ユーザに適応する対話システムのための LLM を用いた自我状態推定

掛川脩人[1] 山田剛一[2] 増田英孝[2]
[1]東京電機大学大学院未来科学研究科 [2]東京電機大学未来科学部
24fmi05@ms.dendai.ac.jp {yamada, masuda}@mail.dendai.ac.jp

概要

対話システムの発展によって，ユーザへのより適応する返答が求められるようになった．この適応する返答を行うために交流分析に注目し，交流分析を用いた対話システム作成を目指している．

しかし,交流分析の対話システムには発話単位でのユーザの自我状態と呼ばれる発話者の心の状態の情報が必要になる．

本研究では,大規模言語モデルを用いて対話中のテキストから自我状態を推定することを検討する．実験の結果として，プロンプトに文脈情報や自我状態の説明文，分類の例を入力することによって精度向上だけでなく，出力の精度向上も確認した．

1 はじめに

大規模言語モデル(Large Language Model 以下 LLM)の登場により，対話応答の精度は向上した．この対話システムの発展に伴い，システムの返答に対する正確さに加え，ユーザを満足させる返答が求められる．ユーザの感情や性格に応じた対話の研究は盛んに行われている[1,2]．

しかし,感情や性格を考慮した対話システムではユーザを満足させる返答に至るまでには課題がある．

まず,感情は得られた感情情報に共感性に注目した研究が多く存在する．共感はコミュニケーションにおいて有用な要素であるが，問題の推定や解決など対話での適切な返答が全て共感で網羅することは難しい．多様な返答を行うには対話意図などの複合的な情報が必要であり，テキストの感情だけでは適切な対話システムを作ることは困難である．

次に，性格はユーザの静的な情報であり，対話システムのペルソナ作成には有効であるが，動的な要素が含まれる対話への影響は発話形式などの限定的な部分に留まる．よって，性格情報では感情のようなリアルタイムにテキストで変化する内容に対応することが難しい．

ここで我々は,心理学のカウセリングで用いられる交流分析の概念に注目する．交流分析は 5 つの自我状態と呼ばれる心の状態が存在する．自我状態の特徴を表 1 に示す．この自我状態が対話で用いられ，交流分析には，対話のパターンを分析する交流パターン分析と呼ばれる概念がある．

表 1 自我状態の特徴

自我状態	特徴
CP (Critical Parent)	批判的・厳格・理想的
NP (Nurturing Parent)	養育的・寛容・保護的
A (Adult)	冷静・理性的・論理的
FC (Free Child)	直感的・衝動的・積極的
AC (Adapted Child)	従順・抑制・消極的・反抗

交流パターン分析は,発話の刺激と反応の関係性として分析される．交流パターンには，期待通りの応答がなされる相補交流，期待と異なる応答による交差交流，表面的な交流と異なる潜在的なメッセージを含む裏面交流の 3 種類が存在する．これらのパターンを分析することで，対人関係における問題の理解が可能となる．

この交流パターン分析を用いた対話システムの実現を目指す．この対話システム実現にはユーザの発話から自我状態を推定することが必要である．

本研究では,交流パターン分析を用いた対話システム実現に向けてユーザの発話から自我状態の推定を行う．

2 関連研究

テキストを用いた自我状態推定の研究は少ない.テキストの分析による類似した研究として，感情分析の研究が存在する．主に BERT による学習を行った研究[3]と LLM を活用した研究[4]がある．感情のテキスト分析では，どちらの手法も有効である．

しかし,自我状態の推定を行うには文脈と発話の両方を考慮する必要がある．BERT では，文脈の情報を入力することで文脈の発話がノイズとなり，精度が落ちることが明らかになっている[5].

一方で，LLM を用いた推定では，文脈と発話の情報の両方を入力して推定の結果を得ることが可能となり，精度向上が期待される．

よって，本研究では，自我状態の推定に LLM を用いて，複数のモデルやプロンプトを用いて推定に関する適切な手法について検討する．

3 提案手法

3.1 概要

本研究では，LLM を用いた自我状態推定機能を対話システムへ組み込む際の妥当性を実証することを目的としている。

目的の達成に向けて,テキストからの自我状態の推定を LLM で行い，分類精度について検証する．この際に，複数のモデルや手法を比較検討し，適切な手法を探索する．

LLM は,主に API を介して利用可能なモデルとローカルで動作するモデルの 2 つに分けられる．これらは学習データ量やモデルサイズに違いがあり，計算コストや応答速度，性能が異なる．

これらの異なる特性を持つモデルでの比較を行い自我状態推定タスクに対してどのような要素が精度向上に影響を与えるのかを明らかにする．

3.2 推定の出力方法

自我状態の推定は,一意に確定することが難しい場合がある．これは感情などと同様に主観が介入することから，個人の解釈によって判断が変わる場合がある．

そこで本研究では,自我状態の曖昧性の可能性に対して推定された各自我状態に対して確率値を付与する形で出力する．推定が難しいものを確率的に評価することでより精度の高い推定が可能になる．

