言語モデルに対するトークンのキャンセルアウト手法の比較

鈴木淳平 [1] 菅原朔 [2] 相澤彰子 [1,2]
[1] 東京大学 [2] 国立情報学研究所
junpei0827kouyou@g.ecc.u-tokyo.ac.jp {saku,aizawa}@nii.ac.jp

概要

言語モデルへの入力文内の特定のトークンの影響を入力から取り除く手法である「キャンセルアウト（cancel out）」は，モデルの推論解釈や得られた解釈の良さの評価のために有効な技術である．しかし既存研究では，キャンセルアウト手法の評価方法が確立されておらず，利点や欠点を相互に比較することが困難であった．本研究ではそれらの性能を直接的に評価するために，1) キャンセルアウトしたトークンの復元可能性，2) 性別情報のキャンセルアウトによるジェンダーバイアスの緩和効果について実験を行う．また，複数のトークンによる置き換えに基づく既存のキャンセルアウト手法を，置き換えをより多様にすることで改善できることを示す．

1 はじめに

言語モデルへの入力文内の特定のトークンの影響を取り除く手法である「キャンセルアウト」は，それらのトークンがある時とない時のモデルの出力を比較することで，その部分の重要性を測って解釈に用いたり [1]，モデルの出力に関する解釈が与えられた時に，その解釈に基づいて文の一部をキャンセルアウトしてモデルの出力の変化を見ることで解釈を評価したりする [2] ことなどに役立つ．図 1 は，理想的なキャンセルアウトの具体例である．映画のレビューの感情分析を学習したモデルに対し，元の “This is a good movie.” を入力すると positive に判定され，“good” をキャンセルアウトした後の入力に対しては，neutral と判定されている．したがって，これらの予測を比較することで “good” が positive のラベルに寄与していると解釈できる．

最も単純なキャンセルアウトは，対象のトークンを文から取り除くことであるが，自然言語においては文法的に不適格な入力を作り得る．そこで既存研究では，特殊なトークンで置き換えたり（Zero Vector, Mask token），注意機構 [3] を持つモデルに対しては，対象のトークンへの注意を抑制する（Attention Mask）手法が考えられてきた．また，キャンセルアウト後の入力が，モデルが学習した訓練データに対して分布外にならないように実際にその箇所に出現しやすいトークンで置き換える手法も提案されている（Weighted Marginalize）[1].

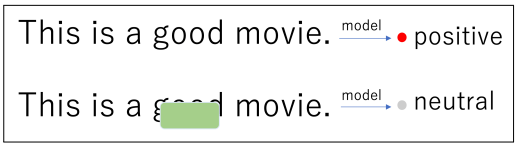

図 1 “good” のキャンセルアウト．映画のレビューの感情分析モデルに元の入力を入れると positive と判定され，“good” をキャンセルアウトすると neural と判定されることから，“good” が positive の根拠であると解釈される．

キャンセルアウト手法の評価は，キャンセルアウトを用いた解釈手法の性能が向上することを示して間接的に行われることがある [1]．しかし，解釈手法の評価自体が，現状確立されたものがなく，何を評価しているかが明示されないことが多いことも指摘されている [4]．また，解釈手法を統一的に評価するためのベンチマークが提案されているが [2]，このベンチマークは評価指標の計算にキャセルアウトを伴うので，キャンセルアウトを含む解釈手法の評価には使えない可能性がある．他にも，キャンセルアウト後の入力がどれくらいタスクの学習データに対して分布外になってしまうかを測るものもあるが [5]，そもそもキャンセルアウトがうまくいっておらず，モデルが周辺文脈から十分に元の情報を復元できている可能性が考慮されていない．例えば，図 1 において，キャンセルアウト後の文をモデルに入力した時に，モデルがそこには元々 “good” があったと予測できてしまう場合にはキャンセルアウトが機能していないと考えられる．

そこで本研究では，モデルに対して対象のトークンをどれだけキャンセルアウトできているかという観点から各手法を比較する．まず 3 節で，文脈から各位置のトークンを予測するように学習されてい

る masked language model （MLM） に対してキャンセルアウト後の入力を与え，キャンセルアウトしたはずのトークンをどれくらい予測できるかを比較する．さらに，対象のトークンの箇所にどれくらい高い確信度を持っているかも比較する．次に 4 節で，キャンセルアウトの応用として，BIOS [6] という個人の経歴から職業を予測するデータセットに対し，どれくらいモデルにジェンダーバイアスをかけずに学習できるかを比較することで，キャンセルアウト手法の性能を確かめる．結果として，2 節で提案する，Weighted Marginalize の周辺化トークンをより多様化した Uniform Marginalize が，3 節と 4 節の実験で最もキャンセルアウトの効果が高いことを示す．

