Automated YouTube Chapter Creation Using Large Language Models: Enhancing User Experience through Transcript-Based Segmentation

Emanuele Sala, Fabio Stefana, Luca Soleri, Ossama Tchina
Group 15

Abstract

The paper focuses on **automating the division of YouTube videos into chapters** based on video transcripts. For this task there isn't a predominant approach in literature, most papers use either topic models or a combination of computer vision and NLP to achieve this goal. Currently, YouTube's chaptering feature requires manual input from uploaders, who often neglect this step, resulting in a fragmented user experience. Our system aims to automate chapter creation to enhance user engagement and convenience. We implemented a method that uses **LLMs** to analyze video transcripts, segment them into chapters, and generate titles, comparing the results with manually created chapters. The results indicate that, while the machine-generated chapters had clearer and more descriptive titles, there remains room for further refinement, promising a more streamlined and user-friendly experience on YouTube in the future.

1 Introduction

YouTube has become a dominant platform for consuming video content. An important aspect of the user experience is the ability to easily browse videos and find specific information. One approach to solving this challenge is to use chapters to allow viewers to navigate to relevant parts of the video. Although YouTube offers chapter creation, it relies on manual input from content creators. Not only do many creators skip this step, but this also means that older videos will rarely benefit from this feature. This can lead to a fragmented user experience for viewers searching for specific content within the video.

This paper explores the potential of **automating chapter generation using natural language processing**. Automatic chapter generation offers several advantages: first, it can reduce the burden on content creators by eliminating the need to create chapters manually. Second, it can ensure the consistency of segments in a video, thereby improving the overall user experience for viewers. Finally, automated chapter generation has the potential to deliver more informative and descriptive chapters, further enhancing value for viewers.

There is no single dominant approach for automatic video chapter generation in the existing literature. Some approaches leverage topic modeling to segment video content based on changes in topic, however, based on our own experiments, this kind of models has lackluster performance. Other approaches combine computer vision, to segment the video into snippets, and NLP (generally language models) to create a fitting title for each snippet.

This paper investigates an approach that uses **large language models** for both the segmentation and the creation of titles. We start from the assumption that the understating of text that LLMs achieve allows them to detect thematic changes in content much more precisely than topic models do, and to generate clearer and more informative chapter titles. An advantage of our model compared to others that use computer vision is the simplification of the whole process, and less stringent requirements on the input data. In fact, **YouTube automatically creates transcripts** for most videos: this avoids the need for additional data processing steps. The main contributions of our work therefore are:

- reducing the mole of data required (video data is not needed) ;
- greater accuracy than simple topic models;
- ease of implementation, because transcripts are automatically generated.

We evaluate the effectiveness of our approach by **comparing automatically generated chapters with manually generated chapters from a benchmark dataset**. Our evaluation looks at two main metrics: the **number of chapters** identified, and the **semantic similarity** between human-generated and machine-generated chapters.

The results of our study indicate that the LLM-based approach yields promising results, offering a significant improvement from our naive model (which we will discuss later). Although there are some differences between chapter numbers determined by LLM and human annotators, LLM chapters generally have more descriptive and precise titles, providing a clearer overview of the video's content and its structure. This enhanced mapping of content not only helps users understand what to expect in each chapter, but it also makes navigation more intuitive and efficient, potentially improving the user experience for viewers browsing the videos.

2 Experiments

The basis for our work was a paper by Yang et al. (2023), where he attempted the same task with a different methodology that also included computer vision. He provided a dataset including around **800k video IDs** and their respective **human made chapters**. From this dataset, we extrapolated a sample of around 20k videos IDs and chapters, we then scraped the transcripts from each video (using the YouTube Transcripts API).

Initially we tried topic modeling on the single transcript, we used various models from an LDA to BERTopic, however since one transcript alone is very short, these methodologies did not work at all, and often only identified one topic per transcript, so we scratched the idea completely.

