否定理解能力を評価するための日本語言語推論データセットの構築

吉田 朝飛 [1]　加藤 芳秀 [2]　小川 泰弘 [3]　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
[3] 名古屋市立大学データサイエンス学部
yoshida.asahi.y7@s.mail.nagoya-u.ac.jp

概要

言語モデルが否定を理解する能力を評価するための様々な英語データセットが構築されているが，日本語においては，そのようなデータセットを構築する取り組みは限られている．本研究では，否定理解能力を評価するための日本語言語推論データセット JNLI-Neg を構築する．さらに JNLI-Neg を用いて既存の言語モデルを評価し，それらの否定理解能力の現状と課題を明らかにする．

1　はじめに

否定 (negation) を正しく理解することは様々な自然言語処理タスクを解く上で重要であるが，言語モデルは否定の処理が苦手であることが報告されている（例えば [1, 2, 3]）．モデルの否定理解能力の適切な評価や，否定に頑健なモデルの開発を目的として，否定に焦点を当てたデータセットが英語で構築されている [4, 5, 6, 7, 8, 9, 10]．しかし，日本語においては，否定の観点でモデルを評価するためのデータセットを構築する取り組みは限られている．英語データセットにおける研究の知見は必ずしも日本語に適用できるとは限らない [11] ため，否定の観点でモデルを評価できる日本語データセットが必要である．本研究では，基本的な言語理解タスクである**自然言語推論** (NLI) を対象として，否定理解能力を評価するための日本語データセットを構築する．

既存の日本語 NLI データセットとして，JSICK [12], JaNLI [13], JNLI [11, 14] などがあるが，これらには否定を含むインスタンスがあまりなく，否定理解能力の評価には不十分であると指摘されている [15, 16]．内田・南條 [15] は，対偶により既存の NLI データセットから否定要素（否定を意味する語）を含むデータセットを自動生成する手法を提案している．しかし，この手法は，否定理解能力のより詳細な評価に必要である，否定の有無に関してのみ異なるデータ対（**否定のミニマルペア**; 表 1 を参照）を含むデータセットを作成できない．また，文末の否定要素のみを対象としており，文の途中に否定要素を含むデータセットは作成できない．

表 1　NLI インスタンスにおける否定のミニマルペアの例．下線は否定要素を，太字はミニマルペア間で有無が異なる否定要素を表す．(1) は「重要でない」否定，(2) は「重要な」否定を含むミニマルペアである．

(1)	*p*: 机の上にいくつかの白い皿がある。 *h*: 机の上に皿がある。 *l*: 含意 (entailment)	*p*: 机の上にいくつかの白く**ない**皿がある。 *h*: 机の上に皿がある。 *l*: 含意 (entailment)
(2)	*p*: 机の上にいくつかの白くない皿がある。 *h*: 机の上に皿がある。 *l*: 含意 (entailment)	*p*: 机の上にいくつかの白くない皿がある。 *h*: 机の上に皿が**ない**。 *l*: 矛盾 (contradiction)

そこで本研究では，否定の理解能力を評価するための新たな日本語 NLI データセット **JNLI-Neg** を構築する[1)]．既存データセット JNLI の前提文や仮説文に否定要素を自動付与し，それに人手で NLI ラベル[2)]を付与することにより，否定を含む新たな NLI インスタンスを得る．JNLI-Neg は十分な数の否定インスタンスを含むだけでなく，否定のミニマルペアを含むという点で既存データセットと性質を異にする．また，JNLI-Neg は文末の否定要素だけでなく文の途中の否定要素も多数含む．これらの特徴により，より詳細に言語モデルの否定理解能力を評価・分析することができる．さらに本研究では，JNLI-Neg を用いて既存の言語モデルの否定理解能力を評価する．幅広い日本語の言語モデルを評価し，否定理解能力の現状と課題を明らかにする．

