

# 複数人同時対話における引継ぎのための 抽象型の対話形式の対話要約の実証

山下紗苗<sup>1</sup> 望月翔太<sup>1</sup> 星牟禮健也<sup>2</sup> 馬場淳<sup>2</sup> 窪田智徳<sup>3</sup> 小川浩平<sup>3</sup> 東中竜一郎<sup>1</sup>

<sup>1</sup> 名古屋大学大学院情報学研究科 <sup>2</sup> 株式会社サイバーエージェント

<sup>3</sup> 名古屋大学大学院工学研究科

yamashita.sanae.w7@es.mail.nagoya-u.ac.jp mochizuki.shota@nagoya-u.jp

{hoshimure\_kenya,baba\_jun}@cyberagent.co.jp

{kubota,k-ogawa}@nuee.nagoya-u.ac.jp higashinaka@i.nagoya-u.ac.jp

## 概要

対話システムと人間が協調して対話サービスを提供する枠組みとして、複数人同時対話が提案されている。この枠組みでは、対話に問題が生じた際に、人間のオペレータが、対話要約を基に進行中の対話を引継ぐ。我々はこれまでに、抽象型の対話形式要約 (Dialogue Format Summary; DFS) を提案し、複数人同時対話における有用性を調査してきた。しかし、先行研究では、オペレータは進行中の各対話を把握する時間的な余裕があり、負荷が高い条件で発揮されうる DFS の利点が表れなかった可能性がある。そこで、本研究では、低負荷条件と高負荷条件を設定した大規模な実証実験を行い、高負荷条件において DFS が対話の引継ぎに適していることを示した。

## 1 はじめに

近年、自律的な対話システムの開発が発展している [1, 2, 3]。この発展を背景として、対話システムと人間が協調して対話サービスを提供する、複数人同時対話の枠組みが提案されている [4, 5]。複数人同時対話では、図 1 のように、対話システムが複数のユーザとの対話を並列に進行し、人間のオペレータがそれらを監視する。問題が生じた際には、オペレータが対話を引継ぎ、対話システムの代わりにユーザと対話することで、限られた人数のオペレータでも多数のユーザに対話サービスを提供可能である。複数人同時対話における引継ぎでは、オペレータは多くの対話のそれぞれについて、対話の流れや重要な点を短時間で把握する必要があるため、対話内容を整理した対話要約が重要となる [6]。

複数人同時対話における引継ぎでは、従来、文章

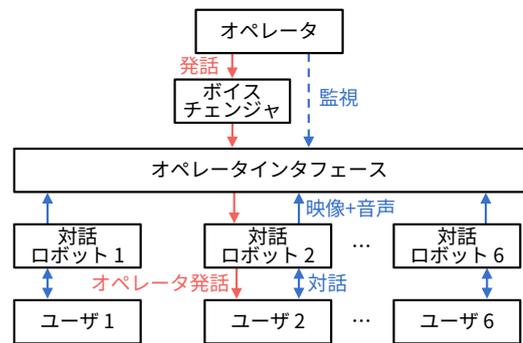


図 1 複数人同時対話引継ぎシステムの概念図。

形式 (Text Format Summary; TFS) [7] やキーワード形式 [8, 9] の対話要約が用いられてきた。これらに対し、我々は、先行研究において、抽象型の対話形式の対話要約 (Dialogue Format Summary; DFS) を提案した [10, 11]。DFS は、対話の流れや発話のニュアンスを保ったまま、誰が何を言ったかが分かる形で、長い対話を短い対話として表現したものである。

我々は、先行研究において、オペレータが対話要約を基に対話を引継ぐという実証実験を実施し、DFS と TFS を比較した [12]。その際、オペレータは DFS に対してポジティブな意見を述べていたものの、対話要約の有用性に関するアンケートでは、DFS と TFS の評価は同程度であり、有意差は見られなかった。その理由として、先行研究では 1 人のオペレータが 3 台の対話ロボットの対話を監視する体制であったため、オペレータは進行中の各対話の内容を把握する時間的な余裕があり、高負荷条件で発揮されうる DFS の利点が表れなかった可能性がある。

