否定の理解への prompt-based finetuning の効果

田代真生 [1]　上垣外英剛 [2]　船越孝太郎 [2]　奥村学 [2]
[1] 東京工業大学工学院　[2] 東京工業大学科学技術創成研究院
{masaki@lr.,kamigaito@lr.,funakoshi@lr.,oku@}pi.titech.ac.jp

概要

自然言語処理において否定の理解は重要な課題の一つである．否定は文章の極性や論理的関係を変える役割がある一方で，その理解は現状の自然言語処理モデルでも難しく，否定の理解の不足を指摘する研究は多くある [1, 2, 3]．否定の理解において，大量のラベルなしコーパスから事前学習を通じて獲得された否定に関する知識は有用と考えられる．そこで本研究では様々な自然言語処理タスクを言語モデリングタスクに変形して解くことにより，事前学習で得られた言語知識をより効率的に活用することを考える．そのための手段として prompt-based finetuning に着目し，テキスト分類タスクにおける否定の理解に与える効果を調査する．

1　はじめに

否定は命題の真偽を変化させる役割があり，自然言語処理において無視できない要素である．そのため，様々な研究において否定を理解するような自然言語モデルが調査されているが，近年開発された事前学習済み言語モデルでも否定の理解が不十分であることが指摘されている．事前学習済み言語モデルは大量のラベルなしコーパス における言語モデルの自己教師あり学習を通して様々な言語的知識を獲得しており，幅広いタスクにおける性能の向上を達成している．Saunshi ら [4] は事前学習の効果を多くの自然言語処理タスクが言語モデリングへの変形によって解けることを用いて説明している．

prompt-based finetuning はこのようなアイデアから生まれたものである．後段のタスクを言語モデリングの形式に変形することで，事前学習タスクと後段タスクとの差異を低減する．それにより訓練事例が少数しか利用できない few-shot な設定下の広範なタスクにおいて性能向上を達成した．ここから prompt-based finetuning は，事前学習済みモデルの利用法として一般的である [CLS] トークンの出力からラベルを予測する従来の head-based な手法 [5] に比べ，事前学習によって得た言語的な知識を学習のより初期段階から引き出せていると考えられる．

そこで，本研究では現状の自然言語処理モデルにおいて課題となっている否定の理解についても prompt-based finetuning を利用することで事前学習により獲得された知識を活用できるかを調べた．具体的には，自然言語処理タスクの一例であるテキスト分類タスクとして，感情分析 (SST-2, SST-3, SST-5) と自然言語推論 (RTE, MNLI) を対象とし，head-based な手法と prompt-based な手法で学習したモデルの性能を，average data points advantage [6](ADPA) を用いて比較する．否定評価用テストデータと否定評価用でないテストデータ両方で調査し比較することによって少数データの条件においては prompt-based finetuning が否定の理解に効果的であることを示す．

本研究は ADPA を用いることで，評価が難しかった自然言語処理モデルの否定理解の度合いを定量化し，それによって今まで理解が進んでいなかった prompt-based finetuning が否定の理解に与える影響を示した. このことは否定を理解する自然言語処理モデルの作成に役立ち，また prompt-based finetunig への理解を促すものであると考える．

2　関連研究

2.1　prompt-based finetuning

prompt-based finetuning は自然言語処理のタスクを言語モデリングのタスクに変形し，分類ラベルと対応づけたトークンを予測する学習方法である [7, 8]. prompt-based finetuning を構成する要素としては元の文章を事前学習時の入力形式に変形するルールである template と，マスク箇所された箇所の出力を対象タスクのラベルに変換するルールである verbalizer の 2 つがあり，その二つは合わせて prompt と呼称される [9]．例えば, 感情分析タスクは表 1 のように元の文章 ‘Best pizza ever!’ を ‘Best pizza ever! It was

表 1: 利用した template と verbalizer

task	template	label words
SST-2	<SentA> It was [MASK].	terrible/great
SST-3	<SentA> It was [MASK].	terrible/okay/great
SST-5	<SentA> It was [MASK].	terrible/bad/okay/good/great
RTE	<SentA> [MASK], I believe <SentB>	Clearly/Yet
MNLI	<SentA> [MASK], you are right, <SentB>	Fine/Plus/Otherwise

[MASK].' のような入力に変形し，[MASK] の部分について，ラベルが正例のときは 'great' を，負例のときは 'terrible' を予測するようにモデルを学習することで，事前学習と同じような形でタスクを解くことができる．

prompt-based finetuning は [CLS] トークンの出力からラベルを予測する従来の head-based finetuning に比べて few-shot で良い性能を発揮することが知られている．Scao ら [6] はこの prompt がもたらす効率性を average data points advantage (ADPA) という指標を用いて定量的に示した．本研究ではこの ADPA の考え方を用いて prompt-based finetuning が否定の考慮に有用であることを示す．

