会議発話間の関係性推定における マルチモーダル情報活用方法の初期検討

大杉康仁　中辻真
エヌ・ティ・ティレゾナント株式会社
y.oosugi@nttr.co.jp

概要

会議において、相手が自分の発言を支持しているかどうかを知ることは、会議を円滑に進める上で重要である。本研究では、対面会議において、発話と応答のペアに対し、応答が発話を支持するものかどうかをマルチモーダル情報を用いて推定することを検討する。発話および応答の書き起こしテキスト・音声・顔動画を入力とする Transformer Encoder に基づくモデルを利用し、各モーダルの効果を検証した。複数人会議コーパス AMI を用いた実験では、テキストモーダルが最も F1 値に影響を与えることが示唆された。

1　はじめに

複数人で行われる会議において、合意が取れている項目を整理し会議を円滑に進めるために、自分の発話と相手の応答との関係性を知ることは重要である。関係性として「納得できたかどうか」[1] や「同意・不同意」「支持・不支持」[2, 3] などが挙げられる。ここで、「『納得』は自分なりに試行錯誤しながら自ら出した解答である」[4] が、「同意」は「自分も同じ意見であるということを態度に表わすこと」であり「同意」に類する「支持」は「同意」に比べ継続的な行為とされている[1)]。そこで本研究では、「同意」や「支持」の方が「納得」よりも外から観察・推定することが容易であり、短時間で行われる発話間の関係性においては「同意」と「支持」はほぼ同一と扱うことができると考え、「支持」に主に着目する。

会議中のある発言に対し別の発言がなされたときの 2 発話間の関係性を分類することは議論マイニングタスクの一つであり、多くの研究がなされている [2, 5, 3]。中でも姫野・嶋田 [5] は、企業で頻繁に行われる会議と同様の形式である複数人議論コーパス AMI[6] について 2 発話間の関係性を 9 種類に分類することを検討している。姫野・嶋田の手法は発話の書き起こしテキストを用いるものであるが、音声の抑揚や表情などパラ言語情報・非言語情報も発話間の支持・不支持の関係性を認識する上で重要である。そこで、本研究では、ある話者の発話と、それに対する別の話者の応答について、2 発話間の関係性をマルチモーダル情報を用いて推定することを検討する。ただし、マルチモーダル情報の効果の測定を簡単にするため、相手の発言を支持する発話かそれ以外かという 2 値分類の問題として検討する。

図 1　マルチモーダルに基づく発話間の関係性推定タスク

マルチモーダル情報を用いて議論マイニングを行う方法として、各モーダルを BERT[7] や Wav2Vec2.0[8] などの事前学習モデルで独立にエンコードした後で連結ベクトルを線形層で変換する方法が提案されている [2, 3]。一方で、マルチモーダル情報を統合する方法として、言語・音声・動画を系列方向に連結して Transformer Encoder[9] に入力することでモーダル間の関係性をより効率よく捉える手法が提案されている [10, 11]。本研究では、この手法を応用し、発話と応答についてそれぞれ Transformer Encoder でマルチモーダル情報を統合した後で線形層を用いて発話と応答の両方を考慮してその関係性を推定する手法について検証する。

