否定の焦点の理解能力を評価する自然言語推論データセット

小島 健太郎 [1]　加藤 芳秀 [2]　松原 茂樹 [1,2]
[1] 名古屋大学大学院情報学研究科 [2] 名古屋大学情報連携推進本部
kojima.kentaro.a0@s.mail.nagoya-u.ac.jp

概要

否定の焦点とは，否定要素が最も顕著に，あるいは明示的に否定する部分である．否定の焦点は重要かつ基本的な言語現象であり，それを理解する能力は LLM の性能を向上させるために不可欠である．しかし，LLM が否定の焦点をどの程度理解しているかについては十分に検証されていない．本稿では，LLM が否定の焦点を理解しているかを検証する方法を提案する．本手法では，否定の焦点を理解していれば解けるような自然言語推論の問題を作成し，それが解けるか否かで LLM を評価する．本手法により，否定の焦点をアノテーションしたデータセットである PB-FOC に基づき自然言語推論データセットを構築した．それを用いて LLM を評価した実験について報告する．また，その実験を通じて明らかになったデータセットの問題点について議論する．

1　はじめに

否定の焦点とは，否定要素が最も顕著に，あるいは明示的に否定する部分である [1]．例えば，以下の文 2 では “by car” が否定要素 “n’t” の焦点である．

1. John came to the party to celebrate Mike’s birthday.
2. He didn’t come to the party by car.
3. Instead, he decided to walk there to enjoy the nice weather.

これは，文 1 と文 3 から，「John がパーティーに来た」ことは否定されておらず，「車で来た」ことが否定されていると解釈できるからである．

否定の焦点は基本的な言語現象であり，大規模言語モデル（LLM）において，否定の焦点を理解する能力は，その性能の向上に不可欠であると考えられる．しかし，LLM が否定の焦点をどの程度理解しているかについては，十分に検証されていない．

本稿では，LLM が否定の焦点をどの程度理解しているかを検証する方法を提案する．本手法では否定の焦点の理解の問題を自然言語推論（NLI）の問題に還元する．これは，Anand ら [2] の人手による焦点アノテーション作業を，NLI の問題と捉え直したものと位置づけられる．NLI は前提（premise）が仮説（hypothesis）を含意するか否かを判定するタスクである．本研究では，否定の焦点を理解していれば解けるような NLI の問題を作成し，それが解けるか否かで LLM を評価する．具体的には，次の 2 種類の NLI の問題を作成する．

- 前提として否定要素を含む文 s とその文脈を与える．仮説として s の否定の焦点だけを明示的に否定するように書き換えた文を与える．これは，「含意」が正解となる問題である．
- 前提として否定要素を含む文 s とその文脈を与える．仮説として s の否定の焦点ではない部分を明示的に否定するように書き換えた文を与える．これは，「非含意」が正解となる問題である．

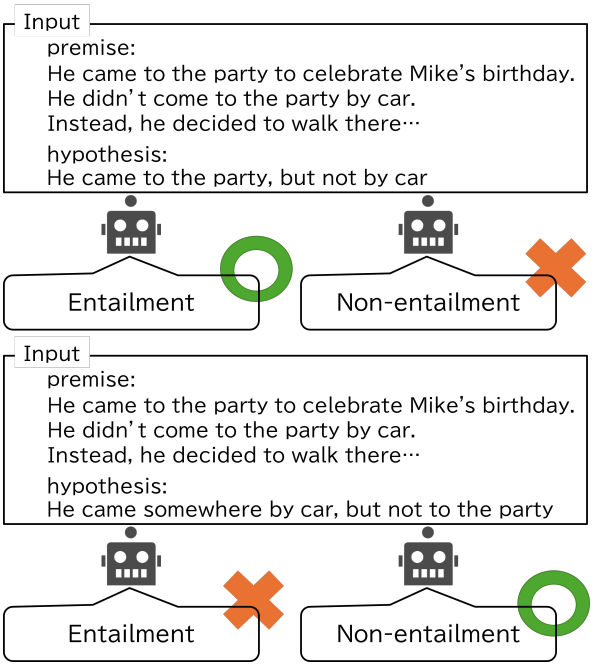

図 1　本手法の概略

本手法の概略を図 1 に示す．NLI の問題に還元する本アプローチには，否定の焦点の位置を同定する直接的なアプローチに対して次のような利点がある．すなわち，後者では，モデルは否定の焦点の理解だけでなく，否定の焦点の言語学的な定義への理解も求められる．一方，本アプローチでは，言語学的な理解といった要因を排して LLM を評価できる．