そこで、本研究では、オペレータの負荷が十分に高い状況において、DFS が対話の引継ぎに効果的であるかを、実フィールドにおける大規模な実証実験

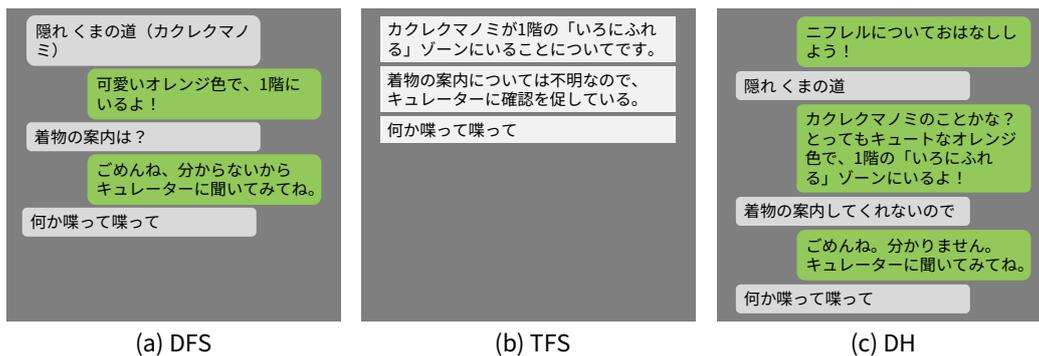


図2 対話要約とその表示の例. DFS と DH については, ユーザの発言は灰色の吹き出し, 対話ロボットの発言は緑色の吹き出しで示す. 略称は 2.2 節の定義を参照.



図3 オペレータインタフェース. 各対話ロボットのカメラ映像, 音声, および, 対話要約が提示される.

により明らかにする. 具体的には, 低負荷条件と高負荷条件を設定し, 両条件を比較することで, 高負荷条件における DFS の効果を実証する.

## 2 実証実験

2024年11月25日から12月22日までの28日間, 水族館と動物園を複合した施設であるニフレルにおいて, 2つのフロアにわたって複数台の対話ロボットを設置し, オペレータの引継ぎにより複数人同時対話を行う実証実験を実施した. なお, 実施にあたっては, 大阪大学において, 収集したデータの取扱いや個人情報の扱いの倫理面について審査を経ている (承認番号: R-1-5-17).

### 2.1 複数人同時対話引継ぎシステム

図1に, 本実験で使用した複数人同時対話引継ぎシステム [13] の概念図を示す. 本実験は, 対話ロボットがユーザと自律的に案内対話を実施し, その様子を, 別室に控える1人のオペレータがオペレータインタフェースに提示される情報を介して監視し, 必要に応じて対話を引継ぎ設定のもとで実施した. 対話ロボットの台数は, 負荷が低いと考えられる4台 (低負荷条件) と, 負荷が高いと考えられる

6台 (高負荷条件) の2条件を用意した.

対話ロボットは施設内に配置され, ユーザに対して生き物や施設の案内を行った. 発言の生成には OpenAI API<sup>1)</sup> の gpt-4o-2024-08-06 を使用し, タスクの指示, 案内対象の情報 (魚の生態など), 対話履歴からなるプロンプトを用いた.

オペレータインタフェースには, 図3のように, 各対話ロボットのカメラ映像, 音声, および, 対話要約を提示した. 引継ぎ時には, オペレータの音声をボイスチェンジャーで変換し, 対話ロボットを通じて出力した.

### 2.2 対話要約

対話要約として, 以下に示す3種類を設定し, 図2のようにオペレータインタフェースに提示した. オペレータは, 1日に1種類の対話要約を使用した.

**DFS (Dialogue Format Summary)** 提案手法である, 抽象型の対話形式の対話要約 [10, 11].

**TFS (Text Format Summary)** 抽象型の文章形式の対話要約. 対話要約として一般的に用いられてきた形式であり, 対話の内容を第三者の観点から俯瞰的かつ包括的に表現したもの.

**DH (Dialogue History)** 音声認識結果. 他2つの対話要約の効果を検証するために使用した.

DFS と TFS は, 我々の先行研究 [14] に基づき作成した. モデルは gpt-4o-2024-08-06 を使用し, タスクの指示, 対話要約の定義, 対話要約の Few-shot [1, 15], DH からなるプロンプトを用いた.