1）https://dictionary.goo.ne.jp/thsrs/4366/meaning/m0u/ 2023/1/4 アクセス

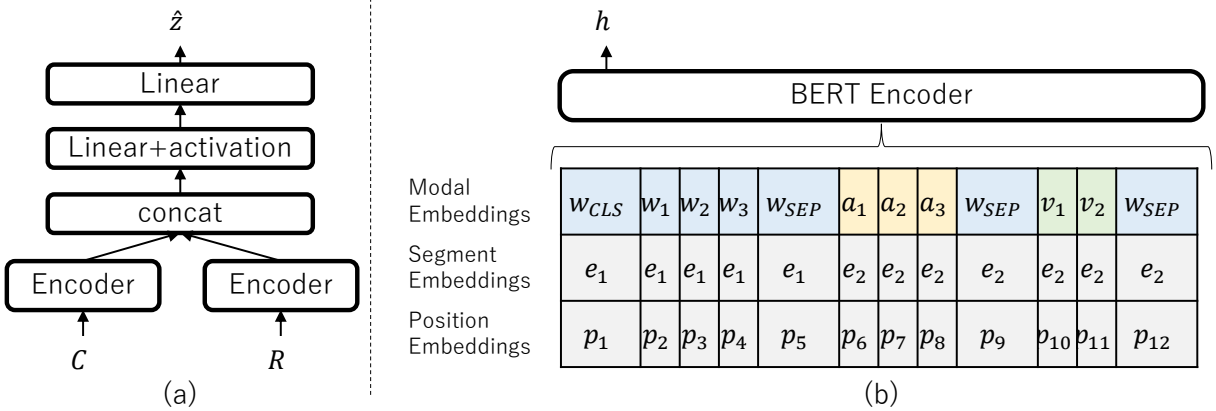

図 2　検討モデル (a) と Encoder の内部構造 (b)。青色部分はテキストの埋め込みを、黄色部分は音声の埋め込みを、緑色部分は動画の埋め込みをそれぞれ表す。

2　問題設定

対面音声会議における発話テキストを対象とした関係性分類タスク [5] では、書き起こしテキストのみを入力として扱っているが、本研究ではそれをマルチモーダルに拡張する。すなわち、図 1 に示す通り、ある話者 S_x の時刻 i の発話 $C = \{\boldsymbol{T}_i^x, \boldsymbol{A}_i^x, \boldsymbol{V}_i^x\}$ に対し、話者 S_y が時刻 j の応答 $R = \{\boldsymbol{T}_j^y, \boldsymbol{A}_j^y, \boldsymbol{V}_j^y\}$ を行ったとき、発話 C と応答 R の関係性 z を推定する。ただし、T_i^* は発話テキストを、A_i^* は音声を、V_i^* は動画をそれぞれ表す。本論文では、話者 S_x と話者 S_y は異なり（$x \neq y$）、かつ、話者 S_x の発話 C を聞いた後、話者 S_y が応答 R を行うものとする（$i < j$）。

3　検討手法

図 2(a) に本論文で検討した手法の全体像を示す。話者 S_x の時刻 i の発話 C と話者 S_y の時刻 j の応答 R を別々に Encoder を用いて特徴量化する。それらを次元方向に連結したベクトルを活性化関数を含む 2 層の線形層で変換し、関係性の予測値 $\hat{z}$ を得る。

$$\hat{z} = \mathrm{Linear}(\sigma(\mathrm{Linear}(\mathrm{Concat}(\boldsymbol{c}, \boldsymbol{r})))) \tag{1}$$

$$\boldsymbol{c} = \mathrm{Encoder}(\boldsymbol{T}_i^x, \boldsymbol{A}_i^x, \boldsymbol{V}_i^x) \tag{2}$$

$$\boldsymbol{r} = \mathrm{Encoder}(\boldsymbol{T}_j^y, \boldsymbol{A}_j^y, \boldsymbol{V}_j^y) \tag{3}$$

ただし、Linear(·) は線形層を、$\sigma(\cdot)$ は活性化関数を、Concat(·) は次元方向の連結を表す。また、Encoder(·) は Transformer Encoder[9] を用いた各モーダル情報の統合を表し、本論文では図 2(b) に示すように MEmoBERT[11] と同じモデル構造を用いた。以下では各モーダルの埋め込みの方法とその統合方法について説明する。

テキストの埋め込み　BERT[7] のトークナイザを用いて発話テキスト $\boldsymbol{T}$ をサブワード系列に分割し、BERT のサブワード埋め込み $\boldsymbol{a} \in \mathbb{R}^{d_t \times L_t}$ をテキスト埋め込みとして用いる。ただし、L_t は発話テキストのサブワード系列長を、d_t は特徴量の次元数をそれぞれ表す。

音声の埋め込み　フレーム長 l の音声 $\boldsymbol{A} = \{a_1, a_2, ..., a_l\}$ について、音響特徴量の埋め込みとして Wav2Vec2.0[8] に基づく特徴量を用いる。ただし、音声のフレーム数は他のモーダルに比べて多いため、フレーム窓 w ごとに平均を取ることで圧縮を行う。また、埋め込み空間の違いを吸収するため線形層による変換を行う。得られる特徴量 $\boldsymbol{a} \in \mathbb{R}^{d_a \times L_a}$ は下記で表される。

$$\boldsymbol{a} = [\boldsymbol{a}_1; \boldsymbol{a}_2; ...; \boldsymbol{a}_{L_a}] \tag{4}$$

$$\boldsymbol{a}_k = \mathrm{Linear}\left(\frac{1}{w}\sum_{t=k}^{k+w}(\mathrm{Wav2Vec2}(\boldsymbol{A}_t))\right) \tag{5}$$

