# 雑談中の発話と文脈から話者情報を抽出する LLM の能力に関する検証

連慎治 [1]　竹下昌志 [1]　伊藤敏彦 [2]
[1] 北海道大学 大学院情報科学院　[2] 北海道大学 大学院情報科学研究院
shinjimuraji@ist.hokudai.ac.jp　takeshita.masashi.68@gmail.com
t-itoh@media.eng.hokudai.ac.jp

## 概要

近年，LLM を用いた雑談システムがユーザやシステム自身の個性を応答生成に活用することで対話の質を向上させる研究が盛んである．しかし，発話者の個性を応答生成に用いるためには対話から発話者の情報を抽出する必要があるが，その抽出手法は十分に研究されていない．また，対話から話者の情報を抽出するためには文脈の考慮も必要だと考えられるが，話者情報抽出における文脈の影響はまだ十分に検証されていない．そこで，本研究では現在最先端とされている LLM がどの程度，雑談発話から文脈を考慮して発話者の情報を抽出できるかを検証する．実験の結果から，最先端の LLM が話者の情報を抽出する際に文脈を十分に考慮できていないことが示された．

## 1　はじめに

近年，大規模言語モデル (LLM) を用いた多くの雑談システムが研究されている．これまで言語モデルを用いた雑談システムの研究において，発話者の個性を自然言語による文の形式[1)]で利用することで，システム自身の発話の一貫性が上昇したり [1]，対話の魅力が上昇することが報告されている [2]．しかし，対話に活用するためには何らかの形で話者の情報を用意する必要があり，事前に用意する方法と対話から抽出する方法が考えられる．事前に用意する方法では対話の中で新しく話された話者の情報に対応できず，対話から抽出する方法はまだ十分に研究されていない．

また，人間は雑談中の発話から発話者の情報を抽出する際に，それまで話された内容を使って目標発話で直接話されていない内容を補完しているはずである．本稿では，対話中の目標発話より前の発話を文脈と呼び，人間と LLM がどの程度文脈を用いて話者情報の抽出を行っているか調べる．

よって，本研究では予備実験として人間の場合はどの程度文脈中の内容を用いて話者情報を抽出するのか確認した後，最先端の LLM が雑談中の発話から話者の情報を抽出する際にどんな手法が有効なのか，どの程度文脈を考慮できているのかを検証する．

## 2　予備実験

予備実験では，人間が目標発話から話者の情報を抽出し，さらに人手で抽出された情報が目標発話だけから分かる内容かどうかを分析する．雑談のデータセットとしては，JPersonaChat データセット [3] を用いる．全データ 61,781 発話からランダムに目標発話を 400 発話をサンプリングし，目標発話とその対話履歴をアノテータに見せる．そして，アノテータが目標発話を聞いたときに想像する発話者の情報を抽出させた．さらに，抽出された話者情報と目標発話のペアを見て目標発話のみから分かる情報か，あるいは何発話前の情報と組み合わせて抽出されたものかをアノテーションした．アノテーションは著者の一人が行った．結果，19%の発話で話者情報の抽出の際に文脈が必要であった．

## 3　関連研究

雑談対話から話者の情報を抽出する研究はいくつか先行研究がある．Xu ら [2] は，発話が行われるごとに対話履歴を要約してデータベースに保存し，応答生成時の入力として対話履歴のみではなく関連する要約文を検索補強して用いることでより相手と話した内容に沿った魅力的な対話を行うことができることを示した．このことから，Xu らは対話の要約

1)　「私は学生です。」のような自己紹介型の文が使われることが多い．ペルソナ文とも呼ばれる．

**表 1**　人間による話者情報抽出の例

| | |
|---|---|
| A の発話 | 大みそかで 20 歳になります。やぎ座ですね。 |
| B の発話 | お若いんですね。わたしはアパレル系で働いているんですけど、あなたは大学生ですか？ |
| A の最後の発話 (目標発話) | そうです。もうすぐ卒業なので将来どうしようかと思って。占いとか好きなんですけど、職業にするには難しいかなと。アパレル系ってどんなことを？ |
| 話者情報 | 大学生。もうすぐ大学を卒業する。占いが好き。就職について悩んでいる。 |

A の最後の発話に含まれる A 自身に関する情報を答えて下さい。
以下の A の最後の発話で新しく分かる A 自身に関する情報を箇条書きでたくさん答えてください。
A 自身に関する情報が無い場合は「なし。」と答えてください。
複数の内容になるときはできるだけ細かく分けて、ひとつずつ答えてください。
指示語や代名詞は使わず、答える情報はそれぞれが具体的で独立した文にしてください。

