とりたて助詞に関する推論データセットの拡張と大規模言語モデルの評価

三上 燿輔 [1,2]　谷中 瞳 [1,2,3]
[1] 東京大学　[2] 理化学研究所　[3] 東北大学
{ymikami, hyanaka}@is.s.u-tokyo.ac.jp

概要

文中の特定の要素を強調する役割を持つとりたて助詞（「も」,「しか」など）を含む自然言語推論 (NLI) は，多層的な意味解釈や，とりたて対象の多様性により，含意関係の判断が複雑になる点で挑戦的である．そのようなとりたて助詞に関する広範な言語現象を扱った既存の日本語 NLI データセットは存在するものの，大規模言語モデル (LLM) の評価については，データのバリエーション，正解ラベルの分布に偏りがあるといった課題がある．そこで本研究では，既存データセットを人手および LLM を用いて拡張する．そしてそのデータセットを用いて LLM の評価・分析を行い，LLM が表層的特徴を中心に推論を行っている可能性を示す．

1　はじめに

とりたて助詞とは，文中の特定の要素を際立たせ，同類の要素との関係を背景として意味的な効果を付加する助詞であり [1]，累加の「も」，限定の「だけ」や「しか」などが存在する．

とりたて助詞を含む表現では，単一の命題だけでなく，多層的な意味解釈が生じる場合がある．例えば，「太郎も来た」という文は，太郎が来たということに加え，太郎以外の人物も来たことを示唆する．また，とりたての対象は名詞句に限らず，節や述語などにも及び，数量表現や比較表現のような他の意味的に複雑な表現と組み合わさる場合も多い．このため，とりたて助詞を含む自然言語推論 (Natural Language Inference; NLI) は容易ではなく，大規模言語モデル (Large Language Model; LLM) がこのような推論をどの程度適切に行えるのかは，十分に明らかにされていない．

先行研究では，日本語の意味理論や含意関係認識システムの評価を目的に，とりたて助詞を含む多様な言語現象を対象とした NLI データセット JSeM [2] が構築されている．しかし，LLM の出力傾向や一貫性といった観点から挙動を分析するには，データのバリエーションや正解ラベルの分布の偏りなどの課題がある．

そこで本研究では，JSeM に対して，人手および LLM による拡張を行い，LLM の挙動分析により適したデータセットを構築する．加えて，日本語に特化したモデルを中心とした LLM の評価を通じて，とりたて助詞を含む推論におけるモデルの特性を分析する．

2　日本語 NLI データセット

日本語における多様な言語現象に着目した NLI データセットとして JSeM [2] が提案されている．JSeM は人手で構築され，意味論的な観点に基づく様々な現象を体系的に扱っている．とりたて助詞は JSeM の中で 1 つの独立したセクションとして扱われており，各種のとりたて助詞に関する問題が全 813 問収録されている．各問題は前提と仮説から構成され，前提から仮説を推論できるかどうかについてのラベル (yes, no, unknown) が付与されている[1)]．(1) は収録されている問題の一例である．

(1)　answer: no
P　太郎しかレポートを出さなかった．
H　太郎はレポートを出さなかった．

JSeM は LLM にとっても挑戦的であると考えられる推論例を扱っているが，LLM の性能評価に用いるにはいくつかの課題が存在する．具体的にはまず，構築方針の都合上，各問題は異なる言語現象や推論タイプを問う形式となっているため，few-shot 設定における性能評価や，内容語を変更した場合の解答の一貫性を評価することは困難である．また，ラベ

1) 一部，unacceptable などのラベルが付されているものも存在する．

ルの分布に偏りが見られ，とりたて詞のセクションでは，yes, no, unknown のうち，およそ 75%が yes に集中している．このような偏りは，モデルの出力傾向や性能を正確に分析する上での妨げとなる可能性がある．

そこで，本研究では，JSeM を LLM の評価に適した形に拡張し，日本語とりたて助詞を含む NLI における LLM の性能評価及び分析を行う．

3 データセット拡張

本節では JSeM データセットの拡張手法について述べる．拡張に際して，yes, no, unknown 以外のラベルを持つ問題や，解釈の曖昧性により複数の正解ラベルがある問題については，LLM の評価で扱うには困難であるため除外した．結果として，正解ラベル分布は (yes/no/unknown) = (508/72/89) となった．

