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

湯浅 令子 [1]　吉田 朝飛 [2]　加藤 芳秀 [3]　松原 茂樹 [2,3]
[1] 名古屋大学情報学部　[2] 名古屋大学大学院情報学研究科
[3] 名古屋大学情報連携推進本部
yuasa.reiko.k5@s.mail.nagoya-u.ac.jp

概要

言語モデルを評価するための英語言語理解ベンチマークが，否定理解能力の評価に有効であるかの分析が行われている．一方，日本語においては，ベンチマークをそのような観点から分析する取り組みはない．本研究では，日本語言語理解ベンチマーク JGLUE を否定の観点から評価する．JGLUE に含まれる否定について現状と課題を明らかにする．

1　はじめに

否定（negation）とは，事態の不成立を表すことであり [1]，自然言語における重要な言語現象である．言語モデルには否定を意味する表現（**否定要素**; negation cue）を正しく理解することが求められる．英語について，言語モデルは否定の理解を苦手としていることが複数の研究により示されている [2, 3, 4]．モデルの言語理解能力を様々な観点から評価するために，既存の**ベンチマーク**（データセット群）が否定理解の評価に有効であるかに関する分析も進められている．その結果，複数の英語言語理解ベンチマークについて，言語モデルが否定を理解しているかの検証には適さないことが示されている [2, 5, 6, 7, 8]．一方，日本語に関しては，ベンチマークが否定の理解の検証に有効であるかは分析されていない．

本研究では，代表的な日本語言語理解ベンチマークの 1 つである JGLUE [9] が，否定理解の検証に有効であるかを評価する．評価においては，以下の観点を考慮する．

1. 否定要素は，量的に十分か．
2. 多様な種類の否定をバランスよく含んでいるか．
3. ベンチマークの各タスクを解くのに，否定の理解を要するか．

3 節で 1. 及び 2. の分析を，4 節で 3. の分析を行う．分析の結果，

1. JGLUE に含まれる否定文の割合は，一般的な日本語の文章のそれよりも小さい
2. JGLUE を構成する多くのデータセットにおいて否定要素の品詞の大半を助動詞が占めており，QA データセットにおいて文末の否定要素が占める割合が小さい
3. JGLUE の一部のデータセットにおいて，タスクを解くために理解が必要な否定要素の占める割合が大きい

ことが明らかになった．

2　関連研究

本研究では，英語言語理解ベンチマークを否定の観点から評価した Hossain ら [2, 10] の手法をベースとして，日本語言語理解ベンチマーク JGLUE を分析する．Hossain らの分析手法と JGLUE について 2.1 節と 2.2 節でそれぞれ説明する．

2.1　否定の観点からの英語言語理解ベンチマークの評価

Hossain らは，英語の言語理解ベンチマークを否定の出現頻度に関して評価し，ベンチマークに含まれる否定の割合は一般的な文章のそれよりも小さいことを明らかにした．また，Hossain らは否定の重要さに関しても評価した．ここで重要さとは，タスクを解くうえで否定が果たす役割であり，次のように定義される．

重要な否定　取り除くとタスクの正解ラベルが変化する否定

重要でない否定　取り除いてもタスクの正解ラベルが変化しない否定

この評価により，ベンチマークに含まれる否定の多くは重要でない否定である，つまり多くの場合は否定を無視してもタスクに正解できることを示した．

2.2 JGLUE

JGLUE は，6 つの言語理解タスクデータセットから構成される．そのうち，本研究では，以下の 5 つを分析対象とする[1)]．

JCoLA 提示された文が容認可能（文法的）か容認不可能（非文法的）かを判定するタスク

JSTS 文ペアの意味的な類似度を推定するタスク

JNLI 前提文と仮説文が与えられたときに，前提文が仮説文に対してもつ推論関係を認識するタスク

JSQuAD 文書を読み，それに関する質問に答えるタスク

JCommonsenseQA 常識推論能力を要する 5 択の問題に答えるタスク

3 ベンチマークに含まれる否定の定量的な分析

本節では，JGLUE に含まれる否定要素を定量的に分析する．分析対象は，JGLUE の学習データ及び検証データとする．手順は以下のとおりである．

1. 否定要素検出器により否定要素を自動的に検出
2. 否定文の出現頻度を分析
3. 否定要素の品詞の分布を分析
4. 否定要素の位置の分布を分析

3.1 否定要素の検出

JGLUE に含まれる否定要素の検出には，否定要素検出器を用いる．蘆田ら [11] と同様に，日本語の否定アノテーション付きコーパス BCCWJ-NEG [12] を用いて日本語 BERT[2)] を fine-tuning することにより，否定要素検出器[3)] を構築した．構築した否定要素検出器を用いて JGLUE から否定要素を検出した．また，一般的な日本語テキストと比較するため，現代日本語書き言葉均衡コーパス（BCCWJ）[13] からも否定要素を検出した．

