可視化による単語埋め込みのジェンダーバイアスとそのデバイアス効果の分析

理学専攻　情報科学コース　2340664　杉野 有咲　(指導教員：伊藤 貴之)

1　はじめに

単語埋め込みとは，単語を連続ベクトル空間での数値に変換することで，単語間の関係を計算可能にする手法である．しかし単語埋め込みには，言語モデルを作成する事前学習に用いたデータセットに内包されたバイアスが，そのモデルに影響を与える欠点がある．モデルがバイアスを有することで，モデルの信頼性の低下や，公平性の欠如などの問題を生じることがある．デバイアス方法には，性別や宗教など，特定の基準に沿って偏りを緩和する方法や，バイアスを持たないデータを用いて再学習する方法などがある．しかし，これらはモデル性能が低下する問題点がある．デバイアスによりモデルの情報が失われ，モデルが単語の意味や文脈を正確に理解できなくなるからである．また自然言語処理におけるバイアスに関する学術研究のほとんどは，バイアス測定・緩和のための技術的アプローチや，バイアスの全容を明らかにすることに集中しており，具体的なユーザ目線が考慮されていない．我々は，ジェンダーバイアスが何を指すのか，またそれがどれほど問題であるかは，個人の価値観や自然言語処理モデルを使用する場面によって異なると考える．以上を踏まえ本研究は，ジェンダーバイアスを緩和する度合いをユーザが指定可能で，さらに特定のカテゴリの単語群ごとにデバイアスの度合いを調整可能なデバイアスを実現するための可視化手法を提案する．単語のカテゴリ分類タスクにもとづいて，デバイアスによる単語埋め込みの性能劣化度を可視化することで，ユーザや用途に適したデバイアスが施され，かつデバイアスによる性能劣化が少ない自然言語処理モデルの実現を目指した．

2　処理手順

2.1　ジェンダーバイアスのデバイアス手法

デバイアス度合いを柔軟に指定可能な手法として，既存手法の Hard Debias[1] を応用する．性別を表す単語 (e.g.「彼」「彼女」) から算出した性別成分を，性別と中立な単語ベクトルから減算することで，全ての単語からジェンダーバイアスが除去される．しかし，Hard Debias は単語ベクトルから特定方向成分を完全に除去するため，デバイアス後の単語埋め込みが本来の意味を十分に保持できず，自然言語モデルの性能が低下する危険性がある．本手法では性別を示す成分に変数 θ を導入し，デバイアス時の特定方向成分除去の割合を調整することで Hard Debias を改良した．単語埋め込みの意味的整合性を保持しつつ，ジェンダーバイアスを効果的に緩和することが可能となり，既存手法では全成分の完全除去により生じる性能低下を回避し，より実用的なデバイアス処理を実現することが期待される．本報告では，東北大学が公開している「日本語 Wikipedia エンティティベクトル」[1]を使用した．

2.2　カテゴリ分類

デバイアスした単語の意味がどのように変化するか確認するために，BERT ベースで学習した単語カテゴリ分類モデルを作成した．ベクトルを受け取ると，HNSW を用いて，入力されたベクトルがどの単語のベクトルと近似するか計算し，最も類似度が高かった単語のカテゴリを分類結果として返す．デバイアス前のカテゴリ類推結果を正解データとし，デバイアス後のベクトルを用いたカテゴリ分類結果と比較することで，デバイアスによる単語の意味変化を確認可能となる．本報告では，単語を entertainment・science・politics・sports・business の 5 種類のカテゴリに分類した．ニュースコーパスの記事の見出しをもとにカテゴリをラベリングして作成した訓練データで，モデルを単語のカテゴリ分類タスク向けにファインチューニングした．次に GPT-4o を用いて作成したテストデータでカテゴリ分類タスクを実施し，出力結果をもとに手動で再ラベリングした新たな訓練データで再学習して，最終的なモデルを作成した．

2.3　最適化問題によるデバイアス度合いの提案

本手法では，モデルの性能劣化抑止とデバイアスがトレードオフの関係にある点を考慮し，多目的最適化の観点からパレート最適解を算出した．モデル性能 (Accuracy と F1 Score) を最大化すると同時に，ジェンダーバイアスを最小化する問題設定のもと，パレートフロントを求めている．

