Topic Segmentation Using Generative Language Models

Pierre Mackenzie [1] **Maya Shah** [1]
[1] Adarga, London

Abstract

Topic segmentation using generative Large Language Models (LLMs) remains relatively unexplored. Previous methods use lexical or semantic similarity between parts of a document to decide on boundaries, but they lack the long range dependency and vast knowledge contained in LLMs. Here, we propose a new prompting strategy and compare to semantic similarity-based methods. Results show that LLMs can be more effective segmenters than existing methods, but issues remain to be solved before they can be relied upon for topic segmentation.

1. Introduction

1.1. The Topic Segmentation Problem

Topic segmentation is the problem of dividing a string of text into constituent parts or 'segments'. Each segment should be semantically self-contained such that it is about one thing. The precise definition of a segment should be dependent upon the specific use cases. For this work, segment boundaries shall always lie on sentence boundaries.

We can think of this as a binary classification task. Given a list of input sentences S, of length n, the model must decide whether there exists a segment boundary between each pair of adjacent sentences. There are $n-1$ possible boundaries, and therefore our solution space is 2^{n-1}. Formally, the model must find a mapping f from the list of sentences to a binary vector of length $n-1$: $f(S) = \mathbf{y}$ where $\mathbf{y} = \{y_1, y_2 \ldots, y_{n+1}\}$ and each of the $y_i \in \{0, 1\}$.

Relative to generative tasks such as summarisation, the space of possible solutions is much smaller, but the problem remains subjective as where a boundary should lie can be ambiguous. Frequently, humans cannot agree on a correct solution (Hearst, 1997).

This task can be important as a processing step before some other NLP task, or can be important in its own right. One might use segments to generate a contents page for a long document, or individual summaries of segments within a document, or generate titles for each segment. Other tasks such as information retrieval, long-document summarisation and classification, can all benefit from first being broken down by topic. For many open source or resource-constrained models, context windows limit the size of input for such tasks which can be constraining (Chung et al., 2022). Although context windows can be increased (Chen et al., 2023) and newer models are consistently increasing context lengths, this alone does not fix all problems as LLMs do not fully utilise long context windows (Liu et al., 2023) (Petroni et al., 2020).

1.2. Related Work

Previous methods of topic segmentation use either lexical or semantic similarity, and a variety of different machine learning approaches. However, there has been no research into the usage of LLMs for topic segmentation. LLMs have been shown to be effective at a variety of NLP tasks due to their vast general knowledge of language, a skill which is also required for topic segmentation.

2. Method

2.1. Datasets

There are 4 types of datasets used in the experiments in this work: a small human-annotated dataset, a scrape of English wikipedia, a 'concatenated' wikipedia scrape and a synthetic GPT3.5 generated dataset.

2.1.1. Human-Annotated Dataset

Lacking the resources to create a large dataset annotated by humans, this work uses a very small manually segmented dataset of 10 documents. These documents are a mix of news articles, wikipedia articles, and miscellaneous documents such as podcast transcripts and scientific reports. This was intended to represent varying difficulties of segmentation, and provide examples which could be manually inspected to interpret segmenter behaviour. Only one annotator was used, and the segments were created by hand. Further work might involve multiple annotators on a much larger dataset, with a more rigorous process to ensure consistency and quality.

2.1.2. Wikipedia Dataset

A plain text English wikipedia scrape[1] was used articles where headings were delimited by special characters. The articles were then automatically segmented based on headings, and filtered to remove articles with very few segments, too short segments, or too much punctuation such as tables and figures. After this process, there remained approximately 1000 segmented articles. We generate two version of this dataset: one with headings removed, and one with headings included. We evaluate models on both versions of the dataset, but include only the version with headings removed in the final results. For full results, see the Appendix A.

2.1.3. Concatenated Wikipedia Dataset

We randomly sampled segments from the previous Wikipedia dataset and concatenated them in order to form new incoherent articles, with segments drawn from completely different domains. Intuitively, this dataset should be easier to segment as there are no semantic links between segments.

2.1.4. Synthetic Dataset

The final dataset used was generated synthetically by GPT-3.5. The source data was a mix of CTC sentinel data[2], UN-Peacekeeping corpus[3] and an internal dataset at Adarga. These documents were segmented by querying OpenAI's API with the model `gpt-3.5-turbo-16k` using the overlapping prompt schema defined in section 2.3. This dataset was exclusively used for fine-tuning the *FlanT5*3.1.5 model. It was not used for evaluation as the results would be biased in favor of the generative models.

