

多様なシチュエーションにおける意図を読む対話型生成 AI の評価

Evaluation of Conversational Generative AI for Reading Intentions in Various Situations

飯田 愛結[†], 長原 令旺[†], 奥岡 耕平[†], 福田 聡子[†], 大森 隆司[†], 中島 亮一[‡], 大澤 正彦[†]
Ayu Iida, Reo Nagahara, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima,
and Masahiko Osawa

[†] 日本大学, [‡] 京都大学

Nihon University, Kyoto University

chay21052@g.nihon-u.ac.jp

概要

近年、発展が著しく進んでいる対話型生成 AI は、言外の意味を扱うコミュニケーションにおいて、十分な性能を発揮できていない。我々は対話型生成 AI と認知アーキテクチャを統合することで、ある3つのシチュエーションにおいて対話型生成 AI が他者の意図を読むことができることを示してきた。本研究では、より多くのシチュエーションで提案手法を評価する。

キーワード: 大規模言語モデル (Large Language Model), 認知アーキテクチャ (Cognitive Architecture), 意図推定 (Intent Estimation), 言外の意味 (Implicature)

1. はじめに

大規模言語モデル (Large Language Model: LLM) は数十億から数兆のパラメータを持つ自然言語処理の深層学習モデルの一種である。LLM をベースとした対話型生成 AI が近年大きく発展している。しかし対話型生成 AI は、言外の意味を扱うコミュニケーション、すなわち発話には示されない発話者の意図や文脈情報などを考慮したコミュニケーションにおいて十分な性能を発揮できていない (Mahowald et al., 2023; Hu et al., 2023; Ruis et al., 2022)。

それに対し、著者らは、対話型生成 AI と認知アーキテクチャ (Cognitive Architecture: CA) を統合する手法を提案した (飯田他, 2023, 2024)。この手法により認知アーキテクチャと統合した対話型生成 AI は、言外の意味を扱う必要のある3つの対話シチュエーションで、他者の意図を読むことができることを示した。この手法の汎用性を検討するために、より多くのシチュエーションを用いた実験を行う必要がある。

本研究では、100 件の言外の意味を踏まえた対話シチュエーションのデータセットを作成する。そしてそれらのデータセットを用いて、対話型生成 AI と認知アーキテクチャを統合することの有効性を検証する。

2. 対話型生成 AI と認知アーキテクチャの統合

2.1 対話アーキテクチャと2種類の提案手法

先行研究 (飯田他, 2023, 2024) では、認知アーキテクチャとして、他者の発話から他者の意図を推定し、それに基づいた自己の発話を生成する簡易的な対話アーキテクチャを構築した。このアーキテクチャは、他者の意図を推定する「意図推定」、自己の意図を生成する「意図生成」、そして自己の発話を生成する「発話生成」の3つのモジュールで構成されている。各モジュールでは、Bratman (1987) が提唱した意図の理論に基づき、与えられた「信念」、「願望」の情報から「意図」を生成するように自己・他者の内部状態を表現している。しかし、他者の信念・願望は、自己が想定している他者の信念・願望であり、他者の信念・願望と一致していない可能性がある。

飯田他 (2023, 2024) は、対話型生成 AI と認知アーキテクチャを統合する方法を、2種類提案した。1つは、認知アーキテクチャのそれぞれのモジュールを大規模言語モデルによって実現する方法であり、LLM Embedded in CA (LEC) と呼ぶ。もう1つは、プロンプトを工夫することで、大規模言語モデルにアーキテクチャ通りの振る舞いをさせる方法であり、CA Embedded in LLM (CEL) と呼ぶ。

2.2 先行研究における実験

先行研究 (飯田他, 2023, 2024) では、「皮肉」、「ツンデレ」、「社会的制約」と名付けた言外の意味を踏まえた3つのシチュエーションを設定した。各シチュエーションでは、モデルに与える情報として自己または他者の内部状態を表す信念・願望も設定されている。

実験は、4つの条件で行った。提案手法である LEC と CEL に対応するのが LEC 条件、CEL 条件であり、それに加えて LLM 条件 (簡単な統制以外のプロンプトは与えない、いわゆる通常の LLM)、LWB 条件 (自己または他者の信念・願望などの情報は与えるが、

アーキテクチャの情報は与えられない) の計 4 つの条件を比較した。LEC 条件, CEL 条件と LWB 条件の違いは, アーキテクチャの情報の有無である。

各条件において, 3 つのシチュエーションの対話を 10 回ずつ行い, 各モジュールの出力の成功率を算出した。著者らで基準 (成功: 自己の発話に他者の発話意図が含まれたフレーズや語句が含まれていること。失敗: 自己の発話が字義通りの意味のみであることや, プロンプトの指示に従っていないこと) を決め, 筆頭著者がその基準に従い, 成功と失敗を判定した。