3.3 プロンプトの内容

自我状態の推定には,文脈と発話の両方の情報が必要となる．これは同じテキストであっても，文脈によって発話の意図が異なり，自我状態も異なる可能性がある．よって，文脈の情報を入力することで精度の向上が考えられる．

近年の LLM を用いたタスクでは，Few-shot プロンプトが広く用いられている．これは，タスクの例を示すことでモデルの理解を促す手法である．自我状態推定においても，適切な例示によってモデルの推定精度が向上する可能性がある．

一方で,不適切な例示はモデルの判断を誤らせる可能性が高く,特徴が極端な場合や曖昧性を持つ場合，性能の低下が考えられる．Few-Shot の数を多くしたことで性能が低下した研究[6]も存在する．

そのため，Zero-shot と Few-shot の比較，および適切な例示の数や内容について検証を行う必要がある．

4 実験設定

4.1 検証データセット

検証データとして，JEmpatheticDialogues[i]の対話データ[7]を用いた. JEmpatheticDialogues は 2 人の対話者の2往復から成る4つの発話を一連の対話として，この対話別に感情と状況文が記述されているデータセットである．

このデータセットに発話と 1 つ前の発話の長さがどちらも 10 文字以上であることを条件として，その中から取り出した 520 発話に対して自我状態のラベル付けを行った．ラベル付けには，1 つ前の発話をリード文として用いた．このリード文は文脈情報として活用を行う．特に，発話単体での自我状態の判断が難しい場合がある．主に NP と A と FC の 3 つの中から 2 つのどちらかで迷うパターンであり，これは相補的交流の観点に注目した．相補的交流は相手と自分の自我状態の一致によって発生する．相補的交流の起こりやすいパターンとして，3 つの中では NP と FC, A と A, FC と FC が挙げられる[8].

このパターンを文脈としての前発話の自我状態と合わせて考えることでより正確なラベル付けを行った．このパターンに該当しない複数の解釈に可能な 19 個の発話はデータセットから除外した．

ラベルの内訳を表 2 に示す．

表 2 ラベルの内訳

CP	NP	A	FC	AC	合計
55 個	106 個	124 個	140 個	76 個	501 個
11.0%	21.2%	24.8%	27.9%	15.2%	100%

[i] https://github.com/nttcslab/japanese-dialog-transformers

4.2 実験条件

4.2.1 推定を行う LLM の種類

本研究の実験では，自我状態の LLM として以下の 3 つのモデルを用いる．

- GPT-4o[ii]
- Claude 3.5 sonnet[iii]
- Llama-3-ELYZA-JP-8B[iv]

これらのモデルの内，GPT-4o(以下 GPT)と Claude 3.5 sonnet(以下 Claude)は API 使用モデルで Llama-3-ELYZA-JP-8B(以下 ELYZA)がローカル環境で動作する LLM となる．

Claude と GPT が様々なタスクにおける推論において，高い性能を持つことが示されている[9]．そのため，良い精度が期待される 2 つのモデルを API 使用モデルとして実験で扱う．

また, 日本語感情分析の精度において日本語と英語の追加学習モデルの効果的だとされている[10].

中でも，GPT-4 に匹敵する評価 [v]を得たモデルの軽量モデルの ELYZA をローカルモデルとして実験で扱う．

また，これらの LLM は確率的に生成が行われるため，推定精度の振れ幅が生じる可能性がある．推定精度への影響を最大限減らすため，API の 2 つのモデルの temperature の値を 0 として，ELYZA では do_sample を False として扱う．

4.2.2 プロンプト設計

提案手法で挙げた内容をプロンプトへ組み込む. ベースで自我状態の推定指示と推定対象のテキスト，出力形式の情報をプロンプトに入力する．

このベース内容に加え, 文脈情報や自我状態の特徴の説明文，例の追加によって，比較分析を行う．文脈情報として，前発話を与えた場合と与えなかった場合の 2 通りと例示として，自我状態の説明文を一切与えない場合と与えた場合，説明文に加えて 3-Shot を行った 3 通りの 6 通りについて実験を行う．プロンプトの詳細は図 2 に示す．

4.3 評価方法

評価方法として，正解率と MAE，F 値の 3 通りの方法を用いる．

正解率は推定結果の確率が最も高いものとラベルの自我状態が一致した割合である．正解率によって単純な推定精度を検証する．

MAE は平均絶対値誤差であり，ラベルの自我状態を 100 として，推定結果との差から値をそれぞれ算出する．これによって，推定結果の単純な正解だけでなく，候補として考慮されているかどうかが明らかになる．