1)　データセット及びソースコードは https://github.com/asahi-y/JNLI-Neg で公開する．
2)　本研究では，JNLI と同様に含意 (entailment), 矛盾 (contradiction), 中立 (neutral) の 3 種類の NLI ラベルを用いる．

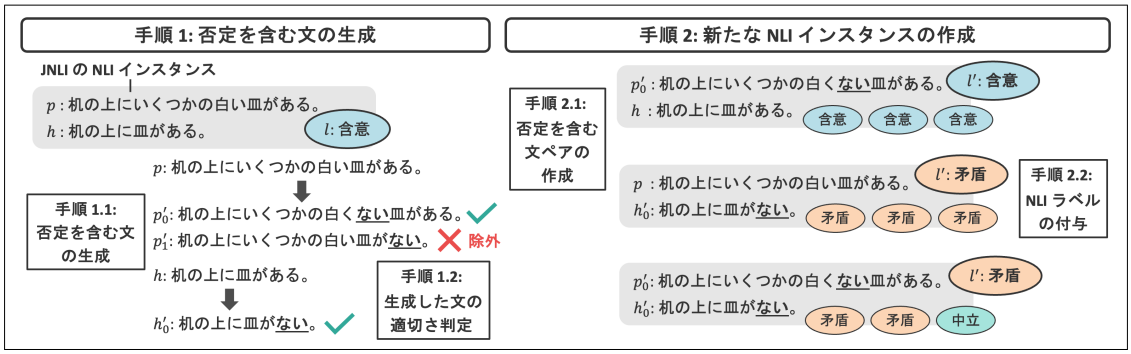

図 1　JNLI-Neg 構築手法の概要

2　関連研究

2.1　否定のミニマルペアを含む NLI データセット

否定に焦点を当てた様々な英語 NLI データセットが構築されている．その中でも否定のミニマルペアを含むように設計されたものは，否定理解能力をより詳細に評価できる．Hossain ら [4] は NLI データセット RTE [17, 18, 19, 20], SNLI [21], MNLI [22] の文に否定要素を付与し，否定のミニマルペアを含むデータセットを構築した．Hartmann ら [5] は XNLI [23] から否定要素を含むインスタンスを抽出してその否定要素を除去し，否定のミニマルペアを含むデータセットを構築した．Hossain らと Hartmann らはミニマルペアにおける否定をその重要性で分類し，否定に対するモデルの振る舞いを分析した．否定の重要性は以下で定義される．

- 否定要素を含むインスタンスと，その否定要素を含まないバージョンのインスタンスにおいて，正解 NLI ラベルが異なるとき，その否定要素が引き起こす否定は重要 (important) である．正解 NLI ラベル が同一のとき，その否定は重要でない．（例を表 1 に示す．）

Hartmann らは，単に否定を含むインスタンスに対するモデルの振る舞いを評価するだけではなく，ミニマルペアを用いて否定に対するモデルの振る舞いをより詳細に分析するべきであると指摘している．

2.2　否定を含むように構築された既存の日本語 NLI データセット

内田・南條 [15] は，対偶により既存の NLI データセットを否定に関して拡張する方法を提案した．この手法で JRTE [24, 25] を拡張し，大半のインスタンスに否定を含むデータセットを作成した．内田・南條 [26] は同様の手法で JSNLI [27] を拡張した．対偶による拡張手法では，前提文と仮説文の両方に否定要素を付与し，2 つの文を入れ替えて新たなインスタンスを得るため，作成されたデータセットは否定のミニマルペアを含まない[3]．さらに，文末に否定要素を含むインスタンスのみが作成され，文の途中に否定要素を含むインスタンスは作成されない．

3　JNLI-Neg の構築

3.1　データセットの構築手法

本節では，否定理解能力を評価するための日本語 NLI データセット JNLI-Neg の構築手法を説明する．本研究では，否定要素「ない」または「ず」1 つによる否定を対象とし，接辞の否定要素（例えば「不」）や二重否定などは対象としない．前節で述べた点を踏まえ，NLI-Neg の要件を以下のように定める．

