映像内テロップ字幕要約技術による教育応用

藤田恭平*1・坂知樹*2・鎌田洋*1
Email: c6301295@planet.kanazawa-it.ac.jp

*1：金沢工業大学工学研究科システム設計工学専攻
*2：東京電機大学システムデザイン工学部デザイン工学科

©Key Words　映像字幕要約，テロップ，大規模言語モデル

1. はじめに

近年，オンライン教育が急速に普及しており，その中心的な要素として教育動画が活用されている。教育動画は従来の教室での授業を補完または代替し，多くの教育機関や企業で採用されている。Carmichael ら[1]によると，教育動画は情報量が多いが，視聴時間が長いため，人々は教育動画を見るのを避ける傾向がある。そのため，教育動画を要約する技術が求められている。

教育動画を要約する技術として字幕を要約する技術が挙げられる。字幕にはクローズドキャプション(以下，CC)とテロップ字幕がある。CC は，字幕の ON/OFF の切り替えができ，音声認識技術で自動生成した字幕と人手で作成した字幕の 2 つがある。一方，テロップ字幕は，字幕の ON/OFF の切り替えが不可能で，映像に常に表示され，人手で作成した字幕である。

従来から様々な手法が提案されてきたが，要約文の精度を改善するため，本研究は新たな要約アプローチとして，「映像内のテロップ字幕から要約文を生成する手法」を提案する。この手法は，CC から生成した要約文より高精度な要約文を生成できる。それに加え，動画にテロップ字幕が存在しない場合，画面に表示された CC からも要約できる。このため，テロップ字幕を含まない場合においても要約文を生成でき，要約対象の範囲が広い手法である。

2. 従来手法と問題点

教育動画を要約する手法として，従来から複数の手法が提案されてきた。

1 つ目は，音声認識技術を用いて自動生成した CC から要約文を生成する手法である。音声認識技術は，動画内の発話音声をすべて抽出するため，教育動画全体の情報を網羅できる。Vybhavi ら[2]は，YouTube 動画を対象に音声認識技術を用いて音声情報をテキスト化し，BERT[3]でテキスト要約を行った。これにより，YouTube に存在する教育動画の要約が可能となった。近年では，ChatGPT[4]を用いた映像要約ツールが開発された。Glasp[5]は「YouTube Summary with ChatGPT & Claude」というツールを公開した。このツールにおいても YouTube 動画を対象として音声認識技術で文字起こしを行っているが，テキスト要約に ChatGPT を用いることで高精度な要約を実現した。このツールによって，多くの学習者は教育動画を要約できるようになった。しかし，この手法の問題点は，要約文の品質が音声認識技術で生成された CC の品質に依存する点である。音声認識技術により文字起こしされた CC には誤字脱字が含まれることがあり，それが要約文に誤った情報として反映される可能性がある。教育動画の重要なキーワードが要約文に正確に含まれていない場合，学習者にとって有用ではない。

2 つ目は，人手で作成した CC から要約文を生成する手法である。音声認識技術を用いて自動生成した CC は，誤字脱字を含み，教育動画の重要なキーワードが誤って変換されて要約文に含まれる可能性がある。これに着目し，Alrumiah ら[6]は，教育動画を対象とし，人手で作成した CC からテキスト要約アルゴリズムを用いて要約文を生成した。これによって，誤ったキーワードの含まれていない要約文を作成できた。しかし，この手法の問題点は，要約対象の文章に無駄な情報が含まれる点である。人手で作成された CC は話者の会話内容をすべて抽出しており，正確な情報ではあるが，重要でない無駄な情報も多く含まれる。

3 つ目は，教育動画における画像情報の手書き文から要約する手法である。教育動画では，解説者がホワイトボードやブラックボードに教科書などの知識を書き込むことがある。Kenny ら[7]は，画像認識技術を用いて講義の手書き内容を抽出し，講義動画の要約を行った。この手法により，画像情報を基にした要約が可能となった。しかし，この手法の問題点は，動画全体の情報を網羅しないことである。講義の手書き文はホワイトボードやブラックボードに記述される内容のみで構成されているため，講義全体の情報を十分に反映しない可能性がある。この点より，講義の手書き文からの要約は適切でない。

