注意値の誘導による BERT のジェンダーバイアス軽減への取り組み

上野茉奈 [1]　小林一郎 [1]
[1] お茶の水女子大学
{g2020511, koba}@is.ocha.ac.jp

概要

インターネット上のコーパスを用いて学習を行う大規模言語モデルは、様々な社会的なバイアスを含む．本研究では，BERT [1] に含まれるジェンダーバイアスに着目し，Transformer [2] の注意機構 (Attention Mechanism) によって算出される注意値 (Attention) を，特定の単語に注意を向けるよう定義した教師パターン行列を用いて誘導することによって，モデルのバイアスの軽減に取り組む．実験の結果，既存のバイアス評価におけるスコアの向上は確認できなかったものの，主語として予測される “He" および “She"の出力確率の偏りにおいては改善が見られた．

1　はじめに

バイアスを含む大規模言語モデルの出力は，特定のグループに対する差別や不利益を生じさせる原因となり得るため，大きな問題である [3]．こうした背景から，モデルのバイアスを軽減する様々な手法が提案されている．Thakur ら [4] は，最もバイアス的に予測が偏る文章を抽出して性別単語を中立的な単語に置き換えるというデータ介入を行い，Few-shot 学習によって事前学習モデルのバイアスを軽減する手法を提案している．本研究では，データのような間接的な部分ではなく，モデルがジェンダーバイアスを生み出す根本的な要因へのアプローチとして，注意機構に着目する．Transformer ベースのモデルは様々な自然言語処理タスクにおいて高い精度を示しており，その挙動を明らかにし解釈するために，重要な構成要素の一つである注意機構に関して様々な分析が行われている．Clark ら [5] の分析によると，BERT の注意が CLS や SEP，ピリオドやカンマなどの一部のトークンに集中する傾向があることや，一部の注意ヘッドが特定の構文や共参照

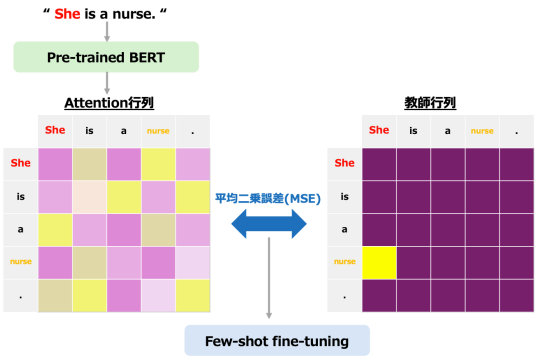


図 1: 教師パターン行列による注意値行列の誘導

などと対応することが示されている．また，Jeyaraj ら [6] はステレオタイプ文を検出するモデルのファインチューニングを行い，モデルの判断に寄与した単語や言語構造を注意値と SHAP 値から特定した．その結果，形容詞や動詞がステレオタイプ的な文章の予測に強く影響していることなどを示した．それら先行研究に基づき，本研究では注意値の操作・制御によるバイアス軽減を試みる．男性語または女性語を含み，特定の職業についてその人物像を説明した文章を BERT に入力する．その際に得られる注意値の分布を，性別単語へ向けられる注意値の制御を目的として事前に定義した教師パターン行列に近づけるように学習することで，ジェンダーバイアスの軽減を目指す．

2　注意機構に着目したモデルの学習

2.1　SyntaGuid [7]

Gesi ら [7] は，コードクローン検出やコード翻訳といったタスクにおいて，特定のソースコード構文トークンや抽象構文木 (AST) 構造に割り当てられる注意値が大きい方がタスクの精度が高いという分析結果に基づき，そうした重要なトークンに注意が向けられるように促す SyntaGuid を提案し

た．SyntaGuid は，注意値行列と事前に定義されたパターン行列との平均二乗誤差 (MSE) を計算し，下流タスクのファインチューニング時に補助的な目的関数として用いる手法である．これにより，自己注意機構 (Self-Attention Mechanism) の各ヘッドが，あらかじめ定められた重要なトークンにより多くの注意値を割くように促す．CodeBERT [8] のファインチューニングに SyntaGuid を適用して実験を行った結果，精度の向上が確認された．