要件 1: データセットを否定のミニマルペアで構成する．
要件 2: 文末，及び文の途中の否定要素をデータセットに含める．
要件 3: 翻訳を用いると日本語としての不自然さが残るため，翻訳を介さず日本語ベースで構築する．

要件 1，及び 2 は JNLI-Neg 独自の特徴である．これらは，以下で説明する否定を含む文の生成，及び NLI インスタンスの作成により達成される．要件 3 は，翻訳を介さずに構築された JNLI [11, 14] をベースとして用いることで満たされる．図 1 に構築手法の概要を示す．以下では，前提文 p，仮説文 h，ラベル l の組 (p, h, l) を NLI インスタンスと呼ぶ．

3)　内田・南條 [26] は，前提文あるいは仮説文の一方にのみ否定要素を含む拡張インスタンスも作成し，評価用データの一部として利用している．この拡張インスタンスと元のインスタンスのペアは，否定のミニマルペアとみなせる．しかし本研究と異なり，内田・南條 [26] はミニマルペアに焦点を当てたモデルの評価・分析を行っていない．

3.1.1 手順 1: 否定を含む文の生成

手順 1 では，JNLI の NLI インスタンスにおける p 及び h を否定を含む文に変換する．

手順 1.1: ルールによる否定を含む文の生成 否定を含む文は，形態素情報を用いたルールベースで生成する[4]．文を形態素解析し[5]，品詞が動詞，形容詞，形状詞である形態素に続く助詞あるいは助動詞の系列に否定要素である形態素「ない」または「ず」を挿入し，必要に応じて前後の形態素の活用形を変更することで，否定を含む文を生成する．1 文の複数箇所に否定要素を挿入できる場合はその数だけ文を複製し，それぞれに対して 1 箇所のみ挿入する．すなわち，1 文に否定要素を挿入できる箇所が n 個存在する場合，文が n 個生成される．以上の方法により生成される文と元の文の対は，ミニマルペアの要件を満たすことが保証される．

手順 1.2: 生成した文の適切さ判定 手順 1.1 では助詞や助動詞の順序関係しか考慮していないため，生成された文が日本語として適切でない場合がある．例えば，「群衆がいて混雑する」に対して生成される「群衆がいて混雑しない」は意味が矛盾し，不適切である．こうした不適切な文を除外するために，手順 1.1 で生成された文に対してその適切さを判定する．判定には，LLM の in-context learning を用いる[6]．変換前の文，すなわち JNLI の文は適切であるという例を与えた上で，変換後の文が適切であるか否かの 2 値分類を解く (1-shot)．変換前の文を適切な文の例として与える理由は，否定を含む文への変換により不適切さが生じた場合のみを考慮するためである．適切でないと判定された文は除外し，手順 2 では適切であると判定された文のみを用いる．

3.1.2 手順 2: 否定を含む NLI インスタンスの作成

手順 2 では，手順 1 で生成した否定を含む文を用いて，否定を含む NLI インスタンスを作成する．

手順 2.1: 否定を含む文ペアの作成 JNLI のインスタンス $i = (p, h, l)$ において，手順 1 で p, h から生成した否定を含む文の集合をそれぞれ P', H' とする．このとき，$|P'| > 0$ かつ $|H'| > 0$ かつ $neg(p) = 0$ かつ $neg(h) = 0$ を満たす i を対象として，否定を含む NLI インスタンスの集合 $D_{\text{neg}}(i)$ を作成する．ここで，$neg(s)$ は文 s に含まれる否定要素の数を表す[7]．$D_{\text{neg}}(i)$ の定義は以下の通りである．

$$
\begin{aligned}
D_{\text{neg}}(i) &= D_{\text{p}}(i) \cup D_{\text{h}}(i) \cup D_{\text{ph}}(i)\,, \\
D_{\text{p}}(i) &= \{(p', h, l') \mid p' \in P'\}\,, \\
D_{\text{h}}(i) &= \{(p, h', l') \mid h' \in H'\}\,, \\
D_{\text{ph}}(i) &= \{(p', h', l') \mid p' \in P' \wedge h' \in H'\}\,.
\end{aligned}
$$