2.2 自然言語処理における否定の処理

否定は命題の真偽を変化させるものであり，文章の意味を正しく理解する上で否定を理解することが重要であるため自然言語処理においても様々な研究で取り組まれてきた課題である [10]．しかし，近年の自然言語処理モデルでも否定を含む文章の処理を適切にできないことを指摘する研究は多くある．

Ettinger や Kassner ら [11, 12] は否定を含む穴埋め式質問応答タスクにおいて，事前学習済み言語モデルの出力が入力中の否定語の有無に影響されないことを示している．また，Ribeiro ら [1] は否定を含んだ人工のテキストのデータセットである checklist を用いて事前学習済み言語モデルや多くの商用の自然言語処理モデルが否定を含んだ入力に対して間違いやすいことを確認している．Hossain ら [2] は自然言語推論タスクにおいて多くのデータが否定を考慮せずに解ける問題を指摘し，否定の考慮が必要なテストセットにおいて既存のモデルが低い性能であることを指摘した．

本研究はこのような現状の自然言語処理における課題を解決することに prompt-based finetuning が有用であるかを調べる．

3 提案手法

本稿では prompt-based finetuning の否定理解の効果を測る手法として ADPA (average data points advantage) を用いることを提案する．ADPA は Scao ら [6] によって提案されたものであり，prompt-based にすることによって得られる利得を学習データの数で表現したものである．本節では 3.1 節で ADPA を紹介し，3.2 節で ADPA を利用することによって，これまで評価が難しかった prompt-based finetuning の否定理解に与える影響を，定量的に調べることができると考える理由を示す．

3.1 Average Data Points Advantage

ADPA は比較対象のモデルがベースラインとなるモデルと比べてどの程度効率的に学習を進めることができるかを定量化した指標である．ベースラインモデルと比較モデルを複数のデータサイズの設定で学習，評価を行い性能を比較することで，比較モデルが平均的にどのくらい少ないデータでベースラインモデルと同程度の性能を達成できるか表現している．

ADPA はデータサイズが x 以下の時のモデルの性能を示す関数 Model(x) を用いて $\frac{1}{X_{max}-X_{min}}\int_{X_{min}}^{X_{max}} \mathrm{Model}_A(x) - \mathrm{Model}_B(x)dx$ と表される．ここで X_{min}, X_{max} は実験を行ったデータサイズの内最小のものと最大のものを示しており，Model(x) は離散的に得られた結果から線形補間されたものとなっている．

3.2 否定理解に関する指標

本研究では ADPA を否定理解を測る指標として，その他の指標に比べて適切であると考えた．以下にその理由を説明する．

否定理解を測るその他の指標としてまず考えられるのが否定を含むデータセットにおける正答率である．これは確かに否定の理解の度合いと評価値が相関するため否定の度合いを測ることにつながると考

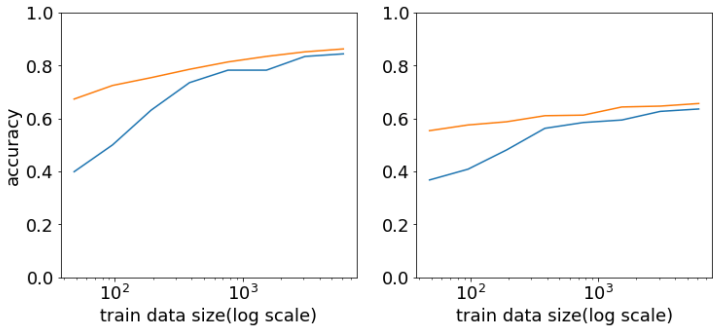

図 1: MNLI におけるデータサイズと性能の関係．右側は否定のテストデータでの性能，左側は一般のテストデータでの性能を示している．赤線（上）が prompt-based で，青線（下）が head-based の正解率を示す．

えられる．しかし，複数モデルを比較する際に否定理解以外の要素によって差が生まれ得るため，この指標では否定理解の能力を直接的に測ることが難しい．

次に考えられる指標としては否定を含まないデータでの評価値と否定を含むデータでの評価値の差分をとる指標である．しかし，この指標は，性質の異なる二つの「評価値の差分」の比較においては解釈が困難であり，手法間の比較には適さない．

