# プロンプトモデルは表面的手がかりを利用するか？

Pride Kavumba[1,2]　高橋 諒[3]　小田 悠介[3,4]
[1] 東北大学 大学院情報科学研究科　[2] 理研 AIP　[3]LegalForce Research
[4] 東北大学 データ駆動科学・AI 教育研究センター
pkavumba@ecei.tohoku.ac.jp
{ryo.takahashi,yusuke.oda}@legalforce.co.jp

## 概要

事前学習済み言語モデルを下流タスクに適応させる手法の一つであるプロンプトに基づくファインチューニングは、少量データでも高精度なモデルを学習できる手法として知られている。自然言語理解の多くのベンチマークデータセットは、タスクとは無関係だが正解ラベルの予測を助ける「表面的手がかり」を持つ。本研究はプロンプトに基づいて学習されたモデルがデータセットの表面的手がかりを利用していることを実験から明らかにする。

## 1　はじめに

事前学習済み言語モデルを下流タスクに適応させる手法として、タスクに応じたヘッドのファインチューニング、すなわち事前学習時に使用した言語モデルのヘッドをタスクに特化したヘッドに置き換える手法が主流となっている [1, 2]。このような手法でタスクに適応されたモデルをここでは「ヘッドモデル」と呼ぶ。

プロンプトに基づくファインチューニングは、少量のデータで言語モデルを適応させるためのより効果的な方法として知られている [3, 4, 5]。この手法では、言語モデルのヘッドはそのままで、モデルが事前学習時に見た入力と一致するようにモデルへの入力を構築する。例えば、「また見たいと思います。」という感情分析タスクのインスタンスが与えられた場合、ヘッドモデルは出力として「ネガティブ」か「ポジティブ」のどちらかを直接予測するように学習する。一方、マスク言語モデルの目的関数 [1] で学習されたプロンプトモデルでは、「{X}。[MASK] 映画でした。」のようなテンプレートを介して「また見たいと思います。[MASK] 映画でした。」という自然なテキストとなるように入力を構成する。モデルは [MASK] に「良い」や「ひどい」などの単語を予測し、それらを事前に定義したヴァーバライザー [4, 5] を使ってポジティブとネガティブに対応させる。

これまでの研究では、SNLI [6]、MNLI [7]、COPA [8] など、多くの自然言語理解データセットにおいて、簡単に利用できる「表面的手がかり」が見つかっている。表面的手がかりは、タスク自体とは無関係だが、タスクの特定のラベルに結びついているインスタンスの言語的または非言語的な特徴と定義できる。例えば、ARC [9] の正解には「not」という単語が含まれているが、これはヘッドモデルが利用することで、タスクを理解していなくても最先端の性能を発揮できる [10]。具体的には、MNLI のマッチしたインスタンスの 90%が文脈なし表面的手がかりを仮説に含んでおり、SNLI の文脈なし表面的手がかりはこれまで知られていたものよりも 4.9 ポイント高いことがわかった [11]。さらに、COPA には文脈なし表面的手がかり [12] だけでなく、78%のインスタンスに文脈あり表面的手がかり（前提と代替の間に存在する表面的手がかり）が含まれていることがわかった。一方、[11, 13] は、ヘッドモデルが SNLI の仮説において表面的手がかりを利用することを示した。[14] は MNLI データセットの前提条件と仮説の間の単語の重複などの表面的手がかりをヘッドモデルが利用することを示し、[12] は COPA の回答の選択肢において表面的手がかりをヘッドモデルが利用することを示した。本研究は、プロンプトモデルにおける表面的手がかりを調査した最初の研究である。我々は以下の問いを立てる。(1) プロンプトモデルは表面的手がかりを利用しているか？ (2) プロンプトモデルは表面的手がかりのないデータセットにどのように汎化するか？

MNLI、SNLI、COPA、および HANS データセット [14] を用いた慎重に設計された実験を通して、プロンプトモデルは表面的手がかりに大きく依存してお


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

り、表面的手がかりのないデータへの汎化に失敗していることを発見した。具体的には、RoBERTa [2] は、表面的手がかりを持つインスタンスでは良好な性能を発揮するが、表面的手がかりを持たないインスタンスではランダムベースラインとほぼ同等以下の性能しか達成できないことがわかった。

