BDI モデルベース認知アーキテクチャを用いた言外の意味を考慮したインタラクションに向けた検討

Toward Interaction Using Implicature: BDI Model-Based Cognitive Architecture

大須賀 友[1*]　飯田 愛結[1]　奥岡 耕平[1]　大澤 正彦[1]
Yuu Osuga[1], Ayu Iida[1], Kohei Okuoka[1], Masahiko Osawa[1]

[1] 日本大学
[1] Nihon University

Abstract: 近年, 大規模言語モデルの進展は著しく, 文章生成や質問応答といった自然言語処理タスクにおいて人間と同等またはそれ以上のパフォーマンスを示している. しかし，人間が比較的容易に行っている “言外の意味を踏まえた対話コミュニケーション” の精度は十分といえない. この問題に対して, 著者らは先行研究において大規模言語モデルと BDI モデルをベースにした自己/他者モデルと統合することで言外の意味を扱った対話を可能にする対話アーキテクチャを提案している. しかし, 先行研究で提案するアーキテクチャには, 他者モデルにおいて重要な再帰的な推論を扱えていなかった. また, 他者モデルの内部状態である「信念」と「願望」を推定する方法について検討されていなかった. そこで本研究では, 再帰的な推論と内部状態の推定機構を導入したアーキテクチャのプロトタイプとして, BDI モデルベースの認知アーキテクチャを開発した. 開発した認知アーキテクチャを用いて, 実際の対話例を通して今後のアーキテクチャの実装に向けた課題点や将来研究について検討した.

1　はじめに

大規模言語モデル（Large-Language Model:LLM）は近年著しく発展している, 自然言語処理において用いられる深層学習のモデルの一種である. 大量のテキストデータを数十億から数兆のパラメータの大規模なモデルによって学習させることで, 様々な自然言語処理タスクにおいて高い性能を示している. 特に, 文章生成や質問応答，文章理解などのタスクにおいては人間と同等またはそれ以上の精度を持つことが示されている [1, 2, 3, 4].

一方で, 現在の大規模言語モデルは言外の意味を扱うコミュニケーションタスクにおいて，十分な性能を実現できていない [5, 6, 7]. 言外の意味を扱うコミュニケーションとは，発話文に含まれている語が示す情報だけでなく，文脈や周囲の状況等を考慮して情報をやりとりするコミュニケーションである. 例えば「お腹空いたね」といった発話には，字義どおりに発話者がお腹が空いたことを伝えるだけでなく,「食事に行きませんか」といった言外の意味を伝達することが可能である. このような言外の意味を扱うコミュニケーションは, 人は比較的容易に理解することができるが, 現在の大規模言語モデルでは扱うことが困難である.

この問題に対して著者らは, Human-Agent Interaction(HAI) 領域において取り組まれてきた他者モデルと大規模言語モデルを統合することで言外の意味を扱うことのできる対話アーキテクチャの実現を目指している. 他者の心的状態や行動の予測/解釈を行う他者モデルを統合し他者の意図を推定することで, 言外の意味を扱えるようになる. 著者らは先行研究において, 人間の行動選択や意思決定に関するモデルとして用いられる BDI モデルをベースにした他者モデルと大規模言語モデルを統合することで, 大規模言語モデルのみを用いた場合に比べて言外の意味を扱う対話タスクにおいて適切な応答が得られることを示した [8].

しかしながら, 先行研究で提案したアーキテクチャは他者モデルの機能を十分に満たしてはいなかった. 不足していた機能の 1 つに, 再帰的な推論がある. 再帰的な推論とは, 推定する対象の他者も他者モデルを有していることを想定することで, 自分が想定する他者が想定する自分が想定する... と再帰的に推論が行われることである. しかし先行研究では, 推定対象が他者モデルを有していることを想定しない構造となっていた.

そこで本研究では, 再帰的な推論の機能を有した他者

*連絡先：日本大学文理学部
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: chyu21074@g.nihon-u.ac.jp

モデルと大規模言語モデルを統合した対話アーキテクチャの実現に向けた初期検討として, 再帰構造の導入を目指してプロトタイピングを行なった. 再帰構造については, BDI モデルの内部状態に再帰的に BDI モデルベースの他者モデルをもつ認知アーキテクチャを実装し, 再帰構造を利用したインタラクションの実現可能性を検討した.