After a lot of trial and error we landed on 2 methodologies, the first we will call the "**Fixed Chapters**", which is the naive methodology that we will use as a baseline, and the second one we will call "**Llama Chapters**" which uses an LLM.

2.1 Fixed Chapters

The first methodology **divided each video transcript into 8 equal segments**, matching the average number of human-made chapters in the dataset; next, we took each segment and **summarized** it using a pre-trained BART language model[1] (Lewis et al., 2019) to create a short title for each one. The Llama Chapters methodology relied on the **Llama LLM**[2] (Touvron et al., 2023) to understand the video's content, divide it into chapters based on topics, and generate coherent titles that clearly explain each chapter.

2.2 LLama Chapters

We initially used meta-llama/Llama-2-7b-chat-hf for its ease of use. However, since this chat model produced varied responses, we switched to **meta-llama/Meta-Llama-3-8B-Instruct**. This instruct version provided consistent output and it is a newer, better-performing model in the small Llama family. The chat version of a Large Language Model is designed to provide varied responses each time, making interactions feel more natural and lifelike. In contrast, the instruct version of a model is engineered to follow precise and detailed instructions, ensuring consistency and adherence to specific guidelines.

While for the Fixed Chapters model we were able to utilize all of the 20k videos, with the LLama Chapter methodology we were only able, due to computational difficulties, to perform our analysis only on about **8500 videos**, because running the LLama LLM took up a lot of resources in the student HPC cluster and was particularly slow.

2.3 Comparison

We obtained **2 different chapter divisions**, then we evaluated them based on 2 metrics:

- the difference between the **number of chapters** identified by the model and the number of human-made chapters;
- the average of the **cosine similarities** between each corresponding human-made and machine-made chapters.

To perform a pairwise comparison when the number of chapters differed between the human-made and machine-made divisions, we introduced empty strings to the shorter division to equalize the total number of chapters. We then considered **all possible permutations** of the combination of original chapters and empty ones. For example, with two original chapters (represented as '1') and one added empty chapter (represented as '0'), the possible permutations would be: 110, 101, and 011. We compared each permutation with the benchmark, assigning a similarity score of 0 to pairs where one of the strings was empty and kept the highest score, as we assumed the highest similarity score from

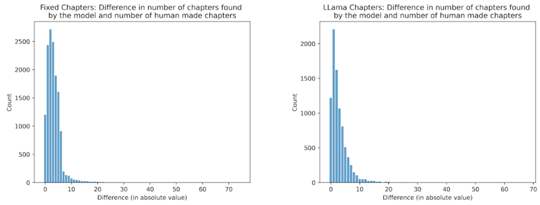
[1] https://huggingface.co/facebook/bart-large.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

these comparisons would be the correct one, indicating the optimal pairing between chapters.
Due to the computational difficulty of comparing all the possible permutations when the total number of chapters increased, we restricted our comparison only to chapter divisions where the difference in the number of chapters was less or equal to 4, resulting in 80% of the results being considered for both LLama Chapters and Fixed Chapters.

3 Results

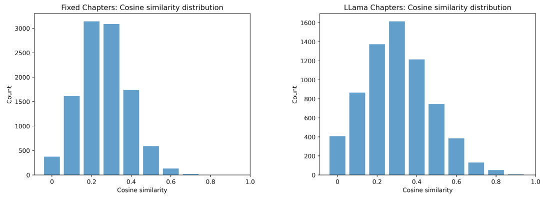
With the first metric we found that **on average the LLama Chapters were off by 2.89 chapters, while the Fixed Chapters were off by 3.3**. Moreover, the median of Llama is 2, while the median of Fixed is 3. Their distributions were also different, as can be seen in Figure 1: the Llama chapters distribution exhibits a more concentrated pattern around 0 and is less tail-heavy.