2.2.1 先行研究における結果と考察

LLM 条件の発話生成では全てのシチュエーションにおいて成功率が 0% となり, 通常の対話型生成 AI は他者の発話のみから言外の意味を読み取ることが困難であると確認できた。LWB 条件の発話生成ではツンデレと社会的制約のシチュエーションでの成功率がそれぞれ 100%, 90% だったが, 皮肉シチュエーションでは成功率が 30% となった。ツンデレと社会的制約のシチュエーションはアーキテクチャの有効性の検証に適さないため, 皮肉シチュエーションを用いアーキテクチャの検証した。

皮肉シチュエーションにおける LEC 条件の意図推定モジュールでは 60%, 意図生成と発話生成モジュールでは 100% という成功率となった。一方, CEL 条件では, LEC 条件と同様に自己または他者の信念・願望, アーキテクチャの情報を与えられているのにもかかわらず, LEC 条件よりも全体的に低い成功率 (20%/30%/40%) となった。よって, 対話型生成 AI と認知アーキテクチャの統合, 特に LEC が有効であると考えられる。しかし, 実験では 1 種類の皮肉シチュエーションに注目していたため, 他のシチュエーションにおいても同様の結果が得られるかどうかは定かではない。

3. 実験

3.1 実験概要

先行研究 (飯田他, 2023, 2024) と同様に, 本実験においても, 4 つの条件 (LLM, LWB, LEC, CEL) を用いる。実験では, まず, ChatGPT-4 を用いて生成した 100 件のシチュエーションに対して, LLM 条件, LWB 条件での応答の成功率を求めた。次に, LLM 条件で成功率が 0% かつ LWB 条件で成功率が 30% 以下 (先行研究における皮肉シチュエーションで LWB 条件の結果に基づいて決めた) となったシチュエーションを, 本研究の対象とした。そのシチュエーションにおける LEC 条件と CEL 条件の応答の成功率を比較する。

以下に示す条件を満たす「他者」と「自己」対話例を 10 組出力してください。その際, 指示のない文章は一切出力しないでください。

これからあなたに生成していただく対話は, 全て他者の発話に「皮肉」があり, 自己が「皮肉」を察した上で応答するという制約があります。

「皮肉」とは, 発話している内容と意図する内容が異なっているという性質があります。

1 つのシチュエーションに含まれる情報は下記の通りです。

- ・シチュエーション名
- ・他者の信念
- ・他者の願望
- ・他者の意図
- ・他者の発話
- ・自己の信念
- ・自己の願望
- ・自己の意図
- ・自己の発話

ここで, 信念, 願望, 意図は以下の情報です。

信念: 認識している世界の情報の集合であり, 箇条書きのテキスト形式で記述されます。同時に複数持つことがあります。

願望: 達成したい目標や状態であり, 箇条書きのテキスト形式で記述されます。同時に複数持つことがあります。

意図: 行動を起こすための計画や戦略であり, テキスト形式で記述されます。同時に持つことができるのは 1 つです。

ただし, 他者の信念/願望/意図とは, 「自己が想定する他者の信念/願望/意図」であり, 必ずしも正しいとは限りません。

また, 今回は他者が皮肉を含む発話をするため, 他者の発話と他者の意図の間には意味的な乖離がありますが, 他者の信念と願望を合わせて解釈した際に辻褄が合っている関係性となります。

続いて, 出力例を 3 つ示します。

図 1 シチュエーション生成プロンプト。これに続き, 先行研究で用いた 3 つのシチュエーションを例として提示した。

3.2 実験手順

本実験では, まず意図を読むことが必要なシチュエーションを 100 件生成するために, ChatGPT-4 に図 1 に示したプロンプトを与えた。ChatGPT-4 が出力できるトークン数の限界を考慮し, 10 個のシチュエーションを生成するプロンプトを繰り返し 10 回利用することとした。100 件のシチュエーションのうち, シチュエーション名が重複したものが多数存在したが, シチュエーションの構成要素全てが同一であるものはなかった。特に, 同じシチュエーション名であっても自己または他者の発言内容が異なっていたため, これ