一方で ADPA の指標は，モデルによってもたらされる利得を両方のテストデータで共通する「訓練データ数」の単位で表現しているため，否定用テストデータと一般のテストデータでの評価値の性質の違いに影響されず，否定を含むテストデータでの利得と否定を含まないテストデータでの利得を比較することができるという利点がある．

4 実験

prompt-based finetuning と head-based finetuning によって少数データの設定で感情分析タスクと自然言語推論タスクを学習し，その結果を ADPA を利用した否定理解の指標を用いて実際に評価することで，prompt-based finetuning がテキスト分類タスクにおける否定理解に与える影響を調査する．

4.1 実験設定

実験では感情分析タスクとして SST-2, SST-3, SST-5 [13] を選択し，自然言語推論タスクとしては RTE [14], MNLI [15] を選択した．これらはタスク中に否定を含む文章が一定割合含まれており，かつ否定の考慮が必要となるテストデータが入手可能であるタスクとして選んだ [13, 2].

実験は，各ラベルごとに N 個ずつ抽出した訓練データと検証データを用いた学習・検証と評価を繰り返しながら，少数データ (数千データ以下) の範囲で N を増加させて行った（付録 表 5 参照）．学習は Gao らの条件 [9] を参考に，学習率 1e-5, 学習回数 1000step で行った．100step ごとに検証データで評価した中で最も性能の良かったモデルを用いてテストデータでの性能を評価した．その評価値で ADPA を計算した．実験は性能の分散を考慮して 10 回ランダムシードを変えて行い，テストデータでの評価について対応のある両側 t 検定を行った．事前学習済み言語モデルとしては Scao らの研究 [6] で prompt-based finetuning にて一貫した結果を示すと報告された roberta-large[1)][16] を利用した．

prompt は表 1 のものを使用した．SST に関しては，Gao らの研究 [9] を参考にし，既存研究で利用されることの多い人手で作成された prompt を利用した．一方で，MNLI，RTE については，人手で作成された prompt に問題があったため，Gao らの研究で自動生成された prompt を利用した．（この点については比較結果を後で示す．）

テストデータで評価をする際には否定評価用のテストデータ (Negation) と否定評価用ではないテストデータ (Not Negation) の両方における prompt-based finetuning の ADPA を比較することによって否定理解の度合いを評価した．

SST では，先行研究 [17] を参考に否定要素 (Negation cue) の候補を用意し（付録 表 6），一般のテストデータのうち否定要素を含む文章を否定用のテストデータに分割して，否定を含むテストデータ (Negation) と否定を含まないテストデータ (Not Negation) を作成した．

RTE・MNLI では既存の否定評価用データ [2] を Negation に用いた．Not Negation には各タスクのオリジナルのテストデータを用いた．Not Negation の

1) https://huggingface.co/transformers/

性質が SST と異なることに注意されたい．

RTE と MNLI の否定用テストデータはオリジナルのデータを元に人手で否定を加えて作成されており，SST では同じデータセットから分割されている．このことから，どちらの実験でも二つのテストデータの性能の差は基本的に否定の理解によってのみ変化することを仮定している．

4.2 結果

prompt-based finetuning の効果

表 2 にテキスト分類タスクにおける prompt-based finetuning の ADPA を示す．この表より少数データの条件の自然言語推論のタスクにおける prompt-based finetuning の否定理解の性能の高さを読み取れる．図 1 を見ても，prompt-based がはじめから比較的高い性能を示している一方で，head-based では否定の理解は比較的緩やかに進み，prompt-based の性能に追いつくために多くの訓練データが必要であることがわかる．ここから，否定理解を事前学習済み言語モデルが事前学習によって獲得しており，言語モデリングの形でタスクを解くことによって，獲得された情報を活用可能であると考えられる．

一方で感情分析タスクにおいては，prompt-based finetuning の否定理解に対する効果を確認することができなかった．この理由としては SST に含まれる否定の理解が言語モデルにとって容易であることが考えられる（付録 表 7）．実際に 1024 データで SST-2 を学習したモデルの Negation を含まないテストセットと含むテストセットでの正答率はそれぞれ 0.949, 0.942 でほとんど違いがなく，head-based でも容易に学習可能な知識であったことが考えられる．