3.1 新規問題の追加

本研究では，JSeM で扱われていないが，挑戦的であると考えられる問題を人手で追加した．ここではその一部を取り上げる．

3.1.1 前提 (presupposition)

前提 (presupposition) とは，ある発話が適切に解釈されるために現在の文脈に含まれていなければならない情報のことである． [3]．JSeM が取り上げているとりたて助詞の中にも前提を伴うものが存在する．累加を表す「も」はその一例である [4]．例えば，(2a) は (2b) を前提とする．

(2) a. 太郎**も**来た．
b. 太郎以外で，来た人がいる．

前提が持つ重要な性質の 1 つに，次の例のように含意を打ち消す操作（否定やモーダルの付加など）に影響を受けないという点がある．

(3) answer: yes
P 太郎も来た**というわけではない**.
H 太郎以外で，来た人がいる．

このような例は JSeM に収録されていないため，本研究で新たに追加した．

3.1.2 演算子のスコープと格

日本語では，可能を表す接辞が動詞に付く場合，目的語に対して対格 (例：太郎は右手**を**挙げられる) と主格 (例：太郎は右手**が**挙げられる) のいずれも許容されることが知られている．ここに限定のとりたて助詞「だけ」を加えた場合，用いる格の違いによって，限定と可能のスコープ関係が変化することが指摘されている [5].

(4) a. 太郎は右手だけ**を**挙げられる．
b. 太郎は右手だけ**が**挙げられる．

対格を用いた (4a) の文は可能の方が高いスコープを取り，「太郎は右手だけを挙げるという行為ができる（左手も挙げられるかもしれない）」と解釈される．一方で，主格を用いた (4b) の文は限定の方が高いスコープを取り，「太郎が挙げられるのは右手だけだ」と解釈される．この結果，(4a) は「太郎は右手しか挙げられない」を含意しないが，(4b) は含意するというような推論上の違いが生じる．このような違いに基づいた推論を問う問題を追加した．

3.1.3 分配性と排他性の相互作用

分配性とは，述語の項が複数の個体を表す場合に，その述語の意味がそれぞれの個体に適用されるという性質である [6]．例えば，「太郎と次郎が来た」は「太郎が来た」を含意する．これは述語「来た」が分配的に解釈されるためである．

一方で，限定のとりたて助詞「だけ」を用いた以下の例では，(5a) と (5b) は互いに矛盾する．

(5) a. 太郎と次郎だけが来た．
b. 太郎だけが来た．

この推論には，(5a) が分配的に含意する「次郎が来た」という命題と，(5b) が含意する「太郎以外は来なかった」という排他的な命題から矛盾を導くことが必要である．このような段階的な推論を必要とする問題を新たに追加した．

3.2 ラベル分布の均一化

3.1 節での問題追加後においても，正解ラベルの分布は (yes/no/unknown) = (559/85/104) であり，依然として大きな偏りがある．本節では，この分布をある程度均一にするために行ったことをステップごとに説明する．

3.2.1 単調性推論のバリエーションの追加

まず，主に unknown ラベルを増やすために，単調性推論に関する問題の拡張を行った．単調性とは，集合の包含関係に従って含意関係が保たれる性質を

指す [7, 8]．とりたて助詞を用いた表現の中にも単調性を示すものがあり，JSeM では以下 (6) のような例が扱われている．

(6) a. この授業には，**文学部の一年生**さえ出席している．

b. この授業には，**一年生**さえ出席している．

太字の名詞句には「文学部の一年生 ⊆ 一年生」という包含関係が成り立ち，それに従って (6a) から (6b) への含意関係が導かれるが，逆は導かれない．

JSeM では，このような名詞句の包含関係を連体助詞「の」によって構成しているが，下位語への置換や形容詞や関係節の付加によっても構成できる [9]．そこで，それぞれの置換パターンについて，GPT-5.1[2] を用いて自然な例文を生成し，新たに問題を追加した．

3.2.2 仮説文の肯定・否定の反転

次に，no ラベルを増やすことを目的とした問題の追加を行った．基本的には，ラベルが yes になる問題に関して，その仮説文が肯定文なら否定文に，否定文なら肯定文に反転させることによってラベルを no にすることができる．ただし，反転させることによって不自然な文となる場合もあったため，そのようなケースは除外した．以降，この段階までで得られた各問題のことを**ベース問題**と呼ぶ．ベース問題の正解ラベル分布は (yes/no/unknown) = (575/321/145) である．