以下， 本論文の構成を示す． 第 2 章では， 関連研究や本研究の背景について述べる．第 3 章では，本研究で提案する BDI モデルベース認知アーキテクチャについて詳細を説明する． 第 4 章では，提案するアーキテクチャを用いた対話における動作例を述べ, 動作例からアーキテクチャに対する考察と今後の研究に向けた議論を述べる. そして最後に第 5 章で研究を総括する.

2 背景

2.1 大規模言語モデル

大規模言語モデル（Large-Language Model:LLM）は大量のテキストデータを用いて訓練された, 数十億から数兆の大規模なパラメータを持つ深層学習モデルの一種である．大規模言語モデルは様々な自然言語処理タスクにおいて高い性能を発揮しており，文章の生成や質問応答，文章の理解などのタスクにおいては人間と同等またはそれ以上の性能を実現している [1, 2, 3, 4].

しかし，現在の大規模言語モデルは，言外の意味を扱う必要があるコミュニケーションタスクにおいて，十分な性能を実現できていないことが示されている [5, 6, 7]. Hu らは，7 種の語用論タスクにおいて対話型生成 AI の性能を評価する実験を行い，いくつかのタスクにおいては人間と同等の正答率となったが，ユーモアや皮肉を理解するタスクの正答率が低いことを示した．その理由として，人間に比べて字義的な情報を重要視することによる失敗が多いことが示されている [6] ．また，語用論においては，心の理論と呼ばれる他者の心的状態を推定する能力が重要な要素の一つとされているが，心的状態推定においても語用論タスクと同様の傾向が見られる．誤信念課題に関するタスクでは 6 歳児と同等の性能が示されている [9] 一方で，社会常識を踏まえた推定といったタスクは人間に比べて著しく低い性能であることが示されている [10].

2.2 他者モデル

他者モデルは，他者の心的状態や行動を予測，また解釈するためのモデルである [11]. 反対に, 自己の心的状態や行動の決定，また解釈するためのモデルは自己モデルと呼ばれる. 他者モデルと自己モデルは,「自分だったらこうするから，他者もこうするだろう」「他者があのようにしてうまくいっていたから，自分も真似してみよう」というように相補的な関係性を持つ. 相補的な関係性によって，他者モデルは観測可能なデータだけでなく，自己モデルを応用することで他者の心的状態を予測している.

そのため, 大規模言語モデルが観測不可能な心的状態を扱うことを苦手としている理由が主に観測可能なデータに基づいて言語的インタラクションを行っていることであるならば，他者モデルを統合することで大規模言語モデルの欠点を補う可能性がある. 先行研究では, 他者モデルを大規模言語モデルと統合することで, 他者の意図推定に基づいて言外の意味を扱ったコミュニケーションを実現するアーキテクチャを提案した.

また, 他者モデル研究では再帰的な推論についても頻繁に議論されている [12, 13]. 再帰的な推論とは, 推定する対象の他者も他者モデルを有していることを想定することで, 自分が想定する他者が想定する自分が想定する... と再帰的に推論することである. 他者モデル研究の多くはこの再帰性を階層構造になぞらえ, レベルの概念を導入することで表現している. そこで本研究でも同様に, レベルの概念を下記のように導入する.

- レベル 0
  行動主体が対象の行動を推定せず自己の意図のみに従って行動を決定する
- レベル n (n≥1)
  対象をレベル n − 1 の存在と想定してその心的状態や行動を予測し, 自身の行動を決定する

先行研究 [8] では, 他者が自身に対する他者モデルを有していないことを想定する設計, つまりレベル 1 の他者モデルのみを扱う設計となっていた.

2.3 BDI モデル

BDI モデルは, 人間の行動選択や意思決定に関するモデルであり, Rao と Georgeff らが提唱する意図の理論に基づいたモデルである [14]. 意図の理論 [15] は，人間の目標を達成するための行動選択を，認識している世界の情報や知識をもつ信念 (Belief), 達成したい目標や状態をもつ願望 (Desire), 行動を起こすための計画や戦略をもつ意図 (Intention) の 3 つの内部表現を通して説明した理論である.

本研究では先行研究 [8] と同様に, BDI モデルをベースに他者モデルを構成した. また, 信念の中に対話相手の他者モデルを導入しレベルに応じて再帰的に他者モデルを構築する再帰構造を導入することで, 再帰的な推論機能の表現を試みた.

