否定表現を伴う文における自然言語理解の性能検証

内田巧 [1] 南條浩輝 [2]
[1] 滋賀大学大学院データサイエンス研究科 [2] 滋賀大学データサイエンス学部
[1] s6022105@st.shiga-u.ac.jp [2] hiroaki-nanjo@biwako.shiga-u.ac.jp

概要

本稿では，対偶により否定表現を伴う含意関係認識の学習・評価データセットを作成し，否定表現の自然言語理解タスクへの影響を調査した．実験の結果，含意関係認識では，前提文に否定表現が伴うことで精度が悪化することを確認した．悪化の原因として，否定表現に過剰に依存する shortcut learning が発生している可能性を示した．STS では，2 文共に否定表現が含まれるだけで，類似度を高めに予測してしまう傾向を確認した．以上より，言語モデルは否定表現による意味の変化を捉えることなく推論していることを示した．

1 はじめに

本稿では，自然言語理解タスク，特に含意関係認識と意味的類似度計算 (STS) への否定表現の影響を調査する．否定表現を伴う文を含むように生成された既存の含意関係認識用データセットとして JSICK [1][2] や JaNLI [3] などがあるが，ラベルに偏りがあり，ルールベースで解けることから，否定表現による文意の変化を問うデータセットになっていないという問題がある [4]．そこで，既存の含意関係認識用データセットから否定表現を伴うデータセットを生成する手法として対偶を提案した [4]．しかし，対偶で生成されたデータだけでは，前提文と仮説文共に否定表現が伴う場合しか評価できない．STS の場合は，既存のデータセットに JSICK などがあるが，否定表現により非含意関係となるデータに対して，2 文の表層表現が類似していれば高い類似度（正解ラベル）が与えられている．したがって，否定表現による文の意味の変化をモデルが捉えているか評価するには適切でない．

そこで，含意関係認識では，否定表現の含意関係認識への影響を網羅的に調査するデータセットを生成して実験を行う．STS では，既存の STS 用データセットに自動で否定表現を付与することで，否定表現の影響を調査するデータセットを作成して実験を行う．また，それぞれのタスクにて，対偶による拡張データを用いて学習する効果も確認する．本稿での否定表現は，文末を否定する「ない」や「なかった」，「ません」に限定し，その他の表現，たとえば文中の否定要素や否定の接頭辞である「非」や「不」などは本研究の対象としない．

2 関連研究

2.1 含意関係認識

既存の言語モデルは含意関係認識において，否定文に対して否定表現による意味の変化を考慮した推論ができていない [5]．その理由に shortcut learning [6] という問題が上げられる．shortcut learning とは，学習データと同分布 (Independent and Identically Distributed: i.i.d.) なデータには高い推論性能を発揮するモデルが，汎化につながる特徴量ではなく表層的な情報しか捉えておらず，学習データと異なる分布のデータ (Out of Distribution: o.o.d.) に対して高い推論性能を発揮できないという問題である．文献 [7] では，非含意の推論において否定表現に強く依存していること，すなわち否定表現だけから非含意と予測する shortcut learning が起きていることが確認された．これは，既存のデータセットによっては否定表現と非含意のラベルが強く共起していることが原因と考えられている [8]．さらに，shortcut learning を引き起こす要因がデータセットに含まれているとそれを早い段階で学習してしまう [9] ことから，既存のデータセットでは shortcut learning を防ぐのは難しい．このように，言語モデルは否定表現を伴う文の意味理解を適切に行えていない．

2.2 STS

否定表現をある文に付与すると，元の文とは意味が大きく異なる一方，表層表現は元の文と否定表現

以外で同一となる．このような元の文と大きく意味が異なる一方，ほぼ同じ表層表現である文は soft negative samples とよばれ [10]，feature suppression により元の文との類似度が高めに予測されることが知られている．この問題に対して，文献 [10] では上記のような文を soft negative sample として扱う対照学習手法 SNCSE が提案されている．否定表現だけを付与した文と元の文のペアの評価はされているものの，元の表層表現が異なる 2 文（文献 [10] では negative sample とよばれている）に対し，それぞれまたは片方に否定表現を付与した場合に言語モデルは 2 文の意味的類似度を正しく予測できるのか，といったことは十分に調査されていない．さらに，このような否定表現を含むデータに対する STS のためのデータセットも著者が知る限り見当たらない．