2 キャンセルアウト手法

本研究では，2.1 節に示す既存の 4 つのキャンセルアウト手法 [5]，および新たに提案する Uniform Marginaize と呼ぶ手法 （2.2 節） を比較する．[1)]

2.1 既存のキャンセルアウト手法

- **Zero Vector**: 対象のトークンの embedding を全要素が 0 のベクトルで置き換える．
- **Mask Token**: 対象のトークンを MASK トークンで置き換える．
- **Attention Mask**: 対象のトークンに対する attention mask を 0 にして，他のトークンの中間ベクトルの計算時に参照されないようにする．
- **Weighted Marginalize**: 以下の式 1 のように，対象のトークンを一旦 MASK トークンで置き換えて MLM に入力し，語彙内の各トークンに対する尤度を得て，上位 k 個のトークンで置き換えた後の入力に対するモデルの出力をそれぞれの置き換えトークンの尤度で周辺化する．[1]

$$p(y \mid \boldsymbol{x}_{-i}) = \sum_{x' \in V} p_M(y, x' \mid \boldsymbol{x}_{-i}) = \sum_{x' \in V} p_M(y \mid x', \boldsymbol{x}_{-i}) p_L(x' \mid \boldsymbol{x}_{-i}). \quad (1)$$

y は予測候補のラベル，$\boldsymbol{x}_{-i}$ は i 番目のトークンをキャンセルアウトしたい入力文，V は尤度の上位 k トークン，p_M はモデルの予測確率，$p_L(x' \mid \boldsymbol{x}_{-i})$ は，$\boldsymbol{x}$ の i 番目を MASK トークンで置き換えて BERT[7] 等の言語モデルに入力した時の x' の尤度である．

1） 以降では，入力文はトークン列に分割され，それぞれ対応する token embedding に変換されることを仮定する．また，MASK トークンを用いた MLM によって事前学習され，注意機構を持つようなモデルを対象にする．[7].

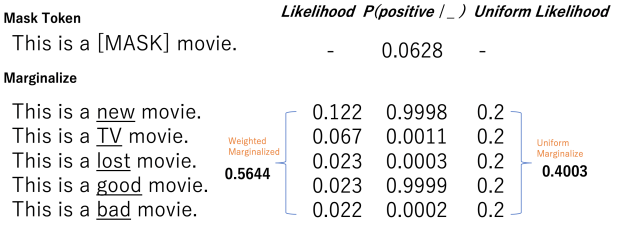

図 2 Weighted Marginalize と Uniform Marginalize の違い．“This is a good movie” の “good” をキャンセルアウトする場合．まず [MASK] で置き換えて，MLM に入力し候補を得る．Weighted では，各候補の尤度とモデルの出力で周辺化し，Uniform では一様分布として周辺化する．

表 1 BIOS の例．下線部は性別を表す代名詞と人物名．

経歴 / 性別 / 職業
Peter also has substantial experience representing clients in government investigations, including criminal and regulatory investigations, and internal investigations conducted on behalf of clients. / **male** / **attorney**
Dianne is registered with the Psychology Board of Australia, the Clinical College of the Australian Psychological Society, and she is a Medicare Provider . / **female** / **psychologist**

2.2 提案手法: Uniform Marginalize

Weighted Marginalize では，キャンセルアウト対象のトークンの MLM による予測確率が 1 に近い時，新しい入力が元の入力と意味的に変わらない可能性がある。そこで，式 1 の $p_L(x' \mid \boldsymbol{x}_{-i})$ を，$\frac{1}{N}$ （N は周辺化に使うトークン数） で置き換えることで，候補トークンを一様に扱う周辺化を提案する （**Uniform Marginalize**） ．図 2 に両者の違いを示す．Weighted Marginalize では尤度の高いものほど重視されて，Uniform Marginalize では一様に扱われる．

3 情報の復元可能性

本節では，キャンセルアウトした情報をモデルがどれくらい復元できてしまうかを尤度を用いて評価し各手法を比較する．実験では，Stanford Sentiment Treebank v2 （SST2） [8] と BIOS [6] を用いる．SST2 は，映画のレビューに対して，positive か negative かのラベルを予測するタスクである．BIOS は，個人の経歴に関する文が入力として与えられ，27 個の職業から 1 つ予測するタスクである．また，表 1 の例のように，経歴内で直接的に性別を表す代名詞と人物名の部分もアノテートされている．

