Vol.2022-NL-253 No.16
2022/9/30

# オンライン会議での自動要約のための マルチモーダル情報を考慮した重要発話抽出に関する検討

新美翔太朗 [1,a)]　後藤啓介 [1]　西田典起 [2]　松本裕治 [2]
荒川智哉 [1]　廣島雅人 [1]

**概要**：近年，ビデオ通話アプリの台頭により，会議を従来のようにオフラインで行うのではなくオンラインで行う機会が増加している．オンライン会議はオフライン会議と比較し，通信システムを介することで生じる視聴覚情報の制限や遅延により議論の流れを把握しづらいという課題がある．以前より画像や音声といったマルチモーダル情報を考慮した自動要約技術も研究が行われてきたが，それらは主にオフラインでの会議を考慮した研究であり，オフライン会議においても同様の結果が得られるとは言えない．
そこで，本研究ではオンライン会議において得られるマルチモーダル情報を考慮した重要発話抽出手法を提案し，自動要約の精度向上によるオンライン会議の理解促進を目指す．

## 1. はじめに

近年，情報通信技術の発展によるビデオ通話アプリの台頭や，Covid-19 の流行により，デバイスを通じて行うオンライン会議が重宝されている．オンライン会議には多くの利点が存在する．例えば，対面で行われるオフライン会議と比較し，参加人数に縛られずに開催できるといった空間上の利点や，会議の録画が容易なため，会議に参加しなくても後ほど状況を把握できるといった時間上の利点，参加者間での発表資料の共有が容易であるという情報共有での利点等が挙げられる．会議内容の要約（議事録）は，オンライン会議のこれらの利便性を享受するのに有用である．そこで本研究では，オンライン会議における会議要約技術の開発に取り組む．

以前より，会議参加者の視線や顔の向き，音声等といったマルチモーダル情報を考慮した会議要約技術が研究されてきた [1, 2] ．しかし，それらはオフライン会議を仮定しており，オンライン会議においても同様の結果が得られるとは限らない．また，オンライン会議を対象にマルチモーダル情報を利用し会議要約を行った研究は存在しない．

本研究では，オンライン会議における自動要約の精度をより向上させるために，文脈を考慮した言語情報とマルチモーダルな非言語情報を用いた，重要発話抽出方法を提案する．

本論文の構成は，まず 2 章において，関連研究を紹介する．3 章において，本稿にて取り組む重要発話の定義について詳述し，4 章において著者らの提案する手法について述べる．最後に，5 章にてまとめと今後の展望について記す．

## 2. 関連研究

前章で述べた通り，会議要約の研究において，長い発話群から要約に必要な部分である重要発話を抽出する工夫が長年研究されてきた [1, 2, 3, 5] ．本章では過去の研究事例と，研究に使用されてきた会議データセットについて述べる．

### 2.1 重要発話抽出

二瓶ら (2016) [1] はスピーチインテンシティ，スピーチスペクトログラム，頭部動作，頭部加速度，言語を入力情報としたマルチモーダルモデルを構築し，言語のみを入力情報としたユニモーダルモデルと比較して高い精度で重要発話が推定可能であることを示した．また，マルチモーダルモデルでは，音声，表情，ジェスチャ，言語のそれぞれ 1 つのモーダルのみを入力情報としたユニモーダルモデルと比較して，それぞれの特徴を考慮したことにより性能が上がったことを推定結果の考察により示している．本研究ではこの知見を参考にし，非言語情報モデルを構築している．二瓶らはオフライン会議を想定しているのに対し，本研究ではオンライン会議を対象としている点で異なっている．また，二瓶らは言語情報モデルの入力情報として skip gram モデルにより得たベクトルを用いているのに対し，本研究では事前学習済みの言語モデルにより生成されたベクトルを用いること検討しており，文脈を考慮した言語情報を利用するという点で先行研究とは異なっている．

Li ら (2019) [2] は言語情報に加え視線情報を用いてトピック毎に注目されている単語の情報を要約モデルに与え，要約精度が向上することを示した．これにより視線情報が会議要約に有効であることが明らかになったが，本研究では利用するマルチモーダル情報の一つとして視線情報は採用していない．本研究では，Li らとは異なりオンライン会議を対象としており，視線移動は会議内容よりもモニターの位置やカメラの起動状況等，参加者の環境に依存しやすいためである．

