ディスカッションの役割分類に基づいたファシリテーション対話システム

藤後英哲 [1] 菊池英明 [1]　藤倉将平 [2]　清水健吾 [3]
[1] 早稲田大学人間科学研究科　[2] 株式会社サイシキ [3]MNTSQ 株式会社
eitetsu@akane.waseda.jp　fujikura@sai-shiki.com
kengo.shimizu@mntsq.com　kikuchi@waseda.jp

概要

会議では，さまざまな問題が生じることが知られている．その問題を解消するアプローチとしてファシリテータの介入があるが，全ての会議にファシリテータを介入させることは困難である．そのため，本研究では会議における問題を解消するため，ファシリテーション対話システムの開発を行なった．開発した対話システムを評価した結果，ファシリテータとしての適切な発話タイミング・発話生成を行える可能性が示唆された．

1　はじめに

多くの組織において，意思決定のために会議が行われている．しかし，会議では発話の不均衡など，様々な問題が生じることが知られている [1]．会議における問題を解消するアプローチの一つとして，ファシリテータの介入がある [1, 2]．ファシリテータの介入は効果的ではあるが，育成や雇用のコストの側面から全ての会議に介入することが難しい．そこで，ファシリテーションを行う対話システムを指向した研究が行われている [3, 4]．ファシリテーション対話システム研究の多くは，音声情報や画像情報に着目しており，ファシリテータの発話内容に着目した研究は十分ではない．

本研究の目的は，テキストベースの会議において，ファシリテーションを行う対話システムを開発することである．また，開発した対話システムを用いることで，会議における問題が解消されるかについても併せて検証する．

2　関連研究

2.1　ファシリテータ

[1] はファシリテータが解消することのできる会議の問題として，「議論の混乱」「議論の対立」「発言の支配」「決定不能」「議論への不参加」の 5 つを挙げている．

また，[5] はグループディスカッションにおける参加者を，グループが行うタスクに関連する行動を行う「タスク遂行役割」，集団が一体として機能するために必要な行動を行う「グループ調整役割」，参加者自身の目標を達成するための行動をする「個人的役割」に分類している．一般的に，ファシリテータにはディスカッションの目的や目標を達成するとともに，グループがグループとして機能するように調整することが期待されている．そのため，本研究では，ファシリテータの重要な機能としてタスク遂行役割・グループ調整役割に着目し，その中でも「オリエンター」「調和者」「ゲートキーパー」を合わせた役割をファシリテータの定義とする．

2.2　テキストベースの複数人対話システム

テキストベースの複数人対話を対象とした対話システムに関する研究 [6] では，LLM の複数人対話能力について評価している．[6] では，複数人対話データセットを構築し，言語モデルをチューニングすることで，複数人対話における言語モデルの発話生成と発話タイミングについて検証した．検証の結果，既存のプロンプトベースの言語モデルと比較して，発話生成と発話タイミングの個々のタスクにおいて優れた結果を示した．しかし，発話タイミングの精度は 50%程度であり，複数人対話システムを実現できているとはいえない．

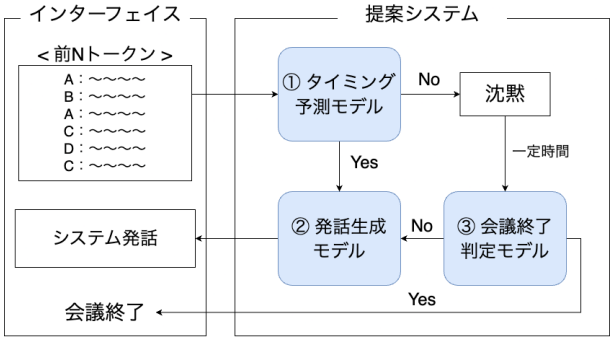

図 1　システム概要図.

