

典型他者モデルを用いた察する対話システムの構想

A concept of an intuitional dialogue system using a typical other model

大信田 和華^{1*} 飯田 愛結¹ 奥岡 耕平¹ 大澤 正彦¹

Kazuha Oshida¹, Ayu Iida¹, Kohei Okuoka¹, and Masahiko Osawa¹

¹ 日本大学

¹ Nihon University

Abstract: 「察し」とは話し手の発話や態度から聞き手がその意図を推しはかる行為である。著者らの一部は、他者の意図を察しながら応答する対話システムを提案した。しかし、意図を形成する元となる信念や願望は、相手にかかわらず固定であった。本研究では、想定される対話相手のタイプごとに信念と願望を設定した典型他者モデルを構築し組み込んだ。対話時は典型他者モデルを切り替えながら対話することで、様々な他者に適応できると考えられる。

1 はじめに

「察し」は、話し手の発話、表情や態度といった言動から聞き手がその意図を推し測る行為であり、重要なスキルである [3]。コミュニケーションという行為は少なからず「察し」のように暗黙的な推論や理解を前提としている [8]。

一方で、近年さまざまな自然言語処理タスクにおいて急速に性能向上をさせている大規模言語モデルは、「察し」が必要な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発揮できていないことが示されている [11, 5, 8]。言外の意味を扱うコミュニケーションとは、発話された言語表現に含まれる情報だけでなく、発話者の情報やこれまでの文脈を考慮したコミュニケーションのことを指す。例えば「のど渴いたな」といった発話には、字義通りののどが渴いたという意味だけでなく、「飲み物がほしい」といった言外の意味を伝達することが可能である。これらのコミュニケーションは語用論と呼ばれる言語学の分野で研究されており、その他の例として比喩や皮肉といった言語表現があげられる [2, 17, 9]。

この問題の解決策の1つに、Human-Agent Interaction(HAI) 領域において取り組まれてきた他者モデルの研究がある。他者モデルとは、他者の心的状態や行動の予測/解釈を行う認知モデルである [6]。著者らの一部は、他者の意図を読むことができる大規模言語モデルの実現を目指して、大規模言語モデルと他者モデルを統合する方法を提案し、その有効性を示した [6]。

しかしながら先行研究 [6] の実験では、他者の心的状態が固定されていた。人の心的状態は、同じシチュエーションにおいても性格や立場、知識などの様々な要因で異なる。より現実的なシチュエーションで意図推定を行うためには、人による違いを扱う必要がある。

そこで本研究では、他者の心的状態に応じて相手の意図を読み、「察する」対話システムを構想した。他者の心的状態に応じる方法として特定の性格ごとに形成される典型他者モデルを使用した。システムでは想定される対話相手の典型他者モデルを複数作成し、選択することで現実的なシチュエーションで意図推定できるようにした。本研究では、構想したシステムを作成し、正しく動作するか検証した。

2 関連研究

2.1 大規模言語モデル

大規模言語モデルは、数十億から数兆のパラメータをもつ自然言語処理の深層学習モデルの一種であり、大量のテキストデータを用いて訓練された結果、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いる。特に、文章の理解や生成、質問応答などのタスクにおいて、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 [15, 16, 12, 4]。

大規模言語モデルの代表例である ChatGPT は、OpenAI によって開発された大規模言語モデル [12] を基盤とするチャットボットである。ChatGPT は、ユーザからのリクエストや質問に対して、自然な言葉で返答することができる。その応答は、膨大な量の訓練データに基

*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: chkz23007@g.nihon-u.ac.jp

づいて生成されるため、広範囲なトピックに対して情報を提供することが可能である。

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た [11, 5, 8]。Huらは、7種の語用論タスクにおいて大規模言語モデルの性能を評価する実験を行った。実験の結果、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してしまうことによる失敗が多いことが示されている [5]。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の推定においても語用論タスクと同様の傾向がみられる。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 [7] 一方で、社会常識を踏まえた推定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

他者モデルとは他者の心的状態や行動を予測するモデルであり [13, 6]、以下の3つに分類できる [18]。

平均他者モデル	これまで関わってきた複数の他者を平均化して構築される他者モデルである。
個別他者モデル	特定の個人に対して構築される他者モデルである。
典型他者モデル	特定の性格ごとに形成される他者モデルであり、平均他者モデルよりも個別化され、個別他者モデルよりも一般化された中間的なモデルである。

