

# GPT-4を用いた配信動画中のQ&A切り抜きを目指して

岡崎 光栄<sup>1,a)</sup> 刀祢 賢志<sup>1,b)</sup> 与那嶺 あきお<sup>1,c)</sup> 北原 鉄朗<sup>1,d)</sup>

**概要:** 本研究では、GPT-4を用いて配信動画から質問とそれに対応する回答を自動的に切り抜き、要約するシステムを開発することを目指す。YouTubeなどのプラットフォームで需要が高まっている切り抜き動画を自動で生成対応するため、動画に対して音声認識技術を使って書き起こしを行い、GPT-4を用いて重要文を抽出した上で、その箇所を抜いた動画を生成する予備実験の結果GPT-4による重要文抽出がうまくいかない例や、抽出された重要文が元のテキストから変化するために最長一致文字列探索がうまく働かない例が示された。

## 1. はじめに

近年、YouTubeなどの動画共有サービスにおける切り抜き動画の需要が急速に高まっている。動画の切り抜きとは、動画から重要な部分を抽出し、それらを繋げて再構築、要約する過程である。この過程では、編集ソフトウェアの操作技術を必要とし、特定の開始点と終了点を見つけるために動画のフレームごとに内容を確認する必要があるため、時間がかかる作業となっている。

本研究の目的は、動画の切り抜きの過程を自動化し、時間を要する編集作業を削減することである。特に、動画編集における専門技術を必要とせず、誰でも簡単に動画を切り抜き、再構築できるシステムの開発を目指す。

動画編集と要約の分野においては、GPT-3や他の技術を用いた研究が存在する。Maazら[1]は、ビデオデータ（動画の会話の内容を自然言語処理による文字起こしと、画像フレームエンコーダによって得られた動画の特徴）とビデオの注釈（人による半自動で作った説明）を組み合わせることで、ビデオに関連する詳細な会話や質問に答える形でのGPT-3でのテキスト生成を出力するシステムを開発している。Sainilら[2]は、深層学習の派生（GPT-4ではない）を用いてビデオデータを入力すると、要約された動画を出力するシステムを実現した。それ以外にも、動画をテキストとして要約する研究[3]、サッカーなどのスポーツの動画を要約する研究[4]も存在するが、単独話者によるトーク動画に対する切り抜き動画の生成はほとんど行われ

ていない。

本研究では、GPT-4を用いた動画切り抜きシステムを試作する。このシステムは、音声認識技術を用いて動画の音声を文字に起こしたテキストと、切り抜いてほしい場所の指定テキストをGPT-4にプロンプトとして与えることで、動画から自動的に意図する部分を識別し、切り抜き動画を作成する。これにより、動画編集の効率化を図り、専門的な編集技術がなくても意図的に動画要約を生成することが可能になる。

## 2. システム概要

この研究で開発されるシステムは、配信動画とプロンプト言語を入力として使用する。まず、自然言語処理技術を用いて、配信動画中の発話をテキストに変換する。次に、このテキストから配信動画内の話者が視聴者の質問を読み上げると、それに対する回答がある動画の部分だけを取り出す。抽出するために、GPT-4にプロンプト言語を与える。GPT-4が抽出した質問と回答の文章から、それぞれの先頭と末尾のインデックスを取得し、この情報を使用して動画内のタイムスタンプを探索、特定する。これにより、動画から特定の部分を切り出すことが可能になる。以下の項では処理の流れを述べる。

### 2.1 配信動画の読み込み

このシステムでは、1時間以上の長さを持つ配信動画を扱う。まず、yt-dlpライブラリを使用して、YouTubeから動画をMP4形式でダウンロードする。次に、Googleコンソールの仕様に基づいて、ダウンロードしたMP4ファイルから音声のみを抽出し、FLACファイル形式に変換する。この変換された音声ファイルは、Google Cloud Storage (GCS)

