應用大語言模型的提示詞工程於影音內容重點提取
(Application of Large Language Model-Based Prompt Engineering for Key Information Extraction from Audio-Visual Content)

Wei Ting Huang, Yu-Chen Liu and Ming-Hsiang Su

Department of Data Science, Soochow University, Taiwan
{bindy.huang, vincent93113, huntfox.su}@gmail.com

摘要

當今科技發達的時代，傳統傳播媒體已逐漸被網路影音所取代，而短影音更成為行銷、傳播資訊的重要媒介。爲協助影音創作者更有效率地將長影片製作成短影音以增加曝光度，本研究利用 BERT 模型探索大型語言模型對於篩選文本精華之能力。首先以人工抓取一百部影片字幕文本之精華，接著運用 GPT-4o 執行 4 種不同指令，蒐集 4 種不同的重要資訊篩選結果，最終將 4 種指令生成與人工篩選之文本進行相似度對比分析，歸納出最準確指令之特點。實驗結果發現給予大型語言模型較為明確地指令的確可以篩選出長影片的重要片段，使創作者能更有效率地將製作好的長影片剪輯為六十秒左右的短影音。

Abstract

In the era of advanced technology, traditional media has gradually been replaced by online video content, with short videos becoming a crucial medium for marketing and information dissemination. To assist content creators in efficiently transforming long videos into short clips to increase exposure, this study leverages the BERT deep learning model to explore the capability of large language models in extracting essential information from text. Initially, we manually curated key excerpts from the transcripts of over a hundred videos. Following this, we utilized GPT-4 to execute four distinct commands, collecting four different sets of important information. Finally, we performed a cosine similarity analysis between the texts generated by the four commands and the manually extracted excerpts, identifying the characteristics of the most accurate command. The experimental results revealed that providing large language models with more specific instructions enhances their ability to identify essential segments of long videos. This enables content creators to edit long-form videos more efficiently into short clips of approximately sixty seconds.

關鍵字：大型語言模型、BERT、短影音
Keywords: LLM, BERT, short video

1 緒論

在現今數位世代，網路已經深刻改變人們的生活習慣。根據 Lyndsi Stafford (2017)在富比士雜誌官網所述，人們觀看影片時能記住95%的訊息，而閱讀文字內容時只能記住 10%。由此可見，影片作為一種生動的資訊呈現方式，不僅在網路媒體中佔據著重要地位，也強化了資訊的傳遞效果。 自 2016 年 TikTok 上市後，短影音變得深受眾人喜愛。而 YouTube 和 Meta 為了因應短影音時代的趨勢，也在近兩年推出短影音服務。

短影音的流行使得觀眾的注意力變短，相較以往更沒耐心觀看完整長 影片。因此，近期許多創作者選擇在上傳長影片的同時，積極發佈短影音。透過短影音的快速傳播，吸引更多觀眾對長影片內容產生興趣，進而 點擊長影片，觀看完整內容。 然而，將長影片轉換為短影音仍然耗時費力。因此本研究將利用 GPT4o 模型挑選 YouTube 影片的重要片段，測試四種不同指令的挑選結果，並與人

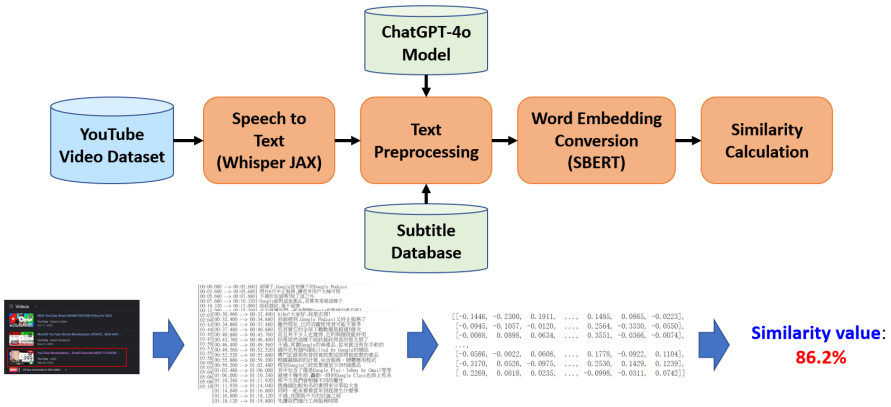

圖 1：研究方法流程圖

工選擇之片段進行比較，以檢測大型語言模型是否能自動剪輯出合適的短影音。

2 文獻回顧

2.1 短影音

短影音(short video)又稱為短影片、短視頻。目前在學術領域的定義仍有一些模糊之處，不過通常被定義為數十秒且豎頻形式呈現的影片。Fahao Chen 等人 (2023)認為短影音的影片時長通常小於一分鐘。另外，戴景麗 (2013)也將短影音定義為一分鐘內的短片。雖然目前對於短影音尚無明確統一的定義，不過考慮到本研究目的為吸引觀眾觀看完整長影片內容，因此本研究將製作目標之短影音定義為五十至六十秒的影片，以貼近實質觀看體驗。