1 京セラ株式会社
KYOCERA Corporation
2 理化学研究所革新知能統合研究センター
RIKEN Center for Advanced Intelligence Project
a) shotaro.niimi.gt@kyocera.jp

中山ら (2021) [3] は後述する Kyutech コーパス [4] に対し，話題毎に対話全体を分割し要約を作成した．また，それぞれの話題における重要発話を推定することにより，重要発話を抽出しない場合と比較して生成された要約の可読性が向上することを示した．この研究では発話を重要か否かの 2 値で捉えているのに対し，本研究では要約元と要約文の類似度に基づき，要約元に対する重要度を算出することを検討している．これにより，相対的な発話の重要度を捉えることが可能となっている．また，中山らは言語情報のみでの重要発話抽出を行ったのに対し，本研究では言語情報に加え非言語情報を利用することで多角的に重要発話を推定することが期待できる．

Kadowaki (2020) [5] は，NTCIR-15 QA-Lab-Poliinfo2 [6]の Dialog summarization task において，テキスト群の前処理を行う Preprocessor，要約のために必要な文の抽出を行う Sentence extractor，要約文を生成する Summary generator の 3 つの処理機構を用いた手法を提案している．この内，文抽出を行う Sentence extractor では，BERT [7] を用いて要約文章に対する各発話の Rouge [8] を推定する回帰モデルを使用している．これにより要約に用いられる可能性の高い文を抽出している．本研究ではこの Sentence extractor を参考に言語情報モデルを構築している．Kadowaki が都議会議録を対象に言語情報のみで重要発話抽出を行っているのに対し，本研究ではオンライン会議を対象とする点，非言語情報を利用する点で Kadowaki の手法と異なっている．

上記手法の特性を表 1 にまとめた．本研究では既存研究と比較し，オンライン会議を対象にしている点， 文脈を考慮している点で大きく異なっていると言える．

表 1 本研究と既存研究の比較

| | 文脈考慮 | マルチモーダル情報考慮 | オンライン会議 |
|---|---|---|---|
| 二瓶ら [1] (2016) | × | ○ | × |
| Liら [2] (2019) | ○ | ○ | × |
| 中山ら [3] (2021) | × | × | × |
| Kadowaki [5] (2020) | ○ | × | × |
| 本研究 | ○ | ○ | ○ |

### 2.2 データセット

会議要約のための大規模な対話要約データセットとして，従来から AMI コーパス [9] や ICSI コーパス [10] が用いられてきた．AMI コーパスはおよそ 100 時間の会議記録から構成されたデータセットである．会議毎に，人手による書き起こし，発話単位で付与されたタイムスタンプ，各参加者の画像，音声情報といったマルチモーダル情報が収録されている．ICSI コーパスは約 70 時間のミーティングを録音して作成された音声データセットである．各参加者の会議中の映像は収集されていないものの，タイムレコード付きの発話と各参加者の音声情報を収集している．

また，上述の 2 つのデータセットについて，新たな視点で再アノテーションがなされた QM Sum コーパス [11] も近年公開されている．これは，AMI コーパスや ICSI データセットに加え，ウェールズ議会の 25 の委員会と，カナダ議会の 11 の委員会についてもアノテーションを行い，全 232 会議を収集している．このデータセットは，会議を収集した会議データセットとしては最大規模となっている．

日本語の会議データセットとしては，Kyutech コーパス [4] が，架空のショッピングモールに出店するレストランを選ぶという議題について 4 名の参加者が討論を行った様子を収録している．

会議の特性の分析や重要発話の抽出を目的としたデータセットとしては，MATRICS コーパス [12] が存在する．これは各会議参加者が架空の議題について 20 分ずつディスカッションを交わした様子について各参加者のマルチモーダル情報や会話内容を記録したコーパスである．

以上の例について，それぞれのデータセットの特性を表 2 にまとめた．なお，QM Sum コーパスについては用意されている情報は要約のみで，マルチモーダル情報については基のコーパスに依存しているため，例外的な記号によって表記している．表 2 に示されるように，会議について収集されたデータセットは全てオフライン会議の形式で収集されており，現状オンライン会議の形式で収集された日本語要約のデータセットで広く利用可能なものは存在しない．

表 2 会議データセットの特性

| | 話題単位要約 | マルチモーダル情報 | 重要発話 | オンライン会議 |
|---|---|---|---|---|
| Kyutech [4] | ○ | ○ | ○ | × |
| AMI [9] | × | ○ | × | × |
| ICSI [10] | × | ○ | × | × |
| QMSum [11] | ○ | — | × | × |
| MATRICS [12] | × | ○ | ○ | × |