F 値は，適合率と再現率の調和平均であり，分類結果の偏りの有無を調査する．

5 実験結果

5.1 出力結果処理

出力結果において，問題が 2 点生じた．

1 点目はテキストが与えられているにも関わらず, 評価を行っていない場合があった．この現象は ELYZA でのみ観測された. この未評価となった出力結果は評価対象から除外した．

2 点目は確率の出力にも関わらず、出力結果の和が 100 を超えるケースが見られた．これはどのモデルにも関わらず発生した．この原因として，プロンプトの指示にそれぞれの自我状態に対して 0-100 の値を出力するように指示した．この際，全体の合計を 100 と解釈するタイプと単純にそれぞれで 0-100 の値を出力するタイプに分かれた．

そのため，MAE は割合から算出したものと，信頼度として値をそのままの状態で算出した 2 つの場合を記載する．

5.2 実験結果詳細

実験結果を表 4 に記載する．ELYZA はベースの正解率が 14.68%と低く，以降の文脈と説明文，3-Shot の入力で 43.23%の改善結果が得られたが記載を省略する．また，対話システムへの展望として正解率が 2 位までの一致率も記載した．

結果からモデル内で最大の記録を出しているものには斜体，最も値が良いものには太字で記入を行った．

[ii] https://platform.openai.com/docs/models
[iii] https://docs.anthropic.com/en/docs/about-claude/models
[iv] https://huggingface.co/elyza/Llama-3-ELYZA-JP-8B
[v] https://huggingface.co/datasets/elyza/ELYZA-tasks-100

表 3 実験結果

評価方法		正解率 (2 位)	MAE (割合)	適合率	再現率	F 値	正解率 (2 位)	MAE (割合)	適合率	再現率	F 値
実験方法		ベース					文脈追加				
モデル	Claude	54.06% (77.64%)	16.69 (0.62)	0.58	0.54	0.52	56.09% *(79.64%)*	15.45 (0.58)	0.61	0.56	0.56
	GPT	54.69% (76.05%)	***11.81*** (0.68)	0.56	0.55	0.54	55.09% (75.25%)	12.72 (0.66)	0.56	0.55	0.55
実験方法		説明文追加					文脈+説明文追加				
モデル	Claude	55.09% *(79.64%)*	16.52 (0.57)	0.59	0.55	0.53	*58.88%* (78.04%)	*14.00* *(0.53)*	0.61	*0.59*	*0.59*
	GPT	58.48% (76.85%)	16.67 (0.59)	0.61	0.59	0.58	58.48% (76.85%)	17.75 (0.55)	0.60	0.58	0.58
実験方法		3-Shot+説明文追加					文脈+3-Shot+説明文追加				
モデル	Claude	51.90% (78.04%)	19.50 (0.56)	0.60	0.52	0.52	57.09% (78.24%)	15.71 (0.54)	***0.64***	0.57	0.58
	GPT	58.68% (75.68%)	19.59 (0.54)	0.60	0.59	0.58	***60.08%*** (***82.24%***)	18.43 (***0.50***)	0.60	***0.60***	***0.60***

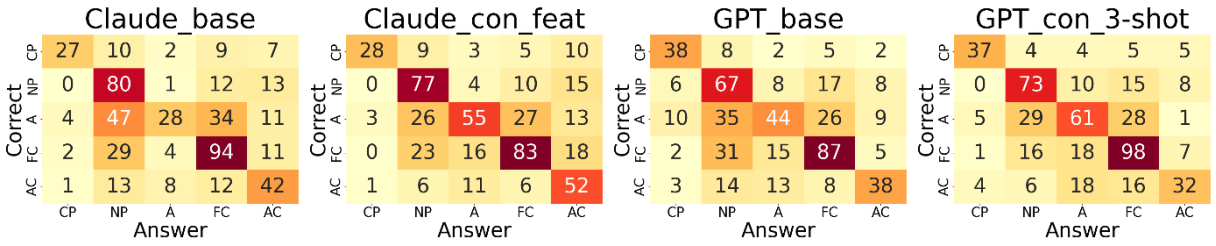

図 1 自我状態別の混同行列(ベースと正解率最大)

6 考察

実験の結果から ELYZA と他の LLM には大きな乖離があった．これはモデルの規模による精度の違いが出ていると考えられる. ELYZA はベース状態での性能は低かったが，情報を与えることで精度が向上したため，自我状態のデータをあまり持っていないと推測される. Claude と GPT も同様に情報の追加による精度の向上の傾向が見られた.

しかし， 3-Shot の Claude に文脈情報を追加した際の精度が低下した. Claude の 3-Shot は与えられた 3-shot の情報が Claude 内の想定する自我状態像と不一致になってしまい，低下した可能性が高い.

