SIPeN: パーソナルナラティブから構築された尺度推意ベンチマーク

佐野朗人 [1] 土井智暉 [1] 綿引周 [1] 谷中瞳 [1]
[1] 東京大学
{akito_sano, doi-tomoki701, amanew, hyanaka}@g.ecc.u-tokyo.ac.jp

概要

心の理論とは, 他者の心の状態や意図を理解する能力である. 近年, 大規模言語モデル（LLM）の発展に伴い, その能力を評価するための分析が進められている. その中でも, 尺度表現に対して聞き手が文字通りの意味を超えた推論を行う「尺度推意」は注目される課題の一つである. しかし, これまでの研究では, 自然なテキストに基づく評価は十分に行われていなかった. そこで本研究では, 尺度推意に関する自然なテキストデータセットを構築し, LLM の尺度推意の推論能力を評価する. 結果から, GPT-4o, Llama 3.1 の尺度推意の推論能力には改善の余地があることを明らかにした.

1 はじめに

心の理論（Theory of Mind; ToM）は, 思考, 信念, 意図や目的といった他者の心の状態を理解する能力である [1]. 人工知能が ToM を持つかどうかという問題は, 認知科学分野を中心に長きにわたり議論されている重要な問題である. 近年の大規模言語モデル（LLM）の急速な発展に伴い, LLM が ToM をどの程度獲得できているかについて, 様々な分析が進められている.

ToM の分析に用いられる推論課題の一つが, 尺度推意 (scalar implicature) という，数量表現などの尺度表現を含む発話が文字通りの意味とは異なる意味を伝達する推意である. 尺度推意は，前提が真であるとき仮説も必ず真なら含意, 偽なら矛盾, どちらともいえないなら中立を判定する, 自然言語推論タスクでも扱われている推論の一つである．例として，次の前提 (1) と仮説 (2) について考えよう.

(1) He ate some of the oranges.

(2) He ate all of the oranges.

(1) について，話者は過不足なく情報を与えるべきという Grice の量の格率に基づくと, わざわざ some と発話した話者の意図に従って「すべてのオレンジを食べたわけではない」ということが推論される. したがって, (1) に対して (2) は矛盾となる.

先行研究 [2] により LLM の尺度推意の性能が低いことが示されている一方で, 既存の尺度推意のベンチマークはテンプレートを用いた人工的なデータが中心であり [2, 3, 4], 自然なテキストデータの尺度推意の推論能力は明らかでない.

そこで本研究では, 日常的な事柄に関する個人の語り（パーソナルナラティブ）のデータから, 尺度推意の推論能力を評価するベンチマーク SIPeN（Scalar Implicature of Personal Narratives）を構築する. このデータセットでは, 尺度表現が品詞に基づいて 3 つのカテゴリに分類されており, それぞれの尺度推移における推論能力の違いを分析できる. また, 尺度表現の強弱や変化の段階数を制御することで, 尺度推意の推論能力の一貫性を測定できる. 実験では SIPeN を用いて, GPT-4o と Llama3.1 を評価する. 結果として, 量化詞と形容詞に関する尺度推意の推論能力が低いこと, また尺度表現の強弱や段階性において一貫性を欠く推論傾向が明らかになった.

2 関連研究

これまで，複数の尺度推意のベンチマークが構築されてきた．Jeretic ら [3] はテンプレートを用いて半自動的に尺度推意ベンチマークを構築し，[4] は形容詞に関する尺度推意データセットを主にクラウドソーシングによって構築した．いずれにおいても，ファインチューニングされた BERT やその派生モデルの評価が行われ，学習後もこれらの尺度推意の推論能力が限定的であることを示した．また，[2] は人間の ToM の獲得に関連する複数の課題を統合した LLM の ToM 評価のベンチマークを構築し，その中で数詞に関する尺度推意の推論能力の評価を

表 1　本研究で扱う尺度表現．左から右にかけて意味的に弱い尺度表現となる．

カテゴリ	尺度	問題数
量化詞	<all, most, many, some, few>	446
数詞	<n,...,2,1>	368
	<ten, ..., two, one>	223
形容詞	<hot, warm>	14
	<excellent, good>	18
	<cold, cool>	7

行った．具体的には [5] の手法に基づいて 40 個の会話テンプレートを設計し, GPT-3 が一貫した予測を行わないことを示した．

しかしながら，既存の尺度推意のベンチマークは，テンプレートによる人工的なデータか品詞が形容詞に限られているデータであり，現実の多様で複雑なテキストデータを反映していない．そこで本研究では自然なテキストデータを使用してベンチマークを構築し, in-context learning の設定で LLM の尺度推意の推論能力を評価する．

