日本語 BERT による否定要素認識についての分析

蘆田 真奈　平澤 寅庄　金子 正弘　小町 守
東京都立大学
{ashida-mana, hirasawa-tosho, kaneko-masahiro}@ed.tmu.ac.jp
komachi@tmu.ac.jp

1 はじめに

言語を理解するさい、文中において否定を表す要素（**否定要素**）を正しく認識する能力は必要不可欠である。一方で、言語モデルにとっては否定要素が否定要素であるということは必ずしも自明ではない。日本語において否定を表す表現は「ない・ん・ず」などがあるが、修辞技法や慣用句に含まれる「ない・ん・ず」は否定要素として使用されないケースも数多くあり、表層形の知識や構文の理解はもちろん、文脈の理解も必要とされる。

近年、自然言語処理では事前学習済み大規模ニューラル言語モデルが多くのタスクにて高い精度を出し、脚光を浴びている [1, 2]。一方でニューラルネットワークを用いたアーキテクチャはブラックボックスなどとも言われ、その振る舞いの解釈は困難である。そのような背景から、ニューラルモデルの内部メカニズムがどれほど人間の言語に関する知識を獲得できているかを調べる研究（Probing）が数多く提案されている [3, 4, 5]。これらの研究を通じて、現時点でのニューラル言語モデルの得手不得手についての新たな知見が得られつつある。

今回の研究ではニューラル言語モデルの中でも、特に Bidirectional Encoder Representations from Transformers (BERT) [1] がどの程度、否定要素を認識できるのかについて、日本語 BERT を用いて否定要素認識タスクを行うことで評価する。また、Probing タスクとして、日本語 BERT に否定要素がマスクされた文を与えた場合に、BERT が否定要素を予測できるか、及びできない場合に何と予測するかについて調査する。BERT による予測結果を分析することで、BERT が何を手がかりにして否定要素を認識しているかについて考察する。最後に、否定に関わる統語情報についての知識が獲得されているかについて、否定極性項目（Negative Polarity Items（NPIs））と否定要素の関係性に着目して分析を行う。

2 関連研究

日本語の否定コーパス　日本語の自然言語処理において、否定要素に限定して行われた研究には 松吉ら [6, 7] などがある。[6] は否定要素のアノテーションを付与したコーパスの作成についての報告である。[7] では [6] のコーパスをデータセットとし、否定要素があらかじめ検出されている文に対して統語情報を与えた上で否定焦点を検出するシステムを構築している。本稿では、否定焦点の検出ではなく否定要素の検出を日本語 BERT を用いて行うためにコーパスを用いる。

BERT を用いた英語の否定要素認識　日本語 BERT を用いた否定要素認識こそ行われていないものの、英語ではいくつかの BERT を用いた否定要素認識の研究がある。多くは*SEM（Joint Conference on Lexical and Computational Semantics）の Shared Task での Sherlock データセット [8] もしくは医療分野のコーパス（the BioScope Corpus[9]）を使用したタスクである。NegBERT [10] は否定要素と否定のスコープのアノテーションがなされたコーパスで fine-tune された BERT であり、文を入力とし、否定要素を検出することができる。[11] は Conversational Question Answering (CoQA) タスクにおいて、NegBERT を用いて系列の否定要素を明示することでモデルの性能が上がることを報告している。

BERT の英語の否定の理解性能　BERT がどのようなメカニズムで色々なタスクにおいて良い性能を出すことができるのかについての研究が盛んに行われており、その研究の中には否定の理解に関するものもいくつかある。例えば、[12] や [13] では空欄補充タスクを用いて BERT が否定の働きについて捉えているかどうかを検証している。「A が X である。」、と「A は X ではない。」という文脈において、X を空

欄とした場合にどちらの場合にも X には A の上位概念（hypernym）を候補としてあげてしまうことから、BERT は否定が含まれた文脈では正しい推論ができていないと結論づけている。今回の研究では、肯定と否定文のペアを作成するのではなく、実際にコーパスに出現する文を用い、BERT が否定要素を正しく予測できるかについて検証する。