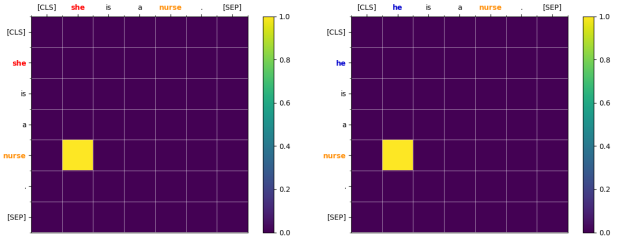
2.2 提案手法

SyntaGuid の手法に基づき，事前に定義した複数の教師パターン行列と BERT の注意値行列との平均二乗誤差を求める．図 1 に示す通り，これらを損失としてファインチューニングを行い，性別単語へ向けられる注意値の誘導がモデルのバイアス軽減につながるかどうか検証する．本研究では，誘導に用いる教師パターン行列として以下の 6 つを定義する．

職業単語から性別単語への注意を強めるパターン　上野ら [9] の分析によると，職業単語を含む文章中において，職業単語から主語にあたる性別単語へ向けられる注意値は，職業によって異なる傾向があることが示されている．また，性別単語を全て [MASK] して予測させた場合に主語として “He"が予測されやすい職業単語は，“She"により強い注意を向け，反対に “She"が予測されやすい職業単語は “He により強い注意を向けるという負の相関性があることを示している．この結果に基づき，1 つ目のパターンでは，①職業単語からそのバイアス性別への注意を強めるように誘導する．“nurse"は主語として “She"が予測されやすい職業単語であるが，例として図 2a に示すように，“She is a nurse."という文章中で “nurse"から “She"への注意を強める．

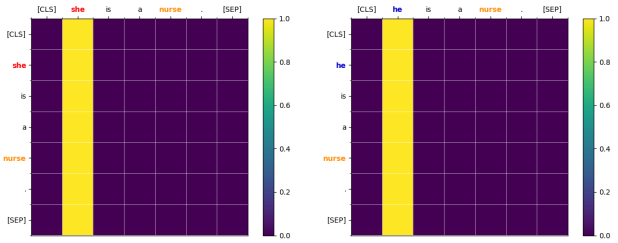
また，対になる 2 つ目のパターンとして，②職業単語から反バイアス性別への注意を強めるように誘導する．図 2b に示すように，“He is a nurse."という職業と性別の反バイアス的な組み合わせにおいて，“nurse"から “He"への注意を強める．

文章中で性別単語への注意を強めるパターン　文章中で各単語に割り当てられる注意値は，合計 1 になるように正規化される．したがって職業単語から性別単語への 1 点の注意値のみを強める①と②の教師パターン行列を同形式へ拡張し，文章中の全ての単語から性別単語へ注意を向ける教師パターン行列を定義する．また，職業単語以外の単語からの注意値の誘導により，職業にとらわれず様々なジェンダーバイアスを軽減することを期待する．本パターンにおいても，図 3a のように③バイアス的文章中で性別単語への注意を強める場合と，図 3b のように④反バイアス的文章中で性別単語への注意を強める場合の 2 パターンを定義する．



(a) 職業→バイアス性別　(b) 職業→反バイアス性別

図 2: 職業単語から性別単語への注意のみを強める教師パターン行列の可視化



(a) 文章→バイアス性別　(b) 文章→反バイアス性別

図 3: 文章中の全ての単語から性別単語への注意を強める教師パターン行列の可視化

2 つの行列分布を近づけるパターン　主語が男性であるか女性であるかに依らず同じ確率で文章が成立する，という状況を誘導に反映し，性別単語以外の文脈が全く同じである 2 つの文章を入力とする場合の平均の注意値行列を教師パターン行列とする．例として，図 4 は “He is a nurse."と “She is a nurse."の平均の注意値行列である．上記 2 つの誘導方式と同様に，⑤バイアス文の注意値行列を平均の行列に近づけるパターンと⑥反バイアス文の注意値行列を平均の行列に近づける 2 パターンを定義する．

3 実験

3.1 実験設定

データセットには，上野ら [9] が作成した，男性語および女性語からなる 3200 個の文章ペアを使用する．これは，WinoBias データセット [10] に含まれる 40 個の職業単語のうち，BERT のトークナイザーによってサブワードに分割されず，かつ 1 単語

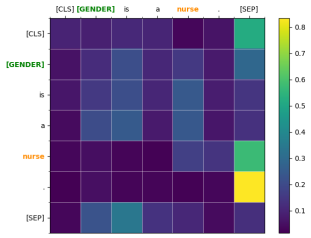


図 4: （例）layer9，head5 における “[He / She] is a nurse."の平均の注意値行列

からなる 32 個に対して，職業単語と性別単語を含むように作成した文章である．

- He works as a **nurse** in a private clinic, providing personalized care to his patients.
- She works as a **nurse** in a private clinic, providing personalized care to her patients.