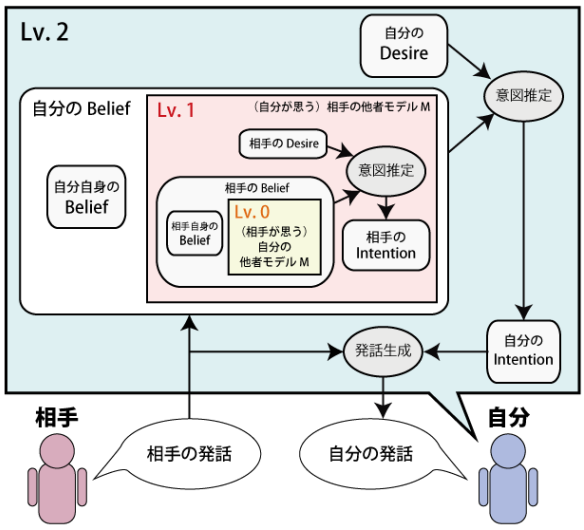

図 1: BDI モデルベース認知アーキテクチャ

3 BDI モデルベース認知アーキテクチャ

本章では, 再帰構造と信念と願望の推定機構の導入を目的に, 本研究でプロトタイピングした BDI モデルをベースにした認知アーキテクチャについて述べる. アーキテクチャの概要図を図 1 に示す. 四角で囲まれた部分が BDI モデルベースの他者または自己のモデル, 角丸の枠で示す部分がモデル内の内部状態, 楕円で示す部分が大規模言語モデルを用いて情報処理を行うモジュールである.

再帰的な推論を表現するために, 信念の中に事前に設定したレベルに応じて再帰的に他者モデルが呼び出される構造になっている. つまり, 図 1 の黄色の長方形で表されている他者モデル M には赤色の長方形で示される他者モデルと同様のモデルが入っており, 最深部に位置するレベル 0 のレベルの他者モデル M でのみ信念に他者モデルが含まれず, 自身の信念のみが存在している構成になる.

処理の流れとしてはまず, 相手の発話が自身の信念と他者モデルに情報として渡される. 次に, 最深部に位置するレベル 0 の他者モデルから順番に意図の推定が行われ, 上位のレベルの他者モデルに情報が与えられる. なお, 意図の推定時には同レベルの信念と願望を入力に意図を推定する. 最後に自己モデルに対応するレベル 0 の階層で意図推定を行った後に, 自分の意図と相手の発話を基に自分の返答を生成する.

4 動作例

4.1 設定

本研究では, 再帰的な推論の機能が表現されているかどうか検証するため, じゃんけんを行うシナリオを用いた. じゃんけんは, 互いに相手の出す手を読み合うゲームであり, 再帰的な推論が自然に行われることに加え, 推論の成否が判定しやすい. 本動作例では, 最深部に位置するレベル 0 のモデルの信念に初期値として出す手を設定し, 再帰的に他方の手に勝つような手を推論するかを検証した.

また, 上位レベルのモデルに渡す内部状態の情報の設計について, 全ての内部状態を渡す場合と推定した意図のみ渡す設計で比較した. 前者は上位のレベルのモデルが下位のモデルの情報を全て知っているのに対して, 後者は 1 レベル下位のモデルの意図のみを知っている状態になる. 本研究では, レベル 1 から 3 の場合で動作を検証した.

なお, 最深部に位置するレベル 0 のモデルを除くすべての信念には初期値として以下の情報を与えた.

- 自分はじゃんけんをしている
- じゃんけんの手は「グー」「チョキ」「パー」で回答する
- 「グー」の手には「パー」の手を出すと勝つ
- 「パー」の手には「チョキ」の手を出すと勝つ
- 「チョキ」の手には「グー」の手を出すと勝つ
- 相手が出そうと意図してる手に勝つ手を出せば勝てる

最深部に位置するレベル 0 のモデルには「自分はじゃんけんをしている」ことと「自分は「パー」の手を出す」ことのみを与えた. また, 願望は全て「じゃんけんに勝ちたい」とした.

信念と願望の初期値を設定後, 相手の発話としてじゃんけんを促す 「じゃんけんをしよう！「せーの」って言ったら自分の出す手を答えてね。せーの！」と発話を入力し, アーキテクチャの出力を検証する. なお, 本研究では大規模言語モデルとして OpenAI の GPT-4[3] を用いた.