<sup>1</sup> 日本大学文理学部情報科学科

a) okazaki@kthrlab.jp

b) tone@kthrlab.jp

c) yonamine@kthrlab.jp

d) kitahara@kthrlab.jp

に URI として自動的に保存され、入力として使用される。

## 2.2 Speech-to-Text を用いた音声書き起こし

前節で得た URI を GCS から呼び出し、Google の Speech-to-Text API を用いて、動画内の発話が全てテキストに変換され、それぞれの単語にタイムスタンプが付与される。Google の Speech-to-Text API では、各認識された単語やフレーズに対して「信頼度スコア」が割り当てられる。この信頼度スコアは、API がその単語またはフレーズを正確に認識したと判断する確信度を数値化したものである。信頼度スコアの計算方法について詳細な情報は Google から直接提供されていないため、その正確な計算方法は一般には公開されていない。ただし、基本的な原則については理解できる。以下では信頼度スコアについて分かっている事を述べる。

- 信頼度スコアは通常、0 から 1 の間の値で表される。0 は最低の信頼度を、1 は最高の信頼度を意味する。
- 音声の明瞭さ、話者のアクセント、背景ノイズのレベル、話された言葉の文脈など、さまざまな要因が考慮される可能性がある。
- Google の API は高度な機械学習アルゴリズムを使用している。これらのアルゴリズムは大量のデータから学習し、音声データの特徴とパターンを識別して信頼度スコアを計算している。
- 信頼度スコアは、システムが新しいデータを学習するにつれて、継続的に調整される。

## 2.3 GPT-4 による質問文と回答文の抽出

OpenAI API の GPT-4 を用いて、プロンプト言語 (日本語の指示) を入力する。これにより、文章を質問とそれに対応する答えの文章を抜き出す。GPT-4 (Generative Pre-trained Transformer 4) は、OpenAI によって開発された最先端の自然言語処理 (NLP) モデルである。GPT-4 は GPT-3 の後継モデルであり、以下のような特徴を持っている。

- 広範囲で多様なデータセットを用いて訓練されている。これにより、より正確で包括的な言語理解と生成能力を持っている。
- 言語理解とテキスト生成において、GPT-3 よりも高い精度を示している。
- GPT-4 は、テキスト生成、翻訳、要約、質問応答、会話システムなど、幅広い応用が可能である。
- 長い文脈を扱う能力が高く、複雑な会話や文書の理解が可能になっている。
- GPT-4 は、異なる種類のテキスト入力に対して柔軟に対応し、様々なスタイルやトピックに適応できる。

与えたプロンプト文を **図 1** に示す。[6] より、プロンプト文をより精度の高いものにするために GPT-4 のプラグ

```
chat_response1=chat_instance1([
    HumanMessage(content=f"%元の文章:{Sentence}%",),
    SystemMessage(content="")
    #目的: 元の文章から質問とそれに対応する答えの両方が含まれているテキストの全てのセクションを特定し、抽出していただけますか?#
    -
    制約 1: 主な焦点は元のテキストにあるべきです。
    制約 2: レスポンスがプロンプトテキストの (元の文字体系を維持する)++ようにしてください。
    制約 3: 元の文章に複数の質問と答えがある場合は、それら全てを元の文章の時系列順にレスポンスに含めてください。
    制約 4: 質問:[ ' テキスト' ] として、それに対応する答え:[ ' テキスト' ] としてフォーマットしてください。
    制約 5: 絶対に余計な説明文は含めないでください。
    -""
    ),
    1)
```

図 1 今回使用したプロンプト

イン PromptPerfect を用いた。このプラグインでは、与えたプロンプト文章の最適化を行った結果を返すことができる。それにより、質問とそれに対応する答えの文章を抜き出すプロンプト文章の最適化を行っている。それ以外に、プロンプト文の目的と制約条件を記号で区切ることで意味の区切りを分かりやすくしている。

本研究では、さらに、GPT-4 を使用しやすくするために、[5] の langchain ライブラリによって openAI API の複雑なプロンプトフォーマットを簡略化している。

## 2.4 不必要な GPT-4 の返信テキストの削除

GPT-4 からレスポンスで返される文章には、プロンプト言語の指定で、質問文の前に「質問:」と答えの文章に「答え:」がつくようになっている。Python の文字列 (String) クラスに属している replace 関数を用いて「質問:」、「答え:」を取り除く。