2 関連研究

本節では，否定の焦点に関する先行研究をデータセットの構築と同定タスクに分けて述べる．特に，本研究と同様の考え方で否定の焦点をアノテーションした Anand ら [2] の研究を詳細に議論する．

2.1 否定の焦点に関するデータセット

否定の焦点をアノテーションしたデータセットとして最も用いられているのは PB-FOC である [3]．PB-FOC では，PropBank [4] のテキストに対して否定の焦点をアノテーションしている．PB-FOC の問題点を指摘して，再アノテーションした研究も行われている [2]．否定の焦点をアノテーションした他のデータセットとして，DT-Neg [5]，SFU Opinion and Comments Corpus [6] などがある．DT-Neg は，対話テキストに否定の焦点をアノテーションしたデータセットであり，SFU Opinion and Comments Corpus は，レビュー文に対して否定の焦点などをアノテーションしたデータセットである．これらは，英語を対象としたデータセットである．英語以外の言語（日本語，イタリア語）に対しても否定の焦点をアノテーションしたデータセットが存在する [7, 8]．

2.2 Anand らの否定の焦点のアノテーション

Anand ら [2] は，PB-FOC の否定の焦点のアノテーションを一貫性のあるものとするために，アノテータに単に否定の焦点にあたる部分をマークするように指示するだけではなく，以下のような手順で確認するように指示している．

1. 文の否定要素を，その文の様々な構成素を否定するような位置に移動する．
2. 1. により得られた文のうち，いずれの文が最も関連があるかを，前後の文脈を参照し決める（最も関連がある文における否定要素の移動先が否定の焦点である）．

例えば，文 2 に対しては，アノテータは各自で以下のような文を作り，どの文が最も関連があるかで否定の焦点の位置を決定している．

2a. Someone but not him came to the party by car.
2b. He came somewhere but not to the party by car.
2c. He came to the party but not by car.

このアノテーションの方法では，アノテータに指示として上述の手順を示しているのみである．一方，本研究では，この手順の結果と同等の NLI の問題を自動的に作成する手法を提案する．否定要素の移動に加えて，LLM を用いた自然な文への修正や人手による確認（3.1.2 節，3.1.3 節）も行う．

2.3 否定の焦点の同定

否定の焦点を同定する研究では，依存構造を用いる手法 [9]，前後の単語とトピックの情報を用いる手法 [10, 11]，談話構造を用いる手法 [12]，エンコーダ型のモデルを用いる手法 [13, 14] などが提案されている．また，DT-Neg [5] を用いて対話テキストにおける否定の焦点を同定する研究 [15, 16] や，否定の焦点を同定して，暗黙的かつ肯定的な主張を取り出す研究 [17] も行われている．

否定の焦点の位置を同定するモデルを評価する研究調査がある一方，デコーダ型の言語モデルが否定の焦点をどの程度理解しているかを検証する研究はこれまでになく、本研究では、その理解度を評価する。

3 否定の焦点の理解能力評価のための NLI データセットの作成

本節では，否定の焦点がアノテーションされたデータセット PB-FOC [3] を用いて，否定の焦点を理解しているか否かを評価する NLI データセットを構成する方法を提案する．データセットのインスタンスは前提，仮説，ラベルからなる．

前提は，PB-FOC の否定の焦点がアノテーションされた文とその前後の文を結合したものである．仮説はアノテーションされた否定の焦点（負例の場合は，焦点以外の一部）のみを明示的に否定した形式に変換した文である．ラベルは「含意」，「非含意」の 2 値であり，仮説において焦点が明示的に否定されている場合は含意，そうでない場合は非含意である．仮説の構成は，Anand ら [2] の手順 1 に相当し，ラベルの決め方は，手順 2（の逆順）に相当する．

```
Please rewrite the following sentence into a natural sentence without changing its meaning and remaining "but":
{sentence}
```

図 2　LLM での修正に用いたプロンプト

3.1　否定の焦点に着目した仮説文の構成

仮説を構成する手順は以下の通りである．

1. ルールベースで文 s を，否定の焦点のみを明示的に否定した文 s' に変換する．
2. s' を LLM（GPT-5）を用いて自然な文 s'' に書き換える．
3. LLM による書き換えにより意味が変わっていないかを人手により確認する．s' と比べて意味が変わっていないと判定された s'' のみを仮説として採用する．