## 3. 重要発話ラベル付きデータセットの構築

本章では本研究において抽出の対象となる重要発話の定義，および抽出のために作成したデータセットについて述べる．

### 3.1 重要発話

会議において，要約に用いられる文章はほぼそのまま対話中に現れているか，対話中に現れた文章を改変して作成されることが殆どである．そのため，要約と要約のもとになっている発話との類似性は高いと考えられる．そこで本研究では，要約との類似性が高い発話を重要発話として抽

出することを目標とする．また，会議においてどの発話が要約として選択されるかは，発話した参加者や発話自体の長さ，参加者の反応等，発話の内容以外の特性に依存することもある．そこで本研究は，重要発話を抽出する際に発話の言語情報のみを用いて判別を行うのではなく，音声特徴や画像特徴等のマルチモーダル情報も併せて学習に用いることで，各モーダル間の特徴を補完しあい，より正確な重要発話抽出を行えるようにするよう試みた．

また，表 2 に示したように，公開されている会議のデータセットについて，重要発話がアノテーションされている例は少ない. 本研究ではアノテーションを自動化するため，重要発話の指標として，BERT Score [13] を利用して疑似ラベルを付与した．対話要約の性能指標としては Rouge が用いられることが多く，2 章で紹介した Kadowaki (2020) [5] の手法においても発話の重要度ラベルとして Rouge-1 スコアが採用されている．しかし対話要約においては発話を適切な表現へ置き換えて作成されることも多い．その点，BERT Score では意味的な類似度を捉えることが出来るため，本研究では BERT Score を重要度の指標とした．

これにより，要約に対する各発話の BERT Score を求め，それを擬似的に発話の重要度のラベルとすることで，半自動的に発話毎の重要度のアノテーションを行うことができる．

### 3.2 データセット

前章で述べた通り，オンライン会議の形式で収集された広く利用可能な日本語要約のデータセットは現状存在しない．そこで本研究では協力者を募り，独自に会議データを収集し，オンライン会議の形式でのデータセットを構築した．データセットの構築にあたり，著者ら，および協力者が行ったことは以下の通りである．

1. 協力者はマイクを装着し，カメラを起動して各々のデバイスの前で Web 会議アプリを用いて会議を行う．
2. 協力者らによる会議はそれぞれ 60~120 分間程度行われ，会議参加者が会議までに行ってきたタスクについての説明と，説明に対する質疑応答を繰り返す形で行われた．
3. 実験協力者からそれぞれの正面からの録画映像および Web 会議アプリの機能により記録された会議全体の記録を受け取り，各発話について 10ms 単位で発話時間を付与した．
4. 受け取った会議全体の映像と各参加者の録画からそれぞれのタイミングで録画を開始したために生じた誤差を算出し，上記 3 で付与した発話毎のタイムスタンプに沿うように修正した．
5. 上記 4 で修正した会議参加者の映像を発話区間毎に分割し，発話時間ごとの各会議参加者の音声情報と画像情報を得た．
6. 受け取った会議映像から人手により各話題を分割し，概ね会議時間 1 分につき 66 文字以内になるようにそれぞれの話題に対して要約を作成した，そして，それぞれの要約に対して小見出しを付与した．
7. 作成した要約に対し，前節で述べた通りの指標として各発話に重要度を付与した．

以上の手順を以て，著者らは最終的に約 10 時間分の会議データを収集し，発話テキスト，参加者毎の映像，音声，重要度ラベルを 5,724 発話分と，75 の話題に対する要約を得た．

## 4. 提案手法

前章では，本研究で用いる重要発話の疑似ラベルと，構築したデータセットについて述べた．本章では重要発話の定義に基づき発話抽出を実現する提案手法で用いた，言語情報モデルと非言語情報モデルについて述べる．なお，言語情報と非言語情報の両方を考慮した言語-非言語統合モデルについては，言語情報モデルと非言語情報モデルより最終層を取り外し，全情報を結合する全結合層を設けることで構築することを検討している．言語-非言語情報の統合モデルの全体像については図 1 に示す．