ただし、L_a は圧縮後の特徴量数を、d_a は特徴量の次元数を表す。

動画の埋め込み　動画 $\boldsymbol{V}$ はフレーム画像 $\{v_1, v_2, ..., v_{L_v}\}$ で構成される。本論文では、各フレーム画像を ViT[12] を用いて特徴量化することで、動画特徴量の埋め込み $\boldsymbol{v} \in \mathbb{R}^{d_v \times L_v}$ を得る。ただし、L_v はフレーム数を、d_v は特徴量の次元数を表す。また、埋め込み空間の違いを吸収するため線形層による変換を行う。

$$\boldsymbol{v} = [\boldsymbol{v}_1; \boldsymbol{v}_2; ...; \boldsymbol{v}_{L_v}] \tag{6}$$

$$\boldsymbol{v}_k = \mathrm{Linear}(\mathrm{ViT}(v_k)) \tag{7}$$

Transformer Encoder を用いた特徴量化　各モーダルの系列長が異なるため、単純な加算は困難で

表 1 TAS データセット

	対話数	発話応答ペア数	Positive	Other
train	79	3020	1643	1377
dev	7	315	159	156
test	8	505	320	185

ある。そこで、図 2(b) に示すように、各モーダルの埋め込みをセパレータトークン [SEP] で連結する。このとき、先頭に特殊トークン [CLS] を追加する。それぞれの埋め込みとして BERT のサブワード埋め込みを用いる。さらに、系列の位置埋め込み p と、モーダルの区別のための BERT のセグメント埋め込み e を足しこむ。得られた埋め込み系列を Transformer Encoder に入力し、[CLS] トークン位置の隠れ状態 $\boldsymbol{h}$ を発話もしくは応答の特徴量として用いる。

4 評価実験

データセット 英語での対面会議を収録した AMI コーパス [6] と、その発話間の関係性を記述した Twente Argument Schema (TAS) データセット [13] を用いた。人手による書き起こしテキストとヘッドセットで収録された音声を用いた。また、動画として参加者の正面に配置されたラップトップ PC からの映像を用いた。TAS では「支持する」「支持しない」を含む 9 種類の関係性ラベルが付与されているが、約 55%もの発話応答ペアに「支持する（Positive）」が付与されておりラベルに偏りが生じている。本研究では相手が自分の発話に同意したかどうかを推定することを目的とするため詳しいラベルは必要ないと考え、TAS の関係性ラベルを「支持する（Positive）」か「それ以外（Other）」の 2 つに分けて実験を行った。

実験条件 TAS の 94 対話を表 1 に示すように train/dev/test に分割した。実装は Huggingface Transformers ライブラリ [14] を使用し、Transformer Encoder とテキスト埋め込みの初期化には BERT base モデル[2)] を用いた。音声の埋め込みには Wav2Vec2.0 の base モデル[3)] を、動画の埋め込みには CLIP[15] の ViT モデル[4)] を、それぞれ用いた。ただし、学習を簡単にするため、Wav2Vec2.0 と ViT のパラメータは固定して用いた。音声について圧縮するためのフレーム窓長 w を 128 に設定した。また、動画については 25fps で収録されているが、差分の少ない画像が不必要に入力されることを避けるため、1 秒区間の先頭フレームのみを用いるサンプリングを行った。パラメータ最適化には Adam[16] を用い、学習率を 2×10^{-5} に、warmup 率を 0.1 に設定して、バッチサイズ 32 で 3 エポック学習した。5 種類の乱数シードで実験し、その平均値をモデル結果とした。

2） bert-base-uncased
3） facebook/wav2vec2-base
4） openai/clip-vit-base-patch16

表 2 テストデータにおける評価結果：T, A, V はそれぞれテキスト、音声、動画のモーダルを表す。

modal	Acc. [%]	F1 [%]
T	**82.6**	**85.9**
A	69.0	72.6
V	65.1	69.6
T, A	82.4	85.8
T, V	82.1	85.6
A, V	69.5	73.1
T, A, V	81.6	85.1