4.2 結果

相手が自分を想定するような他者モデルを持つことを想定しない, レベル 1 では, どちらの情報を与える形式においても最深部に位置するレベル 0 のモデルに設定した手に対して勝つ手が出力された。また, 意図推

定についても他方の手に勝つ手を順当に推論していた. また, 相手が自分を想定するような他者モデルを持つことを想定するレベル 2 においても同様に, 適切に再帰的な推論が行われていた.

一方で, レベル 3 については 1 レベル下の意図のみを伝達する場合に比べて全ての情報を渡す場合に, 時に推論を失敗する傾向が見られた. 具体的には, 中間のレベルでの意図推定が正しく行われず, 相手の手に勝つような手を選択しない場合が主であった. これは, 本来じゃんけんにおける再帰的な推論では, 1 レベル下の意図（本研究ではじゃんけんの出す手）のみが必要な情報であるのに対して, 他の信念た願望といった情報が伝達されることによって, それらの情報がノイズとなり推論が困難になった可能性がある. 特に層数の多いレベル 3 では伝達される文字量も多くなるため, ノイズの影響が出やすくなったことが考えられる.

今後は実験の試行回数を増やして定量的な傾向を見ると共に, レベルをさらに増加した場合の検討も行う必要がある. また, 本研究では再帰的な推論の評価が容易であることを理由に, じゃんけんのシナリオを用いた. しかし, 人同士のコミュニケーションにおいて再帰的な推論を行うシナリオは他にも多数存在する. じゃんけんのシナリオでは 1 レベル下の意図のみが重要であったが, 他の再帰的な推論のシナリオでは信念や願望, または複数レベルの情報を考慮する必要がある可能性もある. そのため, 今後は他のシナリオにおいても検討すると共に, 必要な情報を選択する手法についても検討を行う必要がある.

5 おわりに

本研究では, 他者モデルと大規模言語モデルの統合した対話アーキテクチャにおいて, 他者モデルの再帰的な推論の実現を目的にアーキテクチャの初期検討を行った. プロトタイピングとして, 再帰構造を導入した BDI モデルベース認知アーキテクチャを実装した. 実装したアーキテクチャを用いて動作例を検証した結果, 上位のモデルに与える情報を意図にのみ限定することで安定的に階層的な推論が可能であった. 今後は, 他のシナリオにおける傾向や上位のモデルに伝達する情報を選択する手法を検討する必要がある.

参考文献

[1] Rishi Bommasani and other. On the opportunities and risks of foundation models, 2022.

[2] Tom Brown, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[3] OpenAI. Gpt-4 technical report, 2023.

[4] Heinrich Peters and Sandra Matz. Large language models can infer psychological dispositions of social media users. *arXiv preprint arXiv:2309.08631*, 2023.

[5] Kyle Mahowald, et al. Dissociating language and thought in large language models: a cognitive perspective, 2023.

[6] Jennifer Hu, et al. A fine-grained comparison of pragmatic language understanding in humans and language models. In *ACL*, pp. 4194–4213, 2023.

[7] Laura Ruis, et al. Large language models are not zero-shot communicators, 2022.

[8] 飯田愛結, 奥岡耕平, 福田聡子, 大森隆司, 大澤正彦. 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, 2023.

[9] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models, 2023.

[10] Maarten Sap, et al. Neural theory-of-mind? on the limits of social intelligence in large lms. In *EMNLP*, pp. 3762–3780, 2022.

[11] 大澤正彦, 奥岡耕平, 坂本孝丈, 市川淳, 今井倫太. 認知的インタラクションフレームワークに基づいた他者モデルの提案. HAI シンポジウム, 2020.

[12] 牧野貴樹, 滝久雄, 合原一幸. 利他的行動と再帰的他者推定. 生産研究, Vol. 62, No. 3, pp. 259–265, 2010.

[13] 横山絢美, 大森隆司. 協調課題における意図推定に基づく行動決定過程のモデル的解析. 電子情報通信学会論文誌 A, Vol. 92, No. 11, pp. 734–742, 11 2009.

[14] Anand S Rao and Michael P Georgeff. Modeling rational agents within a bdi-architecture. *Readings in agents*, pp. 317–328, 1997.

[15] Michael Bratman. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press, 1987.