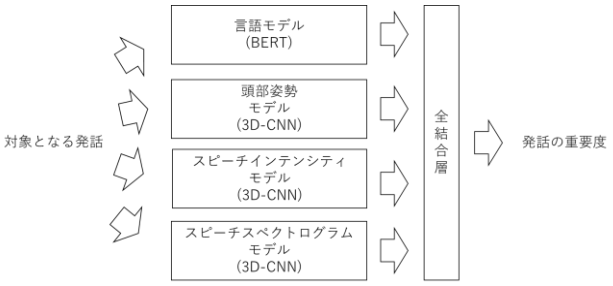


図 1 提案手法の概要

### 4.1 言語情報モデル

2 章で紹介した Kadowaki の手法の Sentence extractor を参考に，各発話の要約に対する重要度を推定する回帰モデルを検討している．

Kadowaki の手法では，与えられた Main Topic と Sub Topic と呼ばれる 2 つのトピック情報と，要約に必要な範囲に存在する各発話文の 2 つを入力情報とし，要約に対する各発話文の重要度を推定している．一般的なオンライン会議において，Main Topic や Sub Topic といった情報は与えられない．また，オンライン会議では一括質問一括応答形式である都議会とは異なり，自由に発話する形式であることが多いと思われる．このような形式の会議において，同じ話題に対する発話が時間を跨いで行われることはあっても，同じ話題に対して全く異なる要約が複数必要になることは稀であると考えた．

以上により，本研究における言語情報モデルでは，発話

群を入力として重要度を推定する BERT ベースのモデルを検討している．

また，本研究において重要度として用いる BERT Score は語彙を文脈上の意味のレベルで考慮していると言える．そのため，対象とする発話のみでなく，その周辺の発話も入力情報とすることで，より抽出精度が高くなることが期待できる．そのため，推定対象とする着目発話の前後 N 発話を含めた発話群を入力とすることを検討している．また，入力する際には，新規に定義した特殊トークン[ATT]を，発話 1 [ATT] 発話 2 [ATT] 発話 3…となるように各発話の切れ目に挿入することで，1 つの発話のように扱う．

言語情報モデルから言語-非言語統合モデルとする際には，回帰推定のための線形層を削除し，[CLS]トークンに対する 768 次元の分散表現を着目発話の言語情報とした．

言語情報モデルでは上記の構成を用いて，各発話についての重要度を推定し，この重要度の高い文を抽出することで重要発話の抽出が可能となる．

### 4.2 非言語情報モデル

非言語情報モデルについては，二瓶ら(2016) [1]を参考に 3D-CNN を用いることを検討している．ここでは，二瓶らと検討中のモデルの差異について述べる．なお，モデルの層等については二瓶らと同様のため，本稿での説明は割愛する．

まず，二瓶らは非言語情報として，頭部スペクトログラム（以下 HS），スピーチスペクトログラム（以下 SS），スピーチインテンシティ（以下 SI），頭部姿勢（以下 HP）をモデルへの入力としたが，本研究ではこの内 HS については用いなかった．二瓶らはオフライン会議における重要発話を推定していたのに対し，本研究ではオンライン会議における重要度を推定する．そのため，オフライン会議で見られる，話者の方へ顔を向ける，話者の発言へのリアクションとして首を振るといった挙動が．オンライン会議では観測が期待できない．よって，HS モデルの有無による推定精度への影響は小さいと考えた．

また，会議の形式について，二瓶らは特定の話題に対するディスカッションを対象としていたのに対し，本研究では発表および質疑応答の形式の会議を収集した．そのため，個人が意見を述べるディスカッション形式と比較し，一回当たりの平均発話時間が長くなる傾向が見られた．そこで本研究では収集した発話より観測できた最大の発話時間である 35 秒を最大発話長とし，35 秒に満たない発話時間については空白で埋める処理を検討している．

以上の点を踏まえ，本研究にて検討中のモデルを二瓶らのモデルと比較し，表 3 にまとめた．

### 4.3 提案手法の考察

提案手法について，本研究では発話の重要度に関する疑似ラベルとして BERT Score を用いた．しかし，BERT Score は要約のために用意された指標ではないため，要約における重要度とは必ずしも一致するとは限らない．

また，BERT Score はマルチモーダル情報を考慮した指標ではない．そのため，言語情報にのみ基づく BERT Score をターゲットラベルとして用いても，マルチモーダル情報を有効に活用するようにモデルが学習できないという可能性がある．

最後に，非言語情報について，今回は HS について精度への影響が少ないと仮定したが，各モーダルがどの程度性能に寄与するかは明らかではない．そのため，今回採用したモーダルの中で必要なモダリティや不必要なモダリティ，また，今回は検討しなかったもののオンライン会議において必要なモダリティの検討を行う必要がある．