表 2: average data points advantage

	Sentiment			NLI	
設定	SST-2	SST-3	SST-5	MNLI	RTE
Not Negation	162	280	-12	850	327
Negation	165	202	-123	**2353**	**820**
p-value	0.965	0.284	0.180	0.015	0.001

template, verbalizer の影響

自然言語推論の template については ‘<SentA>? [MASK], <SentB>’ [7] というものが考案されている（付録 表 8）．しかし，この template では<SentB>が否定を含む際に，<SentA>と<SentB>の関係に関わらず [MASK] 部が No となってしまうという問題が生じることがある．例えば SentA=‘That man is young.’, SentB=‘That man is not old.’ という条件において，事前学習済み言語モデルが前の文は後の文を含意していること理解していても，英語の文法的には矛盾の label と対応する ‘No’ を予測することが正しいため，不正解の ‘No’ を予測してしまう．このような問題に対処するため本研究では NLI に関しては自動生成された prompt [9] を用いたが，表 3 に自動生成された prompt と人手で作成された prompt の比較結果を示す．表の結果は，上記の template が否定の考慮に不適であることを示しており，prompt 設計の重要性に関する先行研究 [18] の主張とも一致している．

表 3: MNLI における template と verbalizer の影響

template	Negation	Not Negation	p-value
人手作成	480	690	0.186
自動生成	2353	850	0.015

モデル比較

roberta-large は Talmor ら [3] や田代ら [19] が主張するように，事前学習によって否定に関する知識を比較的多く獲得していると考えられるが，否定をあまり理解していないと考えられるモデルにおいても prompt-based finetuning による効果が観測されるかを調べるために，そのようなモデルの一つである roberta-base を選択し，MNLI データを利用して表 4 に示すように roberta-large モデルとの性能比較を実施した．

結果を見ると roberta-base では prompt-based finetuning が否定の理解に与える影響を確認することができなかった，ここから roberta-base の様に事前学習において多くの否定知識を獲得していないモデルにおいては prompt-based finetuning を用いても性能が向上するとは言えないことがわかる．

表 4: モデルサイズの影響

model	Negation	Not Negation	p-value
roberta-base	567	539	0.810
roberta-large	2353	850	0.015

5 おわりに

本研究では prompt-based finetuning の事前学習知識を生かしやすい特性に着目し，テキスト分類タスクの否定を含むテストデータと一般のテストデータにおける average data point advantage を比較することによって，prompt-based finetuning が少数データの設定において否定の理解へ有用であることを示した．

参考文献

[1] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics.

[2] 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)**, pp. 9106–9118, Online, November 2020. Association for Computational Linguistics.

[3] Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. oLMpics-on what language model pre-training captures. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 743–758, 2020.

[4] Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. A mathematical exploration of why language models help solve downstream tasks. In **International Conference on Learning Representations**, 2021.

[5] 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.

[6] Teven Le Scao and Alexander Rush. How many data points is a prompt worth? In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 2627–2636, Online, June 2021. Association for Computational Linguistics.

[7] 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.

[8] 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.

[9] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 3816–3830, Online, August 2021. Association for Computational Linguistics.

[10] Roser Morante and Caroline Sporleder. Modality and negation: An introduction to the special issue. **Computational Linguistics**, Vol. 38, No. 2, pp. 223–260, June 2012.

[11] 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.

[12] 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.

[13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[14] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In **Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment**, MLCW'05, p. 177–190, Berlin, Heidelberg, 2005. Springer-Verlag.

[15] 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.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Ro{bert}a: A robustly optimized {bert} pretraining approach, 2020.

[17] Prakash Kumar Singh and Sanchita Paul. Deep learning approach for negation handling in sentiment analysis. **IEEE Access**, Vol. 9, pp. 102579–102592, 2021.

[18] Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? **CoRR**, Vol. abs/2109.01247, , 2021.

[19] 田代真生, 上垣外英剛, 船越孝太郎, 高村大也, 奥村学. 事前学習済み言語モデルにおける否定の理解能力の調査. 情報処理学会研究報告, 第 2021-NL-249 巻, jul 2021.

A 付録 : 実験条件の詳細

表 5: 各実験の X_{min} と X_{max}

	Sentiment			NLI	
	SST-2	SST-3	SST-5	MNLI	RTE
X_{min}	32	48	80	48	32
X_{max}	1024	1536	2560	6144	1024

表 6: 否定要素の候補

否定要素
no not n′t never nor none without nothing neither hardly

表 7: 否定用テストデータの例

タスク	入力例	ラベル
SST-2	no movement , no yuks , not much of anything . it represents better-than-average movie-making that does n't demand a dumb , distracted audience .	negative positive
MNLI	P : The United States falls somewhere between these extremes. H : The United States does not conform to only one of these extremes.	entailment
	P : These laws are not uncoordinated and often inconsistent. H : The laws are not coordinated and applied consistently.	contradiction
	P : The United States does not fall somewhere between these extremes. H : The United States conforms to only one of these extremes.	neutral

表 8: 人手で作成された自然言語推論用 prompt

template	label words
<SentA>? [MASK], <SentB>	Yes/Maybe/No