大規模言語モデルが対話を行う際、に対話相手の言動に適した典型他者モデルを選択することで部分的な個人適応の向上と学習の早さが期待される。

一方で、自己の心的状態や行動の決定、解釈するためのモデルを自己モデルと呼ぶ。自己モデルと他者モデルは相補的な関係性であり、自己モデルの訓練結果を他者モデルに応用したり、その逆を行ったりすることを前提にしている。具体的には「他者があのようになってしまうから、自分も真似してみよう」「自分だったらこうするから、他者もこうするだろう」という判断が有効であるように、自己と他者がある程度共通した知的システムを持つことがインタラクションの前提となっている。つまり、他者モデルは観測可能なデータというだけではなく、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。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からであれば、他者モデルは大規模言語モデルの欠点を補う可能性がある。

2.3 意図を読む AI

著者らの一部による先行研究は、大規模言語モデルと他者モデルを統合した、他者の発話意図に基づく自己/他者モデル付き対話認知モデルを提案した [6]。また当該研究では、他者モデルを認知モデルの一種として位置付けている。

この他者モデルでは、2者間対話を想定しているため自己と他者が存在し、両者の内部表現として、信念 (Belief)、願望 (Desire)、意図 (Intention) を使用している。これらの内部表現は Bratman が提唱した人間の目標を達成するための行動選択を説明した意図の理論 [1] に基づいた、BDI モデル [14] から着想を得ている。ここで信念 (Belief) とは対象の人間の認識している世界の情報や知識であり、願望 (Desire) とは対象の人間の達成したい目標や状態を表す。また、意図 (Intention) とは信念と願望から形成されるもので、行動を起こすための計画や戦略のことを指す。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もある。

この他者モデルは、意図推定、意図生成、発話生成の3つのシステムで構成される。意図推定システムでは、他者の発話と自己が想定する他者の信念及び願望から、他者の意図を推定する。意図生成システムでは、自己が想定する他者の意図と自己の信念及び願望から自己の意図を生成する。発話生成システムでは、自己の意図と他者の発話から自己の発話を生成する。

また、大規模言語モデルと他者モデルを統合する2種類の手法も提案している [6]。1つは、LLM Embedded in Cognitive Model (LEC) という大規模言語モデルを認知モデルに組み込む手法である。もう1つは Cognitive Model Embedded in LLM (CEL) という認知モデルを大規模言語モデルに組み込む手法である。

相手の意図を踏まえた発話を生成できるかを検証する実験では、言外の意味を扱う3つのシチュエーションにおいて LEC が高い性能を発揮し、大規模言語モデルと他者モデルを統合することの有効性を示した。

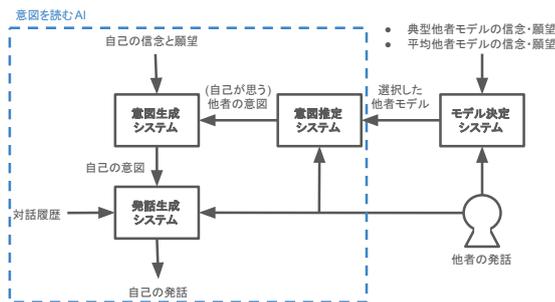


図 1: 対話システムの構成図

3 意図を読む AI と 典型他者モデルの統合

シチュエーションを固定せず相手の意図を察する対話システムを作成するために、先行研究で提案した対話システム [6] に典型他者モデルを組み合わせた対話システムを提案する。以降このシステムを説明するにあたって、搭載されているエージェントを自己、そのエージェントの対話相手を他者と呼ぶ。

対話システムの構成図を図 1 に示す。この対話システムは、先行研究が提案を構成する「意図推定システム」「意図生成システム」「発話生成システム」に、新たにモデル決定システムを新たに加えた 4 つのサブシステムから構成される。図 1 の青い点線に囲まれているの 3 つのシステムは、先行研究が提案した対話システムと同様の仕様にするために、先行研究のプロンプト [6] を用いた大規模言語モデルで実装した。