以下の節では，各手順について詳述する．

3.1.1　ルールに基づく変換

本研究では，Blanco and Sarabi [18] の手法により否定の焦点のみを明示的に否定した文を求める．この手法は次の手順で文を変換する．

1. 文から否定要素を除去する．
   例えば，文 2 は，He came to the party by car. となる．
2. 否定の焦点 f を抽象度の高い表現に置換する．
   例えば，上の文では，否定の焦点 “by car” を抽象的な表現 “in some manner” に置換し，He came to the party in some manner. を得る．
3. “but not $\{f\}$” を節の末尾に追加する．
   例えば，上の文では，“but not by car” を追加し，He came to the party in some manner, but not by car. を得る．

負例については，手順 2 で否定の焦点ではない部分をランダムに選択する．手順 1，2 の詳細については，付録 A で説明する．

3.1.2　LLM による書き換え

ルールベースで書き換えられた文は不自然な場合があるため，LLM を用いて自然な文に修正する．本研究では，GPT-5 を用いて図 2 のプロンプトで修正する．このプロンプトにおいて，単に自然な文に書き換えることだけでなく，but を残すように指示しているのは，否定の焦点のみを明示的に否定する文の形式を維持することが本研究において重要であるからである．LLM による書き換えの結果に but が含まれていない場合，そのインスタンスは除外する．

表 1　データセットの統計量

	正例数	負例数
トレーニングセット	1392	1389
開発用セット	322	320
テストセット	375	316

3.1.3　人手による確認

LLM による修正において意味を変えないように指示しているが，文の意味が変わってしまう可能性があるため，意味が変わっていないかを人手で確認する．アノテータに，書き換える前の文と書き換えた後の文を提示し，それらの文が「同じ意味（same meaning）」，「異なる意味（different meaning）」，「その他（otherwise）」のいずれであるかを選択してもらう．「その他」はどちらかの文の意味がわからないなど，提示された 2 つの文の意味が同じかどうか判断できないときに選択する．ネイティブの英語話者 3 名によってアノテーションを行い，3 名のうち 2 名以上が「同じ意味」と判断したインスタンスのみを採用した．

3.2　データセットの構築

本研究では，PB-FOC のトレーニングセット，開発用セット，テストセットからそれぞれデータセットを構築した．本研究で構築したデータセットの最終的な統計量を表 1 に示す．3.1.3 節の人手による確認の処理は，トレーニングセット，開発用セットに対しては実施せず，テストセットに対してのみ実施した．この確認で，テストセットから，正例は約 6%，負例は約 19%のインスタンスが除外された．人手による確認の一致度を求めるために，Gwet’s AC1 [1)] を各アノテータ間で計測した．その平均は約 0.66 であり，中〜高程度の一致度であった．

4　実験

本研究で構築したデータセットを用いて，いくつかのデコーダ型の言語モデルを評価した結果を報告

[1)] アノテータ間の一致度を測る指標としてよく使われるのは Cohen’s κ 係数であるが，これはラベルが不均衡なデータに対しては不当に結果を低くする場合がある [19]．不均衡データに対しては，Gwet’s AC1 がより頑健であることが示されている [20] ため，今回はこの指標を用いた．

表 2　実験結果

モデル	zero-shot	few-shot
Majority	54.3	-
Llama-3.1-8B-Instruct [21]	51.1	56.2
Mistral-Small-3.1-Instruct [22]	55.6	50.8
Qwen3-32B [23]	61.9	62.1
GPT-4.1-mini [24]	63.0	61.2
GPT-4.1 [24]	65.4	65.6
GPT-5-mini [25]	**66.3**	**67.6**

する．また，構築したデータセットについて，本実験を通じて明らかになった問題点を議論する．

4.1　実験設定

本実験では，3 節で構築したデータセットのテストセットを用いた．オープンモデルとして，Llama-3.1-8B-Instruct [21], Mistral-Small-3.1-Instruct [22], Qwen3-32B [23]，クローズモデルとして，GPT-4.1-mini [24], GPT-4.1 [24]，GPT-5-mini [25] を評価対象とした．プロンプトは先行研究 [26] を参考に一部修正したものを用いた．修正は，先行研究における NLI のタスクが 3 値分類であるのに対して，本研究のタスクは 2 値分類であることによるものである．プロンプトは zero-shot と few-shot の 2 種類を用意し，few-shot では正例・負例それぞれ 2 つずつの例を，評価に用いるインスタンス以外のテストセットからランダムに選択して与えた．今回のデータセットでは，正例と負例の偏りが少ないことから，正解率 (accuracy) を評価指標として用いる．

