

対話に対する共感のアノテーションと 共感制御可能な対話モデルの構築

鈴江万碧 堀尾海斗 折田奈甫 河原大輔

早稲田大学理工学術院

{m.suzue@asagi., kakakakakaito@akane, orita@, dkw@.}waseda.jp

概要

本研究は、Bloom [1] が提案する情動的共感と認知的共感の二軸を用いて、既存の対話コーパスに共感に関するアノテーションを付与した。アノテーション手法の検討においてクラウドソーシングと GPT-4o による自動アノテーションを比較したところ、後者の方が評価用に作成した人手によるアノテーションとより高い一致率を示した。この GPT-4o によるアノテーションを用いて対話モデルの学習を行い、共感を制御した応答生成を評価した結果、ベースラインモデルを大幅に上回る精度を達成した。

1 はじめに

対話システムにおいて、ユーザーの感情を理解し適切に応答するための共感能力は、自然な対話を実現する上で重要な要素である。近年、ChatGPT をはじめとした大規模言語モデルの発展により、対話システムの応答の自然さは大きく向上している。しかし、これらのシステムは必ずしも適切な共感を示せているわけではない。既存の研究では、共感対話データセットやプロンプト手法が開発されてきたが、共感していない状態については十分な検討が行われていない。対話システムにおける共感をより深く理解し、より適切な対話生成を実現するためには、共感している状態と共感していない状態の両方の特徴を分析する必要がある。

本研究では、Bloom の反共感論 [1] に基づき、共感を情動的共感と認知的共感の二軸で捉える、日本語の共感アノテーションを提案・分析する。具体的には、既存の日本語対話データセットに対し、認知的共感と情動的共感の二軸を用いてアノテーション基準を設計し、クラウドソーシングと GPT-4o の2つの手法でアノテーションを実施した。GPT-4o に

よるアノテーションは評価用に作成したアノテーション結果と高い一致率を示し、この手法の有効性を確認した。さらに、構築したアノテーションデータセットを用いて、共感の制御が可能な日本語対話モデルを学習した。生成された応答の共感を評価したところ、提案手法が対話システムにおける共感的応答の生成に一定の効果を示すことがわかった。

2 関連研究

2.1 共感的対話に関するデータセット

対話システムにおける感情理解と共感的応答の生成に関して、複数の重要なデータセットが構築されている。Rashkin ら [2] は、共感を「感情に基づく個人的な状況に適切に応答する能力」として捉え、32種類の感情状態に基づく約 25,000 件の対話からなる大規模な対話データセット EMPATHETICDIALOGUES を構築した。評価実験では、このデータセットで学習したモデルがより適切な共感的応答を生成できることを示した。Liu ら [3] は感情に対するサポートに特化した ESConv を提案し、「励ます」「共感を示す」などの感情サポート戦略を定義しアノテーションを行った。他にも、Welivita ら [4] は、まず少量の対話データに人手で感情や意図のラベルを付け、それを元に機械学習を用いて、32種類の感情ラベルと「相手の気持ちに寄り添う」「アドバイスを提供する」といった応答の意図がラベル付けされた 100 万対話規模のデータセットを作成している。

日本語のデータセットでは、Sugiyama ら [5] が EMPATHETICDIALOGUES の手法を応用して日本語版 EMPATHETICDIALOGUES を構築した。Saito ら [6] は音声対話システムのためのコーパス STUDIES を構築し、対話相手の感情に対する明示的な共感表現の分析とラベル付けを行っている。

これらの先行研究では、対話相手の感情状態に合

わせた応答や感情に寄り添ったサポートなど、様々な観点から共感度の高い応答生成を目指す試みが提案されているが、共感自体を評価する明確な指標は確立されていない。また、共感をしていない応答を統一的にとらえる枠組みは提案されていない。