2.2. Evaluation

The methodology in evaluating the quality of a model's segmentation in this work follows Chris Fournier's *Evaluating Text Segmentation using Boundary Edit Distance* (Fournier, 2013). This work uses edited versions of the Boundary Edit Distance proposed in this work, along with the associated information recall metrics 'boundary precision' and 'boundary recall'.

The boundary edit distance algorithm pairs matches of segment boundaries between a reference segmentation and a hypothesised segmentation, regardless of distance between matches. Exact matches score 1 and boundaries without matches score 0, whilst matches within some distance n score partial points based on a weighting function. For this work, the function always decreases score linearly as the boundary gets farther away from the true boundary. Boundary Edit Distance (B) is the mean score for all these matches, while Boundary Precision/Recall (BP/BR) measure the proportion of the hypothesis/reference boundaries which are matched, respectively.

A model is punished for both missing a boundary and for placing a boundary where there is none in a symmetric and intuitive manner. For further justification of the usage of the boundary edit distance algorithm rather than relying on more traditional metrics such as WD and Pk (Pevzner & Hearst, 2002), see *Evaluating Text Segmentation using Boundary Edit Distance* (Fournier, 2013), or see our own investigations at segmenter evaluation metrics.

2.3. LLM-Based Text Segmentation

How can we get LLMs to output segment boundaries?

We might first consider passing in the input text as a prompt and asking the LLM to copy out the text, adding markers indicating where it has placed boundaries as is suggested by (Xing, 2024). However, not only is this wasteful of tokens, especially if the input text is long, but crucially, the GLM may fail to copy the input accurately, may change the formatting, and we found through qualitative experimentation that boundary placement was not any better than our final method. These problems are addressed by (Xing, 2024) through repeated prompting until the input and output sequence lengths match, but this still does not guarantee integrity of the data. In our use case, guaranteeing that the input data would remain the same was of the utmost importance, so we used a different strategy.

2.3.1. Prompting Method

We first annotate the text with indices between each sentence. For example 'Hello World. [1] The sky is blue. [2] The sun is is yellow. ' We then ask the GLM to return a list of indices corresponding to boundaries. In the previous example, the ideal response might be '1'. In practice, the texts and list of segment boundaries are much longer.

We add a system prompt which describes the segmentation task, desired output format and primes the model to be a talented linguist. We also add a variety of short examples in line with the few shot prompting technique (Brown et al., 2020), which increased performance. This did not necessarily reflect the fact that the GLM learnt how to segment better. Instead, through manual testing, we suspect that it learnt the ideal segment length and amount of information that should be contained within a segment, which was implicitly contained in the few shot prompt (and also the testing datasets, therefore increasing performance). This suggests that different implicit definitions of a segment could be imposed by a few shot prompt to a GLM, dependent upon use case.

[1] `https://www.kaggle.com/datasets/ltcmdrdata/plain-text-wikipedia-202011/data`
[2] `https://ctc.westpoint.edu/ctc-sentinel/`
[3] `https://peacekeeping.un.org/en/reports`

An example prompt is contained in [XXX Appendix A].

2.3.2. Overlapping Prompts

This prompting method works so long as the input text is within the context window of the LLM. At the time of experimentation, and with the use of gpt-3.5-turbo, we had a limit of 16k tokens. Many of our input texts exceeded this limit. Therefore, we needed to split up these long documents into smaller chunks that can be processed by the GLM. However, this cannot be done by simply splitting every at the nearest sentence before every 16k new tokens for two reasons. Firstly, we do not know whether this sentence boundary should serve as a segment boundary, and secondly, the GLM loses valuable context which helps to choose where to place boundaries at the extremes of 16k tokens.

Therefore, we instead send 16k prompts that overlap. The overlap is calculated as twice maximum segment length. In our experiments, we set a maximum segment length of 750 tokens, thus there is an overlap of 1500 tokens between prompts. We must then decide which boundaries to accept in this overlapping region. Given two generations which were prompted by 1500 overlapping tokens, we choose to accept the segment boundaries contained within the first 750 tokens of the overlapping section from the first generation, and the boundaries in the final 750 tokens from the second prompt. We did not experiment with involving the responses from both outputs, but found that there seemed to be no sign of degrading performance towards segment boundaries.

While this method is wasteful of up to 1500 tokens per prompt, this is a small enough fraction of the 16k context that we were satisfied with the solution. If maximum segment lengths were much longer, say 5k tokens, the context may need to be limited more severely.

2.3.3. Segment Validation

We also performed some validation on the segments returned by the GLM. This primarily involved verify that the returned segments are within a maximum and minimum segment length. Segments that are too short (for example, a model would often return just a heading), were concatenated with another segment, and segments that are too long were recursively segmented by the same model, but with another prompt that asks the model to generate only one boundary at a time, which uses a similar few-shot prompting strategy to above. This way, a segment that is too long will be split in 2 recursively until all segments are within the specified lengths.