4.2　実験結果

実験結果を表 2 に示す．GPT-5-mini の正解率が最も高く，zero-shot で 66.3%，few-shot で 67.6%を達成した．一方で，Llama-3.1-8B-Instruct の zero-shot では 51.1%，Mistral-Small-3.1-Instruct の few-shot では 50.8%と，Majority（すべてのインスタンスで「含意」と判定した場合の正解率）よりも低い正解率となった．

一部のモデルがベースラインを下回る結果を受けて，データセットの問題が人手でも解けるかを確認した．著者 1 名が開発用セットの一部（正例と負例の比率がおなじになるようにランダムにサンプリングした 100 インスタンス）に対してラベルを与えた．人手の正解率は 65.0%であった．今回構築したデータセットは，原理的には否定の焦点を理解していれば解けるように設計したが，人手でもそれほど高い正解率を達成できなかった．

4.3　考察

実験結果について，全体的に（人手によるラベル付けも含めて）正解率が低くなった理由を考察する．理由の一つとして，PB-FOC [3] のアノテーションの定義が漠然としている [2] 点が考えられる．

PB-FOC のアノテーションガイドラインでは，否定の焦点は「最も顕著に否定される意味役割（動詞から見たときの句の役割）」と定義されており，顕著かどうかはアノテータの主観的な判断に依存する．その判断の結果は，「否定の焦点だけを否定した文は，全体が否定されている文と同じ意味になる」という性質を持つとは限らない．

2.2 節で議論したように，PB-FOC のアノテーションガイドラインの問題点を指摘し，明確な基準でアノテーションを行った先行研究 [2] が存在する．[2] では，様々な構成素の前に否定要素を移動し，どの位置が最も適切かで判断する基準で PB-FOC のトレーニングセットを再アノテーションした．アノテータ間の一致率は κ 係数で 0.82 と高い一方で，元の PB-FOC のアノテーションとの一致率は 0.48 となり，両者は大きく異なる結果になったと報告されている．すなわち，本実験の正解率が低くなった原因の一つは PB-FOC のアノテーションが，上述の性質を必ずしも持たないことだと考えられる．

5　おわりに

本研究では，LLM が否定の焦点を理解しているか検証するための NLI データセットを構築する手法を提案した．その手法で構築したデータセットを用いて，複数のデコーダ型言語モデルを評価した結果，GPT-5-mini が最も高い正解率を達成した一方で，一部のモデルはベースラインを下回る結果となった．実験結果を踏まえて，PB-FOC のアノテーションに原因がある可能性を指摘した．

今後の課題として，適切に LLM を検証するために，PB-FOC のアノテーションを見直すことが挙げられる．文献 [2] のように新しくアノテーションを施し，再度データセットを構築して，LLM が否定の焦点を理解しているかを検証したい．

謝辞

本研究は JSPS 科研費 JP23K18506 の助成を受けたものです.

参考文献

[1] Rodney D. Huddleston and Geoffrey K. Pullum. **The Cambridge grammar of the English language**. Cambridge University Press, Cambridge, 2002.

[2] Pranav Anand and Craig Martell. Annotating the focus of negation in terms of questions under discussion. In **Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics**, pp. 65–69, Jeju, Republic of Korea, July 2012. Association for Computational Linguistics.

[3] Eduardo Blanco and Dan Moldovan. Semantic representation of negation using focus detection. In **Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies**, pp. 581–589, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[4] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. **Computational Linguistics**, Vol. 31, No. 1, pp. 71–106, 2005.

[5] Rajendra Banjade and Vasile Rus. DT-Neg: Tutorial dialogues annotated for negation scope and focus in context. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation**, pp. 3768–3771, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[6] Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. **Corpus Pragmatics**, Vol. 4, pp. 155 – 190, 2019.

[7] Suguru Matsuyoshi, Ryo Otsuki, and Fumiyo Fukumoto. Annotating the focus of negation in Japanese text. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation**, pp. 1743–1750, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[8] Begoña Altuna, Anne-Lyse Minard, and Manuela Speranza. The scope and focus of negation: A complete annotation framework for Italian. In **Proceedings of the Workshop Computational Semantics Beyond Events and Roles**, pp. 34–42, Valencia, Spain, April 2017. Association for Computational Linguistics.