3.2.3 内容語のバリエーションの拡充

ラベル分布の均一化と同時に，内容語のみを変えた際に LLM が一貫した解答をするかどうかを検証するために，内容語のバリエーション拡充を行った．具体的には，3.2.2 節までで得られた正解ラベルが [yes, no, unknown] のベース問題に対して，その内容語のみを変えた問題をそれぞれ [4, 7, 15] 問ずつ GPT-5.1 によって生成した．生成された問題の妥当性については，ランダムに抽出した一部の問題を著者が人手で確認した．以上の手順により，最終的な正解ラベルの分布は (yes/no/unknown)=(2875/2568/2320) となった．

2) https://openai.com/ja-JP/index/gpt-5-system-card

4 実験

3 節で拡張したデータセットを用いた実験を行った．

4.1 Zero-shot 設定

4.1.1 設定

モデル 本実験では，Llama-3.1-Swallow-[8B/70B]-Instruct[3]（以降 Swallow-[8B/70B]） [10], Llama-3.1-[8B/70B]-Instruct[4]（以降 Llama-[8B/70B]） [11], GPT-5.2[5] を評価対象とする．Swallow-[8B/70B] は Llama-[8B/70B] を日本語コーパスで継続事前学習したモデルであり，日本語特化モデルとして評価する．また，Llama-[8B/70B] はベースラインとして，GPT-5.2 は商用モデルの参考値として評価する．使用したプロンプトは付録 A に示す．

評価指標 データセット全体に対する性能を測る指標として正答率を用いる．加えて，内容語のみを変更した問題に対して LLM が一貫したラベルを予測するかを評価するため，一貫性指標として APA を導入する．具体的には，各ベース問題とそこから生成された問題群に対する予測について，以下で定義される組ごとのラベル一致率 PA を計算し，それを各ベース問題にわたって平均した値を APA とする．

$$\mathrm{PA} = \frac{2}{n(n-1)} \sum_{i<j} \mathbf{1}\left[p_i = p_j\right]$$

ここで，n は問題数，p_k は問題 k に対する予測ラベルを示す．

4.1.2 結果・考察

表 1 zero-shot 設定における各モデルの正答率と APA

	Llama		Swallow		GPT-5.2
	8B	70B	8B	70B	
正答率	.389	.636	.597	.582	**.719**
APA	**.964**	.817	.835	.890	.909

実験結果を表 1 に示す．GPT-5.2 は，正答率および APA のいずれにおいても高い水準を示した．オープンモデルについては，特に Llama-70B が最も高い正答率を達成した．一方，Swallow は，小規模モデ

3) https://huggingface.co/collections/tokyotech-llm/llama-31-swallow-66fd4f7da32705cadd1d5bc6

4) https://huggingface.co/collections/meta-llama/llama-31-669fc079a0c406a149a5738f

5) https://openai.com/ja-JP/index/gpt-5-system-card

ルであっても比較的高い正答率および APA を示した．なお，Llama-8B で APA が最大となっているが，これは解答の大半が yes に偏っていたことによるものである（cf. 付録 B）．

3.1 節で追加した問題では正答率が低く，特に前提に関する問題では，否定やモーダルに引きずられて前提を適切に捉えられていない可能性が示唆された．また，演算子のスコープと格に関しては，いずれの格を用いた場合も同一のラベルを解答する傾向が見られた．さらに，3.1.3 節の段階的な推論に関しては unknown の予測が多く，Chain of Thought を用いたところ，一方からの含意は導いているが，もう一方からの含意と関連付けて矛盾を導けていない傾向が見られた．これらの結果から，モーダルが出現する時は unknown と答えるなどのように，表層的特徴を中心に解答を決定している可能性が示唆される．加えて，単調性については，下位語への置換（例：学生 → 大学生）を含む unknown ラベルの問題に yes と誤答する場合が多く，語彙的な包含関係に関する知識が十分に活用されていない可能性がある．

4.2 Few-shot 設定

4.2.1 設定

Llama-[8B/70B] と Swallow-[8B/70B] について，以下の 2 種類の few-shot 実験を行う．

Few_same 各問題に対して，ベース問題が同じ例をランダムに 1 つ与える．

Few_all 各問題に対して，同一のとりたて助詞を含む yes, no, unknown の例を各 1 例ずつ，ランダムな順序で与える．

4.2.2 結果・考察

表 2 few-shot 設定における各モデルの正答率と APA

		Llama		Swallow	
		8B	70B	8B	70B
Few_same	正答率	.415	.752	.898	**.940**
	APA	.874	.904	.888	**.933**
Few_all	正答率	.552	.679	.544	**.695**
	APA	.753	**.840**	.554	.797