3　データセットの構築

現実の多様で複雑なテキストデータにおけるモデルの尺度推意の推論能力を評価するために, 自然なテキストデータを用いた尺度推意に関するベンチマーク SIPeN を構築する．構築にあたっては, リストアップした尺度表現を含む文に対し, 対象の尺度表現の意味を強めるまたは弱めることで, 前提と仮説のペアを作成する．

本データセットの特徴として, 尺度表現を強めた場合と弱めた場合の比較, また強弱の段階を変化させた際の比較を行える点がある．例として表 2 の量化詞のケースを考える．たとえば前提 Throughout the day, **many** birds drink out of it and bathe in it. の量化詞 many を表 1 に基づいて書き換えることで, $H_{+1}, H_{+2}, H_{-1}, H_{-2}$ という 4 つの仮説が作成できる．このとき H_{+1} と H_{-1}, H_{+2} と H_{-2} の比較により, 尺度表現を強めた場合と弱めた場合のモデルの予測の違いを観察できる．また, H_{+1} と H_{+2}, H_{-1} と H_{-2} の比較により, 尺度を同じ方向に 1 段階変化させた場合と 2 段階変化させた場合の予測の違いを分析できる．

3.1　自然なテキストデータ

自然なテキストデータとして, 個人的な出来事に関するパーソナルナラティブのコーパスである PersonaBank[6] を用いる．PersonaBank は, ウェブ上から収集された健康や天気, 野生動物といった日常的かつ個人的な題材についての 108 の英語ブログ記事から構成されている．1 記事あたりの平均の単語数は 269 単語となっている．PersonaBank に含まれる記事から, 尺度表現を用いたテキストを抽出することで, 既存の人工的なテキストデータセットよりも長く, かつ自然な文脈で尺度推意のデータを構築できる．

3.2　尺度表現のリスト

尺度表現は Horn scale[7] に基づいてリスト化し, 表 1 のように, 量化詞, 数詞, 形容詞の 3 カテゴリを対象とした．尺度表現は他に副詞なども存在するが, ここでは PersonaBank 中に多く含まれており, 4 節で後述する尺度表現の強弱変化の段階数での分析が可能になる表現を優先した．結果として, 表 1 に示す 6 個の尺度を対象とし, 21 個の英単語, そしてアラビア数字のみからなる 1 以上の数値の計 22 パターンを尺度表現として検出対象とした．

3.3　前提・仮説ペアの生成

本研究では, 3.2 節でリストアップした 22 件の尺度表現に対し, PersonaBank を構成する 108 のテキストデータを対象に筆者がアノテーションを行った．イディオム的用法を除くため, 尺度表現は名詞句をなすものに限定して検出した．

前提はアノテーションを行った文を用い, 仮説は前提中の尺度表現について意味を強める・弱める方向の変更を行った文を用いた．また可能なものに関しては 2 段階の変更を行った．

その後, 前提の先頭が大文字のものを疑問文・命令文とみなして除外するなど, いくつかの条件で自動処理を行い除外した．結果, 387 件の前提が残り, 1,076 件の仮説が生成された．前提・仮説ペアの例を表 2 に示す．

3.4　ラベル付与

前提・仮説ペアに対し（意味論的）含意ラベルと推意ラベルの 2 種類の推論ラベルを付与した．含意ラベルと推意ラベルの違いとして，次の前提 (3) と仮説 (4) を考える．

(3)　図書館で本を 3 冊借りた．

(4)　図書館で本を 4 冊借りた．

表 2　PersonaBank から構築した推論テストセットの例. 太字が尺度表現, H_{+1}, H_{+2} は Horn Scale に基づいて T の尺度表現の意味を (1 段階, 2 段階) 強めた場合, H_{-1}, H_{-2} は同様に意味を弱めた場合を表す.

カテゴリ	前提（T）	仮説（H_+, H_-）
量化詞	Throughout the day, **many** birds drink out of it and bathe in it.	H_{+1}: Throughout the day, **most** birds drink out of it and bathe in it. H_{+2}: Throughout the day, **all** birds drink out of it and bathe in it. H_{-1}: Throughout the day, **some** birds drink out of it and bathe in it. H_{-2}: Throughout the day, **few** birds drink out of it and bathe in it.
数詞	We ate a hasty meal then crawled into our sleeping bags for an **11** hour slumber.	H_{+1}: We ate a hasty meal then crawled into our sleeping bags for an **12** hour slumber. H_{+2}: We ate a hasty meal then crawled into our sleeping bags for an **13** hour slumber. H_{-1}: We ate a hasty meal then crawled into our sleeping bags for an **10** hour slumber. H_{-2}: We ate a hasty meal then crawled into our sleeping bags for an **9** hour slumber.
形容詞	A gigantic medallion and **warm** soup to all finishers, and I made good use of the showers and changing rooms.	H_+: A gigantic medallion and **hot** soup to all finishers, and I made good use of the showers and changing rooms.