(a) Fixed chapters (b) LLama chapters

Figure 1: Distribution of the difference in number of chapters

Regarding the second metric, we found that **LLama chapters boasts a better cosine similarity** than the Fixed Chapters, with the first getting an average cosine similarity score of around 0.35, and the second one of around 0.29.
By looking at the cosine similarity distribution we can also see that that of Llama chapters is more skewed to the right with both a higher median and mean. (Figure 2)

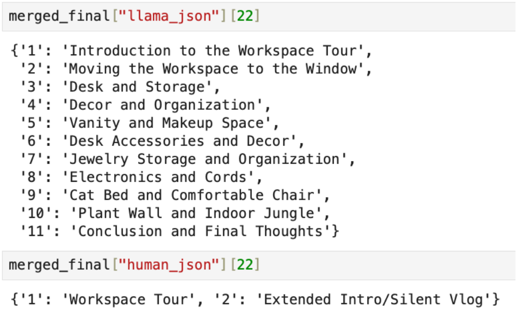


(a) Fixed Chapters (b) LLama Chapters

Figure 2: Distribution of the difference in number of chapters

4 Discussion

We observed that, in some cases, there was a big discrepancy between the number of chapters recognized by LLama and the one by the human annotators.
A quick **visual inspection** of those cases, comparing the results with the original video, proved that the difference was often due to different philosophies in the division of the content, where LLama was able to pick up the structural division of the video but not its objective. An important category where significant discrepancies were often observed was that of home tours, where LLama often was not able to recognize the separation between different rooms. For this kind of problem, an approach such as that of Yang might be better, where visual cues are employed to aid in the separation between different segments.
Moreover, part of the **difference between human-and machine-generated annotations was due to the more descriptive nature of the annotations created by LLama**, compared to the more schematic human responses.
What we see in Figure 3 is an example where the Llama Chapters seem to not perform well, if we only consider the metrics we discussed before; in this example LLama was very specific and detailed both in the segmentation of the chapters and the titles given, while the human annotated ones were more general. In this case the difference in length is 9 and the cosine similarity score is 0.57.



```
merged_final["llama_json"][22]
```

```
{'1': 'Introduction to the Workspace Tour',
 '2': 'Moving the Workspace to the Window',
 '3': 'Desk and Storage',
 '4': 'Decor and Organization',
 '5': 'Vanity and Makeup Space',
 '6': 'Desk Accessories and Decor',
 '7': 'Jewelry Storage and Organization',
 '8': 'Electronics and Cords',
 '9': 'Cat Bed and Comfortable Chair',
 '10': 'Plant Wall and Indoor Jungle',
 '11': 'Conclusion and Final Thoughts'}
```

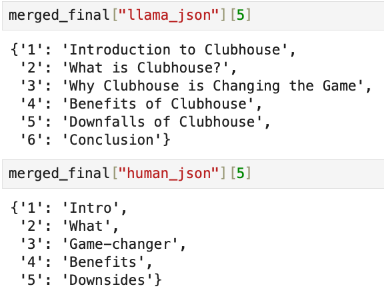
```
merged_final["human_json"][22]
```

```
{'1': 'Workspace Tour', '2': 'Extended Intro/Silent Vlog'}
```

Figure 3: Example where it didn’t work well

Instead in Figure 4, we see an example where LLama chapters worked extremely well; here the difference in length is 1, but the cosine similarity is 0.57 once again.

It is essential to note that relying solely on metrics like cosine similarity may not offer a comprehensive understanding. Indeed, while metrics may suggest one model outperforms the other, they



```
merged_final["llama_json"][5]

{'1': 'Introduction to Clubhouse',
 '2': 'What is Clubhouse?',
 '3': 'Why Clubhouse is Changing the Game',
 '4': 'Benefits of Clubhouse',
 '5': 'Downfalls of Clubhouse',
 '6': 'Conclusion'}

merged_final["human_json"][5]

{'1': 'Intro',
 '2': 'What',
 '3': 'Game-changer',
 '4': 'Benefits',
 '5': 'Downsides'}
```

Figure 4: Example where it worked well

often fail to encapsulate the entirety of the task's complexity. This task lacks determinism, and there exists **no single correct solution**. Something that we could improve in the future is the quality of the human annotated chapters: if we were to hand annotate them with a rigorous standard of quality and precision, the performance metrics would show a greater improvement using the LLama model compared to what we have seen so far.