4.1 評価結果

マルチモーダル情報を考慮することで精度は向上したか？ 表 2 より、テキストと音声の 2 つのモーダルを考慮する場合は、テキストのみを考慮する場合と同程度の精度を達成できたが、動画を加えた 3 つのモーダルを考慮する場合は、テキストモーダルのみを考慮する場合に比べて精度が劣化した。そのため、本検討手法ではマルチモーダル情報の効果は強く見られなかった。

どのモーダルが最も効果的か？ 表 2 に示す通り、テキスト情報が最も効果的であった。このことから、明示的に話者が発言することで応答している場合には、発話テキストを考慮することが最も効果的と言える。一方で、実際の企業で行われる会議では、発言をせずに態度で支持・不支持を示す場合も存在するため、テキスト情報に依存しない推定方法が必要と考えられる。これは今後の課題として対応していく予定である。

検討手法は音声情報を効果的に扱えているか？ 音声情報のみを入力した場合、テキスト情報のみを用いた場合に比べて精度が劣化したため、本検討手法では音声情報は効果的に活用できていない可能性がある。その原因として、Transformer Encoder を BERT のパラメータ（言語ドメイン）を用いて初期化しため、埋め込みの音声ドメインとの乖離が発生し、それを Linear 層のみでは吸収しきれなかったこ

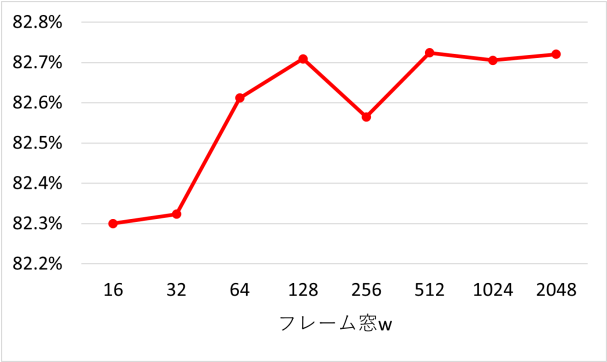

図 3 フレーム窓 w を変更した時の T・A・V のマルチモーダルモデルの F1 値（dev セット）

表 3 TAS データセットの平均フレーム数、および、検出された正面顔のフレーム単位での平均数

	発話		応答	
	フレーム数	顔の数	フレーム数	顔の数
train	9.49	0.957	5.26	0.984
dev	8.67	0.787	4.97	0.763
test	7.27	0.992	4.66	1.13

とが挙げられる。

音声埋め込みについて短い時間で平均を取ることで精度は向上するか？ テキスト・音声・動画の 3 つのマルチモーダルモデルについて、窓幅 w を変更したときの dev セットにおける F1 値を図 3 に示す。w を 128 より小さくして短時間で平均を取ったところ、128 の時よりも精度がわずかではあるが劣化した。これは、w が小さいと言いよどみやポーズなどの情報量の少ないフレームが平均値に悪影響を及ぼす割合が大きくなるため、広い窓幅の方が精度が向上したと考えられる。

検討手法は動画情報を効果的に扱えているか？ 動画情報のみを入力した場合は他の条件に比べて最も精度が低かった。検討手法では、動画の時系列情報は位置埋め込みのみで表現されているため、動画のフレーム間の関係性を捉えきれなかった可能性がある。また、表 3 に示す通り、利用した動画のフレーム数と、OpenCV[5] のカスケード分類器で検出された正面顔のフレーム単位での平均数を調査した。利用した動画は、各参加者のラップトップ PC から撮影されたものであり、本来はその話者の顔を写し続けるはずであるが、検出された顔の数は 1 を下回るものが多く、検出誤りは存在するものの、参加者の正面の顔を正常に撮影できていないフレームが多かったと言える。そのため、支持・不支持の推定に有用と考えられる表情の特徴を捉えきれずノイズになってしまった可能性がある。

5 関連研究

マルチモーダル情報を用いた主張の支持・不支持の推定手法としてディベート対話を用いた研究が挙げられる [2, 3]。しかし、これらは音声とテキストの情報のみを考慮しており、話者の顔動画は考慮できていない。本研究の Encoder の構造は、感情推定の手法として提案された MEmoBERT[11] を参考とした。MEmoBERT は著者の Zhao らが独自に収集した 351 個の映画・テレビ動画を用いて事前学習されているが、初期検討のため本論文では事前学習をしないモデルで評価を行った。