手順 2.2: 人手によるラベル付与 作成した NLI インスタンスのラベル l' は，人手で付与する．3 人の作業者が各インスタンスに対してラベルを付与し，2 人以上が一致したラベルを採用する．作業者に与える指示は，JNLI のそれと同様である[8]．

3.2 データセットの構築

JNLI の学習セット及び検証セット[9] に対して，3.1 節で説明した手順 1, 手順 2.1 を適用し，その一部に対して手順 2.2 に従ってラベルを付与した．JNLI の学習セット及び検証セットそれぞれに含まれる NLI インスタンスをランダムに並び替え，先頭から順に手順 2.1 までを適用して NLI インスタンスを作成した．そのうち，学習セットから作成した 4,803，検証セットから作成した 1,205 の NLI インスタンス[10] に対して手順 2.2 のラベル付与を行った．アノテーションで 2 人以上のラベルが一致したものを採用した結果，学習セット 4,671 インスタンス，検証セット 1,177 インスタンスの NLI ラベルを得た[11]．

JNLI-Neg を構成する NLI インスタンスの集合 $D_{\text{JNLI-Neg}}$ は，以下で定義する．

$$
D_{\text{JNLI-Neg}} = D_{\text{orig}} \cup D_{\text{neg}}\,,\quad D_{\text{neg}} = \bigcup_{i \in D_{\text{orig}}} D_{\text{neg}}(i)\,.
$$

ここで，D_{orig} はアノテーション対象としたインスタンスのもととなった JNLI の NLI インスタンスの集合である．JNLI-Neg は，以下で定義する否定のミニマルペアの集合 M から構成される．

$$
\begin{aligned}
M &= M_{\text{p}} \cup M_{\text{h}} \cup M_{\text{p,ph}} \cup M_{\text{h,ph}}\,, \\
M_{\text{p}} &= \{(i, i_{\text{p}}) \mid i \in D_{\text{orig}} \wedge i_{\text{p}} \in D_{\text{p}}(i)\}, \\
M_{\text{h}} &= \{(i, i_{\text{h}}) \mid i \in D_{\text{orig}} \wedge i_{\text{h}} \in D_{\text{h}}(i)\},
\end{aligned}
$$

4) ルールの詳細は，ソースコードを参照されたい．
5) 形態素解析器 MeCab [28] 及び辞書 UniDic [29] を用いた．
6) モデルは，OpenAI API (https://openai.com/index/openai-api/) の gpt-4o-2024-05-13 を用いた．設定の詳細や判定性能の評価については，付録 A を参照されたい．
7) 湯浅ら [16] の否定要素検出器を用いて $neg(s)$ を求めた．
8) JNLI のガイドライン (https://github.com/yahoojapan/JGLUE/blob/main/task_guidelines.md) を作業者に提示した．
9) JNLI の評価セットは，本論文執筆時に公開されていないため，学習セット及び検証セットのみを用いた．
10) インスタンス数が切りのよい数でないのは，JNLI のインスタンス単位でアノテーション対象を抽出したためである．
11) アノテーション一致度の評価は付録 B を，JNLI-Neg の統計情報は付録 C をそれぞれ参照されたい．