### 1.1 NLI と COPA の表面的手がかり

我々の研究課題「(1) プロンプトモデルは表面的手がかりを利用するか？ (2) プロンプトモデルは表面的手がかりのないデータセットにどのように汎化するか？」に答えるためには、表面的手がかりを含むデータセットが必要となる。本研究では表面的手がかりを含む新しいデータセットを一から作るのではなく、英語の MNLI、SNLI、COPA の既存のデータセットを分析することにした。具体的には、表面的手がかりを利用して正解できる部分集合と、表面的手がかりを利用しても解けない部分集合があるように、データセットを分割する。

我々は、表面的手がかりを、その利用可能性に基づいて、「文脈あり表面的手がかり」と「文脈なし表面的手がかり」の 2 つのカテゴリーに分ける。このように表面的手がかりを分けることで、タスクの種類を抽象化し、考えられているタスクに基づいて文脈を自由に定義することができる。例えば、機械読解や質問応答では、質問を解くために必要な関連する文書を文脈と定義することができるが、常識的な多肢選択式文章完成タスクでは質問を文脈と定義することができ、自然言語推論タスクでは前提を文脈と定義することができる。[14] が見つけた語彙的重複のような文脈あり表面的手がかりは、文脈があるときにしか利用できない。一方、[10] が見つけた正解の選択肢の中に「not」が含まれていることなどの文脈なし表面的手がかりは、タスクの関連する文脈がない場合に利用可能である。これまでの研究では、主に 1 種類の表面的手がかりのみを調査しており、ほとんどの研究が文脈なし表面的手がかりに焦点を当てていた [11, 13, 10, 12, 15]。本論文では、「あるモデルが表面的手がかりを利用するかどうか」という質問に答えるためには両方のタイプの表面的手がかりについて調査する必要があり、少なくとも 1 つのタイプの表面的手がかりを利用する場合にその答えは「Yes」であると主張する。

### 1.2 MNLI の表面的手がかり

**文脈あり表面的手がかり：**[14] は、MNLI において 3 つの文脈あり表面的手がかり（語彙的重複、部分列、構成素）を特定した。これらの表面的手がかりは含意ラベルを予測する。部分列と構成素の表面的手がかりは、いずれも語彙的重複の特殊なケースである。これらの表面的手がかりに対抗するために、[14] は HANS データセットを作成し、表面的手がかりでは情報が得られないインスタンスを含んでいる。HANS データセットには、表面的手がかりを持つインスタンスも含まれており、これらのインスタンスで高い性能を発揮することで、表面的手がかりが利用されていることが確認できる。本論文では、HANS データセットを用いて、プロンプトモデルが文脈あり表面的手がかりを利用しているかどうかを評価する。

**文脈なし表面的手がかり：**プロンプトモデルが文脈なし表面的手がかりを利用する能力を調べるために、元のデータセットを文脈なし表面的手がかりを持つインスタンスと持たないインスタンスに分割し、仮説のみで RoBERTa ヘッドモデルを訓練する。この分析は、より性能の低いモデルを使用しているが、[11] によって行われたものと似ている。仮説が文脈なし表面的手がかりを持たない場合、モデルはランダムベースラインの精度（33%）を超えることは期待できない。しかし、複数の RoBERTa モデルの平均精度は、ドメイン内事例で 90 %、ドメイン外事例で 90 %を達成しており、MNLI の文脈なし表面的手がかりはこれまで知られていたよりもはるかに悪いことが示されている（ドメイン内で 53.9 %、ドメイン外で 52.3 % [11]）。この結果を受けて、データセットを表面的手がかりを持つインスタンスと持たないインスタンスに分けた。表面的手がかりを持つインスタンスは、文脈なしの設定で大多数のモデルが正しく予測したインスタンスを含んでいる。

### 1.3 SNLI における表面的手がかり

**文脈あり表面的手がかり：**HANS データセットは MNLI の表面的手がかりに基づいて構築されているが、先行研究では、SNLI で訓練されたモデルのベンチマークとしても使用できることが示されている。先行研究に倣い、本研究では SNLI で学習されたプロンプトモデルを HANS データセットで評価する。

**文脈なし表面的手がかり：**SNLI において文脈な


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

し表面的手がかりを利用するプロンプトモデルの能力を調べるために、元のデータセットを文脈なし表面的手がかりを持つインスタンスと持たないインスタンスに分割した。また、MNLI で行ったように、RoBERTa ヘッドモデルを仮説のみで訓練する。仮説が文脈なし表面的手がかりを持たない場合、モデルはランダムベースラインの精度（33%）を超えることは期待できない。しかし、RoBERTa の平均精度は 71.9%であり、これは [11] が fastText[16] で得た結果を 4.9 ポイント上回っている。この結果を受けて、データセットを表面的手がかりを持つインスタンスと持たないインスタンスに分けた。表面的手がかりを持つインスタンスには、文脈なしの設定で大多数のモデルが正しく予測したインスタンスが含まれている。