3.1 モデル決定システム

新たに追加したモデル決定システムでは他者の発話と N 個の他者モデルを入力として受け取る。ここで本システムにおける他者モデルは、あらかじめ設定された信念と願望の組み合わせで表現されている。また、ここで受け取る他者モデルは複数個の典型他者モデルと 1 個の平均他者モデルである。

この対話システム上での典型他者モデルと平均他者モデルを以下に定義する。

典型他者モデル

想定する特定の性格の他者の信念と願望を設定したモデルである。特定の他者の性格の数 (N 個) だけ存在する。

平均他者モデル

対話を求める他者を想定したモデルである。他の典型他者モデルの他者どれにも当てはまるモデル

なため、矛盾度は他のモデルよりも低くなる想定である。

他者の発話が、各他者モデルを構成する信念・願望に基づいて生成されたと仮定した場合の矛盾度を 0 から 100 の値としてそれぞれ計算する。矛盾度の測定は LLM を用いて行い、LLM に与えるプロンプトは大須賀ら [19] が提案した矛盾チェックシステムのプロンプトを参考に作成した。測定した矛盾度に基づいて N 個の典型他者モデルと 1 個の平均他者モデルから、適するモデルを選択する。この際、原則として一番矛盾度が低いモデルを選択するが、全ての典型他者モデルの矛盾度が同じだった場合には平均他者モデルを選択することとした。決定したモデルの信念と願望を出力し、意図推定システムに入力する。

4 ケーススタディ

モデルが正しく選択され遷移するか検証するために、作成した対話システムに対して 2 種類の典型他者モデルを追加し、対話を行った。2 種類の典型他者モデルと平均他者モデルの内容を示す。

問題解決したいという願望を持つ典型他者モデル

問題解決したいという願望を持ち、対話型 AI とのインタラクションのモチベーションは、問題解決にある他者モデルである。想定する他者の信念を「問題解決 AI と対話をしている」、願望を「共感ではなく有意義な問題解決をしてくれる返答を得たい」と設定した。本論文では、問題解決モデルと呼ぶ。

共感してほしいという願望を持つ典型他者モデル

共感してほしいという願望を持ち、対話型 AI とのインタラクションのモチベーションは、共感にある他者モデルである。想定する他者の信念を「共感 AI と対話をしている」、願望を「問題解決ではなく有意義な共感をしてくれる返答を得たい」と設定した。本論文では、共感モデルと呼ぶ。

平均他者モデル

どの他者に対して適用しても矛盾を生みづらい、抽象的な他者モデルである。想定する他者の信念を「AI と対話をしている」、願望を「有意義な返答を得たい」と設定した。

本検証では、2 つのシナリオを設定した。1 つは、平均他者モデルから問題解決モデルに遷移する対話を扱うシナリオである。もう 1 つは、平均他者モデルから共感モデルに遷移するような対話を扱うシナリオである。それぞれを、シナリオ 1、シナリオ 2 とする。これ

表 1: シナリオ 1 における発話と選択モデル

話者	発話内容	モデル
U	最近将来のこと考えてるんだよね	-
A	将来について考えるのは大切ですね. 具体的にはどんなことを考えているのですか?	平均
U	就活とかなにから準備すればいいのかなって	-
A	就職活動の準備に関する具体的なアドバイスや情報を提供しますね. まずは自己分析を行い, 自分の強みや興味を明確にすることが重要です. その上で, 業界研究や企業研究を行い, 志望する企業の情報を集めると良いでしょう. また, 履歴書や職務経歴書の作成, 面接対策も忘れずに行っておくことが大切です. 何か特に気になる点があれば教えてください.	問題解決

表 2: シナリオ 2 における発話と選択モデル

話者	発話内容	モデル
U	最近将来のこと考えてるんだよね	-
A	どんなことを考えているの?具体的に聞かせてくれる?	平均
U	将来のこと考えてると不安になってきちゃうんだ	-
A	そうなんだ, その気持ちとてもよくわかるよ. 将来について考えると, 不安になることって多いよね. 何が一番不安なのかな?	共感

らのシナリオを対象として, 第一著者がケーススタディを行った.

4.1 結果

シナリオ 1 とシナリオ 2 の結果を表 1, 表 2 に示す. 表の話者はユーザが U, エージェントが A を示す. またモデルは選択された他者モデルを示す.