## 2.5 テキストのひらがな変換による一貫性の維持

テキストをひらがなに変換し、その一貫性を保つ方法について説明する。ひらがなへの変換は、テキストの処理を簡素化し、特にテキスト探索において重要な役割を果たす。ひらがなに変換することで、テキスト内の数字、漢字やカタカナ等のバリエーションを減らし、処理の複雑さを低減する。この過程は、特に多様な表記が存在する日本語において、一貫性を保ちながら情報を処理する上で効果的である。ただ、数字は日本語として認識されないためひらがなに変換する前に kanjize ライブラリを用いて、数字を漢数字に変換して置く。これによって、前後の文字が日本語の文脈として認識できるようになり漢字をひらがなの読みにす

る場合の一貫性が保たれる。

## 2.6 最長一致文字列の識別

タイムスタンプのインデックスを見つけるために、自然言語処理で文字を起こしたときに得られる全文章の一連の単語リストから、GPT-4 のレスポンスである質問と答えの文から成るリストの各文章の先頭末尾を探索する必要がある。そこで、文字列と単語のリストが最長一致したアルゴリズムを用いて先頭末尾のインデックスを探索した。アルゴリズムは、以下である。

- (1) 最長の一致の長さ (max\_length), 開始インデックス (start\_index), 終了インデックス (end\_index) を初期化。
- (2) 入力として与えられた各文に対して、以下のステップを繰り返す。
- (3) 単語リストを走査し、現在の文に存在する単語を探索する。
- (4) 一致する単語が見つかった場合、隣接する単語が文に含まれているかを確認し、文字列を可能な限り拡張 (隣接する単語列を順次つなげる)。
- (5) より長い一致するシーケンスが見つかった場合は、その長さと開始・終了インデックスを更新する。
- (6) 各文に対して見つかった最長文字列の開始と終了のインデックスを記録する。

## 2.7 切り抜き動画の出力

前節の方法と moviepy ライブラリにより、動画の複数の質問と答えの文章先頭末尾のタイムスタンプに基づく動画を複数切り出し、それを結合する。これにより質問と答えのみが含まれる切り抜き動画が生成される。これは mp4 形式で出力する。

## 3. 予備実験

### 3.1 実験方法

YouTube 上で公開されている、ひろゆきの動画 2 つを用いて切り抜き動画を作成する予備実験を行った。用いた動画は、次の 2 つである。

- (1) 「独立より副業のほうが成功率が高い。Bière du Vexin を呑みながら 2023/10/14 S19」(60 分) [7]
- (2) 「男女給与格差の原因は差別ではなく母親。小諸のワインを呑みながら 2023/11/05 L23」(60 分) [8]

これらはいずれも、YouTube Live のチャット欄に寄せられた質問を読み上げた上で、ひろゆきが自身の意見を述べる形式の動画である。

### 3.2 実験結果

#### 3.2.1 1 つ目の動画 [7]

提案手法に基づいて得られた GPT-4 のレスポンス文を抜粋する。

質問: [ '高2ですバイトバックレたことを親に言ったら怒られてしまい親を悲しませてしまいましたどうしたらいいですか' ]

答え: [ 'もう普通親は怒るし悲しむので別に来なくていいんじゃないかね。まあ、バイトばっくれるのはよくないけど、若いうちの特権だから、これを心に刻んでまともな大人になって。ただ、バイトばっくれることによって迷惑をかけるのは良くないから、今後は気をつけて。' ]

質問: [ 'au か今日はどこかの動画で知ったのですが自衛隊が日本全土を守る軍事力を持つことをアメリカが許されてきましたもし本当であれ日本やっばり独立できないんでしょうか' ]

答え: [ '日本全土を守る軍事力というのがどういう意味かっていうのが問題。日本全国全土にそういった装備を配備するというのは残念ながら自衛隊の人数とか足りませんし、実際に中国やロシアの戦車とかが上陸してきたら、日本は負けてるんですよ。アメリカは本土を踏ませない戦い方をしている、日本が昔は名前を抜擢したってのは珍しいこと。日本はアメリカに依存しているというより、日本人が自ら独立を望まないからです。' ]...