2.2 大型語言模型(LLM)

由亞馬遜服務公司的《什麼是大型語言模型(LLM)？》一文可得知大型語言模型亦即 Large Language Model (LLM)，是一種大型的深度學習模型，可以用於生成式 AI，根據使用者提問回答相對應的內容。而 Csaba Veres (2022)指出大型神經語言模型在學習單字時是以連續向量表示機率函數，而非離散的詞彙，這有助於語言模型能學習更精確地語意表達模式，提供更正確的回應。由此推測，透過大型語言模型，應能自動判斷擷取，進而篩選出影片中的重點片段。

Raisa Islam 和 Owana Marzia Moushi (2024) 表示，GPT-4o 是目前 OpenAI 最新推出的模型，此模型相較於先前幾個版本的 ChatGPT 更精準，且效率也更高。且 GPT-4o 模型可以處理更廣泛知識的互動式問答，並引入了記憶功能，能記得使用者給予的特定細節。而 GPT-4o 模型也在 DROP 閱讀理解基準測試中拿到了 83.4%的分數，顯示出該模型應能判斷出長影片的重點片段，因此本研究使用 GPT-4o 作為篩選片段之語言模型。

2.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) 是一種深度學習模型，根據 Koroteev M.V. (2021) 所述，BERT 模型在自動化文字處理展現了極高的精確度，該模型會使用向量的方式表示文本中的每個單字，以進行文字分類。該模型旨在對雙向文字表示進行深度初步學習，以便隨後在機器學習模型中使用。而相較於其他模型，BERT 更易於使用，只需在現有的神經架構中添加一個輸出層即可獲得在準確處理文本。

Aubrey Condor 等人(2021)表示儘管可以使用平均 BERT 輸出層從原始 BERT 模型導出句子向量，但此類方法會產生較差的結果。相較之下，Sentence-BERT (SBERT)所產生的句子向量在 SentEval 基準上優於其他最先進的方法，因此本研究選用以 BERT 為基礎設計的 SBERT 為文章轉向量之工具。

3 研究方法

本研究先將 YouTube 影片之語音內容轉換為文字檔，接著使用 Python 程式語言撰寫，利用 SBERT 模型將影片字幕檔之文字轉換為向量，再計算其餘弦相似度以比較不同提示詞(Prompt)之間的差異，過程如圖 1 所示。

本研究蒐集一百部 YouTube 之中文影片，透過 Whisper JAX (2024) 語音轉文字網站，將影

片中的語音轉換成文字。接著本研究以人工的方式進行資料預處理，將辨識錯誤的字詞更正，再自行建立字幕段落資料庫，用以分析大型語言模型篩選重要段落之能力。

接著使用 Python 程式語言撰寫程式碼，以 SBERT 模型將字幕文本轉換為文字向量。SBERT 是一個強大的語句嵌入工具，能夠將文本轉換成高維度向量，可用以比較不同文本之間的相似性或進行文本分類。最終本研究計算大型語言模型篩選出之重要文本段落與人工篩選之重要文本段落的餘弦相似度，以找出大型語言模型篩選重要段落之最佳提示詞樣板。

3.1 重要段落擷取

為分析大型語言模型篩選重要段落之能力，本研究除利用人工方式將字幕段落資料庫中之段落文本資料擷取出重要段落文本資料，此外，本研究亦將相同資料餵入(feed into) ChatGPT-4o 大語言模型中，利用本研究設計之四種提示詞用以篩選重要段落，如圖 2 所示。本研究使用之四種提示詞說明如下。

1. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出一段可以剪輯 60 秒短影音的段落。
2. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出 5 個段落，最終可以合在一起剪輯出 60 秒短影音的段落。
3. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這份字幕檔的內容挑選出多個段落，最終可以合在一起剪輯出 60 秒（根據字幕檔的時間格式計算，相加後要介於 50~60 秒之間）短影音的段落。
4. 你是一位專業的影片剪輯師。這份檔案是一部影片的字幕檔，請根據這一整份字幕檔的內容挑選出多個段落。最終剪輯出 60 秒（根據字幕檔的時間格式計算，相加後要介於 50~60 秒之間）、包含影片重點的短影音，讓觀眾可以透過這支短影音就了解原本影片的重點。

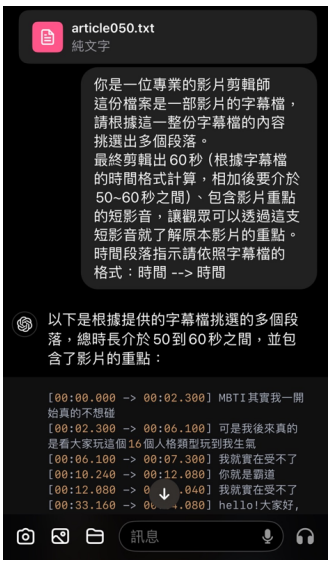

圖 2：ChatGPT-4o 對話示意圖

4 實驗結果與討論

4.1 資料集

本研究蒐集了一百部片長介於五到十分鐘之間的 YouTube 中文影片字幕檔，而影片主題類型包括電影解說、旅遊分享、美妝保養等各式內容。文本蒐集方式是利用以 OpernAI 所開發之 Whisper 製作的 Whisper JAX (2024) 語音轉文字網站讀取 YouTube 影片網址，使其自動將語音內容轉換為文字檔。為確保文字檔內容均與影片內容完全相符，透過網站將語音轉為文字後，再以人工方式進行核對，以確保文字檔與原始影片完全一致。