表 2　評価結果（単位はいずれも%）．Fine-tuning は異なるシード値を用いた 5 回の試行の平均値を示す．

Setting	Model	NLI インスタンス単位			否定のミニマルペア単位					
		D_{JNLI}	$D_{\text{JNLI-Neg}}$		M_{i}			M_{u}		
			D_{orig}	D_{neg}	Acc	Acc′	AccChg	Acc	Acc′	AccChg
Fine-tuning on D_{JNLI}	東北大 $\text{BERT}_{\text{BASE}}$	90.24	87.63	52.30	57.94	28.56	−29.38	72.90	72.64	−0.26
	東北大 $\text{BERT}_{\text{LARGE}}$	92.47	90.32	57.11	66.10	35.55	−30.55	75.12	74.67	−0.45
	早稲田大 $\text{RoBERTa}_{\text{BASE}}$	86.48	82.58	52.20	58.04	29.53	−28.51	71.27	71.21	−0.06
	早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	90.24	88.28	55.87	63.35	34.89	−28.46	75.04	73.78	−1.26
Zero-shot	LLM-jp-3-1.8B-instruct	35.50	35.48	33.56	24.52	25.03	0.51	40.43	42.03	1.60
	LLM-jp-3-3.7B-instruct	65.82	60.22	58.11	49.68	44.98	−4.70	67.17	67.27	0.11
	LLM-jp-3-13B-instruct	79.29	78.49	66.44	71.79	58.96	−12.83	75.83	70.80	−5.03
	Swallow 8B Instruct	45.44	41.40	55.23	42.69	46.63	3.94	53.80	62.03	8.24
Majority Baseline		55.34	55.91	47.66	—			—		

$$M_{\text{p,ph}} = \{(i_{\text{p}}, i_{\text{ph}}) \mid \exists i \in D_{\text{orig}}(i_{\text{p}} \in D_{\text{p}}(i) \wedge i_{\text{ph}} \in D_{\text{ph}}(i))\},$$
$$M_{\text{h,ph}} = \{(i_{\text{h}}, i_{\text{ph}}) \mid \exists i \in D_{\text{orig}}(i_{\text{h}} \in D_{\text{h}}(i) \wedge i_{\text{ph}} \in D_{\text{ph}}(i))\}.$$

M は，以下の式により「重要な」否定を含むミニマルペアの集合 M_{i} と「重要でない」否定を含むミニマルペアの集合 M_{u} に分けられる ($M = M_{\text{i}} \cup M_{\text{u}}$)．この分類は，2.1 節で述べた否定の重要性に関する定義に基づく．

$$M_{\text{i}} = \{((p,h,l),(p',h',l')) \in M \mid l \neq l'\},$$
$$M_{\text{u}} = \{((p,h,l),(p',h',l')) \in M \mid l = l'\}.$$

4　JNLI-Neg を用いた言語モデルの評価

4.1　実験設定

既存の言語モデルの否定理解能力に関する現状と課題を明らかにするために，JNLI-Neg を用いて日本語事前学習済みモデルの性能を評価した．NLI タスクを解く実験設定の基本は，MLM (masked language model) は栗原ら [11] に，LLM は Han ら [30] に従った[12)]．MLM は，JNLI のインスタンス集合 D_{JNLI} を用いてモデルを fine-tuning した[13)]．LLM は，zero-shot で実験を行った．すべての実験において，検証セット上での性能を評価した．

4.2　評価指標

JNLI に従い，正解率 (accuracy) を評価指標のベースとした．これに加え，ミニマルペアに基づき評価を行うため，以下で定義する accuracy change (AccChg) を評価指標として用いた．

$$\text{AccChg} = \text{Acc}' - \text{Acc},$$
$$\text{Acc} = \frac{1}{|M|} \sum_{((p,h,l),(p',h',l')) \in M} \mathbf{1}[\hat{l} = l],$$
$$\text{Acc}' = \frac{1}{|M|} \sum_{((p,h,l),(p',h',l')) \in M} \mathbf{1}[\hat{l}' = l'].$$

ここで，$\hat{l}, \hat{l}'$ はそれぞれ，インスタンス (p,h,l), (p',h',l') に対するモデルの予測ラベルである．AccChg は，ミニマルペア中のインスタンス間における正解率の変化を示す指標であり，否定の有無によるモデル性能の変化を分析する上で有用である．