3. 提案手法

従来手法で利用されていない要約対象として，テロップ字幕が挙げられる。テロップ字幕は，映像内に常に固定され，人手で作成された字幕であるため，音声認識技術で自動生成された CC よりも正確な情報が含まれている。したがって，テロップ字幕から要約することで，正確な情報を含む要約文の生成が期待できる。

また，テロップ字幕は視聴者に重要な情報を強調するために表示される。この字幕は動画全体の重要な情報のみを含むため，人手で作成された CC よりも精査された字幕である。したがって，テロップ字幕から要約文を生成する手法は，重要度の高い情報のみを抽出し，高品質な要約文の生成が期待できる。

さらに，テロップ字幕は，講義の手書き内容に比べて動画全体の情報を網羅している。そのため，手書き内容から要約するのではなく，テロップ字幕から要約文を生成すべきである。

以上のテロップ字幕の有用性より，本研究は「映像内テロップ字幕から要約文を生成する手法」を提案する。本研

表 1　アンケート調査に使用した 5 種類の評価尺度

項目	判断基準
文法性	誤字・文法的に正しくない
非冗長性	同じ情報が繰り返されていない
忠実度	視聴した動画と要約文を比較して、動画内容の重要部分を抽出していない
焦点	視聴した動画と要約文を比較して、動画と無関係な情報が含まれていない
自然さ	文章に違和感がない

究は，従来手法の要約文よりも，高精度な要約文を生成することを目的とする。提案手法の処理について図 1 に示す。提案手法は主に 3 段階の処理を行う。第一に，物体検出モデルを用いてテロップ領域を検出する。第二に，OCR（光学文字認識）モデルを用いて，テロップ文をテキスト化する。最後に，文章要約モデルを用いて抽出した文章を要約する。この手法により，映像内のテロップ字幕から高品質な要約文の生成が可能となる。

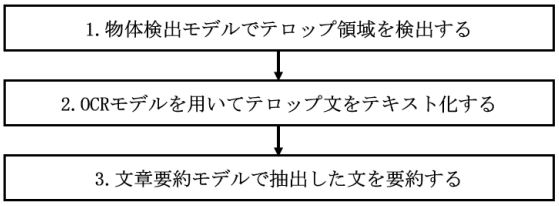

図 1 提案手法の処理内容

4.　試作システム

従来手法と提案手法の要約文の比較実験を行うため，提案手法を搭載した試作システムを開発した。試作システムの処理手順は 3 段階あり，以下の手順で行う。

1 ステップ目では，YOLOv8[8]を用いてテロップ画像を抽出する。YOLOv8 は，Ultralytics が公開した物体検出モデルである。試作システムには，Web 上で収集したテロップ画像を学習済みモデル yolov8l でファインチューニングしたモデルを使用する。動画ファイルから 1 秒に 1 回の頻度で，各フレームに対してテロップ領域を検出し，領域部分の画像抽出を行う。

2 ステップ目では，Google-Drive-OCR[9]を用いて文字起こしを行う。1 ステップ目で抽出したテロップ画像に対して文字認識を行い，テロップ文をテキスト化する。

3 ステップ目では，ChatGPT を用いて要約文を生成する。2 ステップ目で文字起こししたテロップ文を用いて ChatGPT で要約文の生成を行う。

5.　評価実験

本研究の評価実験では，音声認識技術で自動生成した CC の要約文と，提案手法で生成した要約文の品質を比較した。提案手法と比較する従来手法の要約文は，音声認識技術で自動生成した CC の要約文のみとした。テロップ字幕は人手で作成されているため，人手で作成した CC との比較は不要である。また，講義の手書き内容からの要約は動画全体を網羅しておらず，動画全体の情報が著しく少ないため，比較の必要はない。よって，評価実験では，「YouTube Summary with ChatGPT & Claude」で生成した要約文と，提案手法の要約文の品質を比較した。要約文には抽象型要約と抽出型要約の 2 種類があり，両方の評価を行うため，人間による評価と ROUGE[10]による評価を実施した。

人間による評価では，抽象型要約文の言語品質の評価を目的とした。抽象型要約は元の文章に含まれていない単語や表現を用いて生成される要約手法である。したがって，その自然さや文法の正確さを人間が判断する必要があるため，人間による評価を実施した。