3.2 否定要素の出現頻度の分析

否定要素の検出結果を用いて，否定文の割合を分析した．ここで，否定文とは，否定要素を 1 つ以上含む文と定義する．

表 1 JGLUE と BCCWJ における否定文の割合

データセット	否定要素数	否定文の数	文の総数	否定文の割合 *
JCoLA	785	769	8,469	9.08%
JSTS	284	281	27,816	1.01%
JNLI	615	609	45,014	1.35%
JSQuAD（読解文）	8,860	7,388	49,157	**15.03%**
JSQuAD（質問文）	3,032	2,815	68,386	4.12%
JCommonsenseQA	348	341	10,058	3.39%
BCCWJ	963,272	827,466	5,480,569	15.01%

* 太字は BCCWJ を上回った値を示す．

表 2 JGLUE と BCCWJ における否定要素の品詞分布

データセット	助動詞	接頭辞	形容詞	名詞	その他
JCoLA	80.13%	2.68%	16.82%	0.38%	0.00%
JSTS	67.61%	7.04%	24.65%	0.35%	0.35%
JNLI	70.99%	4.38%	23.99%	0.32%	0.32%
JSQuAD（読解文）	56.04%	11.29%	30.67%	1.53%	0.47%
JSQuAD（質問文）	59.10%	13.79%	25.69%	1.29%	0.13%
JCommonsenseQA	60.63%	2.01%	37.07%	0.29%	0.00%
BCCWJ	62.19%	4.63%	31.76%	1.24%	0.18%

JGLUE と BCCWJ における否定文の割合を表 1 に示す．JSQuAD の読解文を除く JGLUE のすべてのデータセットは，一般的なテキストである BCCWJ と比べて否定文の割合が少ないことを確認した．特に，JSTS と JNLI は否定文の割合がかなり小さいことが明らかになった．そもそも否定文の割合が小さいことから，JSQuAD 以外のデータセットについては否定理解の評価には適さないと考えられる．

3.3 否定要素の品詞の分布の分析

JGLUE と BCCWJ における否定要素の品詞の分布を表 2 に示す．JGLUE の多くのデータセットにおいて，否定要素の品詞の大半を助動詞が占めていることを確認した．一方，「不」や「非」などの接頭辞の否定要素が占める割合は小さいことを確認した．

3.4 否定要素の位置の分布の分析

検出された日本語の否定要素をその出現位置に従って分類し分析した．具体的には，否定要素が文末に出現するか，文の途中で出現するかで分類した．ここでは，否定文において，ある否定要素以降に内容語[4)] が一度も現れない，つまり否定要素以降がすべて機能語で構成されているとき，その否定要素を文末の否定要素と呼び，それ以外の否定要素を文の途中の否定要素と呼ぶ．これは，Pullum ら [14] による否定スコープ（否定要素の影響が及ぶ範囲）

1） 残る MARC-ja は本稿執筆時に利用不可能であったため，対象外とした．これ以降，特に断りがない限り MARC-ja 以外の 5 つのデータセットを指して JGLUE と呼ぶ．

2） https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

3） BCCWJ-NEG における 10 分割交差検証による検出性能は F1 値において 94.36%であった．

4） 名詞，形状詞，副詞，動詞，形容詞を内容語とした．

表 3　JGLUE における否定要素の位置分布

データセット	文末の否定要素	文の途中の否定要素
JCoLA	57.83%	42.17%
JSTS	42.25%	57.75%
JNLI	51.86%	48.14%
JSQuAD（読解文）	20.41%	79.59%
JSQuAD（質問文）	11.31%	88.69%
JCommonsenseQA	7.76%	92.24%

に基づく以下の分類を日本語で擬似的に再現したものである．

clausal negation　スコープが文全体である否定要素

sub-clausal negation　スコープが文の一部である否定要素

日本語（特に単文）において，否定スコープが文全体となるときは否定要素が文末にあることが多いため，文末の否定要素は近似的に clausal negation とみなせる．同様に，文の途中の否定要素は sub-clausal negation とみなせる．

JGLUE における否定要素の出現位置に関する分布を表 3 に示す．QA データセットである JSQuAD 及び JCommonsenseQA において文末の否定要素が占める割合が小さいことを確認した．それ以外のデータセットにおいては，文末の否定要素と文の途中の否定要素がほぼ同程度の割合で含まれていることが明らかになった．

4　否定の重要さに関する分析

本節では，JSTS 及び JCommonsenseQA の学習データに含まれる否定要素を重要さの観点から分析する[5]．重要さの定義は 2.1 節で述べた Hossain ら [2, 10] の手法に従う．分析のためのデータ作成手順は以下のとおりである．