言語モデルの英語の NPIs についての分析　NPIs とは文の否定要素と共起することを必要とする項目のことであり、その制約ゆえ、言語学者の間では広く議論されてきた [14]。NPIs には自然言語処理の文脈では、NPIs が否定要素を伴わない場合、その文が非文法的になることを用いて、言語モデルに文の文法性を判断させることで言語モデルが NPIs に関わる統語的制約を学習しているかを検証している [15, 16]。日本語の NPI の例としては『しか』などがあげられる。以下は**しか**が文法的である場合とそうでない場合の例である。

- 太郎は 野菜『しか』 食べ**ない**。
- *太郎は 野菜『しか』 食べる。

ここで『しか』は Licensor と呼ばれる文の否定要素**ない**のスコープのなかに存在するときにのみ文法的である。

3　否定要素認識

否定要素認識タスクは医療分野での言語処理（Biomedical NLP）に端を発する [17]。自然言語処理における否定認識（negation detection）は 2 段階のタスクからなる。まず、文中の否定要素（negation trigger, negation cue）を検知する。日本語では「ない」や「ず」に当たる。次に、否定が及ぶ範囲を同定する。これは否定のスコープの認識（negation scope detection, negation scope resolution）とも言われる。さらに、否定の焦点（focus of negation）は否定のスコープの中でも、特に否定されている要素のことをいう。否定の焦点検知についての研究は [8] での Shared Task に関するものがある。今回の研究では否定要素の認識に取り組む。

3.1　データセット

今回のタスクで使用するコーパス [6] は「現代日本語書き言葉均衡コーパス」（BCCWJ）[1] のコアデータの新聞のうち A グループを対象に否定要素と否定の焦点をアノテーションしたものである。

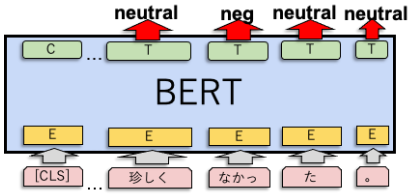

図 1　BERT による否定要素認識

- 十七日まで選手にも協会関係者にも明かさ**ない**。[PN2f_00002]

という例文において、否定要素は「ない」であり、否定の焦点は「十七日まで」である。また、否定のスコープは「十七日まで選手にも協会関係者にも明かさ」で表現される事象であると考えられる [6]。ただし、今回の研究では否定要素のみに注目する。

このコーパスの全 2807 文のうち、80%を訓練データ、10%を開発データ、10%を評価データとして用いた。コーパスの全文に対して mecab-ipadic-neologd を用いて単語分割を行った。これは使用した日本語 BERT の訓練時の分割方式と一致させるためである。ラベルはコーパスに従って付与した。

3.2　否定要素認識器の作成方法

否定要素認識器の作成にあたっては、否定要素を *neg*、それ以外を *neutral* とラベル付けする系列ラベリングのタスクを設定し、日本語 BERT の fine-tune を行なった。今回、*neutral* 以外のラベルは複数のトークンにまたがっていないため、先頭や終わりをマークせず上述の 2 つのラベルのみを使用した。図 1 は BERT を用いた認識器及び入力のイメージである。BERT の中間層の上にランダムに初期化された出力層を重ね、系列ラベリングを訓練させた。

3.3　実験

実験設定　本稿では東北大が公開した日本語 BERT モデル [2] を用いて実験を行った。BERT の dense layer と出力層の両方を fine-tune したモデル（fine-tuned）と BERT の embedding の重みを更新しないモデル（frozen）を用いた。訓練の epoch 数は 10 回とし、F_1 スコアをモデル選択時の指標として、推論時には開発データの F_1 スコアが一番良いモデルを使用した。

実験結果　図 1 は 5 回シードを変えて実験を行って得られた精度・再現率・F_1 スコアの平均を示したものである。テストデータ内に「ない・なく・な

1）https://pj.ninjal.ac.jp/corpus_center/bccwj/

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

表 1　否定要素認識の実験結果

	Pre.	Rec.	F_1
BERT fine-tuned	0.752	0.917	0.835
BERT frozen	0.771	0.924	0.840

かっ・なけれ・ぬ・ん・ず」は計 28 回出現し、そのうち *neg* ラベルが付与されていたのは 16 個であるので、どちらのモデルについても単純な表層形のみについて分類を行うよりは精度が高くなっている。frozen モデルは fine-tuned と比べてわずかに性能が良く、これには fine-tuned モデルが過学習している可能性が考えられる。

