\downarrow$	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

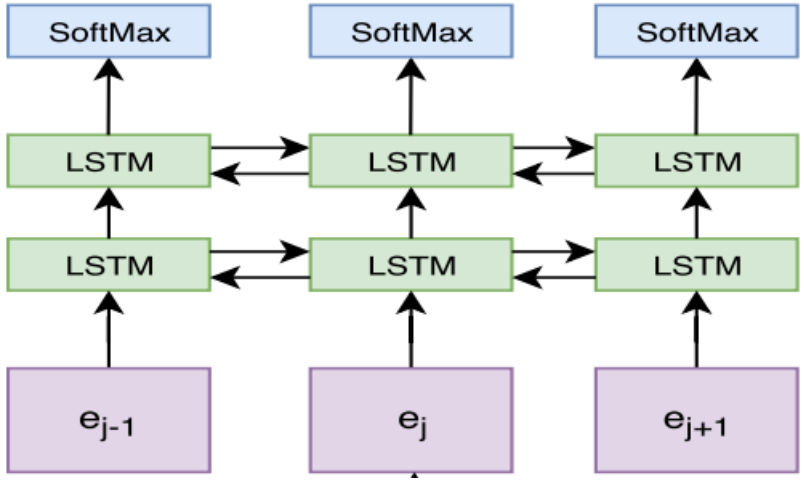
The auto evaluation in topic segmentation

(1) ChatGPT's performance in topic segmentation on text still lags far behind other fine-tuned pre-trained models due to 0-shot setting [2].

(2) The fine-tuned BERT+Bi-LSTM and Hier. BERT [3] achieve the best performance with the two-layer architecture.

[1] Koshorek O, Cohen A, Mor N, et al. Text Segmentation as a Supervised Learning Task[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 469-473.
[2] Fan Y, Jiang F. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study[J]. arXiv preprint arXiv:2305.08391, 2023.
[3] Lukasik M, Dadachev B, Papineni K, et al. Text Segmentation by Cross Segment Attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4707-4716.

Experiments on Corpus Evaluation



Supervised Topic Segmentation [1]

- Topic Segmentation Task

Model	$P_k\downarrow$	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

The auto evaluation in topic segmentation

(1) ChatGPT's performance in topic segmentation on text still lags far behind other fine-tuned pre-trained models due to 0-shot setting [2].

(2) The fine-tuned BERT+Bi-LSTM and Hier. BERT [3] achieve the best performance with the two-layer architecture.

[1] Koshorek O, Cohen A, Mor N, et al. Text Segmentation as a Supervised Learning Task[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 469-473.
[2] Fan Y, Jiang F. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study[J]. arXiv preprint arXiv:2305.08391, 2023.
[3] Lukasik M, Dadachev B, Papineni K, et al. Text Segmentation by Cross Segment Attention[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 4707-4716.

Experiments on Corpus Evaluation

Findings:

(1) Outline generation is still challenging due to poor performance.
(2) ChatGPT-generated subheadings are more friendly for humans, even though they may not align precisely with the original subheadings.
(3) The fine-tuned models show the consistency of orders between manual ranking and auto evaluations.
(4) The performance of the models in title generation is similar to that in outline generation task.

- Outline Generation Task

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

The auto evaluation and manual evaluation (Rank) in outline generation

- Title Generation Task

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**

The auto evaluation in title generation

Experiments on Corpus Evaluation

Findings:

(1) Outline generation is still challenging due to poor performance.
(2) ChatGPT-generated subheadings are more friendly for humans, even though they may not align precisely with the original subheadings.
(3) The fine-tuned models show the consistency of orders between manual ranking and auto evaluations.
(4) The performance of the models in title generation is similar to that in outline generation task.

- Outline Generation Task

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

The auto evaluation and manual evaluation (Rank) in outline generation

- Title Generation Task

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**

The auto evaluation in title generation

Experiments on Corpus Evaluation

Findings:

(1) Outline generation is still challenging due to poor performance.
(2) ChatGPT-generated subheadings are more friendly for humans, even though they may not align precisely with the original subheadings.
(3) The fine-tuned models show the consistency of orders between manual ranking and auto evaluations.
(4) The performance of the models in title generation is similar to that in outline generation task.

- Outline Generation Task

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

The auto evaluation and manual evaluation (Rank) in outline generation

- Title Generation Task

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**

The auto evaluation in title generation

Experiments on Corpus Evaluation

Findings:

(1) Outline generation is still challenging due to poor performance.
(2) ChatGPT-generated subheadings are more friendly for humans, even though they may not align precisely with the original subheadings.
(3) The fine-tuned models show the consistency of orders between manual ranking and auto evaluations.
(4) The performance of the models in title generation is similar to that in outline generation task.