1. 否定要素を含むインスタンス I_{neg} を抽出
2. I_{neg} から否定要素を除去し，新たに正解ラベルを人手で付与（このインスタンスを $I_{\text{rm\_neg}}$ とする）
3. I_{neg} の正解ラベルと $I_{\text{rm\_neg}}$ の正解ラベルを比較し否定要素の重要さを決定

ここでは，各データセットの要素，つまり提示されるテキストと正解ラベルの組のことをインスタンスと呼ぶ．例えば，文ペア分類タスクでは文ペアと正解ラベルの組を，QA タスクでは質問文（及び読解文）と正解ラベルの組をインスタンスと呼ぶ．図 1 に分析用データ作成の概要を示す．

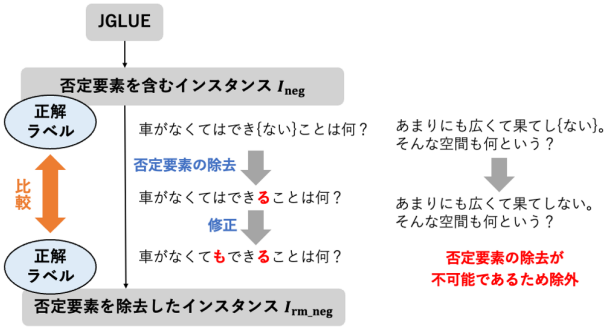

図 1　分析用データ作成の概要

4.1　分析用データの作成

4.1.1　否定要素を含むインスタンスの抽出

3.1 節で構築した否定要素検出器を用いて否定要素を含むインスタンス I_{neg} を抽出した．

4.1.2　否定要素の除去

否定要素検出器により検出された否定要素を除去した文を作成した．否定要素の除去には LLM の in-context learning を用いた[6]．その後，否定要素が適切に除去されているかを作業者（著者 1 名）が確認し，

- LLM の出力文が日本語として文法的・意味的に不自然な場合は，日本語として自然な文になるように修正
- そもそも否定要素の除去が不可能な場合や，否定要素を除去すると一文の中で意味が矛盾する場合は，データから除外

を行った．修正例を表 4 に示す．(1), (2) の LLM の出力文は日本語として文法的に不自然であるため，自然な文になるよう修正した．(3) の LLM の出力文は意味的に不自然であるため，日本語として自然な文になるよう重複する部分を削除した．次に，除外例を表 5 に示す．(4) は慣用的に否定形でしか用いられない表現であるため，(5) は「名前だけで，実質が伴う」という矛盾した状況であるため，データから除外した．(6) は否定要素が誤検出されていたケースである．

以上の確認を行った後の文の数を表 6 に示す．前

5）本稿執筆時点で，JNLI 及び JSQuAD の正解ラベル付与作業は完了していないため，JSTS と JCommonsenseQA の分析結果のみ報告する．

6）モデルは，OpenAI API（https://openai.com/api/）で提供されている gpt-4o を用いた．プロンプトの詳細及びハイパーパラメータについては，付録 A を参照されたい．

表 4　否定要素を除去した文の修正例

	元の文 *	LLM の出力文 **	修正後の文 **
(1)	車がなくてはでき｛ない｝ことは何？	車がなくてはできることは何？	車がなくてもできることは何？
(2)	｛未｝舗装の土地に白色のピックアップトラックが停まっています。	舗装の土地に白色のピックアップトラックが停まっています。	舗装済みの土地に白色のピックアップトラックが停まっています。
(3)	まな板に切ったブロッコリーと切られてい｛ない｝ブロッコリーがある。	まな板に切ったブロッコリーと切られているブロッコリーがある。	まな板に切ったブロッコリーがある。

* {} で囲まれた部分が除去対象の否定要素である．
** 太字は元の文と比較して変更のあった箇所を表す．

表 5　否定要素を除去した文の除外例

	元の文 *	LLM の出力文 **
(4)	あまりにも広くて果てし｛ない｝。そんな空間も何という？	あまりにも広くて果てしない。そんな空間も何という？
(5)	名前だけで，実質が伴わ｛ない｝ことをなんという？	名前だけで，実質が伴うことをなんという？
(6)	ゲコゲコ｛なく｝ものと言えば？	ゲコゲコ鳴くものと言えば？

* {} で囲まれた部分が除去対象の否定要素である．
** 太字は元の文と比較して変更のあった箇所を表す．

述の基準により文を除外すると，いずれのデータセットにおいても全体の 1 割前後が除外された．

4.1.3　正解ラベルの付与

否定要素を除去したテキストに対し，3 人のアノテータが新たにラベルを付与した．アノテータに与えた指示は，JGLUE のガイドライン[7]に基づく．ただし，元の文から否定要素を除去したことによる影響を考慮し，JCommonsenseQA の指示のみ以下の変更を加えた．