2.3　LLM を用いたデータ作成

2022 年に ChatGPT[1)]が発表されて以降，GPT-4[2)]などの LLM を使用したデータ作成を行う研究が増加している [7, 8]．複数人対話においても LLM を用いてデータセットを作成した研究がある [9]．[9] によると，作成されたデータセットは，全ての評価項目において，人手で収集したデータセットより優れているという結果になった．しかし，同研究では複数人の会話データセットを作成するに留まっており，その後の発話生成などのモデル構築には至っていない．そのため本研究では，LLM を使用して複数人会話のデータセットを作成し，発話生成などのモデル構築することで，ファシリテーション対話システムを実現する．

3　提案手法

図 1 は，本研究のシステム概要図である．まず，タイミング予測モデルではシステムが発話をするべきかを 2 値分類する．タイミング予測モデルの結果，発話をするべきと判断された場合は発話を生成し，そうでない場合はシステムは沈黙とする．タイミング予測モデルで沈黙と判断された後，沈黙が一定時間続いた場合，会議終了判定モデルでは会議が終わったかを 2 値分類する．会議が終わったと判断された場合は会議を終了し，そうでなかった場合は発話生成モデルで発話を生成する．

3.1　データセットの作成

学習用のデータセットを構築するため，大規模言語モデル（本研究では GPT-4）を用いて会議の発言録テキスト（以降「会議データ」とする.）を生成した．

1）https://openai.com/blog/chatgpt
2）https://openai.com/research/gpt-4

まず，GPT-4 に与えるプロンプトが適切かを評価するため，20 件の会議データを作成した．20 件の会議データは，プロンプトテンプレートに，事前に用意した会議のトピックと，4 人目の話者をファシリテータと固定し，それ以外の話者に役割をランダムに与えることで作成した．作成した 20 件の会議データに対し，ファシリテーション経験者 4 名に評価を行ってもらった．その際，評価項目は本研究におけるファシリテータの要素である「オリエンター」「調和者」「ゲートキーパー」とし，それぞれ 7 段階評価とした．その結果，各項目ともに平均して 4 以上であったため，本研究でのプロンプトの設定が妥当であると判断した．

表 1　ファシリテーション経験者による評価結果.

項目	平均値	サンプル数
オリエンター	4.9	80
調和者	4.7	80
ゲートキーパー	4.7	80

次に，プロンプトテンプレートに，事前に用意した会議のトピックと参加者の役割を与えることで 485 件のデータセットを作成した．以降のモデル学習では，この 485 件のデータセットを使用した．

3.2　タイミング予測モデル

本研究では，タイミング予測モデルとして BERT を使用した．日本語の事前学習済みモデルとして，東北大 BERT[3)]，京大 DeBERTaV2[4)]，早大 BigBird[5)]を比較した．データセットは，全体の 70%の会議を訓練データ，15%の会議を検証データ，15%の会議をテストデータとし，fine-tuning を行った．モデルの学習では，バッチサイズを 16，学習率を 2.0×10^{-5}，重み減衰率を 0.01 として AdamW[10] を使用し，5 エポック学習を行った．

各モデルのテストデータに対する評価結果を表 2 に示す．正解率と F1 スコアの精度が最も高かった京大 DeBERTaV2 with topic を採用した．

3）https://huggingface.co/cl-tohoku/bert-base-japanese-v3
4）https://huggingface.co/ku-nlp/deberta-v2-large-japanese
5）https://huggingface.co/nlp-waseda/bigbird-base-japanese

表 2 タイミング予測モデルの結果．”with topic”は BERT への入力に会議トピックを含めた学習結果である．

モデル名	Accuracy	F1
東北大 BERT	0.84	0.84
京大 DeBERTaV2	0.66	0.57
早大 BigBird	0.82	0.83
東北大 BERT with topic	0.84	0.84
京大 DeBERTaV2 with topic	**0.86**	**0.86**
早大 BigBird with topic	0.71	0.68

※ *chance rate* = 0.64

3.3 発話生成モデル

本研究では，発話生成モデルとして指示調整言語モデルを使用した．日本語の事前学習済みモデルとして，ELYZA[6)]，stablelm-beta[7)]，StableBeluga[8)]，weblab[9)]，youri[10)]を比較した．データセットは，全体の 80%を訓練データ，10%を検証データ，10%をテストデータとし，LoRA-tuning[11] を行った．モデルの学習では，学習率を 3×10^{-4}，バッチサイズを 16，LoRA の γ を 8，LoRA の α を 16，Drop 率を 0.05 として AdamW を使用し，3 エポック学習を行った．また，チューニングする層は Llama ベースのモデルでは”q_proj”, ”k_proj”, ”v_proj”とし，gpt_neox ベースのモデルは”query_key_value”とした．

各モデルのテストデータに対する評価結果を表 3 に示す．推論速度以外の全ての自動評価指標の精度が最も高かった stablelm-beta を採用した．stablelm-beta が生成した実際の対話例を表 4 に示す．