### 1.4 COPA における表面的手がかり

**文脈あり表面的手がかり：**COPA における文脈なし表面的手がかりはこれまでにも分析されているが [12]、文脈あり表面的手がかりはまだ分析されていない。共通のパターンを見つけるためにすべてのインスタンスを目視することは困難であり、エラーが発生しやすいことがわかる。文脈あり表面的手がかりを分析するために、我々はプロンプトモデルを採用し、表面的手がかりでしか解けないように入力を変更する。具体的には、答えの選択肢に含まれる単語をランダムに並び替えて、文として意味が通らないようにする。この設定では、インスタンスに表面的手がかりがない場合、モデルの性能はランダムベースラインの精度（50%）に一致すると予想される。驚くべきことに、複数の RoBERTa モデルは、78%の平均精度を達成しており、文脈あり表面的手がかりが存在することを示している。この結果を受けて、データセットを、大多数のモデルが解答したインスタンスを含む表面的手がかりを持つ部分集合と、残りのすべてのインスタンスを含む表面的手がかりを持たない部分集合に分割した。

**文脈なし表面的手がかり：**[12] は、RoBERTa を用いて COPA における文脈なし表面的手がかりを特定した。彼らの分析は本研究と同一の設定であるため、本研究では再分析を行わず、彼らが公開しているデータセットの分割を使用する。

## 2 実験と結果

本論文の研究課題に答えるために、文脈あり表面的手がかりのあるデータセットと文脈なし表面的手がかりのあるデータセットでプロンプトモデルを評価する。表面的手がかりを利用しないモデルは、表面的手がかりがある場合とない場合の両方で同等の性能を発揮すると予想される。表面的手がかりがある場合の方が、表面的手がかりがない場合よりも高い性能を発揮するということは、そのモデルが表面的手がかりを利用していることを示している。

### 2.1 MNLI

**文脈あり表面的手がかり：**プロンプトモデルが文脈あり表面的手がかりを利用するかを確かめるために、MNLI 上で複数のプロンプトベースの RoBERTa モデルを学習し、HANS データセットの表面的手がかりを持つインスタンスと持たないインスタンスで評価した。表面的手がかりを利用しないモデルは、両方のインスタンスで同等の性能を発揮することが期待される。

表 1 によると、RoBERTa は表面的手がかりを持つインスタンスではかなりの性能を発揮するが、表面的手がかりを持たないインスタンスでは 50%のランダムな精度に達することができない。このモデルは 3 種類の表面的手がかりのすべてにおいて性能が低く、3 種類の表面的手がかり（語彙的重複、部分列、構成素）のいずれにおいても劣っている。この結果から、プロンプトモデルは表面的手がかりに依存しており、表面的手がかりのないインスタンスへの汎化に失敗していることがわかる。

**文脈なし表面的手がかり：**プロンプトモデルが文脈なし表面的手がかりも利用するかを確かめるために、MNLI 上で複数のプロンプトベースの RoBERTa モデルを学習し、文脈なし表面的手がかりがある場合とない場合のインスタンスで評価した。表面的手がかりを利用しないモデルであれば、表面的手がかりがある場合もない場合も、同等の性能が得られると考えられる。

表 1 によると、RoBERTa は表面的手がかりを持つインスタンスで、表面的手がかりを持たないインスタンスよりもかなり良い性能を示しており、モデルが文脈なし表面的手がかりを利用していることを示している。


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

| データ | w/ Cues | 文脈あり表面的手がかり | | | 文脈なし表面的手がかり |
|---|---|---|---|---|---|
| | | 語彙的重複 | 部分列 | 構成素 | |
| MNLI | Yes | 99.0±0.4 | 97.4±1.0 | 97.4±1.0 | 69.5±0.1/72.0±0.2 |
| | No | 9.3±2.9 | 8.9±2.2 | 8.9±2.2 | 38.6±0.6/41.5±0.7 |
| SNLI | Yes | 87.8±8.4 | 95.3±3.9 | 95.3±3.9 | 82.2±0.9 |
| | No | 57.0±17.8 | 17.6±15.9 | 17.6±15.9 | 67.1±1.0 |