3 実験 1. 含意関係認識

本章では，含意関係認識において，文に否定表現が伴うことで言語モデルの推論性能に変化が見られるかを検証する．また，推論時に shortcut learning が発生していないかを調査する．

3.1 対偶による学習・評価データの拡張

含意関係認識の実験では，JSNLI[11] に対して対偶によるデータ拡張を適用することで，否定表現を伴う学習・評価データセットを生成して使用する．前提文あるいは仮説文に否定表現を伴う割合は 1.7%程度であるため，対偶により生成したデータセットの大半は，前提文または仮説文に否定表現を伴うことになる．文献 [4] に基づき，JSNLI 学習データセット（533,005 件）に対して対偶処理を適用したところ，482,715 件を対偶処理対象として抽出でき，これらに対して対偶をとり否定表現を伴うデータからなるデータセットを生成した．本稿では以降このデータセットを N-JSNLI とよぶ．なお，正確な比較を行うため，元の JSNLI 学習データからも対偶処理の対象となったデータと同数を取り出して使用する．評価データセットに対しても同様の処理を行い，JSNLI 評価データセット（3,916 件）から，N-JSNLI 評価データセット（3,593 件）を生成した．

3.2 網羅的な評価データの生成

既存の含意関係認識のデータセットは，前提文と仮説文ともに肯定文であるペアが多く，対偶処理により従来のデータセットを拡張したものは，前提文と仮説文が共に肯定文または否定文のペアが大半を占める．いずれか一方が肯定文でもう一方が否定文であるデータはほとんど存在しない．したがって，拡張データセットのままでは，このようなデータに対する性能の評価を確認することができない．そこで，JSNLI と N-JSNLI を用いて，前提文あるいは仮説文の一方にのみ否定表現が伴うデータも含むような，網羅的に否定表現の影響を調査できる評価用データセットを作成する．その方法を以下に示す．

- PP（含意/非含意）：JSNLI をそのまま使用
- PN/NP（含意）：JSNLI（含意）の片文を二重否定
- PN/NP（非含意）：JSNLI（非含意）の片文を否定
- NN（含意/非含意）：N-JSNLI を使用

ここで P, N はそれぞれ肯定文，否定文を表す．PP, PN などは，含意関係認識における前提文と仮説文のそれぞれが肯定文か否定文かを表す．PN を例にとって説明すると，これは前提文が肯定文 (P)，仮説文が否定文 (N) あることを表している．

これにより，両文が肯定文のデータ，仮説文のみ否定文のデータ，前提文のみ否定文のデータ，両文が否定文のデータの 4 種類に，それぞれ含意・非含意のデータを作成した．例を付録表 A.1 に示す．片文が否定文かつ含意関係のデータ（PN/NP：含意）には二重否定が含まれている．これは，片文を否定しつつ含意のデータを作成することが難しいことが原因である．

N-JSNLI と同数の JSNLI 評価データセット（3,593 件）に上述した処理を適用して，網羅的な評価データセット（3,526 件）を生成した．本稿では以降このデータセットを NegEval とよぶ．なお，本研究における含意関係認識は，対偶により付与された否定表現による意味の変化を捉えて明確に含意あるいは非含意と判定できるか調査するために 2 値分類タスクとする．そのため，JSNLI 内の中立 (neutral) ラベルのデータは使用しない．

3.3 実験設定

含意関係認識器には，東北大学が公開している事前学習済み BERT[1)] を使用し，既存データセット (JSNLI)，もしくは拡張データセット (JSNLI+N-JSNLI) で学習したモデルを，NegEval で評価する．性能評価指標は Accuracy を使用する．

1) https://huggingface.co/cl-tohoku/bert-base-japanese-v3

表 1　NegEval に対する推論性能の比較

	Evaluation data(NegEval)			
Fine-tune data	PP	PN	NP	NN
JSNLI	0.98	0.90	**0.55**	**0.60**
JSNLI + N-JSNLI	0.98	0.91	0.55	0.96