### 2.3 オペレータによる対話の引継ぎ

オペレータは, 事前に, オペレータインタフェース上で引継ぎを訓練した. その上で, 実証実験にお

1) <https://platform.openai.com/docs/models/gpt-4o>

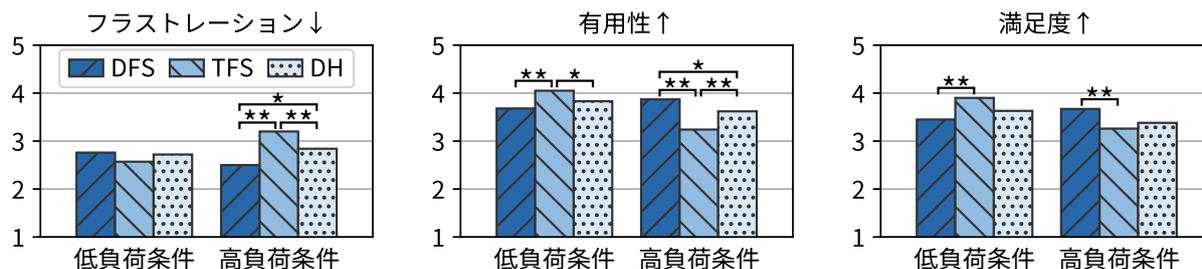


図4 アンケート結果. 丸は平均値を示す. \*\*と\*は, Mann-Whitney の U 検定 (Holm 法で補正) で 3 群を比較した結果, 有意水準 1% と 5% で有意差が見られたことを示す.

表1 実証実験の統計量. 要約率は, 対話要約の文字数を DH の文字数で割った値である.

	引継ぎ回数			対話要約の文字数 (要約率)		
	DFS	TFS	DH	DFS	TFS	DH
低負荷条件	92	114	109	60 (65%)	50 (54%)	93
高負荷条件	132	111	96	68 (61%)	58 (52%)	111

いては, 対話を常に監視し, 対話ロボットが答えられない質問などへの対応, 停滞した対話への対応, 対話ロボットの誤答や聞き間違いへの対応, ユーザーがジェスチャをしたり物を見せたりするといった, 対話ロボットが対処できない非言語行動への対応が必要な場合に, 対話を引継ぐよう指示された.

また, オペレータは引継ぎ毎に以下を記録した.

- アンケート (5 段階リッカート尺度)
  - 対話要約の有用性 (やり取りの情報は引継ぎに有用でしたか?)
  - フラストレーション (やり取りの情報を用いて対話内容を把握するのにフラストレーションを感じましたか?)
  - オペレータが思うユーザーの満足度 (ユーザーはどのくらい満足したと思いますか?)
- 引継ぎ理由 (自由記述)

さらに, 1 日の作業終了後と実験全体の終了後に, フィードバックとして, 対話要約に関する良かった点や改善点を自由記述で回答した.

## 2.4 統計量

実験を通して, 男性 1 人, 女性 3 人の計 4 人のオペレータが対話の引継ぎを行った. オペレータは, 最も少ない場合で 3 日, 最も多い場合で 9 日参加した. 低負荷条件と高負荷条件は 14 日ずつ実施した. 3 種類の対話要約はほぼ同じ日数ずつ使用した.

表 1 に, 実証実験の統計量を示す. 合計で 654 回の引継ぎが行われ, どの対話要約を用いた場合でも 100 回前後の引継ぎが行われた. DH の文字数は 100

文字程度であり, DFS と TFS の文字数はその 5 割から 6 割程度であった.

## 3 DFS の効果の検証

低負荷条件と高負荷条件を比較し, その上で, DFS が高負荷条件において効果的かを検証した.

### 3.1 アンケート結果の分析

図 4 に, オペレータのアンケート結果を示す.

フラストレーション (低い方が望ましい) について, 低負荷条件では, Mann-Whitney の U 検定 (Holm 法で補正) で 3 群を比較したところ, どの対話要約間でも有意差は見られなかった. 一方で, 高負荷条件では, DFS のフラストレーションが最も低く, TFS のフラストレーションが最も高かった. これは, TFS が対話の流れを把握しにくいためだと考えられる. 有意差は, DFS-TFS 間と TFS-DH 間に有意水準 1% で, DFS-DH 間に有意水準 5% で見られた.

低負荷条件では, TFS の有用性が最も高く, DFS の有用性が最も低かった. 有意差は, DFS-TFS 間に有意水準 1% で, TFS-DH 間に有意水準 5% で見られた. 一方で, 高負荷条件では, DFS の有用性が最も高く, TFS の有用性が最も低かった. DFS は, 対話の流れに加えて, 元の発話のニュアンス (口調など) を保っており, 状況の理解やどのように話せばよいかの参考になるため, 有用性が高かったと考えられる. 有意差は, DFS-TFS 間と TFS-DH 間に有意水準 1% で, DFS-DH 間に有意水準 5% で見られた.

低負荷条件では, TFS の満足度が最も高く, DFS の満足度が最も低かった. 有意差は, DFS-TFS 間に有意水準 1% で見られた. 一方で, 高負荷条件では, DFS の満足度が最も高く, TFS の満足度が最も低かった. DFS は対話の流れを短時間で把握しやすく, それによりオペレータが迅速に対応できたため, 満足度が高まったと考えられる. 有意差は, DFS-TFS 間に有意水準 1% で見られた.