- Outline Generation Task

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

The auto evaluation and manual evaluation (Rank) in outline generation

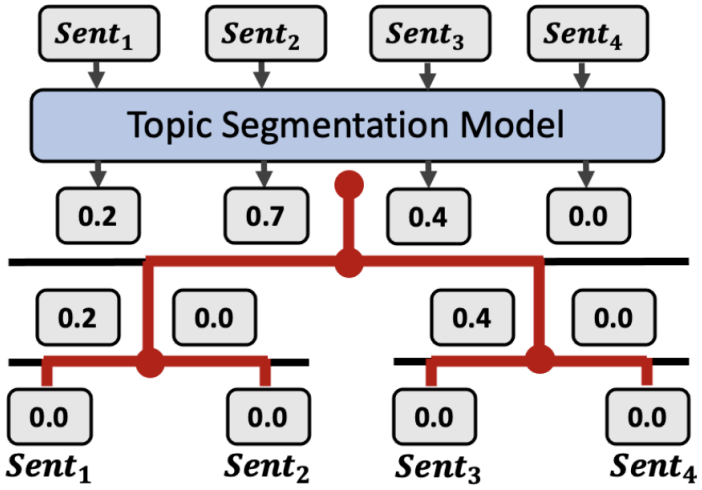
- Title Generation Task

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**

The auto evaluation in title generation

Experiments on Corpus Evaluation

- Application in Discourse Parsing



Distant-supervised Discourse parsing based on Topic Segmentation [1]

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

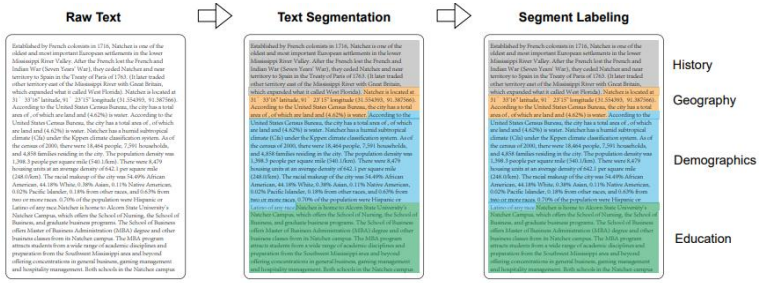
The performance on MCDTB [2]

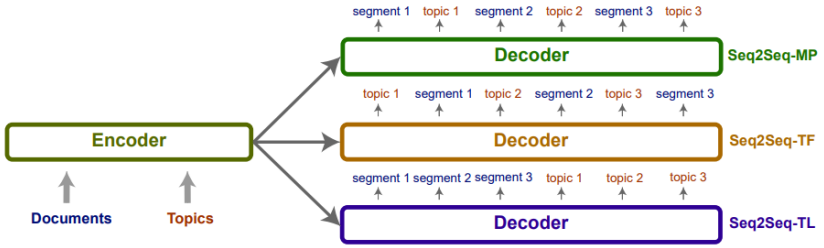
Using topic boundary is better than using paragraph boundary for discourse parsing

[1] Huber P, Xing L, Carenini G. Predicting above-sentence discourse structure using distant supervision from topic segmentation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(10): 10794-10802.
[2] Jiang F, Xu S, Chu X, et al. Mcdtb: a macro-level chinese discourse treebank[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3493-3504.

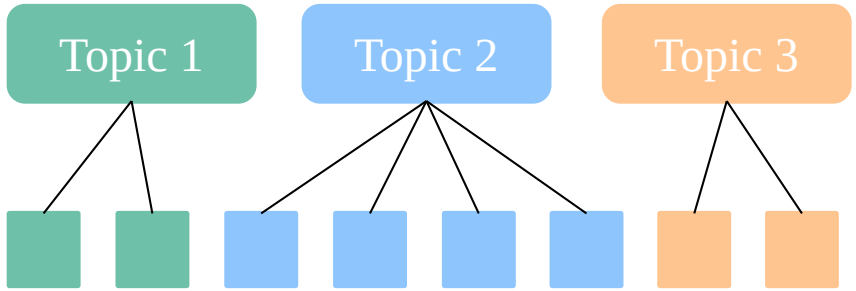
Potential Challenges

- Joint modeling topic segmentation and outline generation

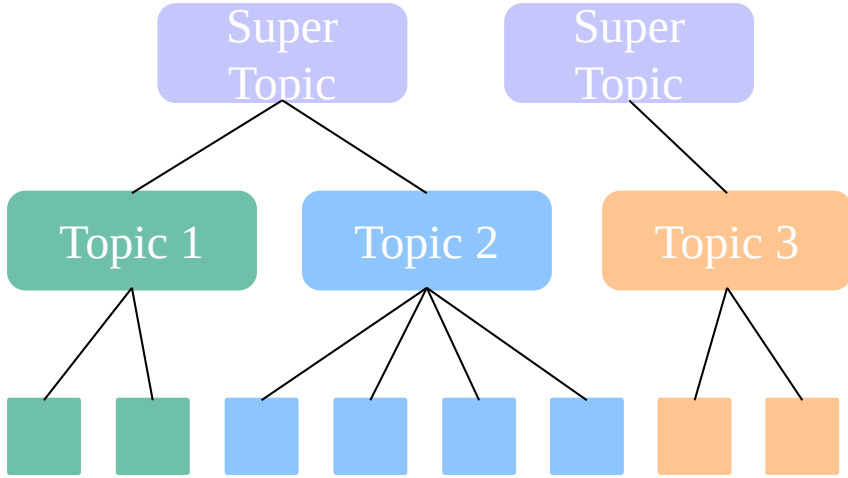




- Hierarchical topic modeling



Flat topic structure



Hierarchical topic structure

Conclusion

Summary:

To fill the gap in Chinese paragraph-level topic structure resources,

(1) **Representation**: Propose a **hierarchical paragraph-level topic structure representation** for modeling the topic structure of documents more comprehensively with three layers.

(2) **Corpus**: 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.

(3) **Benchmark**: Construct **several strong baselines to verify the computability of the CPTS** on two basic tasks: topic segmentation and outline generation, plus a preliminary experiments in the downstream task (discourse parsing).

Next Step: 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.

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

Thanks

Feng Jiang

jeffreyjiang@cuhk.edu.cn

School of Data Science

The Chinese University of Hong Kong, Shenzhen