(多数のため、以下省略)

提案手法に基づいて得られた GPT-4 のレスポンス文において、質問と答えの解釈は大部分が正確であるように見える。これは GPT-4 が質問の意図を理解し、適切な回答を生成していることが分かる。例えば、高校生がバイトをバックレたことを親に言ったら怒られてどうすればいいかという趣旨の質問に対し、バイトをバックレるのはよくないが、これを教訓にしてまともな大人になるべきといった回答が得られている。

次に、レスポンス文から生成した文章の先頭末尾単語インデックスの探索した後にできた文章を抜粋する。

質問: [ '高二ですバイトバックレたことを親に言ったら怒られてしまい親を悲しませてしまいましたどうしたらいいですか' ]

答え: [ 'バックレて迷惑をかけたりにしてこれよくないなっていう心に刻んでまともな大人になって' ]

質問: [ 'か今日はどこかの動画で知ったのですが自衛隊が日本全土を守る軍事力を持つことをアメリカが許されてきましたもし本当であれ日本やっばり独立できないんでしょうか日本全土を' ]

答え: [ 'てるしねもう負けてるよねとアメリカって本当を踏ませないという戦い方をしてるよ日本が昔は名前を抜擢したって' ]...

(多数のため、以下省略)

レスポンス文から生成された文章の先頭と末尾の単語インデックス特定は上手くいっていない。これは、質問と答えの単語が一致しているインデックスの特定が困難であることが分かる。つまり、質問と回答の文脈が完全には捉えられていない。

この結果に基づいて動画を生成したところ、7分26秒の

切り抜き動画が得られた。この切り抜き動画には、8組の質問と回答が含まれていた。

### 3.2.2 二つ目の動画 [8]

提案手法に基づいて得られた GPT-4 のレスポンス文を抜粋する。

質問: ['国民年金で国家ぐるみのねずみ講ですか?']

答え: ['ネズミ講とはなんぞやっていて気になると思うけどネズミ講って基本的に払ったら得をする  
...  
までもねその頃はまだ6本も元気もなくずっと搾取され続けるんじゃないかなと思いますけど結婚にすき家ありますかいりません']

質問: ['結婚に付き合い楽しいのは無理そうかみんなかっこいいとかなりますかでしょうか飲みたくもなくすぐ結婚する秘訣が知りたいですけど']

答え: ['あの恋愛結婚よりもお見合い結構の方が長く続くということ統計上出ていますなので結構長く続かせるのであれば  
...  
なのでこのまま生活を築きますっていう形で離婚しないで最後の行くってパターンだった']...  
(多数のため、以下省略)

GPT-4 のレスポンス文は質問と答えの解釈を適切に行っているように見える。例えば、「国民年金は国家ぐるみのねずみ講ですか?」という質問に対して、年金制度の性質とその機能についての解説が回答されている。

次に、レスポンス文から生成した文章の先頭末尾単語インデックスの探索した後にできた文章を抜粋する。

質問: ["で国民年金で国家ぐるみのねずみ講ですかネズミ講"]

回答: ["で死んでいくっていうので多分えーとで今二十代くらいの人がどう考えても払ってるから降りしてこれおかしいでしょって言うので  
...  
んじゃないかなと思いますけど結婚にすき家ありますかいり"]

質問: ["とかないです結婚に付き合い楽しいのは無理そうかみんなかっこいいとかなりますかでしょうか飲みたくもなくすぐ結婚する秘訣が知りたいです"]

回答: ["あの恋愛結婚よりもお見合い結構の方が長く続くということ統計上出ていますなので結構長く続かせるのであれば  
...  
ただあのじゃあ小学校の時の友達とか親友とか中国とかの大学の話親友って言うてるや"]...  
(多数のため、以下省略)

この動画でもレスポンス文から生成された文章の先頭と末尾の単語インデックスの特定は適切に行えていないようである。従い、質問と回答の内容が完全に一致していないといえる。