3.4 会議終了判定モデル

本研究では，会議終了判定モデルとして BERT を使用した．比較する事前学習済みモデル，学習データの分割，学習の設定は全てタイミング予測モデル（3.2 項）と同様にした．

各モデルのテストデータに対する評価結果を表 5 に示す．正解率と F1 スコアの精度が最も高かった東北大 BERT を採用した．

6） https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b-instruct
7） https://huggingface.co/stabilityai/japanese-stablelm-instruct-beta-7b
8） https://huggingface.co/stabilityai/StableBeluga-13B
9） https://huggingface.co/matsuo-lab/weblab-10b-instruction-sft
10） https://huggingface.co/rinna/youri-7b-instruction

表 5 会議終了判定モデルの結果．”with topic”は BERT への入力に会議トピックを含めた学習結果である．

モデル名	Accuracy	F1
東北大 BERT	**0.96**	**0.96**
京大 DeBERTaV2	0.95	0.94
早大 BigBird	0.94	0.94
東北大 BERT with topic	**0.96**	0.95
京大 DeBERTaV2 with topic	0.95	0.94
早大 BigBird with topic	0.94	0.94

※ *chance rate* = 0.93

4 評価実験

4.1 実験目的

以下の 2 点を目的に評価実験を行った．

1. 提案システムがファシリテータとして適切に振る舞えているかを検証する．
2. 提案システムによって，会議における問題が生じた際，その問題が解消することができるかを検証する．

4.2 実験設定

提案システムの有効性を検証するため，被験者 3 名ずつのグループを 8 つ用意し，提案システム介入会議（被験者 3 名と提案システムによる会議），システム非介入会議（被験者 3 名による会議），ベースラインシステム介入会議（被験者 3 名とベースラインシステムによる会議）の 3 つパターンで被験者内実験を行った．すべての会議は Slack を使用し，テキストベースで行われた．

被験者には各パターンでの会議後，対話システムの性能（対話システムの自然性・一貫性・応答速度・ファシリテータとしての発話タイミング・ファシリテータとしての発話内容）と会議における問題の解消（議論の混乱・議論の対立・発言の支配・決定不能・議論への不参加の 5 つの問題が生じた場合に，その問題が適切に解消されたか）を評価してもらった．

また，事後アンケートとして，普段の会議経験を 5 段階で回答してもらった．

4.3 実験結果

対話システムの性能　対話システムの性能におけ

表 3 発話生成モデルの結果.

モデル名	BLEU	BERTScore	ROUGE-1	ROUGE-2	ROUGE-l	METEOR	推論速度 (秒)
ELYZA	35.83	**0.83**	63.55	45.04	57.46	54.80	16.07
stablelm-beta	**36.56**	**0.83**	**64.17**	**45.88**	**58.34**	**55.24**	15.27
StableBeluga	35.63	**0.83**	63.47	44.87	57.41	54.93	21.24
weblab	33.68	0.82	62.07	43.19	55.92	52.48	**14.04**
youri	34.87	0.82	62.95	44.14	57.20	53.70	15.69

表 4 stablelm-beta の対話例．D が生成された発話である．

A	実社会での，金融周りでの詐欺の被害や機会損失など．
C	それもそうです
D	C さんは貯蓄と投資について，A さんは金融の現実的な問題点を指摘していますね．B さん，何か意見はありますか？
A	B さん，早くしてください．
B	なるほど，確かにそのような点が挙げられるのですね．リスクを恐れて投資に回さないといった保守的な人が日本に多いからのように感じます．
D	A さん，B さんに対するプレッシャーは控えてください．B さん，何か意見はありますか？

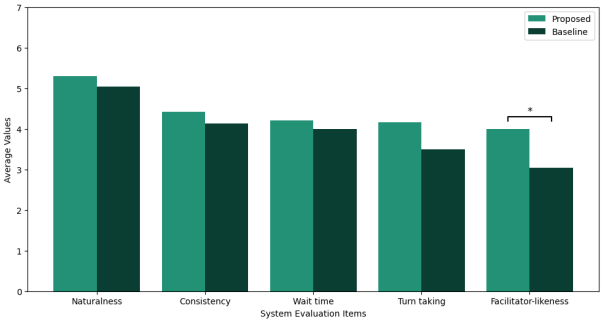

図 2 システム性能における評価結果．左から順に自然性，一貫性，応答速度，ファシリテータとしての発話タイミング，ファシリテータとしての発話内容を表す．

る評価結果を図 2 に示す．正規性が確認された評価項目については対応のある t 検定を行い，そうでない評価項目については Wilcoxon の符号順位検定を行った．検定の結果，発話タイミングでは有意傾向が確認され ($t(23) = -1.46$, $p = .079$)，発話内容では有意差が確認された（$W = 60.50$, $p = .028$）．したがって，提案システムは，ベースラインシステムと比較して，適切なタイミングで発話をしている可能性が示唆され，適切な発話を生成していることが確認された．

