BERT の注意機構に着目したジェンダーバイアスの一考察

上野 茉奈　(指導教員：小林 一郎)

1　はじめに

大規模言語モデルの学習に用いられているインターネット上のコーパスには様々な社会的なバイアスが含まれており，これを用いて学習を行った言語モデルも同様なバイアスを含んでいる．言語モデルにバイアスが含まれていることは，その出力によって特定のグループへの差別や不利益が生じる要因となり得るため，大きな問題である [1]．本研究では，BERT モデルが捉えるジェンダーバイアスの一指標として，性別が表現される単語への注意機構 (Attention Mechanism) によって付与される値に着目する．データセットから，男性語を用いて記述された文章と女性語を用いて記述された文章をそれぞれ作成し，性別の単語への Attention の値を比較することで，文章内の文脈情報が与えるジェンダーバイアスについて考察する．

2　研究概要

図 1に研究の全体像を示す．

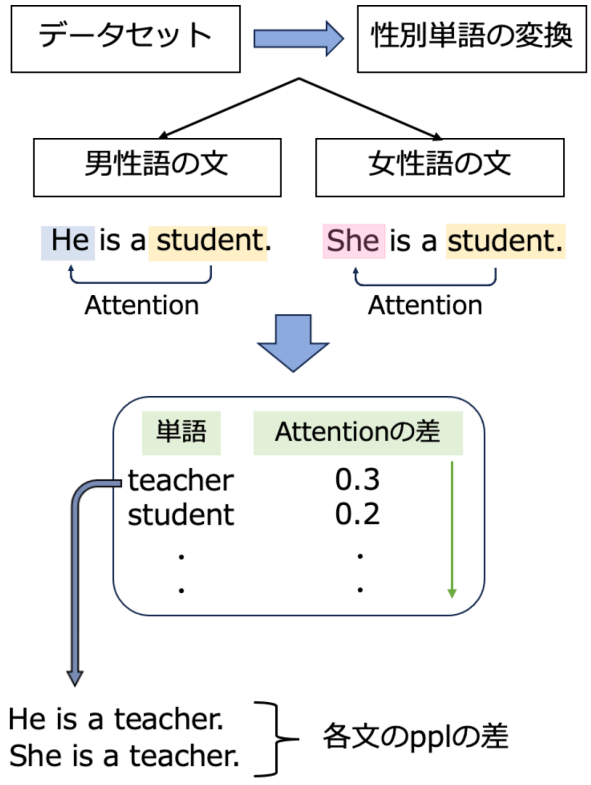

図 1: 研究の全体像

性別の単語に対する注意 (Attention) の値の差が大きい単語を含む文章に，ジェンダーバイアスが見られるかを検証することを目的とする．初めに，男性語を含む文章と女性語を含む文章の 2 つのデータセットで，各文においてそれぞれ性別の単語のみが異なるものを作成する．次に，BERT モデルの自己注意 (Self-Attention) を用いて，各文の性別の単語に当たっている Attention の値の総和を求め単語ごとに平均する．最後に，2 つのデータセットでそれぞれ求めた各単語の Attention の値の差をとり，差が大きい単語から順にその単語を含む文章のペア（性別の単語のみが異なる）を取り出す．ジェンダーバイアスとして，文章のペアに対し，文の自然さを図る指標であるパープレキシティの値の差を求めることにより，性別を表す単語への注意の差が大きい文は言語モデルの観点から自然であるのかについて検証を行う．

3　データセットの作成

3.1　StereoSet

StereoSet [2] は，クラウドソーシングによって作成された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．性別，職業，人種，宗教の 4 つの領域について，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．本研究で新たに作成するデータセットに使用するテキストとして，StereoSet の検証データとテストデータから，The Intersentence Context Association Test 中の全 context 文である 2,123 文を抽出した．また対象とする性別の単語として，ドメインが性別である対象の単語 40 個を，著者が男性語と女性語にそれぞれ 20 個ずつ分類した．

3.2　性別単語の変換

性別の単語のみが異なる文章を作成するために，OpenAI の API を用いて文章変換を行った．モデルを gpt-3.5-turbo-1106，プロンプトを「性別だけを反対にした文章を出力してください。」として，上記で抽出したテキスト中の男性語と女性語を含む文を男性語で表現された文は性別を変えて女性語で表現される文に，またその逆の変換も行い，文意を保存しながら性別が異なるペア文からなるデータセットを作成した．さらに，データセット中で重複している文章を取り除き，それぞれ 2,399 文を含むデータセットとした．

4　実験

4.1　実験設定

対象とする性別の単語　StereoSet から分類した単語に，作成したデータセット中の著者が性別の単語と判断できる単語を新たに加えた．また，ここでは先行研究 [3] で使用されていたデータセット[1] に含まれる男女における典型的な名前の単語を加えそれぞれの性別の単語とした．

対象とする文章　正確に性別の単語の違いによる比較を行うために，データセットにさらなる変更を加えた．文章中の ma'am という単語はトークナイザによってさらに分割されてしまったため，全て madam に変更した．また，文頭を大文字にする，文末にピリオドがないものにピリオドを加える，として全文を統一させた．最後に，意図しない変換が行われた文章や元のデータセットで綴りに誤りを含んでいた文章などを取り除