3.4　実験結果

結果を表 1 に示す．否定表現をほとんど含まない既存データセット (JSNLI) で学習した BERT は NP と NN パターン，すなわち前提文に否定表現を含むデータに対して精度が低い（表 1 上段）．一方で，対偶で生成したデータ (N-JSNLI) も共に学習させた BERT では，NN（前提文と仮説文の両方に否定表現を含む）パターンのデータに対して精度が改善したものの，NP（前提文のみ否定表現を含む）パターンのデータに対しては改善しなかった．これは対偶によって生成されたデータのほとんどが NN パターンで構成されていることが原因である．否定表現をほとんど含まないデータセット (JSNLI) で学習した BERT の誤分類の傾向を図 1 の上段に示す．NP パターンのデータに対しては非含意の事例を含意と誤分類する傾向が見られ，NN パターンのデータに対しては含意の事例を非含意と誤分類する傾向が見られた．この結果は既存の JaNLI を用いても確認できるが，JaNLI のテストデータには NP/NN がそれぞれ 15 件しか含まれておらず，ラベルも偏っており，さらにテンプレートをもとに生成したため，前提文と仮説文がほぼ同じ表層表現という問題点がある．ラベルのバランスがよく，表層表現も異なる 3,500 件以上のデータで実施した本実験でも同様の現象が観測されたことから，NP/NN を学習していない日本語モデルは否定表現ばかり着目して推論する shortcut learning が行われた可能性が考えられる．

そこで，誤分類した事例における pooler_output への Attention を付録図 A.1 のように可視化した．文末の否定表現に Attention がかかっているようにも見えることから shortcut learning の可能性を否定できない．今後詳細に調査する必要がある．

4　実験 2.STS

本章では，STS において，文に否定表現が伴うことで言語モデルの推論性能に変化が見られるか検証する．また，否定表現を伴う学習データを学習する効果も確認する．STS の評価データセットには日本語言語理解ベンチマーク (JGLUE)[12][13] に含まれる JSTS を使用する．

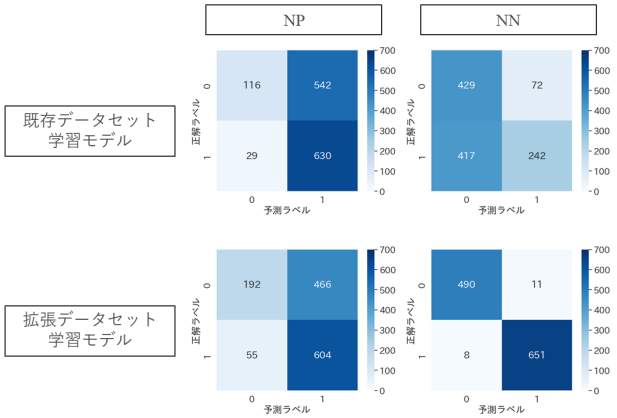

図 1　NP/NN に対する含意関係認識の分類結果

4.1　データ生成

既存の STS 用データセットは否定表現をほとんど含んでいないため，STS への否定表現の影響を調査することができない．そこで，JSTS の評価データセットに含まれる二文共に自動で否定表現を付与することで，否定表現を含む STS 評価データセットを生成した．その際，評価データセットの両文に否定表現を付与する前後で二文間の意味的類似度は変化しないと仮定した．本稿では以降このデータセットを N-JSTS とよぶ．

4.2　実験設定

STS への否定表現の影響を調査する実験方法について述べる．1 文ずつ文をモデルに入力して文埋め込みを獲得し，その cosine 類似度を計算することで類似度を予測する．事前学習済み BERT[1] と RoBERTa[2][14] に対して，既存のデータセット (JSNLI) と拡張データセット (JSNLI+N-JSNLI) を用いて，教師あり SimCSE による対照学習を行ったモデルを使用する．そして，既存のデータセットである JSTS と，JSTS に含まれる 2 文それぞれに否定表現を付与した N-JSTS で評価し，精度の比較を行う．さらに，アーキテクチャによって結果に違いが見られるかを確認するため，Encoder-Decoder モデルを使用して実験を行う．Encoder-Decoder モデルには T5[3][15] と Sentence-T5[4][16] を使用する．性能評

2）https://huggingface.co/nlp-waseda/roberta-base-japanese
3）https://huggingface.co/sonoisa/t5-base-japanese
4）https://huggingface.co/sonoisa/sentence-t5-base-ja-mean-tokens

表 2　否定表現の有無による STS タスクの精度比較

Model	Fine-tune data	Evaluation data JSTS	N-JSTS
BERT	N/A	0.178	0.194
RoBERTa	N/A	0.220	0.237
sup-SimCSE	JSNLI	0.774	0.656
+BERT	JSNLI+N-JSNLI	0.766	0.776
sup-SimCSE	JSNLI	0.770	0.697
+RoBERTa	JSNLI+N-JSNLI	0.775	0.774
T5	N/A	0.646	0.627
ST5-Enc(mean)	N/A	0.809	0.786