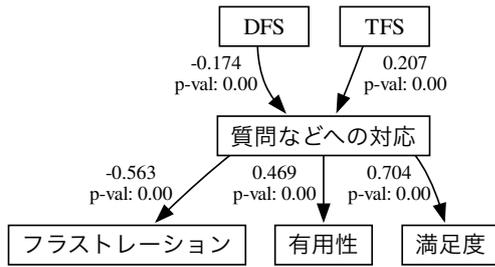


図5 低負荷条件のSEMの結果.

まとめると、低負荷条件において、TFSはDFSと比較して、有用性と満足度で優れていることが示された。高負荷条件においては、低負荷条件とは反対の傾向が見られ、DFSはTFSと比較して、全項目で優れていることが示された。DHはいずれの条件においてもDFSとTFSの中間の評価を得た。

### 3.2 対話要約の効果のモデル化

構造方程式モデリング (Structural Equation Modeling; SEM) [16] を用いて、対話要約の種類、引継ぎ理由、および、アンケート結果の関係を明らかにした。

SEMの構造としては、階層関係を持つ3層を仮定し、1層目に対話要約の種類、2層目に引継ぎ理由、3層目にアンケートの各項目の結果を配置した。対話要約の種類は、DFS、TFS、DHの3変数をワンホットで表現した。引継ぎ理由は、2.3節でオペレータに指示した4つの場面（質問などへの対応、停滞した対話への対応、対話ロボットの誤答や聞き間違いへの対応、および、非言語行動への対応）に従って著者の一人が目視で分類し、これらの4変数をワンホットで表現した。アンケート結果は、フラストレーション、有用性、満足度について、それぞれの5段階の主観評価の値を使用した。

SEMは、低負荷条件と高負荷条件のそれぞれで実施した。具体的には、文献[17]に基づき、全ての変数を用いて初期モデルを構築した後、統計的に有意でないパスを削除する処理を反復し、全ての残存パスが有意となるモデルを採用した。

図5に、低負荷条件の結果を示す。低負荷条件では、TFS使用時に質問などへの対応を理由とした引継ぎが多く、それに伴い、アンケート結果が向上する関係が確認された。低負荷条件では、オペレータは、オペレータインタフェース上で、特定の対話の映像を拡大表示する画面を積極的に使用しており、対話要約に加えて映像や音声で直接確認する余裕があったと考えられる。そのため、質問などへの対応が必要であることを早期に認識できた可能性がある。

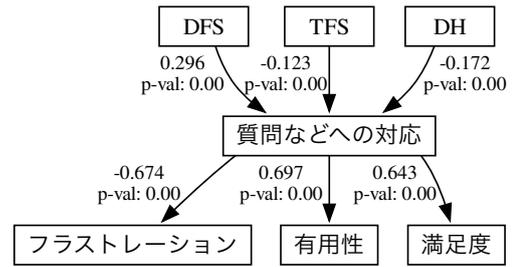


図6 高負荷条件のSEMの結果.

る。このような状況では、TFSは、映像や音声で状況を把握した上で話題を簡潔に再確認する用途に適しており、映像や音声と相補的に機能した結果、主観評価が改善したと考えられる。一方で、この状況では、誰が何を言ったかはDFSに頼らなくても理解できるため、DFSの利点は相対的に小さくなったと考えられる。

図6に、高負荷条件の結果を示す。高負荷条件では、DFS使用時に質問などへの対応を理由とした引継ぎが多く、それに伴い、フラストレーションの低下や、満足度と有用性の向上が確認された。DFSは、誰が何を言ったかを簡潔に表現する形式であり、短時間で対話の流れを把握しやすい。そのため、高負荷で即時的な判断を要する、質問などへの対応においても状況を把握しやすく、引継ぎ時の負担が軽減された結果、主観評価の改善につながったと考えられる。一方で、TFSやDH使用時には、質問などへの対応以外の理由で引継がれる傾向が強くなり、主観評価も改善しなかった。

## 4 おわりに

本研究では、複数人同時対話において、オペレータの負荷の大きさについて、低負荷条件と高負荷条件を設定した大規模な実証実験を行い、DFSの効果を実証した。その結果、高負荷条件では、DFSが対話の引継ぎに適していることが示された。