2.2 共感的対話の生成

共感的な対話を生成するためのモデルやアプローチも提案されている。EmpSOA [7] は、従来の共感的対話モデルが相手の感情を理解し模倣することのみ焦点を当てていた点を課題とし、システム自身の視点も考慮した対話生成を提案している。Chain-of-Empathy [8] は、Chain-of-Thought 手法を活用し、心理療法の手法を思考段階に取り入れることで、感情状態に関する論理的な推論を行う。また、Qian ら [9] は大規模言語モデルを用いた共感的応答生成に対して、文脈内学習、二段階生成、知識ベースの活用という3手法を提案し、その有効性を検証している。Zhan ら [10] は、心理療法で用いられる認知的再評価に基づくプロンプト手法を提案し、臨床心理士による評価実験を通じてメンタルヘルス支援での有効性を示している。

3 共感のアノテーション

本研究は、Bloom [1] の理論に基づく認知的共感と情動的共感という二軸の基準を導入し、共感的対話生成に新たな側面からアプローチする。共感を二軸で定義することで、共感を示さない応答も同一の枠組みでとらえることが可能となり、より自然で多様な対話システムの実現に貢献することを目指す。

3.1 既存の対話データセット分析

既存の日本語対話データセット (Japanese Daily Dialogue [11] と日本語版 EMPATHETIC DIALOGUES [5]) を質的に分析し、以下の特徴を観察した。

1点目は、感情表現を含まない会話の存在である。例えば(1)のような対話では、感情的な要素がほとんど含まれていない。このような会話で、話者間の感情の共有・共感を評価することは困難である。

- (1) A: 生ごみはいつ収集場に出せばよいですか?
B: 生ごみの収集日は毎週火曜日です。
A: 何時までにごみを出せばよいですか?
B: 7時までです。収集場はわかりますか?
A: 分からないので、教えてくださいませんか?

2点目は、感情が不一致であるにもかかわらず共感的な会話の存在である。例えば(2)の会話では、BはAと同じ感情を共有してはいないものの、Aの気持ちを理解した上で適切な声かけを行っている。

- (2) A: 暇で何人か誘ったのにみんなに断られた。
B: それは残念だったね。
A: 一人ぼっちの気分だよ。
B: そんなことない、みんな忙しかっただけよ。

これらの例が示すように、会話における共感を適切に評価するためには、感情の一致・不一致とは独立した、相手の立場や気持ちを理解し尊重しているかという観点も必要である。本研究は、共感的応答における後者の特徴をとらえるために、次節で説明する認知的共感と情動的共感という二軸を用いたアノテーション方法を提案する。

3.2 認知的共感と情動的共感

本研究では、対話における共感を明確に定義・評価するため、Bloom [1] が提案する認知的共感 (cognitive empathy) と情動的共感 (emotional empathy) という二軸的定義を採用した。前節の会話データの質的分析で観察された感情表現を含まない会話や感情は一致しないが共感性のある会話を適切に説明できる理論的枠組みとして選択した。

認知的共感とは、他者の心理状態を知的に理解し、その立場に立って考える能力を指す。対話においては、相手の発話内容に対して感情的な共鳴を示さなくとも、状況を理解した上で適切な応答をする場合が該当する。情動的共感とは、他者の感情状態に共鳴し、同様の感情を自身も体験する能力である。対話においては、相手の感情表現に対して同様の感情を示す応答として現れる。これら2種類の共感はずしも同時に同じ強度で生じるわけではなく、図1のようにグラデーションがあると仮定する。