また，GPT での説明文の情報追加から文脈と説明文の 2 つの情報の追加では精度は低下しているが，差は 0.2%で生成の振れ幅の範囲と捉えられる.

Claude で精度が最も高くなったのは文脈と説明文の情報追加となった．文脈と説明文，3-Shot は性能低下に影響している裏付けとなっている.

図 1 の混同行列から Claude と GPT の回答に傾向の違いが見られる．正解率最大のものでも同様に傾向が残っているが精度が改善している.

また, 2 位までの正解率を考慮すると, 最大 82.24%となる．これによって，対話システムへの一定の精度が期待できる.

7 おわりに

本研究では, 交流パターン分析を用いた対話システム実現に向け, LLM を用いてユーザの発話から自我状態の推定を行った．結果として，文脈や説明文などの情報を追加することで精度の向上が見られた．プロンプトのより厳密な構成によって更なる精度向上が見込まれるため，今後の課題として検討を行いユーザに適応する対話システムの実装を目指す.

参考文献

[1] Qiang Zhang, Jason Naradowsky and Yusuke Miyao: Self-Emotion Blended Dialogue Generation in Social Simulation Agents, In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp.228–247, 2024.

[2] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy and Jad Kabbara: PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits, In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp.3605–3627, 2024.

[3] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima and Hajime Nagahara: WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations, In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp.2095–2104, 2021.

[4] Xin Hong, Yuan Gong, Vidhyasaharan Sethu and Ting Dang: AER-LLM: Ambiguity-aware Emotion Recognition Leveraging Large Language Models, *arXiv preprint arXiv:2409.18339,* 2024.

[5] AoGuo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori and Ryuichiro Higashinaka: Personality prediction from task-oriented and open-domain human–machine dialogues, In *Scientific Reports*, Vol.14, Article Number.3868, 2024.

[6] 齋藤 太我, 藤田 智弘: テキストベース感情推定のための大規模言語モデルによる学習データ生成における few-shot 学習の影響, 情報処理学会研究報告, Vol.2024-NL-259, No.5, pp.1-5, 2024.

[7] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba and Hideharu Nakajima: Empirical Analysis of Training Strategies of Transformer-Based Japanese Chit-Chat Systems, In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp.685-691, 2023.

[8] 杉田 峰康: 講座サイコセラピー 第 8 巻 交流分析, 日本文化科学社, 1985.

[9] Introducing Claude 3.5 Sonnet, https://www.anthropic.com/news/claude-3-5-sonnet, 2024.

[10] 近藤 里咲, 大塚 琢生, 梶原 智之, 二宮 崇, 早志 英朗, 中島 悠太, 長原 一: 大規模言語モデルによる日本語感情分析の性能評価, 情報処理学会第 86 回全国大会公演論文集, pp.859-860, 2024.

A 付録

System プロンプト
[ベース]テキストの交流分析における 5 つの自我状態を分析し、
[文脈情報の場合](文脈情報として 1 つ前の発話と分析対象のテキストが与えられます。¥n 文脈情報を参考に与えられた分析対象のテキストの交流分析における 5 つの自我状態を分析し、)
JSON 形式でのみ出力してください。
[説明文(例:3-Shot)](分析の際は以下の点に注目してください:

1. Critical Parent (CP): 批判的、命令的な表現
   例:「どんな理由があろうと、規則は規則なんだから弁解の余地はありません！」「俺の目の黒いうちは、誰にもそんなことはさせん！」「この部署の新人は仕事が遅くて困るな」
2. Nurturing Parent (NP): 保護的、養育的な表現
   例:「あいつ、叱られてばかりいて、可哀そうで見ていられないよ。」「風邪をひかないように気をつけるんだよ。」「よくできたね。君ならやれると思っていたよ。」
3. Adult (A): 事実に基づく論理的な表現
   例:「今回は日程が厳しいので、次回を待ってみます」「何時頃なら診ていただけますか」「この書類をチェックしていただけますか」
4. Free Child (FC): 自由な感情表現、創造的な表現
   例:「わあ、この山の景色は素晴らしいわね！」「この前見たの映画、最高だったよ！」「全く…..調子いいなあ」
5. Adapted Child (AC): 順応的、依存的な表現
   例:「気づかなくて…」「まったく同感だね。本当にそうだよ～」「え～っ、とにかく今買います！」)

[ベース]出力フォーマット:

```
 "ego-states": [
        {"ego": "CP", "score": 0-100 の数値},
        {"ego": "NP", "score": 0-100 の数値},
        {"ego": "A", "score": 0-100 の数値},
        {"ego": "FC", "score": 0-100 の数値},
        {"ego": "AC", "score": 0-100 の数値}
    ]
  }
```

User プロンプト
[ベース]
発話テキストのみ
[文脈]
″文脈情報:″ b
″分析対象:″ c

図 2 プロンプトの書式