- 否定要素を除去した結果，5 択の選択肢の中に正解が存在しない場合がある．正解が選択肢の中にない場合も，その否定要素が重要であるとみなせるため，「存在しない」という選択肢を追加した．また，「存在しない」「わからない」以外の選択肢については複数選択を許容した．

ラベルをアノテーションしたインスタンス $I_{\text{rm\_neg}}$ のに対して，以下の基準に基づき最終的な正解ラベルを決定した．

7）https://github.com/yahoojapan/JGLUE/blob/main/task_guidelines.md で公開されている．

表 6　否定要素除去後の文の数

データセット	修正した文の数	除外した文の数	否定要素除去後の文の数
JSTS	87	23	244
JCommonsenseQA	88	41	266

表 7　JSTS 及び JCommonsenseQA における重要な否定要素と重要でない否定要素の分布

データセット	重要な否定要素	重要でない否定要素
JSTS	63.37%	36.63%
JCommonsenseQA	94.21%	5.79%

JSTS　アノテータは 0 から 5 の整数値を付与する．最終的な正解ラベルは，平均値とする．ただし，アノテータの回答の分散が 1.0 以上の場合は，データから除外する．

JCommonsenseQA　2 人以上が選択した選択肢を正解ラベルとする．

4.1.4　否定要素の重要さの分析

I_{neg} の正解ラベルと $I_{\text{rm\_neg}}$ の正解ラベルを比較し，2.1 節で述べた Hossain らによる定義に従って否定要素の重要さを決定した．JSTS は正解ラベルが連続値をとるため，I_{neg} の正解ラベルと $I_{\text{rm\_neg}}$ の正解ラベルとの差の絶対値が 1 以上のとき，正解ラベルが変化したとみなした．

4.2　分析結果

JSTS 及び JCommonsenseQA における重要な否定要素と重要でない否定要素の分布を表 7 に示す．いずれのデータセットにおいても，重要な否定要素が占める割合が大きいことを確認した．特に，JCommonsenseQA には重要でない否定要素がほとんど含まれないことが明らかになった．

5　おわりに

本研究では，日本語ベンチマーク JGLUE を否定の観点から分析した．否定要素の出現頻度に関する分析により，JGLUE に含まれる否定文の割合は一般的な日本語の文章のそれよりも小さいことを示した．一方で，英語ベンチマークと異なり，JSTS 及び JCommonsenseQA において重要な否定要素が占める割合が大きいことを確認した．

否定要素の重要さに関して，JNLI 及び JSQuAD についても今後分析する予定である．

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．本研究で利用した JGLUE 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，柴田知秀氏に感謝いたします．

参考文献

[1] 日本語記述文法研究会. 現代日本語文法 3. くろしお出版, 2012.

[2] Md Mosharaf Hossain, Dhivya Chinnappa, and Eduardo Blanco. An analysis of negation in natural language understanding corpora. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, 2022.

[3] 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 (*SEM 2023)**, 2023.

[4] 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**, 2023.

[5] 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**, 2022.

[6] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, 2018.

[7] Alice Lai and Julia Hockenmaier. Illinois-LH: A denotational and distributional approach to semantics. In **Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)**, 2014.

[8] Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. New protocols and negative results for textual entailment data collection. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2020.

[9] 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.

[10] 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 (EMNLP)**, 2020.

[11] 蘆田真奈, 平澤寅庄, 金子正弘, 小町守. 日本語 bert による否定要素認識についての分析. 言語処理学会第 29 回年次大会 発表論文集, 2021.

[12] 松吉俊. 否定の焦点情報アノテーション. 自然言語処理, Vol. 21, No. 2, pp. 249–270, 2014.

[13] K. Maekawa, M. Yamazaki, T. Ogiso, et al. Balanced corpus of contemporary written japanese. **Lang Resources & Evaluation**, Vol. 48, pp. 345–371, 2014.

[14] Geoffrey K. Pullum, Rodney Huddleston, Rodney Huddleston, and Geoffrey K. Pullum. **Negation**. Cambridge University Press, 2002.

user: 否定要素（「ない」や「ず」）を {} で囲った日本語の文を与えるので、その否定要素を取り除いて否定を含まない文に書き換えてください。
書き換えた後の文のみを出力することを遵守してください。
user: [否定要素除外対象の文]

図 2　LLM による否定要素除去のプロンプト

A　否定要素除去の詳細

本節では，$I_{\text{rm\_neg}}$ の作成における LLM による否定要素の除去の詳細を説明する．LLM に与えたプロンプトを図 2 に示す．プロンプトは，4 節の分析とは独立の予備分析により作成したものである．

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