4.3　実験結果

表 2 に実験結果を示す．D_{JNLI} で学習した MLM の D_{neg} における正解率は，D_{JNLI} や D_{orig} におけるそれと比べて大きく下回った．また，AccChg が −30% 程度と小さい値であった．これらの結果より，D_{JNLI} で学習した MLM は，否定を無視して推論を行っている可能性が示唆される．LLM の zero-shot においては，モデルのサイズや種類により結果に異なる傾向が見られた．より詳細な分析や，モデル間における性能差異の原因の考察は今後の課題とする．

5　おわりに

本研究では，否定の理解能力を評価するための日本語 NLI データセット JNLI-Neg を構築した．さらに，JNLI-Neg を用いてモデルの性能を評価した．今後の課題として，JNLI-Neg を用いて，否定に対するモデルの振る舞いをより詳細に分析することが挙げられる．本研究で対象外とした，接辞の否定要素や二重否定なども含むようにデータセットを拡張することも今後の課題である．

12)　詳細な実験設定は，付録 D.1, D.2, D.3 を参照されたい．
13)　$D_{\text{JNLI}} \cup D_{\text{neg}}$ を用いて MLM を fine-tuning する実験も実施した．その結果は，付録 D.4 を参照されたい．

謝辞

本研究は，一部，科学研究費補助金基盤研究（C）（No. 22K12148）により実施しました．また，実験の一部は，名古屋大学のスーパーコンピュータ「不老」を利用して実施しました．本研究で利用したJNLIデータセット及びそのアノテーションガイドラインを提供いただいた栗原健太郎氏，河原大輔氏，柴田和秀氏に感謝いたします．

参考文献

[1] Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. Say what you mean! large language models speak too positively about negative commonsense knowledge. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 9890–9908, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In **Proceedings of the 12th Joint Conference on Lexical and Computational Semantics**, pp. 101–114, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Mengyu Ye, Tatsuki Kuribayashi, Jun Suzuki, Goro Kobayashi, and Hiroaki Funayama. Assessing step-by-step reasoning against lexical negation: A case study on syllogism. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 14753–14773, Singapore, December 2023. Association for Computational Linguistics.

[4] Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. An analysis of natural language inference benchmarks through the lens of negation. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing**, pp. 9106–9118, Online, November 2020. Association for Computational Linguistics.

[5] Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. A multilingual benchmark for probing negation-awareness with minimal pairs. In **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 244–257, Online, November 2021. Association for Computational Linguistics.

[6] Thinh Hung Truong, Yulia Otmakhova, Timothy Baldwin, Trevor Cohn, Jey Han Lau, and Karin Verspoor. Not another negation benchmark: The NaN-NLI test suite for sub-clausal negation. In **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing**, pp. 883–894, Online, November 2022. Association for Computational Linguistics.

[7] Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 8729–8755, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[8] Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics**, pp. 1803–1821, Toronto, Canada, July 2023. Association for Computational Linguistics.

[9] Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. This is not a dataset: A large negation benchmark to challenge large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 8596–8615, Singapore, December 2023. Association for Computational Linguistics.

[10] Orion Weller, Dawn Lawrie, and Benjamin Van Durme. NevIR: Negation in neural information retrieval. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 2274–2287, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[11] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. 自然言語処理, Vol. 30, No. 1, pp. 63–87, 2023.

[12] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. **Transactions of the Association for Computational Linguistics**, Vol. 10, pp. 1266–1284, 2022.

[13] Hitomi Yanaka and Koji Mineshima. Assessing the generalization capacity of pre-trained language models through Japanese adversarial natural language inference. In **Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP**, pp. 337–349, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[14] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[15] 内田巧, 南條浩輝. 否定表現を伴う文における含意関係認識のための対偶によるデータ拡張. 情報処理学会研究報告, 自然言語処理, 2023.

[16] 湯浅令子, 吉田朝飛, 加藤芳秀, 松原茂樹. 否定の観点からみた日本語言語理解ベンチマークの評価. 言語処理学会第 31 回年次大会論文集, 2025.

[17] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In **Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment**, pp. 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[18] Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In **Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment**, 2006.

[19] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In **Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing**, pp. 1–9, Prague, June 2007. Association for Computational Linguistics.