どちらのシナリオもユーザの「最近将来のこと考えてるんだよね」という発話から始まり, この発話からは気持ちの共有や問題の提示といった具体的な意図を読み取ることは困難である. その上で, シナリオ 1 での矛盾値は, 平均他者モデルが 0, 問題解決モデルが 0, 共感モデルが 0 であり, シナリオ 2 での矛盾値もシナリオ 1 と同一であった. 平均他者モデルが選択されたエージェントは発話内容に細かな相違は見受けられるが, どちらもユーザがどういった点で将来のことを考えているかについて具体的な内容を追求する発話を行った. その後, シナリオ 1 とシナリオ 2 で別の典型他者モデルが選択されるようにユーザの発話を入力した.

シナリオ 1 では, ユーザは「就活とかなにから準備すればいいのかなって」といった就活に対する具体的な対策を求める発話を行った. 矛盾値は平均他者モデルが 0, 問題解決モデルが 20, 共感モデルが 80 であった. この結果, 問題解決モデルが選択され, エージェントは就活に対する解決策を述べた.

シナリオ 2 では, ユーザは「将来のこと考えてると不安になってきちゃうんだ」といった将来に対する不安な気持ちを共有する発話を行った. 矛盾値は平均他者モデルが 20, 問題解決モデルが 30, 共感モデルが 20 であった. この結果, 共感モデルが選択され, エージェントは将来の不安に関して共感し慰めるような発話を行った.

以上より, 提案システムは意図が不明確な初期状態では平均他者モデルを用い, その後の発話内容に応じて適切なモデルへと繊維することが確認された.

5 おわりに

本研究では, シチュエーションによらず「察する」対話システムの構想として, 意図を読む AI と典型他者モデルを統合した. ケーススタディを通して, 矛盾値に基づくモデル選択および遷移が設計通りに動作することを確認した.

本検証は実際の対話においてどの程度有効であるかは十分に明らかになっていない. 今後は実際の被験者を対象とした対人実験を行い, 本提案の有効性および対話品質への影響を評価する予定である.

参考文献

- [1] Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, Cambridge, MA, 1987.
- [2] P. Grice, H. Logic and conversation. In Peter Cole and L. Morgan, Jerry editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pp. 41-58. Academic Press, San Diego, CA, 1975.
- [3] Chiga Hayashi. 「察する」とは何か、その発話解釈のメカニズムを探る-関連性理論からの試み-. 城西国際大学紀要 = Josai International University bulletin / 城西国際大学紀要委員会 編, Vol. 26, No. 2, pp. 1-22, 2018.
- [4] Sandra Matz Heinrich Peters. Large language models can infer psychological dispositions of social media users. *arXiv preprint arXiv:2309.08631*, 2023.

- [5] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [6] Ayu Iida, Kohei Okuoka, Satoko Fukuda, Takashi Omori, Ryoichi Nakashima, and Masahiko Osawa. Integrating large language model and mental model of others: Studies on dialogue communication based on implicature. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pp. 260–269. ACM, 2024.
- [7] Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.
- [8] Stella Biderman Sara Hooker Tim Rocktäschel Laura Ruis, Akbir Khan and Edward Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *arXiv preprint arXiv:2210.14986*, 2023.
- [9] Stephen C. Levinson. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press, Cambridge, MA, 2000.
- [10] Daniel Fried Yejin Choi Maarten Sap, Ronan Le Bras. Neural theory-of-mind? on the limits of social intelligence in large LMs. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [11] Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *arXiv preprint arXiv:2301.06627*, 2023.
- [12] Steven Adler Sandhini Agarwal Lama Ahmad Ilge Akkaya Florencia Leoni Aleman Diogo Almeida Janko Altschmidt Sam Altman Shyamal Anadkat Red Avila Igor Babuschkin Suchir Balaji Valerie Balcom Paul Baltescu Haiming Bao Mohammad Bavarian Jeff Belgum Irwan Bello Jake Berdine Gabriel Bernadett-Shapiro Christopher Berner Lenny Bogdonoff Oleg Boiko Madelaine Boyd Anna-Luisa Brakman Greg Brockman Tim Brooks Miles Brundage Kevin Button Trevor Cai Rosie Campbell Andrew Cann Brittany Carey Chelsea Carlson Rory Carmichael Brooke Chan Che Chang Fotis Chantzis Derek Chen Sully Chen Ruby Chen Jason Chen Mark Chen Ben Chess Chester Cho Casey Chu Hyung Won Chung Dave Cummings Jeremiah Currier Yunxing Dai Cory Decareaux Thomas Degry Noah Deutsch Damien Deville Arka Dhar David Dohan Steve Dowling Sheila Dunning Adrien Ecoffet Atty Eleti Tyna Eloundou David Farhi Liam Fedus Niko Felix Simón Posada Fishman Juston Forte Isabella Fulford Leo Gao Elie Georges Christian Gibson Vik Goel Tarun Gogineni Gabriel Goh Rapha Gontijo-Lopes Jonathan Gordon Morgan Grafstein Scott Gray Ryan Greene Joshua Gross Shixiang Shane Gu Yufei Guo Chris Hallacy Jesse Han Jeff Harris Yuchen He Mike Heaton Johannes Heidecke Chris Hesse Alan Hickey Wade Hickey Peter Hoeschele Brandon Houghton Kenny Hsu Shengli Hu Xin Hu Joost Huizinga Shantanu Jain Shawn Jain et al. OpenAI, Josh Achiam. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [13] 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.
- [14] Anand S. Rao and Michael P. Georgeff. Modeling rational agents within a bdi-architecture. In Michael Wooldridge and Nicholas R. Jennings, editors, *Readings in Agents*, pp. 317–328. Morgan Kaufmann, 1997.
- [15] Ehsan Adeli Russ Altman Simran Arora Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji Annie Chen Kathleen Creel Jared Quincy Davis Dora Demszky Chris Donahue-Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh-Li Fei-Fei Chelsea Finn Trevor Gale