3.1 元のトークンの予測確率

SST2 と BIOS のテストデータそれぞれから，ランダムに 9,000, 5,000 トークンずつ選んで，それぞれの

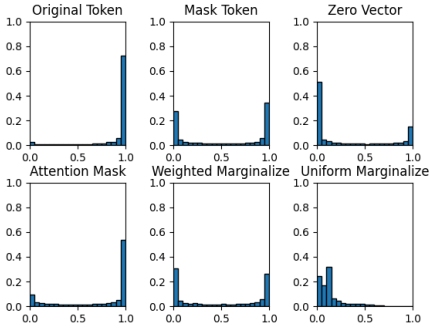

図 3 SST2 におけるキャンセルアウトしたトークンの尤度の分布．横軸が尤度，縦軸が相対度数．

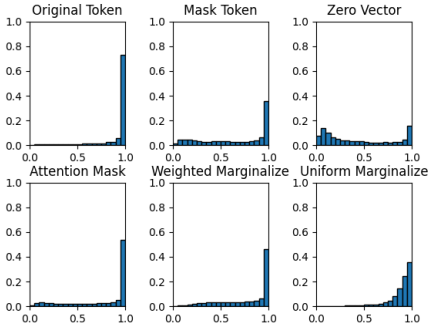

図 4 SST2 におけるキャンセルアウト箇所の最大尤度の分布．横軸が尤度，縦軸が相対度数．

手法でキャンセルアウトし，MLM に入力して元のトークンの尤度を測定する．Weighted Marginalize と Uniform Marginalize に関しては，式 1 の p_M を MLM の元のトークンの予測確率として周辺化（計算コストの都合上，周辺化候補のトークン数は 10）を行った．Original Token は，元の入力をそのまま入れたときのベースラインである．

SST2 に関しての結果が図 3 である．BIOS に対しての結果は付録 A の図 5 である．同様の傾向が観察されたため，以下の考察は SST2，BIOS に共通である．まず Original Token は，入れたトークン自身の尤度の値の分布は高い値に偏っている．Attention Mask も同様の傾向である．Mask Token, Weighted Marginalize は元のトークンを完全に復元できる場合と全くできない場合で左右に割れた．二つの分布の形が似ているのは以下の理由だと考えられる．Mask Token において高い尤度で元のトークンが予測できた例に対しては，Weighted Marginalize においても，周辺化時に元のトークンが高い尤度を持つ．さらに，その時の式 1 の p_M は，MLM に元のトークンをそのまま入れたときの自身の確率であるため，結局元のトークンの確率が高くなるためである．つまり，Mask Token で元のトークンが予測しやすい状況においては（ヒストグラムの右端），Mask Token と Weighted Marginalize はほとんど同じ処理をすることになる．対して，Uniform Marginalize, Zero Vector の順で，元のトークンは予測しづらくなっている．このことから，2 節で述べた狙い通り，Weighted Marginalize より周辺化候補を多様化させた Uniform Marginalize の方が，元のトークンをモデルが復元できてしまう（キャンセルアウトが機能しない）ことを防ぐことができることがわかった．

3.2 キャンセルアウト箇所の最大尤度の分布

キャンセルアウトされた箇所に，モデルがどれくらい確信を持って推論しているのかを比較するために，3.1 節と同様にトークンをそれぞれの手法でキャンセルアウトして，MLM に入力した時の一番尤度の高いトークンの尤度を測定する．SST2 の結果が図 4 で BIOS の結果が付録 A の図 6 である．Weighted Marginalize と Uniform Marginalize に関しては，式 1 の $p_M(y \mid x', \boldsymbol{x}_{-i})$ を，x' で置き換えた時のその箇所の一番高い尤度とした[2)]．結果としては，Zero Vector 以外は，かなり高い尤度に偏っており，Mask Token, Attention Mask はキャンセルアウトしても，そこには何らかのトークンであるとモデルが確信を持って計算していることになる．その高い尤度を持つトークンが元のトークン自身や，類義語である場合にはキャンセルアウトが機能しない．Marginalize の二手法に関しては，候補トークンで周辺化した結果高い尤度に偏っている．つまり一回のキャンセルアウトを考えると，どの候補に対しても最も高い尤度が 1 に近く，さらに 3.1 節の Original Token の結果を踏まえると，置き換えた候補のトークン自身の尤度がそれぞれ一番高くなっていることが多いと考えられる．

3.1 節と 3.2 節の結果をまとめると，Uniform Marginalize は，各置き換え後の文に対しては置き換えたトークンの情報を強く持たせることができ（モデルが置き換えられたトークンに高い確信を持つ），置き換えの多様性によって元のトークンが容易に予測できてしまうことを防ぐといえる．