**図 1**　指示文プロンプトの形式

が話者情報の拡張として機能すると主張している．また，Xu らは要約を保存する動機として，それぞれの発話を個別に保存してしまうと文脈依存な情報が多く保存されてしまい，保存された情報の検索や応答生成の際に負担が大きくなることを挙げている．我々の動機も同様であり，文脈に依存しない話者情報の抽出を目指すが，対話履歴の要約は行わない．それには二つ理由があり，ひとつは雑談において発話から得られる話者の情報は多いほどよく，単純な要約では残したい情報が削れてしまう可能性がある．もうひとつは，発話から推論を通して得られる話者の情報はしばしば単純な要約とは異なるため，我々は推論を必要とするような話者の情報も抽出したいからである．また，Xu らは Transformer のエンコーダデコーダモデルを用いて対話履歴から要約を作成したが，作成された要約に対しては分析を行っておらず，データセットでの perplexity を測るに留まっている．さらに，川本ら [4] や Ribeiro ら [5] も発話から話者情報の抽出を行っているが，文脈の影響は調査されていない．

## 4　本研究で扱うタスク

本研究では，二名の人間同士のテキストの雑談から発話者の情報を文の形式で抽出することを考える．発話する二名を A と B と仮定し，交替で発話を行ったものとする．最後に A が発話した内容を聞いたときに，人間が想定する A の情報を文で書き出すというタスクで LLM の話者情報を抽出する能力を検証する．人手での話者情報の抽出例を表 1 に示す．この例では，4 つの話者情報が抽出されている．以下の点に注意して抽出を行う．

- それまでの対話で判明している必要な情報が欠けていない文で抽出する
- 常識的なものは抽出しない
- 目標発話と関係のない情報は抽出しない

1 点目に関しては，表 1 の二つ目の話者情報が必要な情報を含んでいる例に該当する．ここで「大学を」という情報が抜け落ちて「もうすぐ卒業する」という文を抽出して保存しても，内容が文脈依存的になり後の応答生成で活用が難しくなる．3 点目に関しては，表 1 における，目標発話とは関係のない「誕生日が大晦日」や「19 歳」，「やぎ座」が該当する．

また，本研究では履歴として用いる文脈の長さには制限をかける．予備実験から，今回用いるデータセットでは履歴として 2 発話分遡れば 98.25%の発話で話者の情報が抽出できることが分かったため，本研究では目標発話に対して表 1 のように文脈として対話履歴の発話を最大 2 発話付与して話者情報の抽出を行う．

## 5　LLM による話者情報の抽出手法

本研究では日本語の雑談対話から話者情報を抽出した先行研究 [4] に倣って LLM にプロンプトを与えることで話者情報の抽出を行うが，LLM で抽出する際の条件を変えて精度を調査した．変更する条件は与える履歴発話の数，プロンプトの内容，温度とした．条件の相互作用で性能が変わる可能性もあるが，全ての組み合わせを調べることは困難なため，今回は文脈の条件以外はそれぞれが独立していると仮定しベースとなる抽出システムの一部を変更することで実験を行った．ベースとする指示文を図 1 に示す．

### 5.1　与える履歴発話数の比較

予測時に与える発話履歴数による精度の違いを調査した．人間の場合は発話履歴が多いほど具体的で正確な話者情報を抽出できると考えられるが，LLM

がどの程度文脈を考慮して話者情報を抽出しているかは未知であるため，文脈として対話履歴を与えないもの，1 発話分の対話履歴を与えるもの，2 発話分の対話履歴を与えるものの 3 つを比較した．

### 5.2 温度の比較

LLM は温度の設定で予測を行う際の出力の多様性を変更できる．本研究ではより確実な出力を行うためにベースの温度を T=0 と設定したが，より高い温度を設定することで多様な話者情報が得られる可能性があるため T=1 で抽出を行い比較を行った．

### 5.3 プロンプトの比較

プロンプトによって抽出できる話者情報が変わる可能性がある．本研究では，先行研究 [4] で用いられているような 5 例のタスク例示を行うプロンプトをベースとして，以下の 7 つの条件を変えて効果を確認した．また，ベースの例はデータセットの開発用に分けた対話から手動で作成し，いくつか試したものの中から抽出精度が最も良かった 5 例を用いた．