実験結果を表 2 に示す（予測分布は付録 B 参照）．Few_same 設定では，すべてのモデルで zero-shot より正答率が向上し，特に Swallow で顕著な改善が見られた．APA についても，Llama-8B を除くすべてのモデルで上昇が確認された．一方，Few_all 設定では，Swallow-8B を除くモデルで正答率は向上したが，Swallow では APA が大きく低下する傾向が確認された．この結果を Few_same 設定と併せて考えると，Swallow は few-shot 例に含まれる正解ラベルの影響を，Llama よりも受けやすいと考えられる．

また，zero-shot 設定で正答率の低かった問題について，Few_same 設定では大きな性能向上が見られた一方で，Few_all 設定では改善は限定的であった．これは，Few_same 設定では同一の推論パターンに基づく例が与えられるため，段階的な推論を明示的に行わなくとも正答可能になった一方，Few_all 設定では必ずしも類似した推論パターンの例が提示されないためと考えられる．

4.3 PromptEOL

PromptEOL [12] とは，文の意味を 1 語で表現させるプロンプトを LLM に与え，その語を出力する際の隠れ状態をその文の埋め込みとして用いる手法である．ここでは，「『[文]』の意味を 1 単語で表すと『」というテンプレートで前提文や仮説文を LLM に与え，得られた埋め込みに基づいて zero-shot 設定での結果に対するエラー分析を行う．

演算子のスコープと格について，対格文に対する主格文のコサイン類似度を算出したところ 0.940 と高く，対格文とランダムに選んだ 1000 文との類似度 (0.450 ± 0.074) を大きく上回った．よって，埋め込みレベルにおいても対格文と主格文がほぼ同一の意味として表現されている可能性が示唆される．

また，単調性の下位語置換を含む例では，前提文と仮説文の類似度が平均 0.951 と高かった．このことから，語彙的な包含関係ではなく，意味的な類似性に基づいて yes と誤答した可能性が考えられる．

5 おわりに

本研究では，日本語とりたて助詞を含む推論に対する LLM の評価および分析を目的として，既存の日本語 NLI データセットの拡張を行った．拡張したデータセットを用いた実験の結果，LLM は段階的な推論よりも表層的特徴を中心に出力を決定している可能性が示唆された．今後の展望として，PromptEOL に基づく分析に加え，モデル内部の表現を観察する他の手法も併用することで，LLM が困難とする推論におけるエラー要因をより詳細に分析していくことが考えられる．

謝辞

本研究は JSPS 科研費 JP24H00809，JST CREST，JPMJCR2565，JST BOOST，JPMJBY24H5 の支援を受けたものです．

本研究において，データセットに追加する問題のアイデアについて有益な助言を頂いた松岡大樹さんに感謝申し上げます．

参考文献

[1] 日本語記述文法研究会（編）. 現代日本語文法 5. くろしお出版, 東京, 2020.

[2] Ai Kawazoe, Ribeka Tanaka, Koji Mineshima, and Daisuke Bekki. An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In Mihoko Otake, Setsuya Kurahashi, Yuiko Ota, Ken Satoh, and Daisuke Bekki, editors, **New Frontiers in Artificial Intelligence**, pp. 58–65, Cham, 2017. Springer International Publishing.

[3] Christopher Potts. Presupposition and implicature. **The handbook of contemporary semantic theory**, pp. 168–202, 2015.

[4] Sachiko Shudo. **The Presupposition and Discourse Functions of the Japanese Particle Mo**. Routledge, 2013.

[5] Hiroaki Tada. Nominative objects in japanese. **Journal of Japanese Linguistics**, Vol. 14, No. 1, pp. 91–108, 1992.

[6] Lucas Champollion. Distributivity in formal semantics. **Annual review of linguistics**, Vol. 5, No. 1, pp. 289–308, 2019.

[7] Johan Van Benthem. Determiners and logic. **Linguistics and Philosophy**, pp. 447–478, 1983.

[8] Thomas F Icard III and Lawrence S Moss. Recent progress on monotonicity. **Linguistic Issues in Language Technology**, Vol. 9, , 2014.