意味論的に (3) を「少なくとも 3 冊借りた」と解釈する場合, この推論の含意ラベルは「中立」となる. 一方で 1 節で述べたように, 推意ラベルは「矛盾」となる.

仮説が前提中の尺度表現を強めたものか, 弱めたものかに応じて, 表 3 の基準に従ってラベルを付与した. 2 種類のラベルでモデルの正答率を比較することで, モデルが意味論的推論を行う傾向か語用論的推論を行う傾向か分析することができる.

表 3　強弱の方向に従ったラベルの付与

強弱の方向	推意ラベル	含意ラベル
強める	矛盾	中立
弱める	矛盾	含意

4　実験

4.1　実験設定

本研究では, SIPeN を用いて LLM が尺度推意の推論能力をどの程度備えているかを評価した. 評価対象モデルは, OpenAI が提供する GPT-4o, GPT-4o-mini[1] および, Hugging Face で公開されている Llama-3.1-instruct-8B（以下, Llama-8B）, Llama-3.1-instruct-70B（以下, Llama-70B）[8] である.

本実験では in-context learning の設定として zero-shot でモデルの尺度推意の推論能力を評価する. この設定では, 文脈情報としてアノテーションが付与された文より前の全ての文をモデルに与えた. 回答形式は ”Yes”, ”No”, ”Maybe” の 3 択であり, それぞれを推意ラベルおよび含意ラベルに割り当てて正答率を算出した. [2]

4.2　結果と分析

本研究の結果を表 4 に示す. 推意ラベルの正答率を分析すると, 今回の実験に用いた 4 種類のモデルのうち Llama-70B が全てのカテゴリにおいて最も正答率が高かった. Llama-70B においては, 小型モデルである Llama-8B と比較して, 全てのカテゴリで推意ラベルの正答率が向上する傾向が見られた. この結果は, パラメータ数の増加によってモデルが尺度推意をより適切に学習できる可能性を示唆している.

次に, 比較対象である含意ラベルに対する正答率を分析すると, 量化詞については含意ラベルに正答しているパターンが高いことが確認された. 特に, Llama-70B 以外の 3 モデルでは, 全体の約 5 割超が含意ラベルに正答していた. この結果は, 量化詞を含むケースについて意味論的推論を行う傾向があることを示唆する.

また, 推意ラベルおよび含意ラベルの双方で誤答する割合を分析すると, Llama-8B では, 全てのカテゴリにおいてこの割合が Llama-70B に比べて高かった. この結果は, パラメータサイズが小さいモデルの方が推論能力が限定的である可能性を示している.

尺度表現のカテゴリ別傾向　数詞に対する推意ラベルの正答率は, GPT-4o において 7 割, Llama-70B では 9 割を超え, 他のカテゴリと比較して最も高い性

1）https://openai.com/index/gpt-4o-system-card/

2）使用したプロンプト及びラベルの割り当ては付録 A に示す.

表 4 尺度表現のカテゴリ別の正答率 (%)

モデル	量化詞		数詞		形容詞	
	推意	含意	**推意**	含意	**推意**	含意
GPT-4o-mini	13.5	**56.7**	76.5	10.5	15.4	20.5
GPT-4o	23.3	56.3	87.3	6.4	17.9	**28.2**
Llama-8B	5.4	54.9	56.0	**18.6**	12.8	17.9
Llama-70B	**49.8**	33.6	**92.6**	2.9	**53.8**	15.4

表 5 尺度表現の強弱変化の段階数での比較

モデル	1 段階	2 段階
Some, 意味を強める		
GPT-4o-mini	16.5	20.3
GPT-4o	17.7	35.4
Llama-8B	1.3	2.5
Llama-70B	49.3	73.4
Two, 意味を強める		
GPT-4o-mini	90.5	95.2
GPT-4o	95.2	95.2
Llama-8B	76.2	90.5
Llama-70B	95.2	95.2
One, 意味を強める		
GPT-4o-mini	65.2	76.1
GPT-4o	78.3	80.4
Llama-8B	37.0	45.7
Llama-70B	87.0	87.0
All, 意味を弱める		
GPT-4o-mini	6.4	6.4
GPT-4o	14.1	20.5
Llama-8B	2.6	1.3
Llama-70B	38.5	21.8

表 6 尺度表現の強弱の方向性での比較

モデル	強める	弱める
Some, 1 段階		
GPT-4o-mini	16.5	13.9
GPT-4o	17.7	22.8
Llama-8B	1.3	13.9
Llama-70B	49.4	69.6
Two, 1 段階		
GPT-4o-mini	90.5	85.7
GPT-4o	95.2	95.2
Llama-8B	76.2	57.1
Llama-70B	95.2	95.2

能を示した. 一方, 量化詞および形容詞の推意ラベルの正答率は, 最も高い Llama-70B でも約 5 割にとどまり, これらのカテゴリでの推論能力に課題があることを示唆している.