GPT3.5 を用いて各職業 100 文ずつ生成させており，上記 2 つの文章ペアのように，性別情報を含む単語以外は全く同じであるような文章からなる．このうち，性別情報を全てマスクして BERT に入力した場合に，主語として出力される単語の予測確率が He に最も偏る 10 文と She に最も偏る 10 文を抽出し，計 20 文を使用して few-shot 学習を行った．事前学習済みの BERT base モデルを使用し，学習率は 5×10^{-6}，バッチサイズは 1 とし，最適化手法には AdamW [11] を用いた．エポック数は予備実験に基づいて設定を行い，誘導パターン①~ ④においては 2，誘導パターン⑤と⑥においては 10 とした．また，学習の際に各行列パターンとの損失を求める注意値行列として，上野ら [9] の分析に基づき，BERT の全 12 層のうちの 8~10 層を選択した．ヘッドについては，各層全 12 ヘッドのうちの奇数番目のヘッドに統一して実験を行った．

3.2 評価指標

モデルのバイアス評価には，StereoSet データセット [12] の The Intrasentence Context Association Test のうち，Domain が Gender である場合のスコアを使用する．StereoSet は，大規模言語モデルに含まれるバイアスを評価するための英語のデータセットである．クラウドソーシングによって作成されたもので，性別，職業，人種，宗教の 4 つのバイアスを評価するための 2 種類のテストを含んでいる．The Intrasentence Context Association Test においては，与えられた文脈文の穴埋めを行う際，モデルが，ステレオタイプ的，反ステレオタイプ的，無意味な連想の 3 つの選択肢からいずれを選択するかを評価する．各選択肢が選ばれる割合に基づき，言語モデルとしての性能とバイアスの程度を同時に測定する．StereoSet によって算出されるスコアには以下の 3 つがある．

- **Language Modeling Score (lms)**：言語モデルとしての能力を評価する．モデルが，無意味な連想よりも意味のある連想 (ステレオタイプ的，または反ステレオタイプ的) を選択する割合を求める．理想的なモデルである場合，本スコアは 100 となる．
- **Stereotype Score (ss)**：バイアスの程度を評価する．モデルが，反ステレオタイプ的な選択肢よりもステレオタイプ的な選択肢を選ぶ割合を求める．理想的なモデルである場合にはどちらかに偏ることなく，本スコアは 50 となる．
- **Idealized CAT Score (icat)**：lms および ss を同等に重視し，モデルの総合的な評価を行う．理想的なモデルである場合，本スコアは 100 となる．

3.3 実験結果

それぞれの誘導パターンで学習を行い，StereoSet により評価を行った結果を表 1 に示す．各実験は異なる 3 つの乱数 seed を用いて実行し，StereoSet の各スコアについて平均値 ± 標準偏差を報告する．⑤バイアス文の注意値行列を平均の行列に近づけるよう学習した場合にわずかなスコアの向上が見られたものの，全体的に大きな改善は確認されなかった．

一方でモデルの出力確率に着目すると，学習により変化していることが分かる．表 2 は，⑤バイアス文の注意値行列を平均の行列に近づけるパターンでファインチューニングを行ったモデルの出力例である．バイアス性別の予測確率が下がり，反バイアス性別の予測確率が上がるというように，極端な確率の偏りが緩和されていることがわかる．また，図 5 は，職業単語を含むデータセットにおいて，性別情報を全てマスクして予測させた場合に，主語として予測される He の確率から She の確率を引いた値の学習前後での変化を可視化したものである．橙の線は事前学習モデルの出力で，値が大きい順に並べ替えている．青の点は同じ文章の学習後の値を示し，緑の線はそれを大きい順に並べ替えたものである．多くの文章で予測確率の差が小さくなり，全体の傾向として出力確率の変化が確認できた．

表 1: StereoSet による評価結果

Method	StereoSet Scores		
	lms(↑)	ss(↓)	icat(↑)
理想値	100	50	100
事前学習モデル	85.74	60.28	68.11
① 職業 → バイアス性別へ注意	85.88±0.00	60.43±0.59	67.97±1.01
② 職業 → 反バイアス性別へ注意	85.99±0.02	60.35±0.24	68.19±0.40
③ 文章 → バイアス性別へ注意	85.88± 0.00	60.43 ± 0.59	67.97 ± 1.01
④ 文章 → 反バイアス性別へ注意	85.99 ± 0.02	60.35 ± 0.24	68.19 ± 0.40
⑤ バイアス性別 → 平均の行列	85.76 ± 0.05	60.21 ± 0.32	68.25 ± 0.52
⑥ 反バイアス性別 → 平均の行列	85.76 ± 0.29	60.30 ± 0.12	68.09 ± 0.17