[20] Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In **Proceedings of the Second Text Analysis Conference**, Gaithersburg, Maryland, USA, November 2009. NIST.

[21] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[22] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[23] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[24] 林部祐太. 知識の整理のための根拠付き自然文間含意関係コーパスの構築. 言語処理学会第 26 回年次大会論文集, pp. 820–823, 2020.

[25] Yuta Hayashibe. Japanese realistic textual entailment corpus. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 6827–6834, Marseille, France, May 2020. European Language Resources Association.

[26] 内田巧, 南條浩輝. 否定表現を伴う文における自然言語理解の性能検証. 言語処理学会第 30 回年次大会論文集, pp. 1581–1586, 2024.

[27] 吉越卓見, 河原大輔, 黒橋禎夫ほか. 機械翻訳を用いた自然言語推論データセットの多言語化. 情報処理学会研究報告, 自然言語処理, Vol. 2020, No. 6, pp. 1–8, 2020.

[28] Takumitsu Kudo. Mecab : Yet another part-of-speech and morphological analyzer. **http://mecab.sourceforge.jp**, 2006.

[29] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源 : 形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–123, 10 2007.

[30] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. 言語処理学会第 30 回年次大会論文集, pp. 2085–2089, 2024.

A　文の適切さ判定の詳細

A.1　プロンプト及びハイパーパラメータ

JNLI-Neg の構築における手順 1.2 で LLM に与えたプロンプトを図 2 に示す．プロンプトは，JNLI とは無関係な文[14]を用いた予備実験により作成したものである．

user: 次の日本語が文法的・意味的に正しいかどうかを判定してください。正しい場合は「正しい」、そうでない場合は「正しくない」と出力してください。「正しい」「正しくない」のいずれかのみを出力することを遵守してください。
user: {JNLI の文 }
assistant: 正しい
user: { 判定対象の文 }

図 2　LLM による文の適切さ判定のプロンプト

OpenAI API のハイパーパラメータとして，temperature は 0，max_tokens は 32 を用いた．

A.2　判定性能の評価

LLM による文の適切さ判定の性能を評価するために，手順 1.1 において JNLI の学習セットから作成された否定を含む文から 200 文をランダムサンプリングし，各文が日本語として適切であるか否かを人手で判定した[15]．LLM による判定と同様に，JNLI の文は正しいという前提のもとで，それを元に生成された否定を含む文の適切さ判定を行った．人手判定の結果を正解として LLM による判定の性能を評価した．「不適切」を positive クラス，「適切」を negative クラスとして[16]2 値分類タスクの評価をした結果，適合率 0.892，再現率 0.815，F_1 値 0.852 であった．

B　アノテーション一致度の評価

JNLI-Neg の構築手順 2.2 におけるアノテーション作業の一致度を評価するため，アノテーション対象とした全インスタンスに対して，3 人の作業者のラベルの Fleiss' Kappa 係数を算出したところ，0.458 であった．JNLI における値は 0.399 であることから，JNLI-Neg のアノテーション作業は JNLI と同程度の一致度で行われたと言える．

C　JNLI-Neg の統計情報

JNLI-Neg の統計情報を表 3 に示す．また，ミニマルペアにおける否定要素の位置の分布を表 4 に示す．ここで，手順 1.1 において否定要素を挿入した位置が，文末の句読点の直前であればその位置を「文末」，そうでなければ「文の途中」とした．

表 3　JNLI-Neg の統計情報

		学習セット	検証セット
NLI インスタンス	D_{orig}	823	186
	D_{neg}	4,671	1,177
	合計	5,494	1,363
否定のミニマルペア	M_{i}	3,475	787
	M_{u}	3,260	935
	合計	6,735	1,722

表 4　JNLI-Neg のミニマルペアにおける否定要素の位置の分布

		文末	文の途中
M_{i}	学習セット	2177	1298
	検証セット	485	302
M_{u}	学習セット	1494	1766
	検証セット	424	511