この結果に基づいて動画を生成したところ、2分45秒の

切り抜き動画が得られた。この切り抜き動画には、3組の質問と回答が含まれていた。

### 3.3 考察

この予備実験は、配信動画から質問と回答を抽出するためのシステムの有効性を評価するものだったが、いくつかの重要な課題が浮き彫りになった。特に、最長一致文字列の識別アルゴリズムの不足が顕著で、質問と回答の単語の先頭と末尾のインデックスを正確に特定することが難しいことがわかった。これは、アルゴリズムが動画内の複雑な言語的構造や文脈を完全に捉えきれていないことを示している。

また、配信動画の発言の文脈が不整合であることも、正確な文章の特定を妨げる要因となっていた。これは、話者の言葉遣いや話し方、さらには背景ノイズや会話の流れなど、多くの要因によって引き起こされる可能性がある。これらの要因が複合することで、システムが正確な質問と回答を特定することが困難になる。

さらに、自然言語処理の音声からのテキスト変換や、GPT-4 によるプロンプトの変形も、文章の連続一致が途切れてしまう原因となり得る。特に、誤字や言葉の途切れなどの小さなエラーが、識別において大きな影響を与えることがある。これは、自然言語処理技術が完全には動画内の発言の複雑さを捉えきれていない可能性が考えられる。

## 4. おわりに

本稿では、GPT-4 を用いて、トーク中心の動画に対する切り抜き動画を自動生成する方法について検討した。視聴者からの質問とそれに対する回答というスタイルの動画に対して、質問と回答の組を GPT-4 により特定して動画として出力することを試みた。

本研究の予備実験を通じて、配信動画から質問と回答の文章を抽出するシステムの潜在能力と限界が明らかになった。最長一致文字列の識別アルゴリズムの不足、動画の発言の文脈の不整合、自然言語処理や GPT-4 によるテキスト変形といった課題は、今後の研究開発において注目すべき点である。これらの課題を克服し、システムの精度を向上させることで、動画内の特定の文章を効率的に抽出し、より使いやすいシステムを実現することが可能となるだろう。

さらなる改善には、アルゴリズムの改良、動画の発言の文脈をより深く理解し整理する方法、およびテキスト前処理技術の向上が求められる。これにより、より正確で信頼性の高いシステムが構築されることが期待される。最終的に、本システムは質問とその答えに対応する動画部分の切り抜きを自動で行えるようになることが望ましい。

## 参考文献

- [1] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models, *Journal of Computer Vision and Language Processing*, Vol. 7, No. 2, pp. 150–165, 2023.
- [2] Parul Saini, Krishan Kumar, Shamal Kashid, Ashray Saini, Alok Negi, Video summarization using deep learning techniques: a detailed analysis and investigation, *Artificial Intelligence Review*, vol. 56, pp. 12347–12385, 2023.
- [3] Mayu Otani, Yuta Nakashima, Tomokazu Sato, Naokazu Yokoya, Video summarization using textual descriptions for authoring video blogs *Multimedia Tools and Applications*, Vol. 76, No. 21, pp. 22267–22286, November 2017.
- [4] John Doe, Jane Smith, Automatic Soccer Video Analysis and Summarization, *Journal of Sports Technology*, vol. 10, pp. 123–130, 2023.
- [5] 入手先 [〈https://zenn.dev/umi\\_mori/books/prompt-engineer/viewer/langchain\\_overview〉](https://zenn.dev/umi_mori/books/prompt-engineer/viewer/langchain_overview)
- [6] 入手先 [〈https://acro-engineer.hatenablog.com/entry/2023/08/24/120000〉](https://acro-engineer.hatenablog.com/entry/2023/08/24/120000)
- [7] 入手先 [〈https://www.youtube.com/watch?v=ZUha6OruO7w〉](https://www.youtube.com/watch?v=ZUha6OruO7w)
- [8] 入手先 [〈https://www.youtube.com/watch?v=NGfjgxXgfHk〉](https://www.youtube.com/watch?v=NGfjgxXgfHk)