上記の事項について検討するための実験は今後随時行っていく予定である．

表 3 非言語情報モデルの各モーダルの入力サイズ，畳み込みカーネルサイズ，プーリングフィルタサイズ

| ユニモーダル 非言語情報モデル | | 入力サイズ | 畳み込み カーネルサイズ | プーリング フィルタサイズ |
|---|---|---|---|---|
| 二瓶ら [1] (2016) | HS Model | 450, 15, 4, 1 | 3, 3, 4 | 2, 2, 1 |
| | SS Model | 750, 32, 4, 1 | 5, 3, 4 | |
| | SI Model | 1500, 1, 4, 1 | 10, 1, 4 | |
| | HP Model | 450, 3, 4, 2 | 3, 3, 4 | |
| 提案手法 | HS Model | － | － | － |
| | SS Model | 24, 32, 6, 1 | 5, 3, 6 | 2, 2, 1 |
| | SI Model | 3500, 1, 6, 1 | 10, 1, 6 | |
| | HP Model | 1050, 3, 6, 2 | 3, 3, 6 | |

## 5. おわりに

本論文では，オンライン会議における要約の精度向上のため，現在検討しているマルチモーダル情報を考慮した重要発話抽出方法と，手法実現に向けたデータセット構築について提案した．

前章で述べた懸念事項を解消するために，以下の実験を検討している．

- 重要度の指標として BERT Score を利用した場合と，BERT Score 以外の指標（Rouge 等）を用いた場合の性能比較．
- マルチモーダル情報を考慮して重要発話の抽出を行うモデルと，言語情報のみで重要発話の抽出を行うモデルとの性能比較．
- 各非言語情報のモーダルについて，モーダル毎の重要発話の抽出精度比較．
- オフライン会議とオンライン会議に現れる非言語特性の違いと，オンライン会議において有用な特徴量の選定．

本論文ではマルチモーダル情報を考慮した重要発話抽出

の手法を提案した．この手法はオンライン会議について，マルチモーダル特徴を考慮した重要発話の抽出が可能であり，オンライン会議における要約精度向上への貢献が期待できる．

## 参考文献

[1] 二瓶芙巳雄, 中野有紀子, 高瀬裕,: 言語・非言語情報の融合に基づく重要発言の推定, 2018 年度人工知能学会全国大会（第 32 回） (2018).
[2] Li, M., Zhang, L., Ji, H. and Radke, R. J.: Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization, *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2190–2196 (2019).
[3] 中山友梨, 塩田　宰, 嶋田和孝,: 複数人対話におけるトピック単位の要約データの構築とその要約, 研究報告知能システム（ICS）, Vol. 2021-ICS-203, No. 6, pp.1–6 (2021).
[4] Shimada, K., Yamamura, T. and Kawahara S.: The Kyutech Corpus and the Topic Segmentation Using a Combined Method, *In Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 95–104 (2016).
[5] Kadowaki, K.: JRIRD at the NTCIR-15 QA Lab-PoliInfo-2 Task: An Abstractive Dialog Summarization System for Japanese Assembly Minutes, *In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies* (2020).
[6] Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K., Ishioroshi, M., Mitamura, T., Yoshioka, M., Akiba, T., Ogawa, Y., Sasaki, M., Yokote, K., Mori, T., Araki, K., Sekine, S. and Kando, N.: Overview of the NTCIR-15 QA Lab-PoliInfo-2 Task, *In Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies* (2020).
[7] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019).
[8] Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries, *In Proceedings of the Text Summarization Branches Out*, pp. 74–81 (2004).
[9] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W. and Kronenthal M.: The AMI Meeting Corpus: A Pre-announcement, *In Proceedings of the International Workshop on Machine Learning for Multimodal Interaction*, pp. 28–39 (2015).
[10] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. The ICSI meeting corpus. *In Proceedings of the IEEE ICASSP*, (2003).
[11] Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X. and Radev D.: QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization, *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* pp. 5905–5921, (2021).
[12] 林佑樹，二瓶芙巳雄，中野有紀子，黄宏軒，岡田将吾，“グループディスカッションコーパスの構築および性格特性との関連性の分析，” 情報処理学会論文誌，vol.56，no.4，pp.1217‐1227，(2015).
[13] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. and Artzi, Y.: BERTScore: Evaluating Text Generation with BERT, *In Proceedings of the 8th International Conference on Learning Representations* (2020).