3.4　エラー分析

BERT は「違い **ない**」「似 て も 似つか **ぬ**」「暴力 に 暴力 で 向かっ て は いけ **ない**」「やる気 を **なく す** 人」などの否定として機能していない要素を否定と捉えるエラーを出していた。「違いない」「似ても似つかぬ」「いけない」のケースは慣用表現であり、「ない」が文否定として働いていないにも関わらず、そのことが捉えられていないからだと考えられる。このことから、BERT はトークンの一部が慣用表現の一部なのか、文の要素として機能しているのかを認識できていないということが示唆される。

一方で、否定を捉えられていないケースには「珍しく **なかっ** た」というケースがあり、5 回試行したうちの 1 度の試行で存在した。ただし、テストデータには「**なかっ**」は 3 回出現していた。

4　日本語 BERT による否定要素予測

3 節の実験では fine-tune された日本語 BERT が否定要素を認識できるかについて検証した。この節では fine-tune を行っていない BERT が否定要素を正しく予測できるかについて分析する。これはニューラル言語モデルの Probing の研究で行われている、言語モデルをブラックボックスとみなし、異なる入力を与えた際の出力からモデルの性質を分析するというアプローチ [18, 16] の系譜を踏むものである。

4.1　分析方法

まず 3 節の実験で用いたコーパスに含まれる 303 個の否定要素について、否定要素を含む文を抽出し、否定要素をマスクした文を BERT にインプットとして与え、BERT にマスクされたトークンを予測させた。予測に際しては Transformers から利用可能な fillmaskpipeline[3] を利用した。予測されたトークンが不正解であった場合について分析する。

4.2　分析結果

303 個の否定要素のうち、予測が不正解であったのは 66 個であった（付録 A）。

動詞との関連性　不正解のうち、11 個について観察された事象は、BERT が「ない」と「ある」の両方に対して比較的高い尤度を与えているというものであった。以下はそのような例の一つである。

- 罪の**ある**（正：ない）記者が次々死ぬのを見ると、戦争の意図がクリーンではないことがよく分かる」[PN3a_00002]

日本語においては「ある」という存在の動詞の否定形が「ない」であり、他の一般動詞の否定形のように「動詞の未然形＋ない」の統語構造を持たず、動詞の活用形から接続を予測することが困難であることが誤答の理由のと考えられる。また、一般動詞の否定形は動詞の未然形に接続するが、同じく未然形に接続する助動詞である「せる」にも比較的高い尤度を付与していたケースもあった。統語構造の観点からは、一段活用の場合には未然形と連用形が同じであるために五段活用と比較してさらに多くの選択肢が存在することから、BERT にとっても困難である可能性がある。このように、動詞の接続から一意に否定が導かれない場合における曖昧性の解消のためには BERT がより正確に文脈を考慮する必要があるだろう。

否定の接頭辞　次に多くみられた不正解の予測は否定の接頭辞を正しく予測することができていないという 16 の事象である。例えば、「就学児」の否定形は「不就学児」であるが BERT は「未就学児」であると予測した。これは「不就学」と比較して「未就学」という言葉が使われる頻度が高いからであると考えられる。また以下の例のように、「不 可能」を「ほとんど 可能」と予測するなど、正反対の予測をする場合も観察された。

- （前略）もはや民間まかせでは過剰債務処理は**ほとんど**（正：不）可能ということだ。[PN1b_00004]

このことからも BERT は否定の働きを理解できていないことが考えられる。

3）https://huggingface.co/transformers/main_classes/pipelines.html#fillmaskpipeline

加えて、「無 得点」を「連続 得点」、「不 干渉」を「相互 干渉」と予測していて、他の結びつきが強い名詞を選んで複合名詞を予測していると考えられる。ただし、「無 得点」に関しては、直前に「依然・まだ」という言葉がある場合には正しく予測することができ、これは BERT がマスクされたトークンの直前の文脈を考慮することができていることを示すものと考えられる。

5 日本語 NPIs に関する分析

この節では日本語 NPIs についての分析を行う。先行研究で示した NPIs に関する研究ではデータセットを構築し、BERT に文法性を判断させていたが、ここでは既存のコーパスを使用し、以下のような方法を用いて分析を行う。