本研究では外部から観察することが比較的容易な「同意・支持」に着目したが、グループディスカッションやカウンセリングの中で話者の納得度合や説得力、コミュニケーション能力などをアノテーション・モデル化する研究も行われている [1, 17, 18]。特に伊藤ら [17] は、言語・音声・動画情報を考慮して会議中の各話者の説得力の高さを他の参加者と比較・推定することを検討しており、異なる話者の情報を比較する点で本研究で検討したモデルと類似しているが、マルチモーダル情報の統合に Transformer Encoder を用いる点で異なる。

6 おわりに

本論文では、対面会議において、発話とその応答のペアに対し、応答が発話を支持するかどうかをマルチモーダル情報を用いて推定することを検討した。先行研究である MEmoBERT を参考に、Transformer Encoder にテキスト・音声・動画の埋め込みを連結して入力するモデルを用いた。AMI コーパスを用いた実験では、テキスト情報のみを考慮した場合が最も精度が高く、検討したモデルではマルチモーダル情報、特に動画情報の考慮が十分にできなかった。ただし、正面の顔を撮影できた動画フレームが少なかったため、データとして動画の持つ情報が少なく、本来支持・不支持の推定に有用な表情の情報を十分に活用できなかった可能性も示唆された。今後は、音声・動画をより効果的に活用する手法について検討する。また、発言を伴わない暗黙的な支持・不支持を推定するため、テキスト情報に依存しないモデル構造についても検討する。

5） https://opencv.org/ 2023/1/5 アクセス

参考文献

[1] 松隈亮太, 岡田将吾, 松本妹子, 中元淳. オンラインカウンセリング対話データコーパスの構築と動作シンクロニーの分析. 人工知能学会 全国大会論文集, pp. 2I6OS9b03–2I6OS9b03, 2022.

[2] Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In **Proceedings of the 8th Workshop on Argument Mining**, pp. 78–88, 2021.

[3] Eleonora Mancini, Federico Ruggeri, Andrea Galassi, and Paolo Torroni. Multimodal argument mining: A case study in political debates. In **Proceedings of the 9th Workshop on Argument Mining**, pp. 158–170, 2022.

[4] 今井芳枝, 雄西智恵美, 板東孝枝. 納得の概念分析. 日本看護研究学会雑誌, 第 39 巻, pp. 73–85, 2016.

[5] 姫野拓未, 嶋田和孝. 複数人議論における発話間の関係を対象とした関係分類. 言語処理学会 第 27 回年次大会 発表論文集, pp. 981–985, 2021.

[6] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In **Proceedings of Machine Learning for Multimodal Interaction**, pp. 28–39, 2006.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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**, pp. 4171–4186, 2019.

[8] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In **Proceedings of Advances in Neural Information Processing Systems**, Vol. 33, pp. 12449–12460, 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[10] Md Kamrul Hasan, James Spann, Masum Hasan, Md Saiful Islam, Kurtis Haut, Rada Mihalcea, and Ehsan Hoque. Hitting your MARQ: Multimodal ARgument quality assessment in long debate video. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6387–6397, 2021.

[11] Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing**, pp. 4703–4707, 2022.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **Proceedings of the 9th International Conference on Learning Representations**, 2021.

[13] Rutger Rienks, Dirk Heylen, and Erik van der Weijden. Argument diagramming of meeting conversations. In **Proceedings of Workshop on Multimodal Multiparty Meeting Processing**, pp. 85–92, 2005.

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 38–45, 2020.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139, pp. 8748–8763, 2021.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In **Proceedings of 3rd International Conference on Learning Representations**, 2015.

[17] 伊藤温志, 坂戸達陽, 中野有紀子, 二瓶芙巳雄, 石井亮, 深山篤, 中村高雄. グループディスカッションにおける説得力推定のためのマルチパーティモデル. 人工知能学会 全国大会論文集, pp. 3H3OS12a02–3H3OS12a02, 2022.

[18] 有岡無敵, 山本賢太, 井上昂治, 河原達也, 中村哲, 吉野幸一郎. 遠隔操作アンドロイドを用いたマルチモーダル説得対話コーパスの収集と分析. 言語処理学会 第 28 回年次大会 発表論文集, pp. 185–190, 2022.