2) 図 2 の例であれば，“This is a new movie.” を MLM に入力した時の new の部分で最も高い尤度，他の候補トークンでも同様に最も高い尤度を計算して，平均を計算する（周辺化する）．

4 バイアスのキャンセルアウト

BIOS は web 上で経歴データを集めており，職業ごとの男女のデータ数の偏りが大きいためモデルがバイアスを学習してしまうことが指摘されている [6]. [3] そこで，BERT を BIOS に fine-tuning する時に，各データで，直接性別を表す代名詞や人物名をキャンセルアウトした状態で学習させた時に，どれくらい学習後のモデルがジェンダーバイアスを持っているかを比較する．比較方法は，BIOS を提案している論文 [6] に従って，テストデータにおいて各職業ごとに，女性のデータと男性のデータでの真陽性率の差分（式 2 の $Gap_{f,y}$）と，訓練データにおける女性率（female rate）との相関をみる．この相関が高いほど訓練データにおける偏りを学習していることになる．

$$\begin{aligned} TPR_{g,y} &= P(Y^{'} = y \mid G = g, Y = y) \\ Gap_{f,y} &= TPR_{f,y} - TPR_{m,y} \end{aligned} \tag{2}$$

Y'，Y は職業で Y' が予測ラベル，Y が正解ラベル，G はジェンダーで男（m）か女（f）かである．

4.1 学習時の Uniform Marginalize

代名詞と人物名の Uniform Mariginalize は次のように行う．まず全ての代名詞と名前を MASK トークンに置き換える．この状態で MLM に入力して 1 個目の MASK に対応する部分の尤度の高い二つのトークン（A, B）を得る．次に，最初の MASK トークンを A で置き換えた状態で MLM に入力して 2 個目の MASK に対応する一番尤度の高いトークンを得てそれで置き換える．以降全ての MASK トークンが埋まるまで繰り返す．さらに，1 個目の MASK トークンを B で置き換えたものに対しても同様に最後まで埋める．学習時は，こうして得た二つの置き換え後の文（s1, s2）をモデルに入れ，それぞれに対し各ラベルの予測確率を出し，平均したものを最終的な予測確率として勾配の計算に使う．置き換えの例は付録 C に示す．

4.2 結果

本実験では，Original Token（元の学習データのまま訓練）, Zero Vector, Mask Token, Attention Mask, Uniform Marginalize を比較対象とした．ここでは，3 節の実験で Mask Token と作用が似ていること，Uniform Marginalize の方が置き換えに多様性があることから，Weighted Marginalize は含めていない．

表 2 手法ごとの相関係数，p 値，テストデータの正解率．

	相関係数	p 値	正解率
Original Token	0.838	5.06×10^{-8}	0.861
Zero Vector	0.804	4.29×10^{-7}	0.862
Mask token	0.701	4.63×10^{-5}	0.837
Attention Mask	0.796	6.91×10^{-7}	0.863
Uniform Marginalize	0.517	5.80×10^{-3}	0.805

各手法について，テストデータにおける各職業の女性率と男女の真陽性率の差（式 2）の相関係数の値を比較した結果を表 2 に示す。まず元の学習データのまま訓練した Original Token は，相関係数が 0.838 となり強くバイアスを学習してしまっている．[4] Zero Vector, Attention Mask に関しては相関係数の値が Original Token の場合と比べてほとんど変化せず，キャンセルアウトが機能していない．Uniform Marginalize, Mask Token の順に相関係数が改善できており，Uniform Marginalize が最もキャンセルアウトの効果が高いことがわかる．両者において，テストデータでの正解率が顕著に下がっているのは，BIOS を解くのに性別情報が有用であり，これが使えなくなると予測難易度が上がるからだと考えられる．今後の課題として，性別情報がないと予測が難しいタスクを新たに構築することで，詳細な分析が可能になることが期待される．

Uniform Marginalize においては，4.1 節における置き換えを具体的に見てみると（付録 C 節），s1，s2 で片方が男性を表す名詞，もう片方が女性を表す名詞になっていることが多く，それによって予測時に性別を使わないように学習できたと考えられる．しかしそれでも相関が出てしまう理由としては，BIOS を提案している研究 [6] で指摘されているように，代名詞や名前以外にも “mother” や “husband” 等ジェンダーを表す単語が存在することが挙げられる．

5 おわりに

本研究では，直接キャンセルアウト手法を比較する方法を提案するとともに，置き換え候補の多様化を狙った Uniform Marginalize がキャンセルアウトに有効であることを示した．今後は，品詞ごとに置き換え方を適切に変えたり，複数箇所同時にキャンセルアウトすることの影響を調べたりして，より正確なキャンセルアウト手法を追求したい．

3） 例えば，訓練データで nurse の女性率は 0.91，rapper は 0.10.