[9] Sabine Rosenberg and Sabine Bergler. UConcordia: CLaC negation focus detection at *Sem 2012. In ***SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)**, pp. 294–300, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[10] Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Negation focus identification with contextual discourse information. In **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 522–530, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[11] Bowei Zou, Guodong Zhou, and Qiaoming Zhu. Unsupervised negation focus identification with word-topic graph model. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 1632–1636, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[12] Zahra Sarabi and Eduardo Blanco. Identifying the focus of negation using discourse structure. In **The Florida AI Research Society**, 2019.

[13] Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. Negative focus detection via contextual attention mechanism. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing**, pp. 2251–2261, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. Predicting the focus of negation: Model and error analysis. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8389–8401, Online, July 2020. Association for Computational Linguistics.

[15] Rajendra Banjade, Nobal B. Niraula, and Vasile Rus. Towards detecting intra- and inter-sentential negation scope and focus in dialogue. In **The Florida AI Research Society**, 2016.

[16] Dipesh Gautam, Nabin Maharjan, Rajendra Banjade, Lasang Jimba Tamang, and Vasile Rus. Long short term memory based models for negation handling in tutorial dialogues. In **The Florida AI Research Society**, 2018.

[17] Eduardo Blanco and Dan Moldovan. Retrieving implicit positive meaning from negated statements. **Natural Language Engineering**, Vol. 20, No. 4, p. 501–535, 2014.

[18] Eduardo Blanco and Zahra Sarabi. Automatic generation and scoring of positive interpretations from negated statements. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1431–1441, San Diego, California, June 2016. Association for Computational Linguistics.

[19] Domenic V. Cicchetti and Alvan R. Feinstein. High agreement but low kappa: II. resolving the paradoxes. **J Clin Epidemiol**, Vol. 43, No. 6, pp. 551–558, 1990.

[20] Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L. Gwet. A comparison of cohen's kappa and gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. **BMC Med Res Methodol**, Vol. 13, p. 61, April 2013.

[21] Aaron Grattafiori et al. The Llama 3 Herd of Models. **arXiv e-prints**, p. arXiv:2407.21783, July 2024. https://ui.adsabs.harvard.edu/abs/2024arXiv240721783G.

[22] Mistral AI. Mistral Small 3.1, 2025. https://mistral.ai/news/mistral-small-3-1.

[23] An Yang et al. Qwen3 technical report, 2025.

[24] OpenAI. Introducing GPT-4.1 in the API, 2025. https://openai.com/index/gpt-4-1/.

[25] OpenAI. GPT-5 System Card, 2025. https://cdn.openai.com/gpt-5-system-card.pdf.

[26] Beiduo Chen, Siyao Peng, Anna Korhonen, and Barbara Plank. A rose by any other name: LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. In **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 10777–10802, Vienna, Austria, July 2025. Association for Computational Linguistics.

表 3　意味役割と置換表

意味役割	置き換わる句
A0〜A4	something / someone / some people
AM-TMP	at some point of time
AM-LOC	somewhere
AM-MNR	in some manner
AM-CAU	because of something
AM-PRP	to do something

A　ルールに基づく変換の詳述

本付録では，3.1.1 節で述べたルールに基づく変換の否定要素の除去と否定の焦点の置換について説明する．

A.1　否定要素の除去

否定される範囲を否定の焦点（負例の場合は焦点以外からランダムに選択された一部）のみとするために，文全体を否定する否定要素を以下の手順で除去する．

1. not や never などの否定要素を除去する
2. 助動詞や動詞の過去形・三人称単数現在形の s などを修正する (e.g., doesn't go → goes, didn't go → went, won't go → will go)
   - 不規則動詞については先行研究と同様の情報[2)]を参照して対応表を作成した
3. 否定極性表現を，対応する肯定極性表現に置き換える (e.g., any → some, yet → already)
   - 対応表は先行研究と同様に [1] を参照して作成した

A.2　否定の焦点の置換

PB-FOC において，否定の焦点は意味役割単位でアノテーションされている [3]．例えば，文 2 の意味役割は以下のようになり，否定の焦点は，意味役割が割り当てられた部分から選択される（この例の場合は，AM-MNR が否定の焦点として選択される）．

- $[\text{He}]_{\text{A0}}$ did$[\text{n't}]_{\text{AM-NEG}}$ come $[\text{to the party}]_{\text{AM-LOC}}$ $[\text{by car}]_{\text{AM-MNR}}$.

本研究では先行研究に基づいて，以下の表 3 をもとに置換を行う．

[2)] https://en.wikipedia.org/wiki/English_irregular_verbs