**タスク例示の有無の影響**　例を用いない方が抽出性能が上がる可能性もあるため，指示文のみでの抽出を実行した．

**指示内容の影響**　ベースでは，LLM が文脈情報として与える別発話から話者情報を抽出するのを避けるために図 1 のように「A の最後の発話で新しく分かる」と記述した．人間の場合は「A の最後の発話で述べられている」と指示すれば最後の発話と無関係な話者情報の抽出は防げると考えられる．LLM において，このような人間的な指示に変更した場合の影響を調査した．

**対象発話の与え方の影響**　ベースとするプロンプトでは，発話を与える際に表 1 のように文脈中の発話を「A の発話」「B の発話」とし，目標発話を「A の最後の発話」と記述した．この記述が LLM にとってどの程度効果的であるかは不明である．よって，文脈と目標発話を LLM がより区別できるようにそれぞれ「A の発話 (C)」「B の発話 (C)」「A の最後の発話 (T)」とマーキングしたものと精度の違いを調査した．

**指示文の位置の影響**　ベースでは，指示文を記述した後に例示を与えたが，例示の後に指示文を置いた場合の精度の違いを調査した．

**タスク例示の選択の影響**　ベースでは手動で選択した抽出するべき話者情報がある場合のタスク例示を用いたが，実際の雑談には話者情報が無い発話も存在する．そのため，雑談中の話者情報がない発話の割合に合わせて例示の割合を調節すると精度が上がる可能性がある．予備実験より，21.75%(87/400) の発話に話者情報が無いことが分かっているので，5 例の内 1 例を話者情報が無い例に変更し精度を調査した．

**負例の有無の影響**　ベースでは正解例を与えているが，失敗例を与えることで精度が上がる可能性があるため，正例 5 例の前に負例を 1 件追加した場合の精度を調査した．

**話者情報の記述形式の影響**　ベースでは表 1 のような人手で抽出した際の話者情報を与えているが，先行研究では話者情報を「私は大学生です」のような自己紹介の形で与えることが多い．また，データを A と B の雑談としてプロンプトに与えているため，自己紹介ではなく「A は大学生です」のような A の紹介として与えることで精度が変わる可能性がある．よって，例示の際の話者情報を自己紹介形式，他者紹介形式に変更した場合の精度を調査した．

## 6 実験設定

LLM としては，最先端のものとして GPT-4 のスナップショット (gpt-4-0613) [2] を用いた．実験はそれぞれの条件で 3 回づつ行い精度の中央値をスコアとした．データセットとしては JPersonaChat(JPC) を用いた．データセット 5000 対話から例示の選択時に用いた 10 対話を除いた 4990 対話からランダムな目標発話 100 件に対し，LLM で話者情報をできるだけたくさん抽出した．抽出された話者情報と目標発話とその文脈をアノテーターに見せ，評価させた．アノテーターには，文脈+目標発話から抽出された文が話者情報として断定できる場合は o, 抽出された話者情報が発話者の情報として可能性が高い場合を p, 間違っているものや余分な情報が含まれているもの，日本語として成り立たないもの，目標発話と関わりがないもの，必要な情報が抜けているものは x とする 3 値で判断してもらった．アノテーションは 20 代男性 1 名に依頼した．

2） https://platform.openai.com/docs/models

**表 2**　話者情報の抽出精度

| 条件 | 文脈無し | 対話履歴 1 発話 | 対話履歴 2 発話 |
|---|---|---|---|
| ベース | 80.52%(124/154) | **84.56%(126/149)** | **75.88%(129/170)** |
| 高温度 | 79.62%(125/157) | 84.35%(124/147) | 70.76%(121/171) |
| 例示なし | 60.39%(154/255) | 66.81%(159/238) | 59.33%(159/268) |
| 正確な指示に変更 | 81.76%(121/148) | 85.03%(125/147) | 64.50%(129/200) |
| 直前指示 | 82.80%(130/157) | 84.21%(128/152) | 67.55%(127/188) |
| マーカ追加 | 83.89%(125/149) | 86.39%(127/147) | **81.70%(125/153)** |
| 例分布正規化 | 82.35%(126/153) | 84.11%(127/151) | 66.67%(128/192) |
| 負例追加 | 78.43%(120/153) | 83.33%(125/150) | 80.77%(126/156) |
| 自己紹介型 | 83.11%(123/148) | 86.13%(118/137) | **81.63%(120/147)** |
| 他者紹介型 | 81.76%(121/148) | 87.77%(122/139) | 72.29%(120/166) |