一方，ROUGE による評価では，抽出型要約文の正確性の評価を目的としている。抽出型要約は元の文章から直接抽出される要約手法である。したがって，要約の正確性や適切さを定量的に評価するため，ROUGE による評価を実施した。ROUGE は，人間が作成した正解要約とシステムが生成した要約の類似性を測定する指標である。主な測定方法には，ROUGE-1，ROUGE-2，ROUGE-L がある。ROUGE-1 は単一単語の一致数，ROUGE-2 は二連続単語の一致数，ROUGE-L は最長共通部分列を評価する。ROUGE-1，ROUGE-2，および ROUGE-L は以下の計算式で算出する。

$$Precision = \frac{\text{共通部分の数}}{\text{システム生成要約の総単語数}} \quad (1)$$

$$Recall = \frac{\text{共通部分の数}}{\text{正解要約の総単語数}} \quad (2)$$

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

共通部分の数は，システムが生成した要約文と正解要約文が共起した数である。ROUGE-1 は単一単語の一致数，ROUGE-2 は二連続単語の一致数，ROUGE-L は最長共通部分列を指す。

5.1　人間による評価

人間による評価の対象動画として，料理分野の教育動画[11]を 1 本使用した。この動画から従来手法と提案手法で抽象型要約文を生成し，その品質を比較した。評価には金沢工業大学の学部 1 年生から修士 2 年生までの学生を対象とし，Web 上で要約文の品質評価アンケートを実施した。調査では，最初に要約対象の動画を呈示した後，評価尺度を用いた調査を行った。評価尺度には DUC-2005[12]で提案された Quality Questions を修正して使用した。修正した評価尺度は，表 1 の通りであった。被験者には各要約文を呈示し，表 1 の 5 種類の評価指標に対して，1～5 の 5 段階(very-poor～very-good)で回答させた。そして，得られた回答データを基に，Wilcoxon の符号順位検定を実施し，提案手法の要約文が従来手法よりも高い品質を有

表 2　各尺度に対する各手法の平均値と標準偏差

手法	文法性		非冗長性		忠実度		焦点		自然さ	
	M	SD	M	SD	M	SD	M	SD	M	SD
従来手法	4.10	1.09	4.45	0.59	2.25	1.37	2.60	1.53	3.50	1.12
提案手法	4.40	0.58	4.40	0.92	4.00	1.00	4.45	0.67	4.15	0.91

表 3　各尺度に対する Wilcoxon の符号順位検定の結果

項目	文法性	非冗長性	忠実度	焦点	自然さ
Z	-1.04	0.14	-3.36	-3.22	-1.94
P	0.15	0.55	3.86×10^{-4}**	6.41×10^{-4}**	0.03*

**p<0.01, *p<0.05

するのかを評価した。有意水準は片側検定で 5%未満とした。

5.2　ROUGE による評価

ROUGE による評価の対象動画として，教育動画データセット EDUVSUM[13]の動画を使用した。EDUVSUM は，98 本の動画とそれぞれの動画の字幕が含まれた教育動画データセットである。このデータセットの教育動画は，科学・工学の歴史，コンピュータサイエンス，Python と Web プログラミング，機械学習とコンピュータビジョン，IoT，ソフトウェアエンジニアリングを扱っている。評価対象の動画には，機械学習・コンピュータビジョン分野の約 7 分の動画と工学・科学の歴史に関する分野の約 10 分の動画の計 2 本を使用した。ROUGE による評価では，正解要約が必要である。EDUVSUM には 1～10 段階で各動画のセグメントごとに重要度をラベル付けした Annotation ファイルが存在する。この Annotation ファイルを参考に，重要度が 8～10 に該当する字幕を抽出し，正解要約を作成した。そして，2 本の動画から従来手法と提案手法の抽出型要約文を生成し，各手法の ROUGE スコアを比較した。

6.　評価実験の結果と考察

6.1　人間による評価結果と考察

1 本の動画から従来手法と提案手法を用いて，抽象型要約文を生成した。従来手法は 3209 字の文を抽出し，134 字の要約文を生成した。また，提案手法は 2947 字の文を抽出し，253 字の要約文を生成した。したがって，抽出した文に対して，従来手法は 96%削減し，提案手法は 91%削減した。