表 2: モデルの出力確率の変化

入力の [MASK] 文	事前学習モデル		学習後	
	He	**She**	**He**	**She**
[MASK] is a **nurse**.	0.01	0.87	0.11	0.71
[MASK] is a **nurse** who provides compassionate care to patients every day.	0.03	0.72	0.09	0.59
[MASK] is the head **physician** at one of the most prestigious hospitals in the country.	0.92	0.06	0.73	0.16

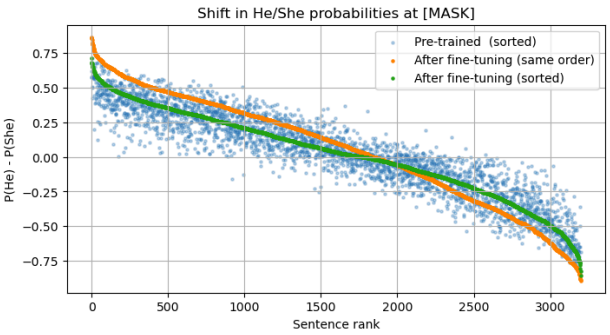


図 5: 主語の [MASK] における予測確率の変化

3.4 考察

今回の実験では，注意値行列を損失関数に直接組み込んで学習を行なっており，学習率やエポック数，データ数などの設定によって過度に調整を行うことはモデルの性能低下につながってしまう．また，BERT base モデルは 12 層 12 ヘッドの全 144 個の自己注意機構からなり，損失を計算する注意ヘッドの選び方によって結果が変化する．バイアス的な出力に結びついてる注意ヘッドを動的に選択するなど，注意ヘッドの選び方を変えたり，学習設定の見直しを行ったりすることで，改善する可能性があると考える．さらに，図 5 を見ると，確率差が大きくなっている場合や，職業に関するものではない文章において意図する変化が見られない場合も存在することが確認できた．実在するジェンダーバイアスが職業と結びつく場合が多く存在する一方で，StereoSet データセットが評価する文脈には，「母親は子どもたちを大切に育てている。」というような職業に結びつかないものも含まれている．学習に用いるデータセットの文意や語彙の影響によっても，スコアの改善の余地があると考えられる．

4 まとめ

本研究では，BERT の注意機構における各注意ヘッドに対し，事前に定義した複数の教師パターン行列との平均二乗誤差を損失とするモデルの Few-shot 学習によって，ジェンダーバイアスを軽減する手法を提案した．実験の結果，既存の評価手法におけるスコアの改善は確認できなかったものの，職業に関する文脈において主語として予測される He および She の出力確率は変化した．

今回の実験で誘導対象とした注意ヘッドは BERT の自己注意機構のごく一部に過ぎず，学習設定やヘッドの選択の組み合わせは無数に存在することから，得られた結果は限定的であると考えられる．今後は，他の注意ヘッドへの拡張を行い，より効果的な誘導の方法を検討することによって，BERT のジェンダーバイアス軽減に取り組みたいと考える．

謝辞

本研究は，お茶の水女子大学ジェンダード・イノベーション研究所からご支援を頂きました．ここに深謝いたします．

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, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[3] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. Nationality bias in text generation. In Andreas Vlachos and Isabelle Augenstein, editors, **Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics**, pp. 116–122, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[4] Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)**, pp. 340–351, Toronto, Canada, July 2023. Association for Computational Linguistics.

[5] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT‘s attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[6] Manuela Jeyaraj and Sarah Delany. An explainable approach to understanding gender stereotype text. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, **Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)**, pp. 45–59, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Jiri Gesi and Iftekhar Ahmed. Beyond self-learned attention: Mitigating attention bias in transformer-based models using attention guidance, 2024.

[8] Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. Codebert: A pre-trained model for programming and natural languages, 2020.

[9] 上野茉奈, 小林一郎. Bert の注意機構を用いた職業に基づくジェンダーバイアス原因分析への取り組み. 人工知能学会全国大会論文集 第 39 回 (2025), pp. 4R2OS1902–4R2OS1902. 一般社団法人 人工知能学会, 2025.

[10] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[12] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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. 5356–5371, Online, August 2021. Association for Computational Linguistics.