14）　JSICK [12] の学習セットから抽出した文を用いた．
15）　判定は著者の 1 人が行った．
16）　不適切な文を除外するフィルタリングタスクとして考えたため，「不適切」を positive クラスとした．

D　評価実験の詳細

D.1　評価したモデルの詳細

評価対象としたモデルは，masked language model (MLM) と生成系の大規模言語モデル (LLM) に分類される．いずれも，Hugging Face (https://huggingface.co/) で公開されているモデルを用いた．モデルの詳細を表 5 に示す．

表 5　利用したモデルの詳細

分類	本論文の表記	Hugging Face 上の名称
MLM	東北大 $\text{BERT}_{\text{BASE}}$	tohoku-nlp/bert-base-japanese-v3
	東北大 $\text{BERT}_{\text{LARGE}}$	tohoku-nlp/bert-large-japanese-v2
	早稲田大 $\text{RoBERTa}_{\text{BASE}}$	nlp-waseda/roberta-base-japanese
	早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	nlp-waseda/roberta-large-japanese-seq512
LLM	LLM-jp-3-1.8B-instruct	llm-jp/llm-jp-3-1.8b-instruct
	LLM-jp-3-3.7B-instruct	llm-jp/llm-jp-3-3.7b-instruct
	LLM-jp-3-13B-instruct	llm-jp/llm-jp-3-13b-instruct
	Swallow 8B Instruct	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3

D.2　学習セットのデータ分割

MLM の fine-tuning においては，学習セットをランダムに並び替えた上で，80% を**学習用**データ，20% を**検証用**データとした．表 2 に示す結果は，データ分割のシード値を変えて 5 回の試行を行った平均値である．

D.3　ハイパーパラメータ

MLM については，表 6 に示すハイパーパラメータを用いた．learning rate 及び epoch については，**検証用**データを用いた探索により最適な値を選択した．

表 6　MLM の実験で用いたハイパーパラメータ

パラメータ名	値
learning rate	{5e-5, 3e-5, 2e-5}
epoch	{3, 4, 5}
warmup ratio	0.1
max seq length	128

LLM の生成におけるハイパーパラメータは，llm-jp-eval[17] のそれを用いた．

D.4　$D_{\text{JNLI}} \cup D_{\text{neg}}$ における fine-tuning

$D_{\text{JNLI}} \cup D_{\text{neg}}$ で学習した MLM の評価結果を表 7, 8 に示す．

表 7　$D_{\text{JNLI}} \cup D_{\text{neg}}$ で学習した MLM の NLI インスタンス単位における評価結果．

Model	D_{JNLI}	$D_{\text{JNLI-Neg}}$	
		D_{orig}	D_{neg}
東北大 $\text{BERT}_{\text{BASE}}$	90.22	90.00	70.14
東北大 $\text{BERT}_{\text{LARGE}}$	92.40	89.68	73.39
早稲田大 $\text{RoBERTa}_{\text{BASE}}$	86.79	86.02	64.33
早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	89.83	88.49	70.03

表 8　$D_{\text{JNLI}} \cup D_{\text{neg}}$ で学習した MLM の否定のミニマルペア単位における評価結果．

Model	M_{i}			M_{u}		
	Acc	Acc′	AccChg	Acc	Acc′	AccChg
東北大 $\text{BERT}_{\text{BASE}}$	79.49	65.69	−13.80	79.68	72.66	−7.02
東北大 $\text{BERT}_{\text{LARGE}}$	82.01	69.35	−12.66	82.35	72.23	−7.12
早稲田大 $\text{RoBERTa}_{\text{BASE}}$	74.99	58.04	−16.95	74.74	67.74	−6.99
早稲田大 $\text{RoBERTa}_{\text{LARGE}}$	79.85	66.89	−12.96	78.95	71.29	−7.66

17）　https://github.com/llm-jp/llm-jp-eval で公開されているコードの generator_kwargs のデフォルト値を用いた．