Spearman の順位相関係数を使用

価指標には文献 [17] を踏まえてスピアマンの順位相関係数を使用する．

4.3　実験結果

実験結果を表 2 に示す．対照学習したモデルは 1，2，7 段目の baseline モデル (BERT/RoBERa/T5) と比較して精度が高くなった．既存データセット (JSNLI) で対照学習した Encoder モデルの結果は表 2 の 3，5 段目である．sup-SimCSE+BERT のスピアマン順位相関係は，JSTS では 0.774，N-JSTS では 0.656，sup-SimCSE+RoBERTa の場合は，JSTS では 0.770，N-JSTS では 0.697 となった．いずれも否定表現が付与されるだけで推論性能が悪化することが確認された．次に，Encoder-Decoder モデルの結果を表 2 の 7，8 段目に示す．T5 のスピアマン順位相関係は，JSTS では 0.636，N-JSTS では 0.616，Sentence-T5 の場合は，JSTS では 0.809，N-JSTS では 0.786 となった．Encoder モデルと同様に，否定表現が付与されるだけで推論性能が悪化することが確認された．

次に予測結果の変化について詳細に分析を行った．分析には，sup-SimCSE+BERT と Sentence-T5 を用いた．sup-SimCSE+BERT の予測類似度とラベル（正解の類似度）の分布を図 2 に示す．ほとんど否定表現を学習していない Sup-SimCSE+BERT（図 2 左）は，テストデータの両文に否定表現が伴うだけで文章の意味的類似度を高めに予測していた．さらに，付録図 A.2 で示すように，Sentence-T5 でも同様の現象が確認された．これは Transformer ベースのモデルが語彙の類似度と意味的な類似度の違いを区別できないという feature suppression が働いたことが原因と考えられる．両文に否定表現を付与したことで類似した語彙（否定表現）が増えたため，二文の

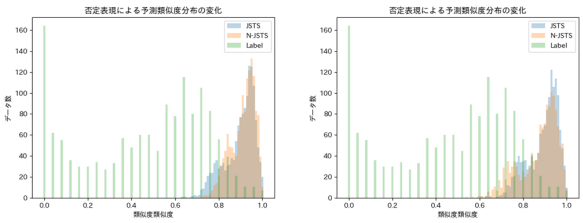

図 2　予測類似度の分布（左: JSNLI で対照学習した BERT，右: JSNLI+N-JSNLI で対照学習した BERT）
※: 目的タスク (JSTS/N-JSTS) で学習していないため，正解ラベルと予測類似度は乖離している．

類似度を高めに予測するようになったと推測する．以上より，既存のデータセットで学習したモデルでは否定文同士の意味的類似度を適切に予測することができないことが明らかになった．

一方で，対偶により生成されたデータを含む拡張データセットを用いて対照学習したモデルの結果を表 2 の 4，6 段目に示す．sup-SimCSE+BERT は JSTS におけるスピアマン順位相関係数が 0.761，N-JSTS においては 0.775，sup-SimCSE+RoBERTa は JSTS におけるスピアマン順位相関係数が 0.775，N-JSTS においては 0.774 となった．いずれも否定表現を含む文に対する推論性能が改善しつつ，肯定文のテストデータ (JSTS) に対しても性能が悪化しないことを確認した．拡張データセットで学習した sup-SimCSE+BERT による類似度分布を図 2（右）に示す．既存データセットで学習した BERT と比較して，拡張データセットで学習した BERT は否定文同士の文ペア（N-JSTS）に対して，低い類似度を予測するできるようになったことがわかる．このことは，対偶データを学習データに追加したことで，否定表現という表層表現の類似度の影響を軽減し，意味的な類似度をより反映した結果と考えられる．以上のように，対偶により生成されたデータも含めて対照学習することで，既存のデータでは解くことができなかった否定表現を伴う文章に対する STS タスクを解くことができるようになることが明らかになった．

5　まとめ

対偶で拡張したデータセットを用いることで，含意関係認識と STS への否定表現の影響と，否定表現を伴う文を多く含むデータセットを学習する効果を確認した．

参考文献

[1] 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.