3. Experiments

3.1. Models

3.1.1. Baselines

We use two naive baselines as a point of reference. First, a segmenter which splits every n sentences. We decided to split every 5 sentences which we call the *Split5Segmenter*. We also define a *RandomF0.1Segmenter* which splits at 10% of boundaries, placed randomly, with each potential boundary equally likely.

3.1.2. BERT Segmenter

Our existing method generates a sequence of sentence similarities using sentence embeddings generated by (Reimers & Gurevych, 2019). Similarities are calculated as a weighted sum of the cosine similarity to the previous n sentences. Ideal boundaries are then generated as troughs in the sequence of similarities, before further processing to ensure there are no segments that are too long or too short, either in sentence length or in token length. We name this the *BERTSegmenter*. Further details are omitted for proprietary reasons.

3.1.3. BERT-Graph Segmenter

(Costacurta, 2023) describes 'text tiling' (topic segmentation) using BERT-generated similarity scores followed by graph clustering to find the best segments. This follows a similar methodology as the previous Exact details of this method can be found in the linked article. This model is called *BERTGraphSegmenter* in our experiments. Code for this model was copied from the repository linked in the article.

3.1.4. GPT-3.5

OpenAI's `gpt-3.5-turbo-16k` was queried using the prompting and segment validation strategy defined in Section 2. We call this model *GPT3.5*. We used the largest model available to us, which is the 16k token context window. Given the nature of the task, we chose to use deterministic outputs from the model, using topk $= 1$, rather than sampling. This makes sense as we are looking for the best possible segmentations, rather than a variety of possible segmentations, and helps with reproducibility of results.

3.1.5. Flan-T5-Finetuned

Our intended application of segmentation requires us to use our own models, therefore we fine-tune Google's Flan-T5 large (Chung et al., 2022) on a combination of wiki, concatenated wiki and synthetic segmentations generated by GPT-3.5. We call this model *FlanT5*. We used the same deterministic output and a similar prompting strategy as with *GPT3.5*. Because the model has been fine-tuned, we use a much shorter instruction section of a prompt with no examples. Additionally, given the model's shorter context window, we have many more overlapping prompt per document.

3.2. Quantitative Results

Models are tested on the human-annotated dataset, the wikipedia dataset, and the concatenated wikipedia dataset. We do not test on the synthetic dataset results would be biased in favor of the generative models which generated the segmentations. Indeed, while performance of the LLMs on the synthetic dataset is far from perfect (less than boundary similarity of 0.5 with $n = 2$), it is disproportionately better than other models on the synthetic dataset.

Due to resource constraints, we could not test *GPT3.5* on the full wikipedia or full concatenated wikipedia datasets. Instead, we took the largest subset that fit within resource constraints. We evaluated all other models on the full datasets to verify that similar results are obtained. Further details on the experimental procedure are witheld for proprietary reasons.

We evaluate primarily using the previously discussed boundary similarity metric with $n = 2$. Results can be seen in Table 1. Evaluation was also run with boundary similarity $n = 5$, but the metric becomes noisier with higher values of n. Indeed, the baselines begin to report better performance with high enough n. Full results for both the smaller and larger datasets and all evalutation metrics can be found in the Appendix A.

We see that *GPT3.5* outperforms all other models on the human-annotated dataset, the wikipedia dataset, and the synthetic dataset, with the other LLM *FlanT5* coming in second in the same datasets. The close performance between the models on the synthetic dataset implies that *FlanT5* is a good approximation of *GPT3.5* for this task, on boundaries generate by *GPT3.5*. However, the biggets performance gap is on the human-annotated dataset. Although this dataset is small, it represents a meaningful distribution shift and, unsurprisingly, the fine-tuned model is unable to generalize as well as the base model.

Interestingly, we see that both BERT-based models are superior in the supposedly easier task posed by the concatenated wikipedia dataset. We theorise that these models are better at finding clear boundaries between different domains, but struggle with more nuanced segment boundaries found in a news article, for example. Whereas, the LLMs are better at finding more nuanced segments that a human might produce, but are not significantly better at finding these easier boundaries.