It's important to recognize that **cosine similarity serves as a valuable metric for distinguishing between poor and good chapter divisions**. However, when comparing high-quality divisions that differ due to varying levels of segmentation detail and title specificity, cosine similarity may not effectively capture these nuances. In such cases, human evaluation becomes indispensable for accurately assessing the quality of LLama chapters. Only through human judgment can we truly evaluate the effectiveness and appropriateness of the LLama model's output in these nuanced scenarios.

5 Related Works

Researchers have made significant advancements in automatic segmentation and tagging of YouTube videos. Morchid and Linarès (2013) used Latent Dirichlet Allocation (LDA) for automatic tagging by extracting keywords from transcripts to map video content into a topic space. Yu-Jin Ha1 (2023) employed LDA for topic-aware video summarization, segmenting videos into coherent parts based on transcript-derived topics.Vybhavi et al. (2022) proposed a model to identify key segments and summarize video content using transcript analysis and machine learning techniques. Liu et al. (2022) explored deep learning approaches for automatic video summarization by analyzing visual and auditory data alongside transcripts. Porwal et al. (2022) developed a transcript summarization application that divides video and audio into chunks and uses extractive text summarization to generate coherent summaries.

Our work differs from the aforementioned ones by focusing on automating YouTube chapter creation using LLMs. To the best of our knowledge, we uniquely integrate LLMs for automatic YouTube chapter creation and evaluate it against human-made benchmarks. This approach fills a gap by offering an automated solution for content creators who lack the time or resources to manually create chapters, improving the overall user experience on the platform.

6 Conclusions

Our work investigates the effectiveness of applying **Large Language Models** to automate the identification of chapters in YouTube. This is a novel approach to the subject, where topic modeling is mainly used for our objective, and it allows us to take advantage of the contextual understanding that LLMs have. Moreover, compared to other approaches relying on large amounts of information about the video, ours **only requires the transcript of a video**, which, on YouTube, is automatically generated. Furthermore, it allows us to have meaningful and informative titles, often even better than the human-generated ones at that. The comparison with the fixed chapters naive approach shows better results, both in terms of number of chapters, and, more importantly, in terms of cosine similarity between the chapters. However, the model also shows some **areas for improvement**, namely, non-textual cues might be incorporated for specific categories of videos to improve the quality of predictions, and optimization might help to make the model run faster and more efficiently. Moreover, evaluation of the chapters that is not only based on the comparison with a human-generated benchmark might give important insights into the areas where the model does not perform well and give a useful map for further corrections. Finally, future efforts should be aimed at determining with precision the boundaries between chapters and transforming them into time stamps, to enable the creation of clickable links to each segment.

References

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. 2022. Umt: Unified multimodal transformers for joint video moment retrieval and highlight detection.

Mohamed Morchid and Georges Linarès. 2013. A lda-based method for automatic tagging of youtube videos. In *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4.

Porwal, Harshit Srivastava, Ritik Gupta, ShiveshPratap Mall, and Nidhi Gupta. 2022. Video transcription and summarization using nlp.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Atluri Naga Sai Sri Vybhavi, Laggisetti Valli Saroja, Jahnavi Duvvuru, and Jayanag Bayana. 2022. Video transcript summarizer. In *2022 International Mobile and Embedded Technology Conference (MECON)*, pages 461–465.

Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vidchapters-7m: Video chapters at scale. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gun-Woo Kim Yu-Jin Ha1. 2023. Topic-aware video summarization technique for product reviews exploiting the bertopic and bart models.