4） 各職業について女性率と男女の真陽性率の差をプロットしたものが付録 B 節の図 7 である．

謝辞

本研究は JSPS 科研費 JP21H03502 および JST さきがけ JPMJPR20C4 の支援を受けたものです.

参考文献

[1] Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. Interpretation of NLP models through input marginalization. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 3154–3167, Online, November 2020. Association for Computational Linguistics.

[2] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[4] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics.

[5] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In **Proceedings of the Advances in Neural Information Processing Systems**, Vol. 35, 2021.

[6] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In **Proceedings of the Conference on Fairness, Accountability, and Transparency**, pp. 120–128, 2019.

[7] 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.

[8] 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.

A　BIOS における復元可能性

以下の図 5 が 3.1 節の実験の BIOS に対する結果である．考察は,3.1 節を参照.

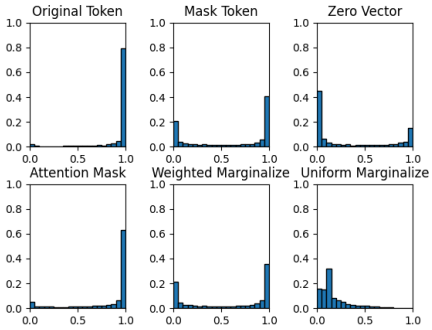

図 5　BIOS におけるキャンセルアウトしたトークンの尤度の分布．横軸が尤度，縦軸が相対度数.

以下の図 6 が 3.2 節の実験の BIOS に対する結果である．考察は,3.2 節を参照.

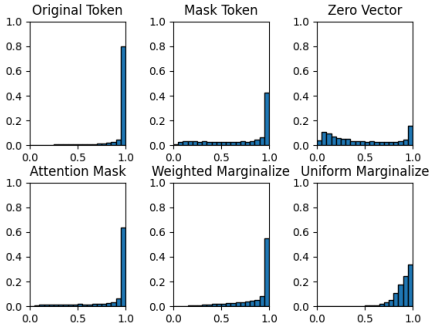

図 6　BIOS におけるキャンセルアウト箇所の最大尤度の分布．横軸が尤度，縦軸が相対度数.

B　BIOS のジェンダーバイアス

図 7 が 4.1 節で，元の学習データで BERT を訓練したときの，テストデータにおける各職業の女性率と男女間の真陽性率の差の関係である．相関係数は 0.838 であり，強くバイアスがかかっている.

C　Uniform Marginalize の置き換え

4.1 節の Uniform Marginalize における置き換えの具体例を表 3 に示す．元の “Dianne” が “She” と “He”，元の “she” が “she” と “he” に置き換えられた．大部分の例に対し，このように男性を表す名詞で置き換えたものと，女性を表す名詞で置き換えたものが一つずつ作られた.

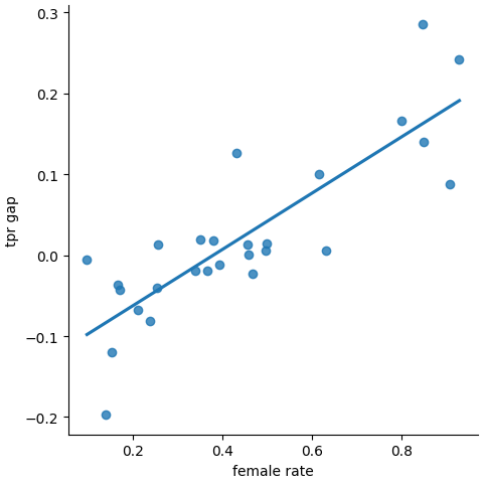

図 7　Original Token における女性率と性別間の真陽性率の差の関係

表 3　Uniform Marginalize における置き換えの例．元の “Dianne” が “She” と “He”，元の “she” が “she” と “he” に置き換えられた.

経歴
Dianne is registered with the Psychology Board of Australia, the Clinical College of the Australian Psychological Society, and she is a Medicare Provider .
She is registered with the Psychology Board of Australia, the Clinical College of the Australian Psychological Society, and she is a Medicare Provider .
He is registered with the Psychology Board of Australia, the Clinical College of the Australian Psychological Society, and he is a Medicare Provider .