この結果の考えられる原因の一つとして, 数詞が表す量は具体的で明確であるのに対し, 量化詞や形容詞が表す量や程度は文脈に依存して変化しやすいため, 推意が曖昧になりやすい可能性がある. ただし, 表 1 に示した通り形容詞は量化詞や数詞に比べてテストセットのデータ数が少なく, データ分布の違いも結果に影響を与えている可能性がある.

強弱変化の段階数の影響 前提中の尺度表現および強弱の方向性を統一し, 強弱変化の段階数が推意ラベルの正答率に与える影響を分析した. 表 5 を見ると, (all → most → many) 以外の尺度では, 2 段階の変化をさせた際に 1 段階の変化よりも正答率が高くなる傾向が見られた. これは, 尺度表現の差が大きい場合に LLM が推意ラベルをより正確に予測できることを示している. 一方で, (all → most → many) においては, (all → most) が (all → many) よりも高い正答率を示した. この結果は, many の曖昧性 [9] が性能を低下させる一因である可能性を示唆する.

強弱の方向性の影響 前提中の尺度表現および強弱変化の段階数を統一した上で, その尺度表現の意味を強めた方向および弱めた方向の推意ラベルの正答率の比較に基づく分析を行った. 表 6 を見ると, (some, 1 段階の変化) の場合, 意味を弱めた方向の方が強めた方向より正答率が高かった. 一方で, (two, 1 段階の変化) の場合には意味を強めた方向の方が高い正答率を示し, 逆の結果となった. このような傾向は, 強弱の方向性そのものよりも, 尺度表現の種類に依存する要因が支配的である可能性を示している.

5 おわりに

本研究では, ToM（Theory of Mind）を分析する推論の一つである尺度推意について, 自然なテキストデータから推論データセットを構築し, GPT-4o をはじめとする LLM の尺度推意の推論能力を評価した. 実験の結果, 一部のモデルが推意ラベルに対して高い正答率を示した一方で, 量化詞や形容詞における推論の正確性に課題があり, 尺度表現の程度の強弱に対して一貫した推論が困難である傾向が確認された. これらの結果は LLM の尺度推意の推論能力に改善の余地があることを示唆する. 今後, 副詞などのより多様な尺度表現を含むようデータセットを拡張し, LLM の分析を進める.

謝辞

本研究は JSPS 科研費学術変革領域研究（B）「ナラティブ意識学」JP24H00809 の支援を受けたものである.

参考文献

[1] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? **Behavioral and Brain Sciences**, Vol. 1, No. 4, p. 515–526, 1978.

[2] Cameron R. Jones, Sean Trott, and Benjamin Bergen. Comparing humans and large language models on an experimental protocol inventory for theory of mind evaluation (EPITOME). **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 803–819, 2024.

[3] Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8690–8705, Online, July 2020. Association for Computational Linguistics.

[4] Rashid Nizamani, Sebastian Schuster, and Vera Demberg. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 14784–14795, Torino, Italia, May 2024. ELRA and ICCL.

[5] Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. **Topics in cognitive science**, Vol. 5, No. 1, pp. 173–184, 2013.

[6] Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. PersonaBank: A corpus of personal narratives and their story intention graphs. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 1026–1033, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[7] Laurence Horn. **A natural history of negation**. University of Chicago Press, 1989. （ローレンス R. ホーン『否定の博物誌』, 河上誓作監訳 濱本秀樹・吉村あき子・加藤泰彦訳, ひつじ書房, 2018）.

[8] Llama Team Ai @ Meta. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[9] Stefan Heim, Corey T. McMillan, Robin Clark, Stephanie Golob, Nam E. Min, Christopher Olm, John Powers, and Murray Grossman. If so many are "few," how few are "many"? **Frontiers in Psychology**, Vol. 6, , 2015.

A　実験設定

A.1　プロンプト

実験で用いたプロンプトは以下の通り.

```
# Your task is to determine whether a given
    premise implies, contradicts, or has no
    clear relationship to a hypothesis.
For each question, evaluate the relationship
    based on the provided premise and hypothesis
    . Your response **must only** use one of the
     following answers: {options}

**Do not provide any additional explanations,
    comments, or answers outside of these
    options.**

Premise: {premise}

Question: Does the given context imply the
    following sentence?

Hypothesis: {hypothesis}
Choices: {options}
Answer: '''
```

A.2　選択肢に対するラベルの付与

実験において, 表 7 のように選択肢に対して推意ラベルと含意ラベルを付与した.

表 7　選択肢に対するラベルの付与

強弱の方向	推意ラベル	含意ラベル
強める	”No”	”Maybe”
弱める	”No”	”Yes”