に表出する認知モデルの検証. 電子情報通信学会
研究会講演論文集 (KEN), August 2025. 講演番
号: 20250821gcm1.

Lauren Gillespie Karan Goel Noah Good-
man Shelby Grossman Neel Guha Tatsunori
Hashimoto Peter Henderson John Hewitt Daniel
E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas
Icard Saahil Jain Dan Jurafsky Pratyusha
Kalluri Siddharth Karamcheti Geoff Keeling
Fereshte Khani Omar Khattab Pang Wei Koh
Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony
Lee Jure Leskovec Isabelle Levent Xiang Lisa Li
Xuechen Li Tengyu Ma Ali Malik Christopher D.
Manning Suvir Mirchandani Eric Mitchell Zanele
Munyikwa Suraj Nair Avanika Narayan Deepak
Narayanan Ben Newman Allen Nie Juan Carlos
Niebles Hamed Nilforoshan-Julian Nyarko Giray
Ogut Laurel Orr Isabel Papadimitriou Joon Sung
Park Chris Piech Eva Portelance Christopher
Potts Aditi Raghunathan Rob Reich Hongyu
Ren Frieda Rong Yusuf Roohani Camilo Ruiz
Jack Ryan Christopher Ré Dorsa Sadigh Shiori
Sagawa Keshav Santhanam Andy Shih Krishnan
Srinivasan Alex Tamkin Rohan Taori Armin W.
Thomas Florian Tramèr Rose E. Wang William
Wang et al. Rishi Bommasani, Drew A. Hud-
son. On the opportunities and risks of foundation
models. *arXiv preprint arXiv:2108.07258*, 2022.

- [16] Nick Ryder Melanie Subbiah Jared 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 M. Ziegler Jeffrey Wu Clemens Win-
ter Christopher Hesse Mark Chen Eric Sigler
Mateusz Litwin Scott Gray Benjamin Chess
Jack Clark Christopher Berner Sam McCand-
lish Alec Radford Ilya Sutskever Dario Amodei
Tom B. Brown, Benjamin Mann. Language
models are few-shot learners. *arXiv preprint
arXiv:2005.14165*, 2020.
- [17] George Yule. *Pragmatics*. Oxford University
Press, Oxford, UK, 1996.
- [18] 阿部将樹, 田足井昇太, 長原令旺, 大森隆司, 大
澤正彦. 繰り返し囚人のジレンマを題材にした
典型他者モデルの獲得. In *Proceedings of the
HAI 2023 Symposium (Human-Agent Interaction
2023)*, 2023. Poster Paper P-73.
- [19] 大須賀友, 飯田愛結, 大澤正彦. 大規模言語モデル
との統合によってエージェントの心的状態を自然