5.1 分析方法

まず 3 節で用いたコーパスのうち、「ない」を含む文を抽出し、その後、日本語の NPIs である副詞の「さほど・それほど・そんなに・ろくに・だてに・二度と・あまり・しか・そんな・ろくな」などを含む文とそうでない文に分類した。用いた文全体に「ない」は 214 個存在しており、そのうち 43 個が否定の機能を持たない「ない」であり、171 個が否定要素の「ない」であった。否定要素の「ない」のうち、NPI と共起するものは 8 つであり、そのうちの 5 つが「しか」との共起であり、残りの 3 つはそれぞれ「そんな・そんなに・それほど」との共起であった。

3.4 で述べたように、コーパスには否定として機能していない「ない」が存在する。BERT が NPIs に関する統語的制約についての知識を持っているとすれば、中間層のレベルで、否定として機能していない「ない」と明確に区別されているのではないかと考えた。

そのことを検証するため、抽出した各文の「ない」というトークンの BERT の中間層の第 1 層を t-SNE アルゴリズムを用いて可視化した。ラベルは *neutral* の「ない」(neutral)、*neg* の「ない」で NPI と共起している場合（neg_npi)、*neg* の「ない」で NPI と共起しない場合（neg_non_npi）である。

5.2 分析結果

図 2 は日本語 BERT の第 1 層目の中間層を t-SNE で可視化したものである。右の方に否定の「ない」のクラスターができているように見えるが、NPI と共起した否定要素の「ない」がクラスターになっていることは観察できなかった。このことから文中の NPI の存在と否定要素の結びつきに関しては特別な情報を共有していないことが示唆される。

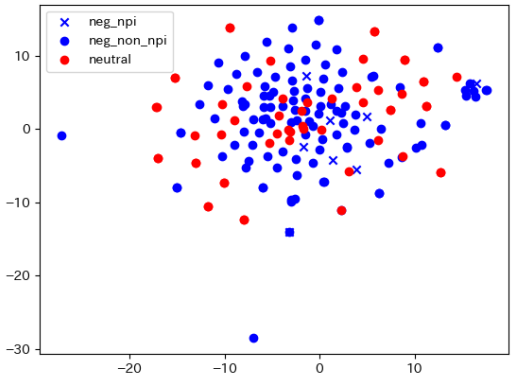

図 2 異なるラベルを付与された「ない」の分布

また、[3] によれば BERT は深い層より浅い層で表層的・統語的な情報を学習しており、[19] によれば、浅い層に比べて深い層で文脈依存の情報が扱われると報告されていて、実際により深い層においては分布にばらつきが存在していた（付録 B)。

今後は日本語 NPIs を用いて非文と文法的な文のペアを用意し BERT に文の尤度を出力させるなどしてより詳細に検証していきたい。

6 おわりに

本稿では事前学習済みニューラル言語モデル BERT の否定要素認識性能に着目した。今後は否定要素だけでなく、否定のスコープをアノテーションすることにより BERT がスコープを同定できるかについても詳しく検証していきたい。

否定要素の予測については、ある熟語に対して反意語を作る際に正しい接頭辞を予測できるかどうかで BERT を評価するタスクなどを行い、BERT が否定の接頭辞についてどのような傾向を示すかについても研究を進めていきたい。また、BERT が動詞の接続などの統語情報以外のどのような知識を用いて否定を認識しているのかについても研究を進めていきたい。

日本語 NPIs については、他言語での先行研究に準じ、データセットの構築を行って BERT による文法性判断をもとに BERT の統語構造に関連する知識についての研究を進めていきたい。

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] 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.

[3] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1073–1094, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] 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.

[7] 大槻諒, 松吉俊, 福本文代. 否定の焦点コーパスの構築と自動検出器の試作. 言語処理学会第 19 回年次大会論文集, pp. 936–939, 2013.

[8] Roser Morante and Eduardo Blanco. *SEM 2012 shared task: Resolving the scope and focus of negation. 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. 265–274, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[9] György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. The BioScope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pp. 38–45, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[10] Aditya Khandelwal and Suraj Sawant. NegBERT: A transfer learning approach for negation detection and scope resolution. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5739–5748, Marseille, France, May 2020. European Language Resources Association.

[11] Ieva Staliūnaitė and Ignacio Iacobacci. Compositional and lexical semantics in RoBERTa, BERT and DistilBERT: A case study on CoQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7046–7056, Online, November 2020. Association for Computational Linguistics.

[12] Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 34–48, 2020.

[13] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–7818, Online, July 2020. Association for Computational Linguistics.