アンケートを収集した結果，20 名から回答を得られた。5 種類の各尺度に対して得られた各手法の回答データを箱ひげ図として図 2 に示す。また，5 種類の各尺度に対する各手法の平均値と標準偏差を表 2 に示す。文法性と非冗長性の尺度において，両手法ともに平均値が高かった。文法性は，従来手法よりも提案手法の方が 0.30 上回ったが，非冗長性は，提案手法よりも従来手法の方が 0.05 上回る結果となった。

次に，5 種類の各尺度に対する Wilcoxon の符号順位検定の結果を表 3 に示す。忠実度と焦点の尺度において，1%水準で有意な向上が見られた。それに加えて，自然さの尺度において，5%水準で有意な向上が見られた。テロップ字幕は，音声認識技術で自動生成された CC よりも，動画の重要な部分だけを含む文章から要約文を生成している。また，音声認識技術で自動生成した CC には誤字脱字が多く，正確な文章から要約文を生成することが困難であるのに対し，提案手法では誤字脱字のない正確な文章から要約文を生成することが可能である。以上の点より，忠実度・焦点・自然さで有意な向上が見られたと推察する。

一方，文法性と非冗長性の尺度において，有意差はみられなかった。この 2 つで有意差がなかったのは，従来手法と提案手法の両方が同じ要約アルゴリズムで要約しているのが原因と考える。

よって，提案手法は，忠実度，焦点，自然さの 3 点で有効性が示された。提案手法は，音声認識技術で自動生成した CC の要約文と比較して，動画の正確な情報と重要部分を含む違和感のない文章を生成可能だと示唆された。

しかし，人間による評価において，評価したのは 1 本の動画のみである。したがって，評価対象の動画の本数を増やし，様々な分野の教育動画に対して評価を行うことが今後の課題として挙げられる。

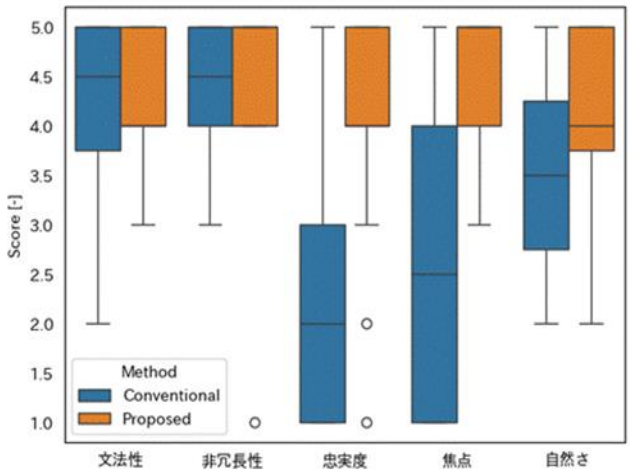

図 2 アンケートの集計結果の箱ひげ図

6.2　ROUGE による評価結果と考察

人間の手で作成した正解要約は 2 本の動画を平均して，4379 字であった。2 本の教育動画から従来手法と提案手法を用いて，抽出型要約文を生成した。2 本の動画を平均して，従来手法は 7761 字の文を抽出し，2287 字の要約文を生成した。また，2 本の動画を平均して，提案手法は 8386 字の文を抽出し，2400 字の要約文を生成した。したがって，両手法ともに抽出した文を 71%削減した。

ROUGE の結果を表 4 に示す。ROUGE-1 において，提案手法は従来手法よりも Recall と F-measure が上回った

表 4　ROUGE による評価結果

手法	ROUGE-1			ROUGE-2			ROUGE-L		
	Pricision	Recall	F-measure	Pricision	Recall	F-measure	Pricision	Recall	F-measure
従来手法	0.71	0.41	0.52	0.32	0.19	0.24	0.39	0.23	0.28
提案手法	0.70	0.45	0.55	0.33	0.21	0.26	0.41	0.26	0.32