**表 3**　履歴 2 発話条件で参照できた文脈

| 条件 | 発話のみ | 履歴 1 発話 | 履歴 2 発話 | 正解数 |
|---|---|---|---|---|
| gold | 84.84% | 12.12% | 3.03% | 165 |
| ベース | 93.02% | 5.43% | 1.55% | 129 |
| マーカ追加 | 96.00% | 2.40% | 1.60% | 125 |
| 自己紹介型 | 97.50% | 2.50% | 0.00% | 120 |
| 例示なし | 89.31% | **9.43%** | 1.26% | 159 |

# 7　実験結果

話者情報を抽出してアノテータによって o または p と判定されたものを正解とし，表 2 に精度を示す．結果から，まずベースにおいて文脈 1 発話と文脈 2 発話で精度に 8.68%の差が見られた (表 2 に太字で示す)．これはどちらの条件でも抽出に成功している話者情報数は大きな差が無いが文脈 2 発話の条件では誤抽出が多いためである．誤抽出の多くが目標発話と無関係な文脈部分の発話から抽出された話者情報であり，対象発話だけに上手く注目できていないことが分かる．この誤抽出はマーカを追加することや負例を追加することで減少がみられ精度の上昇が見られた (表 2 に太字で示す) が，ベースと有意になるような精度の差は見られなかった (検定にはフィッシャーの正確確率検定で片側検定し，p 値の補正をボンフェローニ法で行った)．また、ゼロショット設定では予測時の文脈発話数に関わらずベースから精度が下がった．このことから，例示の効果が分かる．全体を通してベースより有意に精度が上がった条件は無かった．精度が少し上がった条件がいくつかあったため，精度の上がった条件を組み合わせて実験をやり直したが，有意な精度の上昇は見られなかった．

## 7.1　LLM が抽出した話者情報の分析

文脈を履歴 2 発話分つけて予測されたモデルは 1 発話分のモデルや文脈無しのモデルと比べ有意に精度が低いが，より文脈から情報を補完して抽出できている可能性がある．これを調べるため，対話履歴を 2 発話分与えたモデルによって抽出された話者情報の内，アノテータが正しいとしたもの (o,p) に対して，予備実験と同様の分析を行った．分析は著者 1 名が行った．比較の対象として，人間が抽出した話者情報，履歴 2 発話のベース，履歴 2 発話を使ったものの中で特に精度が良かった 2 つの条件 (マーカ追加，自己紹介型)，抽出が最も多かった例示なしの条件の 5 つを使う．結果を表 3 に示す．表より，LLM が抽出した話者情報は人手で抽出したものよりも文脈から情報を補完できている割合が低い．さらに，ベースよりも精度が良かった 2 条件は文脈から情報を補完する割合が下がっている．また，例示の有無を比較 (ベース vs 例示無し) すると，例示なしの条件はベースより精度が低いものの文脈から情報を補完する割合は高いことが分かった (表 3 に太字で示す)．このことから，このタスクにおいて LLM は与えた例を真似て表面的な抽出を行うことで，抽出の精度が上がる代わりに文脈からの情報の補完ができなくなってしまうことが分かった．

# 8　まとめ

本論文では，近年雑談システムにおいて活用されている話者の情報を，最先端とされている LLM がどの程度雑談から抽出できるか，抽出時にどの程度文脈を参照しているかを検証した．結果から，LLM は文脈が長くなると目標発話に注目して話者の情報を抽出することが困難となり，目標発話とは無関係な話者情報の抽出を多くしてしまうことが分かる．目標発話に注目させるための工夫を行っても，精度は上がるものの文脈の考慮を犠牲にしてしまうことが分かった．今後は，目標発話と話者情報のペアを集めたデータセットを作成し，精度と文脈からの補完を両立させた話者情報の抽出を目指す．また，推論が必要な話者情報の分析等は今後の課題とする．

## 謝辞

本研究の一部は，JST 科学技術イノベーション創出に向けた大学フェローシップ創設事業 JPMJFS2101 の支援を受けたものです.

## 参考文献

[1] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[3] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chit-chat systems, 2021.

[4] 川本稔己, 山崎天, 佐藤敏紀, 奥村学. 大規模汎用言語モデルによるペルソナを考慮した応答生成, May 2022.

[5] Rui Ribeiro, Joao Paulo Carvalho, and Luisa Coheur. PG-Task: Introducing the task of profile generation from dialogues. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, **Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 183–189, Prague, Czechia, September 2023. Association for Computational Linguistics.