3.3 アノテーション基準

3.2節で定義した認知的共感と情動的共感の概念を、実際のアノテーション作業に落とし込むため、以下の評価基準を定めた。具体的なアノテーション例を図2に示す。

認知的共感については、「相手の感情を理解し、適切に応答しているか」という観点からアノテーションを行った。

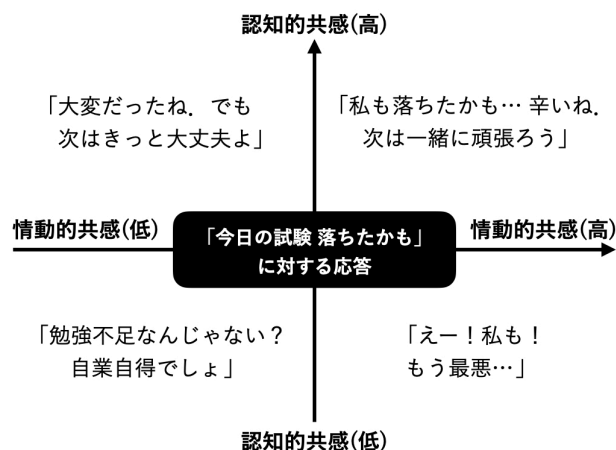


図1 認知的共感と情動的共感の組み合わせによる応答パターンの例

- 相手の気持ちを理解し、寄り添った返答をしている、または相手のことを尊重しているか（「はい」または「いいえ」）¹⁾

情動的共感については、「話者同士の感情がどの程度一致しているか」という観点からアノテーションを行った。

- はい：相手の感情に同調している（同じ感情や気持ちを抱いていると考えられる）場合
- いいえ：相手とは異なる感情を示している場合
- わからない：感情の推測が困難な場合

3.4 使用したデータセット

本研究では、Japanese Daily Dialogue [11] と日本語版 EMPATHETIC DIALOGUES [5] の2つのデータセットに含まれる計25,261対話に対してアノテーションを実施した。情動的共感の評価で「わからない」とアノテートされた対話（全体の約79.6%）は除外し、最終的な学習データは5,150対話となった。

3.5 アノテーション手法の比較実験

本研究では、効率的かつ高品質なアノテーション手法を検討するため、クラウドソーシングとGPT-4oによる自動アノテーションを比較した。

クラウドソーシング Yahoo!クラウドソーシングを用いて、対話コーパスに対して提案のアノテーションを実施した。ワーカーには各発話に対して、認知的共感と情動的共感の2点についてアノテーションを依頼した。アノテーション時には、各選択

1) 「わからない」という選択肢も検討したが、該当する事例が既存のコーパス上で見当たらず、二択とした。

A: 最近、何をやっても上手くいなくて落ち込みます。
B: Aさんはいつも頑張ってますよ。
A: でも、周りの人と比べると自分が劣っている気がして…
B: そんなことはないですよ、自分のペースで進めば大丈夫です。

Bの最後の発話に対するアノテーション

→ ✓ 認知的 × 情動的

理由: Bは相手の感情を理解した上で励ましを示しており、認知的共感が見られる。一方で、Aの感情には同調せず異なる態度を示しており、情動的共感は見られない。

図2 認知的共感と情動的共感に基づくアノテーションの例

肢の判断基準を明示し、具体的な対話例とその評価理由を提示することで、判断基準の統一を図った。各発話に対して5名のワーカーから回答を収集し、最頻値を当該発話のラベルとした。

GPT-4oによる自動アノテーション GPT-4oを用いたアノテーションでは、クラウドソーシングと同一の評価基準に基づくプロンプトを設計した。各対話の話者AとB双方の発話について、認知的共感と情動的共感のアノテーションを行った。評価をできるだけ一貫的にするため、プロンプトには判断基準の詳細な説明と、異なる性質を持つ対話例（感情的な対話と実務的な対話）を含めた。

評価用データセットの作成 アノテーション手法の精度を評価するため、共著者による人手でのアノテーションを実施した。アノテーションの基準について十分な説明を行った上で、Japanese Daily Dialogue [11] からランダムに選択した153対話に対してアノテーションを実施した。