**表 1** MNLI（ドメイン内／ドメイン外）および SNLI における RoBERTa プロンプトモデルの精度。“w/ Cues” の列は評価セットに表面的手がかりを含むか否かの二値を示す。

| Train Size | 文脈あり | | 文脈なし | |
|---|---|---|---|---|
| | No | Yes | No | Yes |
| 8 | 83.0±0.5 | 55.7±1.4 | 79.1±0.9 | 77.4±0.0 |
| 16 | 81.6±1.9 | 57.9±1.0 | 77.0±1.3 | 77.4±1.6 |
| 32 | 82.4±2.0 | 53.5±0.5 | 77.7±3.1 | 76.8±0.9 |
| 64 | 84.4±1.4 | 53.8±4.1 | 80.2±0.5 | 78.1±1.6 |
| 96 | 87.0±1.7 | 57.9±4.1 | 83.7±1.1 | 80.4±0.7 |
| 100 | 87.9±1.7 | 54.2±1.9 | 84.6±1.1 | 80.0±1.6 |

**表 2** COPA における RoBERTa プロンプトモデルの精度。

## 2.2 SNLI

**文脈あり表面的手がかり：**プロンプトモデルが SNLI 上の文脈あり表面的手がかりを利用するかどうかを調査するために、SNLI 上でプロンプトモデルを訓練し、HANS データセットからの表面的手がかりを持つインスタンスと持たないインスタンスで評価する。このデータセットは MNLI 特有の表面的手がかりのために設計されたものだが、先行研究では表面的手がかりを利用する SNLI モデルを明らかにできることが示されている。

表 1 によると、RoBERTa は表面的手がかりを持つインスタンスではかなりの性能を発揮するが、表面的手がかりを持たないインスタンスでは同じ性能を達成できない。この高い分散はヘッドモデルについて調査した先行研究で報告された分散と似ている。SNLI で学習したモデルは MNLI モデルよりもはるかに良い性能を発揮するが、それでも性能は悪いままである。この結果はプロンプトモデルが文脈あり表面的手がかりを利用していることを示している。

**文脈なし表面的手がかり：**プロンプトモデルが SNLI 上で文脈なし表面的手がかりを利用する能力を調べるために、SNLI で学習し、文脈なし表面的手がかりを持つインスタンスと持たないインスタンスで評価した。

表 1 によると、文脈なし表面的手がかりがある場合とない場合では精度に大きな差があり、プロンプトモデルが文脈なし表面的手がかりを利用することを示している。

## 2.3 COPA

**文脈あり表面的手がかり：**COPA データセットにおいてモデルが文脈あり表面的手がかりを利用しているかどうかを調べるために、8 から 100 までの様々なサイズで訓練を行い、1.4 節 で説明されている文脈ありと文脈なし表面的手がかりを持つインスタンスで評価を行った。

表 2 によると、RoBERTa は、表面的手がかりを持つインスタンスで非常に優れた性能を発揮するが、表面的手がかりを持たないインスタンスでは 50%のランダムな精度をほとんど超えない。これは表面的手がかりに過度に依存し、表面的手がかりのないインスタンスへの汎化に失敗していることを示している。またこのモデルは文の意味に敏感ではないことも示されている。

**文脈なし表面的手がかり：**プロンプトモデルが COPA 上の文脈なし表面的手がかりを利用しているかどうかを調べるために、COPA データセットの 8 から 100 までの異なるサイズで RoBERTa を学習し、文脈なし表面的手がかりを持つインスタンスと持たないインスタンスで評価した。

表 2 は、32 以下の小さい学習サイズでは RoBERTa が表面的手がかりを利用しないことを示している。しかしサイズを大きくすると、文脈のない表面的手がかりを持つインスタンスと持たないインスタンスの間の性能の差がさらに大きくなる。このモデルが数ショットの設定で一般的に使用されるサイズにおいて、文脈なし表面的手がかりを利用しないことは心強いことである。


This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

# 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. **CoRR**, Vol. abs/1907.11692, , 2019.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[4] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume**, pp. 255–269, Online, April 2021. Association for Computational Linguistics.

[5] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2339–2352, Online, June 2021. Association for Computational Linguistics.

[6] 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.

[7] 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, Volume 1 (Long Papers)**, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning**, Stanford University, 2011.

[9] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. SemEval-2018 task 12: The argument reasoning comprehension task. In **Proceedings of The 12th International Workshop on Semantic Evaluation**, pp. 763–772, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. When choosing plausible alternatives, clever hans can be clever. In **Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing**, pp. 33–42, Hong Kong, China, November 2019. Association for Computational Linguistics.

[13] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In **Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics**, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[14] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[15] Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. 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. 3410–3416, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 427–431. Association for Computational Linguistics, April 2017.




This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).