[2] 谷中瞳, 峯島宏次. JSICK: 日本語構成的推論・類似度データセットの構築. 人工知能学会全国大会論文集 第 35 回, 2021.

[3] Hitomi Yanaka and Koji Mineshima. Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference. In **Proceedings of the 2021 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP2021)**, 2021.

[4] 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. Technical report, 第 257 回自然言語処理研究会, 2023.

[5] 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)**, pp. 716–723. Association for Computational Linguistics, May 2022.

[6] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. **Nature Machine Intelligence**, Vol. 2, No. 11, pp. 665–673, 2020.

[7] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut Learning of Large Language Models in Natural Language Understanding. **Communications of the ACM**, Vol. 67, No. 1, pp. 110–120, 2023.

[8] 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)**, pp. 107–112. Association for Computational Linguistics, June 2018.

[9] Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 915–929, 2021.

[10] Hao Wang and Yong Dou. SNCSE: Contrastive Learning for Unsupervised Sentence Embedding with Soft Negative Samples. In **International Conference on Intelligent Computing**, pp. 419–431. Springer, 2023.

[11] 吉越卓見, 河原大輔, 黒橋禎夫. 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. Technical report, 第 244 回自然言語処理研究会, 2020.

[12] 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. 言語処理学会 第 28 回年次大会, 2022.

[13] 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. European Language Resources Association, June 2022.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv preprint arXiv:1907.11692**, 2019.

[15] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. **The Journal of Machine Learning Research**, Vol. 21, No. 1, pp. 5485–5551, 2020.

[16] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 1864–1874. Association for Computational Linguistics, May 2022.

[17] Nils Reimers, Philip Beyer, and Iryna Gurevych. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 87–96. The COLING 2016 Organizing Committee, December 2016.

A　付録

表 A.1　NegEval の例

Pattern	Label	Role	Sentence
両文肯定 (PP)	含意	前提文	人が２頭の馬の間にひざまずいている。
		仮説文	人と二頭の馬がいます。
	非含意	前提文	男と女が丘の上を歩きます。
		仮説文	男と女が座っています。
片文否定 (PN)	含意	前提文	人が２頭の馬の間にひざまずいている。
		仮説文	人と二頭の馬が**いないわけではありません**。
	非含意	前提文	人が２頭の馬の間にひざまずいている。
		仮説文	人と二頭の馬が**いません**。
片文否定 (NP)	含意	前提文	人が２頭の馬の間にひざまずいて**いないわけではない**。
		仮説文	人と二頭の馬がいます。
	非含意	前提文	人が２頭の馬の間にひざまずいて**いない**。
		仮説文	人と二頭の馬がいます。
両文否定 (NN)	含意	前提文	人と二頭の馬が**いません**。
		仮説文	人が２頭の馬の間にひざまずいて**いない**。
	非含意	前提文	男と女が座って**いません**。
		仮説文	男と女が丘の上を歩き**ません**。

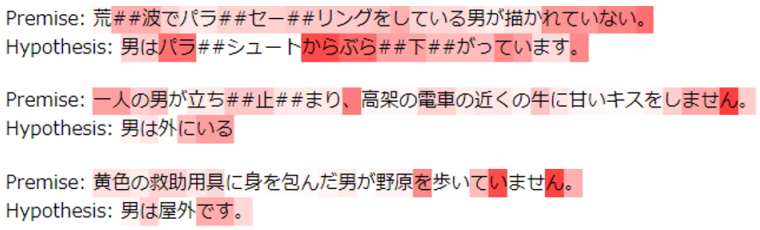

図 A.1　NP データに対して誤分類した事例の推論根拠を Attention で可視化

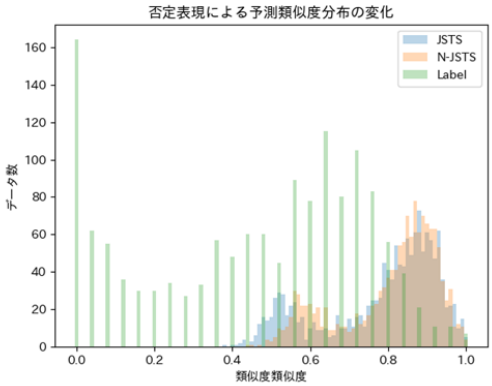

図 A.2　Sentence-T5 の予測類似度の分布

※: 目的タスク (JSTS/N-JSTS) で学習していないため,
正解ラベルと予測類似度は乖離している.