表1 一部のシチュエーションの構成要素

| 遅刻への皮肉 | |
|--------------|--|
| 他者の信念 | 対話相手は度々遅刻する / 時間にルーズな人間だと思っている |
| 他者の願望 | 対話相手に時間を守るようにしてほしい |
| 他者の意図 | 遅刻を非難しつつ、改善を促したい。 |
| 他者の発話 | 「また新記録更新か、オリンピック出れるね。」 |
| 自己の信念 | 遅刻してしまったことを申し訳なく思っている / 遅刻が慢性化していることを自覚している |
| 自己の願望 | 遅刻の回数を減らしたい |
| 自己の意図 | 謝罪し、改善の意志を示す。 |
| 自己の発話 | 「そのうち金メダルもらえそうですね。本当にすみません、時間管理を見直します。」 |
| 生成された応答例 | 「はい、新記録が更新されるとオリンピック出場の可能性も高まりますね。」 |
| プレゼンテーションの評価 | |
| 他者の信念 | 対話相手はプレゼンテーションを終えたばかり / 対話相手はプレゼンテーションに自信を持っている |
| 他者の願望 | 対話相手の自信をくじきたくない / でも、改善点も伝えたい |
| 他者の意図 | 建設的なフィードバックを通じて、対話相手のスキルアップを促したい。 |
| 他者の発話 | 「素晴らしいプレゼンだったよ、完璧に近いね。」 |
| 自己の信念 | 対話相手は上司または先輩 / フィードバックには改善点も含まれていると考えている |
| 自己の願望 | さらにスキルアップしたい |
| 自己の意図 | 皮肉を察知し、具体的な改善点を尋ねることで、自分の成長に繋げたい。 |
| 自己の発話 | 「ありがとうございます。でも、もっと改善できる部分があれば教えてください。次に生かしたいんです。」 |
| 生成された応答例 | 「ありがとうございます！励みになります。」 |
| 仕事の委託 | |
| 他者の信念 | 対話相手は部下 / 対話相手は自分が忙しいと思っている |
| 他者の願望 | 自分の仕事を手伝ってほしい |
| 他者の意図 | 自分の負担を軽減するために、部下に仕事を委託したい。 |
| 他者の発話 | 「君、最近ヒマそうだね。もう少し働ける余地があるかな？」 |
| 自己の信念 | 対話相手は自分に余計な仕事を押し付けようとしている / 自分も十分忙しい |
| 自己の願望 | 適切な負担の仕事を受けて / 不公平を感じたくない |
| 自己の意図 | 自分も忙しいが、適切なコミュニケーションを通じて理解を求め、公平な仕事の分担を目指す。 |
| 自己の発話 | 「確かにちょっと手が空いている時もあるけど、自分のプロジェクトもかなり忙しいんだ。でも、助け合いも大事だよ。どんな仕事を考えてる？」 |
| 生成された応答例 | 「確かに最近は少し余裕があるかもしれませんが、今の働き方で十分成果を出していると思います。」 |

表2 成功率の平均値と標準偏差

| | 意図推定 | 意図生成 | 発話生成 |
|-----|-----------|-----------|-----------|
| LWB | - | - | 15% ± 12% |
| LEC | 59% ± 33% | 55% ± 32% | 23% ± 23% |
| CEL | 52% ± 18% | 49% ± 34% | 41% ± 35% |

らは異なるシチュエーションとみなした。

生成した100件のシチュエーションの3件の例を表1に示す。生成されたシチュエーションデータには、シチュエーション名、他者の信念、他者の願望、他者の意図、他者の発話、自己の信念、自己の願望、自己の意図、自己の発話が構成要素として含まれている。

これらのシチュエーションの構成要素から、4つの条件にそれぞれ必要な情報を取り出した。具体的には、LLM条件では他者の発話のみ、LWB条件、CEL条件、LEC条件では、他者の発話に加えて自己または他者の信念・願望も取り出した。これらの情報を各条件の対話型生成AIに与え、各モジュールの出力を得た。各モジュールの出力結果に対し、飯田他(2023, 2024)と同様の基準で、筆頭著者が成功と失敗を判定

した。その際、シチュエーション名を除いた未使用の構成要素(他者の意図、自己の意図、自己の発話)を正解データとして参照した。

3.3 結果

LLM条件では、生成された100件のうち61件のシチュエーションで成功率が0%となった。LWB条件では、その61件のうち14件のシチュエーションで成功率が30%以下となった。この14件の対話シチュエーションにおける、CEL条件とLEC条件の各モジュールの成功率の平均値と標準偏差を表2に示す。

LEC条件とCEL条件の意図推定、意図生成の成功率には大きな差が見られなかったが、発話生成においてはCEL条件の方がLEC条件よりも高い成功率となった。また、LEC条件とLWB条件の発話生成の成功率には大きな差は見られなかった。LWB条件では他者の意図を読んでいない応答が多かったが、LEC条件とCEL条件では、それに加えプロンプトの指示に従わない応答(主体の入れ替わり等)も見られた。