アノテーション精度の比較 表1は、評価用データセットを正解データとした際の各アノテーション手法の評価結果を示している。対話における各発話方向（A → B, B → A）において共感の種類（認知的共感・情動的共感）が正解データと一致するかを評価した。GPT-4oによる自動アノテーションは全評価項目において高い精度を示した。特に認知的共感では双方向で85%前後の高い一致率を達成した。一方、クラウドソーシングでは、B → Aの認知的共感（81.2%）を除き、特に情動的共感において30-40%台の低い一致率となった。

4 共感を制御する対話モデル

前節でアノテーションを実施した共感ラベル付き対話データセットの有効性を検証するため、認知的共感と情動的共感の二軸を制御可能な対話モデルの

表 1 各アノテーション手法の精度比較

評価項目	クラウドソーシング	GPT-4o
A → B 認知的共感	52.6%	84.4%
A → B 情動的共感	29.9%	68.2%
B → A 認知的共感	81.2%	89.0%
B → A 情動的共感	40.3%	74.0%

学習を行う。

4.1 実験設定

モデル 本研究では、以下の 2 つの日本語大規模言語モデルをファインチューニングする。

- llm-jp-3-1.8b²⁾
- rinna/japanese-gpt2-medium³⁾

学習手法 特定の属性に基づいて言語モデルの出力を制御する SteerLM [12] の手法を参考に、3.3 節で述べたアノテーション基準によって GPT-4o でラベルを付与した 5,150 対話のデータセットを用いて、認知的共感と情動的共感を制御可能な形でファインチューニングを実施した。このラベル付きデータを訓練・検証・評価用に 8:1:1 の比率で分割し、各発話に対して話者情報、発話直前の文脈、および 2 種類の共感それぞれについて 2 値のラベル（あり: 1, なし: 0）を付与した形式で学習を行った。学習データの具体的な形式については付録 A に示す。

評価 以下の手順で学習したモデルを評価した。

1. テストデータから認知的共感・情動的共感の有/無の組み合わせ 4 通りを 25 対話ずつ、計 100 対話を無作為に抽出
2. 抽出した対話に対し、認知的共感・情動的共感の有無（制御目標）を指定して応答を生成
3. 生成された応答に対し、アノテーション基準に基づき GPT-4o による共感の評価を実施
4. 指定した制御目標と評価結果の一致率を算出し、モデルの共感制御性能を評価

4.2 実験結果と考察

評価の結果を表 2 に、生成例を付録 B に示す。評価結果から、以下の三点が明らかになった。

第一に、ファインチューニング済みモデルは両方の共感軸においてベースラインを大きく上回る性能を示した。特に認知的共感では、LLM-JP で 38 ポイント、Rinna で 43 ポイントの改善が見られ、提案手

表 2 共感制御対話モデルの性能評価

モデル	認知的共感	情動的共感
LLM-JP (FT)	57.0%	41.0%
LLM-JP (Base)	19.0%	20.0%
Rinna (FT)	60.0%	39.0%
Rinna (Base)	17.0%	18.0%

注: FT=ファインチューニング済み, Base=ベースライン. 数値は共感制御目標と GPT-4o 評価の一致率。

法の有効性が確認された。

第二に、認知的共感（57-60%）と情動的共感（39-41%）の制御精度に顕著な差が見られた。この差は複数の要因が考えられる。(1) 情動的共感のアノテーションでは約 80%が「わからない」と判定された。学習データからこれらを除外し「はいいいえ」の 2 値のみを使用したことは、実運用時の性能評価において考慮すべき制約となる。(2) 情動的共感、話者の感情状態の一致という条件を必要とするため、モデルにとってより困難なタスクとなった可能性がある。(3) 言語モデルが学習した感情表現パターンと、本研究で定義した情動的共感の基準との間にギャップが存在する可能性がある。

第三に、LLM-JP と Rinna の性能差は比較的小さく、両モデルとも同様の傾向を示した。このことは、提案手法が特定のモデルに依存せず、一般的な日本語大規模言語モデルに適用可能であることを示唆している。