We also looked at precision and recall to better characterize the behavior of each model. Results can be seen in Table 2. *GPT3.5* has the highest precision and recall in the human-annotated dataset and the highest recall on both the wiki and concatenated wiki datasets. This property of having high recall is true in general for the LLMs, even in the concatenated wiki dataset where the BERT models have higher overall boundary similarity. By contrast, precision scores are closer in general and much better for the BERT models on the concatenated wiki dataset. This suggests that, in general, the BERT models are more hesitant in placing boundaries, but when they do, they are more likely to be correct. It should be noted that this behaviour in both the LLMs and BERT models is subject to the prompt engineering and segment processing procedures used.

	Human	Wiki	Conc-Wiki	Synthetic
GPT3.5	**0.38**	**0.25**	0.29	**0.35**
FlanT5	0.25	0.24	0.41	0.33
BERTGraph	0.20	0.15	0.45	0.21
BERT	0.18	0.09	**0.46**	0.18
RandomF0.1	0.09	0.11	0.10	0.10
Split5	0.13	0.19	0.19	0.23

Table 1. Boundary similarity scores with $n = 2$ for each model and the human, wiki, concatenated wiki and synthetic datasets.

	Human		Wiki		Conc-Wiki	
	BP	BR	BP	BR	BP	BR
GPT3.5	**0.51**	**0.60**	0.36	**0.55**	0.42	**0.63**
FlanT5	0.38	0.46	**0.43**	0.37	0.65	**0.63**
BERTGraph	0.47	0.25	0.39	0.21	0.79	0.54
BERT	0.33	0.39	0.23	0.37	**0.91**	0.50
RandomF0.1	0.16	0.21	0.34	0.16	0.20	0.16
Split5	0.17	0.48	0.33	0.39	0.25	0.52

Table 2. Boundary Precision (BP) and Recall (BR) with $n = 2$ for each model and the human, wiki, and concatenated wiki datasets.

3.3. Qualitative Evaluation

Limited manual testing and inspection of the segmentations produced by the models was also conducted, both during prompt engineering and after the models were trained.

Through manual inspection of segmentations with the human-annotated dataset, we found that *GPT3.5* generally found the boundaries which seemed reasonable from a human perspective, especially for simple documents like short news articles. The fine-tuned *FlanT5* model imitated this behavior, but was less consistent. BERT segmenters would find reasonable segments, but after the manual gluing and splitting procedure, would often lead to off-by-1 errors.

For documents with far more complex documents such as a podcast transcript, or with messy data like tables or artefacts from pdf to text conversions, *GPT3.5* would sometimes return indices with a regular pattern. For example, the LLM might return '$[1, 15, 22, \ldots, 76, 79, 82, 85, 88, \ldots, 184, 187, \ldots]$'. Often, the pattern would continue far beyond the number of sentences in the input indicating that the model became stuck in a regular pattern. Perhaps better prompt engineering, a more rigorous data-processing procedure, the use of newer models would help or the use of better generation parameters, but our current approach was resource constrained and required the ability to pass noisy documents to the model. A more thorough investigation of the logits computed by the model is required to understand how and when this occurs, and how to mitigate it.

4. Conclusion

Our work empirically compares generative LLMs with previous methods which use BERT embeddings and cosine similarity. In order to be more token-efficient and to provide guarantees that the original document will be unedited, we propose a new overlapping prompt schema, which centres on asking the LLM to return a list of indices corresponding to segment boundaries. We also support the use of boundary similarity and its associated information recall metrics as an evaluation for topic segmentation. Results indicate that LLMs can be more effective segmenters than existing methods where more nuanced segmentations are required, but that when the input is noisy or the segment boundaries are clear, BERT-based methods are more reliable. Future work should focus on addressing highlighted issues with LLMs, such as the regular patterns found in segmentations, and on comparing models on larger human-annotated datasets to better assess generalisation capabilities.

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation, 2023. URL https://arxiv.org/abs/2306.15595.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

Costacurta, M. Text tiling done right: Building solid foundations for your personal llm, 2023.

Fournier, C. Evaluating text segmentation using boundary edit distance. In Schuetze, H., Fung, P., and Poesio, M. (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1702–1712, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1167.

Hearst, M. A. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1): 33–64, mar 1997. ISSN 0891-2017.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts, 2023.

Petroni, F., Lewis, P., Piktus, A., Rocktäschel, T., Wu, Y., Miller, A. H., and Riedel, S. How context affects language models' factual predictions, 2020. URL https://arxiv.org/abs/2005.04611.

Pevzner, L. and Hearst, M. A. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002. doi: 10.1162/089120102317341756. URL https://aclanthology.org/J02-1002.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Xing, L. *Versatile neural approaches to more accurate and robust topic segmentation*. PhD thesis, University of British Columbia, 2024. URL https://open.library.ubc.ca/collections/ubctheses/24/items/1.0440128.

A. Appendix