会議における問題の解消　会議における問題の解消に関する評価結果を図 3 に示す．正規性が確認された評価項目については分散分析，そうでない評価項目については Kruskal-Wallis 検定を行った．検定の結果，議論の混乱・議論の対立・発言の支配・決定不能・議論への不参加の全ての問題において，問題を解消していることが確認されなかった（$H(2) = 1.36$, $p = .506$; $F(2, 63) = 1.33$, $p = .273$; $H(2) = 0.54$, $p = .763$; $F(2, 64) = 1.93$, $p = .153$; $H(2) = 0.19$, $p = .909$）．

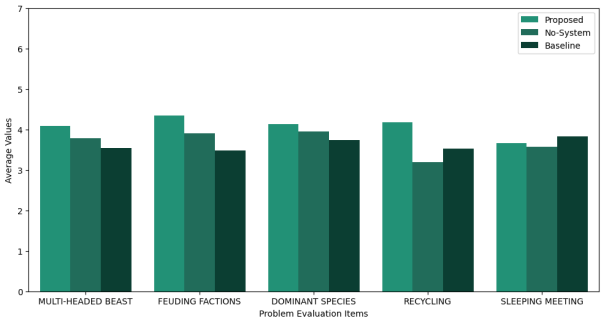

図 3 会議における問題の解消に関する評価結果．左から順に議論の混乱，議論の対立，発言の支配，決定不能，議論への不参加を表す．

続いて，提案システムのファシリテータとしての発話内容の評価値が 7 段階中 5 以上だった被験者について分析した．その結果，議論の対立では有意差が確認された（$F(2, 29) = 7.22$, $p = .003$）．

最後に，事後アンケートをもとに，会議経験の浅い被験者について分析した．その結果，議論の対立では有意差が確認された（$F(2, 23) = 4.58$, $p = .021$）．

5 おわりに

本研究では，LLM を用いて会議のデータセットを作成し，ファシリテーション対話システムを開発した．開発したシステムでは，適切な発話タイミングと発話生成が行える可能性が示唆された．また，全ての被験者の評価では，全ての問題において，提案システムは，ベースラインシステムと人間だけの会議と比較し，問題を解消することが確認されなかった．しかし，発話が適切に生成されたと回答した被験者と，会議の経験が浅い被験者による評価では，議論の対立の解消に効果があることが示された．

参考文献

[1] Frances Westley and James A Waters. Group facilitation skills for managers. **Management Education and Development**, Vol. 19, No. 2, pp. 134–143, 1988.

[2] Liam Bannon, Mike Robinson, and Kjeld Schmidt. Proceedings of the Second European Conference on Computer-Supported Cooperative Work: ECSCW'91. Springer Science & Business Media, 2012.

[3] Tsukasa Shiota, Takashi Yamamura, and Kazutaka Shimada. Analysis of facilitators' behaviors in multi-party conversations for constructing a digital facilitator system. In **Collaboration Technologies and Social Computing: 10th International Conference, CollabTech 2018, Costa de Caparica, Portugal, September 5-7, 2018, Proceedings 10**, pp. 145–158. Springer, 2018.

[4] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie, and Tetsunori Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. **Computer Speech & Language**, Vol. 33, No. 1, pp. 1–24, 2015.

[5] Kenneth D Benne and Paul Sheats. Functional roles of group members. **Journal of social issues**, Vol. 4, No. 2, pp. 41–49, 1948.

[6] Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models. **arXiv preprint arXiv:2304.13835**, 2023.

[7] Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. Targen: Targeted data generation with large language models. **arXiv preprint arXiv:2310.17876**, 2023.

[8] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. **arXiv preprint arXiv:2310.07849**, 2023.

[9] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. Places: Prompting language models for social conversation synthesis. **arXiv preprint arXiv:2302.03269**, 2023.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. **arXiv preprint arXiv:1711.05101**, 2019.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. **arXiv preprint arXiv:2106.09685**, 2021.