ことが確認できる。また，ROUGE-2 および ROUGE-L において，提案手法は従来手法よりも全ての指標で優れていた。以上より，ROUGE-1 の Pricision を除いた全ての指標で，提案手法が従来手法を上回ったことが示された。これは，音声認識技術で生成した字幕を利用するのではなく，映像内の字幕を利用することで，正確な文章が抽出できたため，要約の精度が向上したことを示している。

よって，映像内のテロップ字幕から生成した要約文は，音声認識技術で自動生成した CC の要約文よりも有効的であることが示された。しかし，ROUGE による評価では，評価対象の動画の本数が少なかったため，今後の研究では，さらに評価対象の動画の本数を増やして評価を行うことが必要である。

7.　おわりに

本研究は教育動画から従来手法よりも高精度な要約文を生成することを目的とし，新たな教育動画の要約アプローチとして「映像内のテロップ字幕から要約文を生成する手法」を提案した。具体的には，テロップ検出モデルに YOLOv8，OCR モデルに Google-Drive-OCR，要約モデルに ChatGPT を採用し，試作システムを開発した。提案手法は，テロップ字幕だけでなく，画面に表示された CC からも要約文を生成できる。

提案手法によって高精度な要約文が生成できたのかを調査するため，従来手法との比較実験を行った。その結果，提案手法が従来の音声認識技術を用いた要約手法よりも高精度な要約文を生成できることが示された。

人間による評価において，提案手法は忠実度，焦点，自然さの点で有意に優れていることが確認された。これにより，テロップ字幕の活用が，誤字脱字のない正確な情報を提供し，学習者にとって有用な要約文を生成する上で効果的であることが示唆された。また，ROUGE による評価でも，提案手法が従来手法を上回るスコアを記録し，映像内のテロップ字幕から抽出された情報がより正確であることが明らかとなった。

本研究では，人間による評価と ROUGE による評価において，評価対象の動画の本数が少なかったことが課題として挙げられた。したがって，今後は，多くの動画の本数を増やした条件で評価を行うことが課題である。

参考文献

(1) M. Carmichael, A. K. Reid and J. D. Karpicke, “Assessing the Impact of Educational Video on Student Engagement, Critical Thinking and Learning: The Current State of Play”, https://us.sagepub.com/sites/default/files/hevideolearning.pdf（最終参照日: 2024.06.13）.

(2) A. N. S. S. Vybhavi, L. V. Saroja, J. Duvvuru and J. Bayana, "Video Transcript Summarizer," 2022 International Mobile and Embedded Technology Conference (MECON), pp.461-465（2022）.

(3) J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 1, pp.4171-4186（2019）.

(4) OpenAI, https://chatgpt.com/（最終参照日: 2024.06.13）.

(5) Glasp, “YouTube Summary with ChatGPT & Claude”, https://glasp.co/youtube-summary（最終参照日: 2024.06.13）.

(6) S. S. Alrumiah and A. A. Al-Shargabi, “Educational videos subtitles’ summarization using latent dirichlet allocation and length enhancement”, Computers, Materials & Continua, 70, 3, pp.6205-6221（2021）.

(7) D. Kenny, X. Fei, S. Srirangaraj and G. Venu. “FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos.”, IEEE Access. pp.1-1（2021）.

(8) Ultralytics, “YOLOv8”, https://docs.ultralytics.com/ja/（最終参照日: 2024.06.13）.

(9) Google-Drive-OCR, https://google-drive-ocr.readthedocs.io/en/latest/（最終参照日: 2024.06.13）.

(10) C-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, In Text Summarization Branches Out, Association for Computational Linguistics, pp.74-81（2004）.

(11) 料理研究家リュウジのバズレシピ, “料理研究家が辿り着いた最高の一皿【至高のペペロンチーノ】『Spaghetti aglio e olio peperoncino』”, https://www.youtube.com/watch?v=Jx-tqntWPCM（最終参照日: 2024.06.13）.

(12) H. T. Dang. “DUC 2005: Evaluation of Question-Focused Summarization Systems”, In Proceedings of the Workshop on Task-Focused Summarization and Question Answering, pp.48-55（2006）.

(13) J. A. Ghauri, S. Hakimov and R. Ewerth, “Classification of important segments in educational videos using multimodal features” in the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020). Proceedings of the CIKM2020Workshops, pp.1-8（2020）.