[9] Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. Can neural networks understand monotonicity reasoning? In Tal Linzen, Grzegorz Chrupa la, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 31–40, Florence, Italy, August 2019. Association for Computational Linguistics.

[10] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. **arXiv preprint arXiv:2404.17790**, 2024.

[11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[12] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In **Findings of the association for computational linguistics: EMNLP 2024**, pp. 3182–3196, 2024.

表 3　zero-shot, few-shot 設定で使用したプロンプト

system	前提文と仮説文が与えられます. 前提文が仮説文を含意しているか答えてください. 「含意」,「中立」,「矛盾」のいずれかで答えてください.
user	前提：[premise] 仮説：[hypothesis]
assistant （few-shot のみ）	解答：[answer]

A　使用したプロンプト

表 3 に実験で使用したプロンプトを示す. 前提が複数ある場合は「前提 1：[premise1], 前提 2：[premise2]」のようにナンバリングして提示した.

B　Zero-shot 設定での LLM の予測分布

図 1 に各モデルの解答に基づく混合行列を示す. 特に Llama-8B では yes の予測が非常に多いことがわかる.

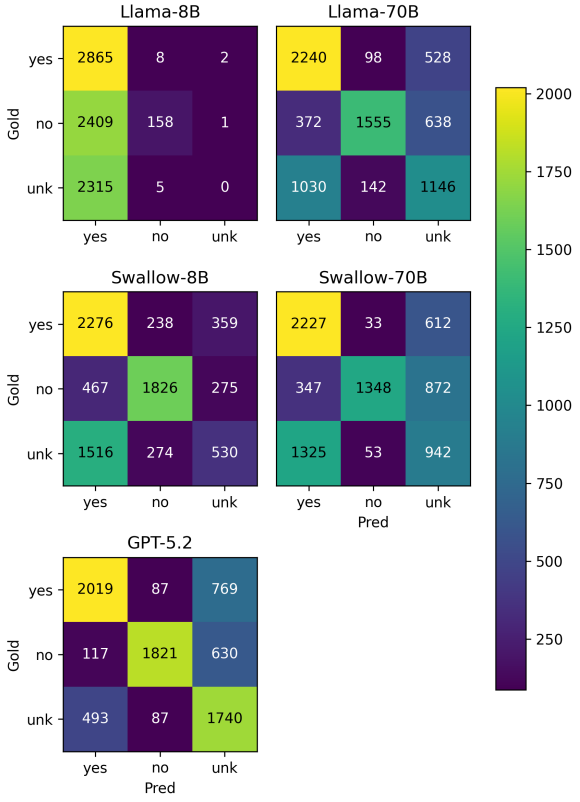

図 1　zero-shot 設定における各 LLM の解答の混合行列（セル内の数字は問題数）

C　Few-shot 設定での LLM の予測分布

図 2, 3 にそれぞれ Few_same, Few_all 設定における各モデルの解答に基づく混合行列を示す.

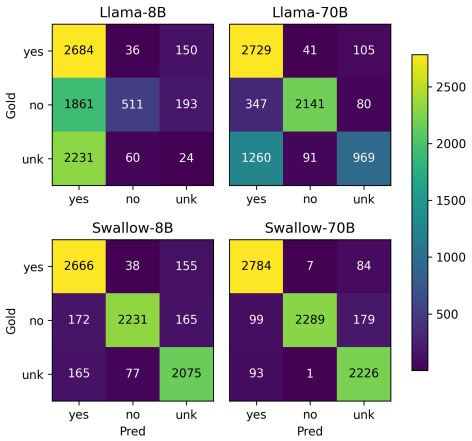

図 2　Few_same 設定における各 LLM の解答の混合行列（セル内の数字は問題数）

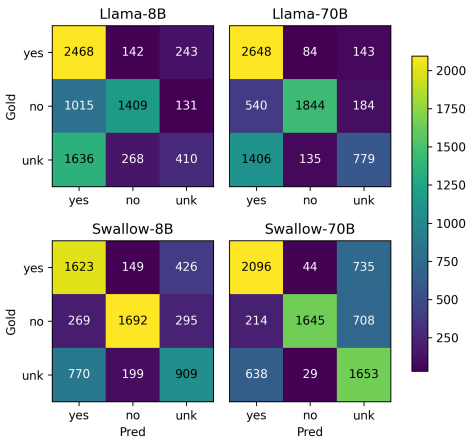

図 3　Few_all 設定における各 LLM の解答の混合行列（セル内の数字は問題数）