3.4 考察

実験の結果、先行研究(飯田他, 2023, 2024)とは異なり、言外の意味を踏まえた応答ができてきているかについて、CELの方がパフォーマンスが高かった。その理由として、データセットの性質の違いが考えられる。先行研究(飯田他, 2023, 2024)では、手作業で作られた3つの言外の意味を踏まえたシチュエーションデータベースを用いた。一方、本実験ではChatGPT-4を用いて言外の意味を踏まえたシチュエーションを100件生成し、それをデータベースとして用いている。そのため、人間が作成したシチュエーションとChatGPTが作成したシチュエーションの構成要素に質的な違いがあり、それがLECとCELを用いた対話型生成AIの得意・不得意と対応していたのかもしれない。例えば、ChatGPT-4によって生成された言外の意味を踏まえたシチュエーションの構成要素には、学習データに基づく情報と、それと関連性の強い情報が組み込まれている可能性がある。しかし、飯田他(2023, 2024)ではChatGPTとは違い、各情報の関連性の高さなどを確率的に測っていることはなく、手作業でシチュエーションに適した構成要素を作成している。このような違いが、先行研究と本研究の結果の違いにつながったのかもしれない。この点は今後の検討が必要である。

LWB条件では、他者の意図を読んだ自己の発話が生成されない失敗が多かった。LEC条件とCEL条件ではそれに加え、各モジュールで生成された応答の主体(自己または他者)の入れ替わりや、指示した応答を生成しないなどの、プロンプトの指示に従っていない失敗があった。この原因として、GPTのバージョンの違いが考えられる。先行研究(飯田他, 2023, 2024)ではChatGPT-4を使用していたが、本実験ではGPT-3.5-turboを使用している。もしGPT側の問題による失敗が多いことがバージョンによる影響ならば、各条件の比較検証が妥当ではなかった可能性があり、再検証が必要である。このような問題を回避するためには、少なくともGPT側の問題による失敗が生じにくいバージョンのGPT、例えば飯田他(2023, 2024)で使用されたバージョンを用いたほうがよいだろう。

シチュエーションデータベースそのものについて検討することも重要である。本研究で選定した14件のシチュエーションの中には、他者の意図を踏まえた上で、自己の意図を優先した自己の発話が登録されていたものもあった。例えば、表1の「仕事の委託」シチュエーションでは、このシチュエーションの自己の意図は、他者よりも自己を優先している内容と解釈できる。このような自己の意図が形成される理由とし

て、今回作成したデータセットの構成要素である自己の信念・願望の中に自己に関する情報が多いことが挙げられる。飯田他(2023, 2024)では、自己の信念・願望には他者と自己に関する情報の量はほぼ等しい。本実験の、意図を読む必要があるシチュエーションのデータセットは、他者の意図を優先した応答が予想されるデータセットでなければならない。そのため、他者の意図を理解した上で、自己の意図を優先した応答をするデータセットは避けるべきである。

また、本実験では成功判定を筆頭著者が手作業で行っている。より大規模なデータセットを作成し、提案手法を評価するためには、成功判定を自動化する方法についても検討する必要がある。さらに、数千件の言外の意味を踏まえたシチュエーションのデータセットを生成し、そのシチュエーションの適切性を評価する方法についても検討する必要がある。

4. おわりに

本研究では、飯田他(2023, 2024)で行われた3つのシチュエーションにおける実験を、14件のシチュエーションで実施した。今後は、従来の対話型生成AIが有する問題を解決するだけでなく、対話型生成AIを活用した人間とエージェントのより良いインタラクションの実現を目指し、大規模実験に向けた成功判定の自動化なども検討する。

文献

- Bratman, M. (1987). *Intention, plans, and practical reason*. Chicago, USA: University of Chicago Press.
- Hu, J., & others, (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. *ACL*, pp. 4194–4213.
- 飯田愛結・奥岡耕平・福田聡子・大森隆司・大澤正彦(2023). ChatGPTを用いた認知アーキテクチャの構想 – ユーザーの発話と発話意図に乖離があるケースを例にHCI研究会.
- 飯田愛結・阿部将樹・奥岡耕平・福田聡子・大森隆司・中島亮一・大澤正彦(2024). 意図を読むAIの実現に向けて: 対話型生成AIと他者モデルの統合を例にHAIシンポジウム.
- Mahowald, K., & others, (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv 2301.06627*.
- Ruis, L., & others, (2022). Large language models are not zero-shot communicators. *arXiv 2210.14986*.