4.2 分析方式

本研究以 Python 3 環境撰寫程式碼，將重要段落文本資料透過 12 層 MiniLM v2 架構之 SBERT 模型轉換為向量。取得向量後再分析人工擷取之重要段落與 ChatGPT-4o 擷取之重要段落的餘弦相似度，並計算一百個文本的平均值與標準差，判斷四種不同提示詞所擷取之重要段落與人類篩選之重要段落最為相似。

實驗結果顯示，提示詞 4 的平均值約為 0.7583，是四者當中最高的，且標準差為 0.1825，相較其餘三種提示詞低，顯示提示詞 4 所擷取之重要段落最近似於人工擷取之重要段落。我們認為提示詞 4 相較其他三種指令給予的需求更為明確，包括秒數限制、必須出現影片重點，以及剪輯該短影音的目的。

表 1：不同提示詞與人工擷取重要段落相似詞評估

	Prompt 1	Prompt 2	Prompt 3	Prompt 4
平均值	0.6817	0.6786	0.7180	**0.7583**
標準差	0.2005	0.1931	0.1890	**0.1825**

5 結論與未來展望

在短影音時代的浪潮下，長影片創作者正面臨前所未有的挑戰，如何在有限時間內吸引觀眾注意力成為了核心問題。大型語言模型的生成技術為創作者提供了一個強大的工具，而如何有效利用這些工具則成為了一個關鍵課題。本研究探索了如何透過不同的指令來優化生成腳本的過程，以提升創作者的效率與作品品質。

我們的研究重點在於通過實驗驗證，不同的 prompt 如何影響影片文本的選取及生成短影音腳本的效果。我們首先以 4 種不同的指令來進行實驗，將影片文本上傳至 GPT-4o 模型，生成對應的 4 種短影音腳本。接著，我們以人工方式閱讀上百篇影片文本，挑選出適合節錄至短影音的片段，並使用 Bert模型對生成的腳本進行相似度分析。通過對比平均數、標準差等數據，我們能夠找到最適合的指令，為創作者提供有力的參考依據。

本研究的未來規劃是為創作者提供一個系統化的方法，以最優化的指令來生成高質量的短影音腳本。我們期望透過這些實驗結果，能夠歸納出一套適用於各種影視及自媒體產業的公式模型，進一步推廣這些模型。再來，我們也計劃在未來研究自動剪輯的技術，使用人工智慧辨識及篩選長影片中之精華片段後，自動編排和剪輯。結合深度學習和情感分析技術，理解影片內容及觀眾偏好，更加靈活的創造出符合受眾的短影片，提高創作效率及市場競爭力。

References

Lyndsi Stafford. 2017. How to incorporate video into your social media strategy. Obtained from: https://www.forbes.com/sites/yec/2017/07/13/how-to-incorporatevideo-into-your-social-media-strategy. Access Date: 2024/07/15.

Fahao Chen, Peng Li, Deze Zeng and Song Guo. 2023. Edge-Assisted Short Video Sharing With Guaranteed Quality-of-Experience. *Journal of IEEE Transactions on Cloud Computing*, 11(1):13-24. https://doi.org/10.1109/TCC.2021.3067834.

戴景麗. 2013. *微視頻的內容定位與盈利模式*. PhD Thesis. 上海師範大學。

亞馬遜網路服務公司. 什麼是大型語言模型(LLM)?. Obtained from: https://aws.amazon.com/tw/what-is/large-language-model/. Access Date: 2024/07/15.

Csaba Veres. 2022. Large Language Models are Not Models of Natural Language: They are Corpus Models. *Journal of IEEE Access*. 10:61970-61979. https://doi.org/10.1109/ACCESS.2022.3182505.

Raisa Islam, Owana Marzia Moushi. 2024. GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. *Authorea Preprints*.

Koroteev M.V. 2021. BERT: A Review of Applications in Natural Language Processing and Understanding. arXiv preprint arXiv:2103.11943. https://doi.org/10.48550/arXiv.2103.11943.

Aubrey Condor, Max Litster, Zachary Pardos. 2021. Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of The 14th International Conference on Educational Data Mining*, pages 345-452. https://eric.ed.gov/?id=ED615495.

Whisper JAX: The Fastest Whisper API. Obtained from: https://huggingface.co/spaces/sanchit-gandhi/whisper-jax. Access date: 2024/07/15.