[1] https://github.com/himansh005/data_debias.git

くため，性別の単語としたもの以外の単語が異なる文章の組み合わせは取り除いた．以上により，使用したデータセットは男性語からなる文章と女性語からなる文章それぞれ 2,111 文となった．

4.2 実験課題

(1) 各単語の性別の単語への Attention 値の差

モデルには事前学習済みの BERT の base モデル[2] を使用した．また，Attention の値として最終層の値を用いた．男性語からなる文章において，男性語にあたっている各単語からの Attention の値を出現回数で平均し，その単語の男性語への Attention の値とする．女性語からなる文章においても同様の値を求め，最後に各単語ごとに対象とした性別を表す単語への Attention 値の差を計算した．

(2) パープレキシティの差

モデルには事前学習済みの GPT-2[3] を用いた．(1)の結果を用いて，全文に含まれるピリオドを除き，Attention 値の差が大きい単語から順に，重複なくその単語を含む文章のペアを取り出す．それぞれの文に対するパープレキシティを GPT-2 による言語モデルの確率によって計算しそれらの差分を求めた．その単語を含む文のペアが複数ある場合には平均した．

(3) Attention 値の差分とパープレキシティの差分との相関関係

性別の単語へあてている Attention の差が大きい単語を含むペア文は，性別に関する内容に特化した文である可能性も想定される．このことから Attention の差分とパープレキシティの差分との間に相関関係の有無を確認する．

4.3 実験結果

4.2節の実験課題に対する結果を(3)の相関関係によりまとめて示す．横軸に性別の単語へあてている Attention の差が大きい単語順に並べ，縦軸をその単語を含みかつ性別の単語のみが異なる文におけるパープレキシティの差とするグラフを図 2に示す．

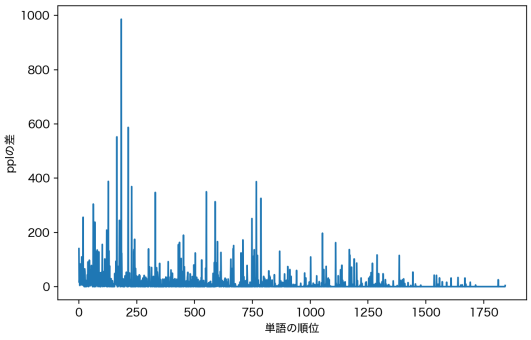

図 2: Attention の差が大きい単語とその単語を含む文のパープレ器シティの差との相関

結果としては，一部やや相関がありそうな部分も見られるもののグラフによる視覚的確認においては全体として明確な相関関係を確認することはできなかった．

4.4 考察

作成したデータセット中の文章のペアとその文におけるパープレキシティの差について，文ごとに個別に分析を行う．

以下に示すように，パープレキシティの差が大きい文のペアには直感的にバイアスがありそうだと感じるものもあり，性別の単語のみが異なることで大きな差となっている．

- The baseball is hers.
- The baseball is his.

一方で，今回は性別の単語として名前を表す単語を含めたが，名前を含む文のペアにおいてはその名前の組み合わせによってパープレキシティの差が変わってしまう．以下の文においては Bart を Bob に変更することでパープレキシティの差が大幅に小さくなった．

- Beth is a mommy.
- Bart is a daddy.

このように，パープレキシティの値は単語の頻度により影響を受けることから，性別の単語のみが異なる場合であっても，その他の要素が差に影響したことが考えられる．データセット中に含める単語に制限をかけることで，より正確な相関となる可能性がある．

また，文自体に誤りがあるものが一部含まれており，こうしたデータセット中のノイズも，相関に影響した可能性がある．

5 まとめ

本研究では，性別の単語のみが異なるような文のペアにおける BERT の注意機構 (Attention Mechanism) から得られる Attention 値の差を用い，ジェンダーバイアスを含むような文を取り出すことを目的として検証を行った．

性別の単語へあてている Attention の差が大きい単語を含むペア文のパープレキシティの差分との間に相関関係があるという仮説の下，検証を行ったが明確な相関性は確認できなかった． データセットおよびジェンダーバイアスの定義において再検討を行い，今後も慎重に相関性をみる必要があると考える．今後は，ジェンダーバイアスを生み出す原因となる部分の解析を進めるとともに，今回求めた Attention に差がみられる単語を用いた，モデルのジェンダーバイアスを減らす手法についても検討していきたいと考える．

参考文献

[1] Pranav Narayanan Venkit, et al. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 116–122, May 2023.

[2] Moin Nadeem, et al. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5356–5371, August 2021.

[3] Himanshu Thakur, et al. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 340–351, July 2023.

[2]https://huggingface.co/bert-base-uncased
[3]https://huggingface.co/openai-community/gpt2

A　性別の単語

男性語							
male	stepfather	man	groom	father	gentlemen	grandfather	himself
boy	boyfriend	brother	gentleman	sir	schoolboy	son	daddy
he	husband	him	his	dad	uncle	boys	men
niece	fiance	boyfriends	##boy	grandson	##man	brothers	sons
女性語							
female	stepmother	woman	bride	mother	ladies	grandmother	herself
girl	girlfriend	sister	lady	madam	schoolgirl	daughter	mommy
she	wife	hers	her	mom	aunt	girls	women
nephew	fiancee	girlfriends	##girl	granddaughter	##woman	sisters	daughters

以上の単語に名前の単語を加えたものとした.