5 おわりに

本研究では、既存の日本語対話データセットに対して認知的共感と情動的共感の二軸でアノテーションを行い、これらの共感ラベルを用いて対話モデルの学習を行った。評価実験では提案手法が対話システムにおける共感的応答の生成に一定の効果を示すことを確認した。今後の課題として、対話相手の感情の種類によって適切な共感を生成する方法の検討が挙げられる。例えば、喜びに対しては情動的共感を伴う応答が、悲しみに対しては認知的共感に基づく慰めが効果的である可能性がある。

2) <https://huggingface.co/llm-jp/llm-jp-3-1.8b>

3) <https://huggingface.co/rinna/japanese-gpt2-medium>

謝辞

本研究は JSPS 科研費 JP23K17641, JP23K22374 の助成を受けて実施した。

参考文献

- [1] Paul Bloom. **Against Empathy: The Case for Rational Compassion**. Ecco Books, 2016.
- [2] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5370–5381, 2019.
- [3] Siyang Liu, Chujie Zheng, Orianna Demasia, Sahand Sabour, Yu Li, Zhou Yu, Yong Jianga, and Minlie Huang. Towards emotional support dialog systems. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3469–3483, 2021.
- [4] Anuradha Welivita, Yubo Xie, and Pearl Pu. A large-scale dataset for empathetic response generation. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 1251–1264, 2021.
- [5] Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. Empirical analysis of training strategies of transformer-based japanese chat systems. In **2022 IEEE Spoken Language Technology Workshop (SLT)**, 2022.
- [6] Yuki Saito, Yuto Nishimura, Shinnosuke Takamichi, Kentaro Tachibana, and Hiroshi Saruwatari. Studies: Corpus of japanese empathetic dialogue speech towards friendly voice agent. In **Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022)**, 2022.
- [7] Weixiang Zhao, Yanyan Zhao, Xing Lu, and Bing Qui. Don't lose yourself! empathetic response generation via explicit self-other awareness. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 13331–13344, 2023.
- [8] Yoon Kyung Lee, Inju Lee, Minjung Shin, Seoyeon Bae, and Sowon Hahn. Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models. **Korean Journal of Cognitive Science**, pp. 23–48, 2023.
- [9] Yushan Qian, Weinan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 6516–6528, 2023.
- [10] Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C. Ong. Large language models are capable of offering cognitive reappraisal, if guided. **arXiv preprint arXiv:2404.01288**, 2024.
- [11] 赤間怜奈, 磯部順子, 鈴木潤, 乾健太郎. 日本語日常対話コーパスの構築. 言語処理学会 第 29 回年次大会発表論文集, pp. 108–113, 2023.
- [12] Yi Dong, Zhilin Wang, Makesh Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In **Findings of the Association for Computational Linguistics: EMNLP 2023**, 2023.

A 学習データ形式の具体例

本研究で学習に使用したデータは以下のような形式である。

【入力】	
User:	高橋さん、一緒にジェットコースターに乗りませんか？
Assistant:	僕は、ジェットコースターが苦手です。
User:	どうして苦手なのですか？
Assistant:	高いところから、ゆっくり上った後の急降下をする瞬間が一番怖いからです。
User:	確かに、その瞬間はドキドキしますね。
認知的共感レベル:	1
情動的共感レベル:	0
Assistant:	
【出力】	
清水さんは、怖くないのですか？	

B 共感制御対話モデルによる応答例

学習したモデル（LLM-JP）による共感制御の具体例を示す。

入力: 最近、なかなか眠れなくて困っています。		
制御目標		生成された応答
認知的	情動的	
0	0	睡眠不足はお肌のトラブルにもつながりますから、早寝早起きを心がけるとよいですよ。
1	0	分かりました。それでは、睡眠に効果があると言われている音楽を聴いてみましょう。
0	1	困りましたね。どうしたら良いのでしょうか？