[14] 片岡喜代子. 日本語否定文の構造: かき混ぜ文と否定呼応表現, 第 18 巻. くろしお出版, 2006.

[15] Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2877–2887, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 521–535, 2016.

[17] Roser Morante, Anthony Liekens, and Walter Daelemans. Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 715–724, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[18] Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5523–5539, Online, July 2020. Association for Computational Linguistics.

[19] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics.

A　否定要素の種類とその頻度

4 節の実験において、予測させたトークンごとに頻度とその不正解個数をまとめたものが以下の表である。

否定要素	ざる	ず	ない	なかっ	なき	なく	なくなっ	なけれ	なし	ぬ	ん	不	未	無	非
頻度	1	23	171	31	1	26	1	3	6	3	4	19	2	2	10
不正解	1	4	32	3	0	5	1	0	2	2	0	8	0	1	7

B　t-SNE による BERT 中間層の可視化

以下は 5 節の実験において異なるラベルを付与された「ない」の日本語 BERT の中間層の第 2（左上）、5（右上）、7（左下）、12（右下）層目の重みを t-SNE で可視化した様子である。

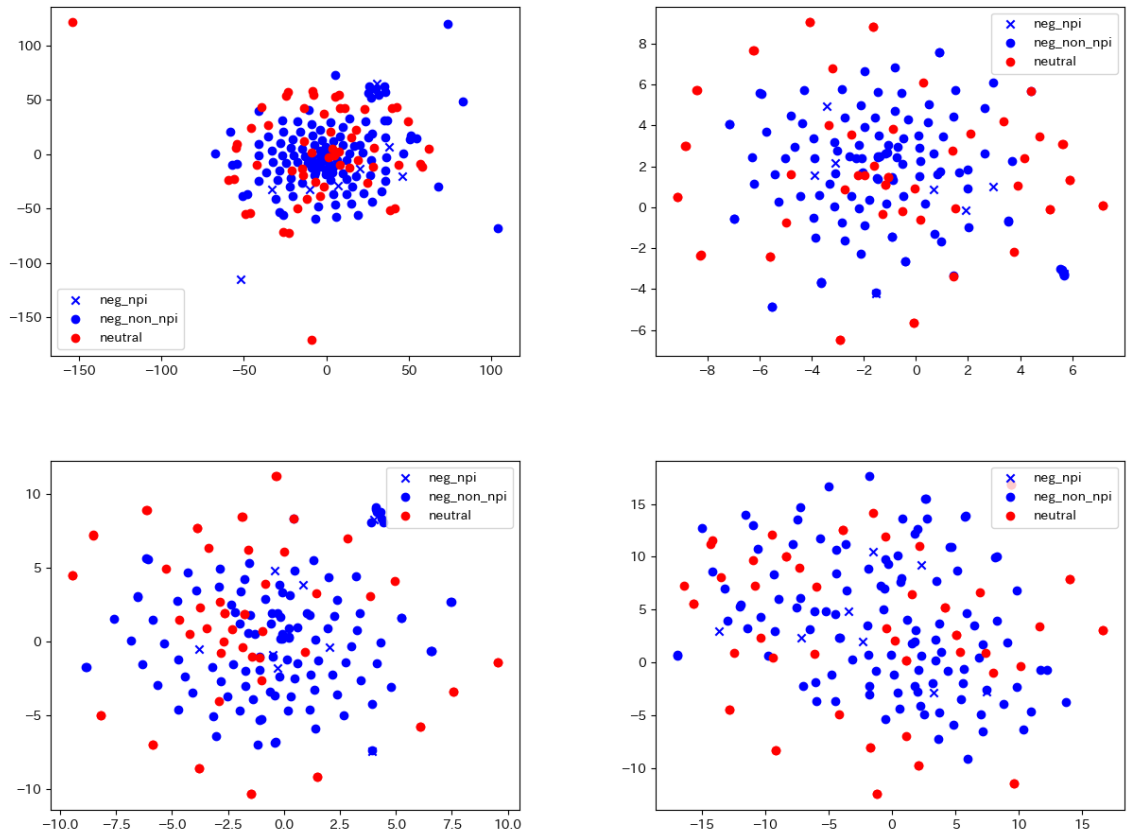

図 3　BERT 中間層における異なるラベルが付与された「ない」の分布