本研究の限界としては、まず、DFSの効果が案内対話のみで実証された点が挙げられる。今後は、より複雑でタスク指向性の高い対話においても、同様の結果が得られるかを検証したい。また、本研究の満足度はオペレータの推定に基づくものであり、ユーザの主観を厳密に反映しているとは限らない。そのため、対話時間や離脱率といった行動指標を用いたり、対話中の表情や動作といったマルチモーダルな行動からユーザの満足度を推定したりすることで、ユーザの満足度をより適切に評価し、最大化するような対話要約を実現していきたい。

## 謝辞

本研究は、JST ムーンショット型研究開発事業、JPMJMS2011 の支援を受けたものである。実証実験の実施にご協力いただいたニフレルのスタッフの皆様感謝する。

## 参考文献

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **Proceedings of Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901, 2020.
- [2] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. Blenderbot 3: A deployed conversational agent that continually learns to responsibly engage. **arXiv preprint arXiv:2208.03188**, 2022.
- [3] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.
- [4] Tatsuya Kawahara, Naoyuki Muramatsu, Kenta Yamamoto, Divesh Lala, and Koji Inoue. Semi-autonomous avatar enabling unconstrained parallel conversations—seamless hybrid of WOZ and autonomous dialogue systems—. **Advanced Robotics**, pp. 1–7, 2021.
- [5] Tatsuya Kawahara, Hiroshi Saruwatari, Ryuichiro Higashinaka, Kazunori Komatani, and Akinobu Lee. Spoken dialogue technology for semi-autonomous cybernetic avatars. In Hiroshi Ishiguro, Fuki Ueno, and Eiki Tachibana, editors, **Cybernetic Avatar**, pp. 71–105, 2024.
- [6] Mathis Poser, Sukhpreet Singh, and Eva Bittner. Hybrid service recovery: Design for seamless inquiry handovers between conversational agents and human service agents. In **Proceedings of the 54th Hawaii International Conference on System Sciences**, pp. 1181–1190, 2021.
- [7] Divesh Lala, Koji Inoue, Haruki Kawai, Zi Haur Pang, Mikey Elmers, and Tatsuya Kawahara. Development and evaluation of a semi-autonomous parallel attentive listening system. In **Proceedings of the 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference**, pp. 1–6, 2024.
- [8] Haruki Kawai, Yusuke Muraki, Kenta Yamamoto, Divesh Lala, Koji Inoue, and Tatsuya Kawahara. Simultaneous job interview system using multiple semi-autonomous agents. In **Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 107–110, 2022.
- [9] 川崎一賢, 小川浩平. 複数エージェントの遠隔操作システムの実現に向けた, 多地点情報の適切な要約表示方法. 第 36 回人工知能学会全国大会論文集, Vol. JSAI2022, p. 2FIGS904, 2022.
- [10] Sanae Yamashita and Ryuichiro Higashinaka. Clarifying characteristics of dialogue summary in dialogue format. In **The 13th International Workshop on Spoken Dialogue Systems Technology**, 2023.
- [11] 山下紗苗, 東中竜一郎. 対話形式の抽象型対話要約の提案とその特徴および対話の引継ぎに及ぼす負荷の分析. **自然言語処理**, Vol. 33, No. 1, 2026.
- [12] Sanae Yamashita, Shota Mochizuki, Kazuyoshi Kawasaki, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. Investigating the effects of dialogue summarization on intervention in human-system collaborative dialogue. In **Proceedings of the 11th International Conference on Human-Agent Interaction**, pp. 316–324, 2023.
- [13] Shota Mochizuki, Sanae Yamashita, Kazuyoshi Kawasaki, Reiko Yuasa, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. Investigating the intervention in parallel conversations. In **Proceedings of the 11th International Conference on Human-Agent Interaction**, pp. 30–38, 2023.
- [14] Sanae Yamashita and Ryuichiro Higashinaka. Multifaceted evaluation of automatically generated dialogue format summary. In **Proceedings of the 14th International Workshop on Spoken Dialogue Systems Technology**, 2024.
- [15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 1107–1128, 2024.
- [16] Yi Fan, Jiquan Chen, Gabriela Shirkey, Ranjeet John, Susie R Wu, Hogeun Park, and Changliang Shao. Applications of structural equation modeling (SEM) in ecological studies: an updated review. **Ecological processes**, Vol. 5, No. 1, 2016.
- [17] Ao Guo, Shota Mochizuki, Sanae Yamashita, Kenya Hoshimure, Jun Baba, and Ryuichiro Higashinaka. A comparative study of human-operated and AI-driven guidance with a teleoperated mobile robot. In **Proceedings of the International Joint Conference on Natural Language Processing & Asia-Pacific Chapter of the Association for Computational Linguistics**, 2025.