3　実行例

3.1　デバイアス前後のカテゴリ分布の差

データセット全体を完全にデバイアスした後のカテゴリ分類の変化を確認するため，図 1 のようにデバイアス後のカテゴリ分類結果からデバイアス前の結果を差し引き，カテゴリ分布割合の差を混同行列で可視化した．図 1 の縦軸は正解ラベル，横軸は予測ラベルを示している．対角線成分に注目すると，デバイアスにより，すべてのカテゴリにおいてカテゴリ誤分類が生じていて，特に politics と science におけるカテゴリ分類の正確性が低下していることがわかる．

[1] https://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

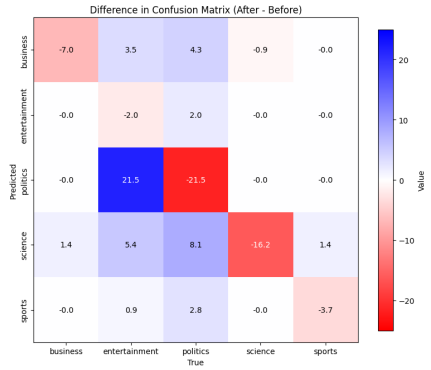

図 1: デバイアス前後のカテゴリ分布の差.

3.2 特定カテゴリに着目したデバイアス

図 1 よりデバイアスの影響が最も大きかった politics のデバイアスの度合いを軽減して，再度デバイアスを実施した．デバイアス時の性別情報の減算の度合い θ を politics カテゴリのみ 0.0 から 1.0 の範囲で変化させ，その他の 4 カテゴリには θ=1.0(完全デバイアス) を適用した結果を図 2 に示す．図 2 の横軸は politics の θ，縦軸左側は Accuracy および F1 Score，右側には politics の単語のうちバイアスが大きい単語群と「男性」を表す単語，および「女性」を表す単語とのコサイン類似度の差をプロットした．この差の値が 0 に近いほど，バイアスが大きい単語がデバイアス操作によって性別からより中立化されていることを意味する．

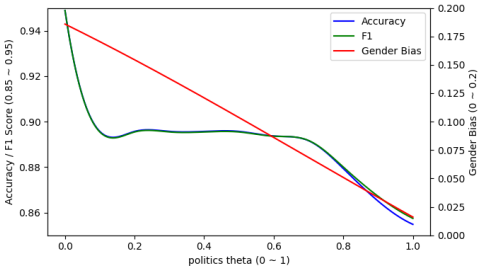

図 2: politics θ 別の分類精度とバイアス.

図 2 に示すように，モデル性能の劣化抑止とデバイアスには明確なトレードオフが存在し，両者を同時に最大化する θ を一意に定めることは困難である．しかし，パレート最適の観点からパレートフロントを算出したところ，図 2 におけるパレートフロントを構成する θ の値は 0.0, 0.6〜1.0 であることがわかった．加えて，モデル性能とバイアスを等しい重みで扱う設定のもと加重和を算出した結果，θ=0.7 のときにモデル性能の劣化とバイアスの影響が同程度となることを確認した． すなわち，デバイアスをより重視したい場合は θ=0.8〜1.0 が適しており，モデル性能の劣化をなるべく抑えたい場合は θ=0.0 もしくは 0.6 を選ぶのが望ましいと言える．

さらに，各カテゴリに対して上述の手順を繰り返し適用し，モデル性能の劣化とバイアスの影響が同程度となる各 θ を算出し，その際の「デバイアス重視」「性能劣化抑止重視」「両者を等しく重視」する結果を比較したものが表 1 である．表 1 の「両立」に対応する θ 値を用いてデバイアスを実施することで，モデルの

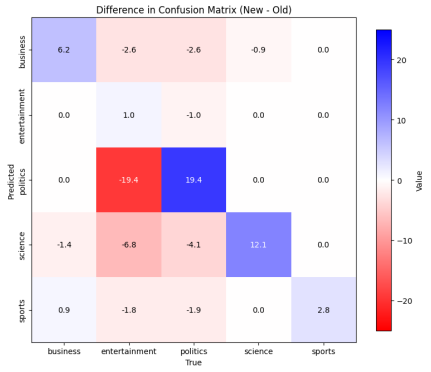

図 3: 本手法と Hard Debias のカテゴリ分布の差.

性能劣化とデバイアスの影響を同程度に抑えることができる．図 3 はモデルの性能劣化とデバイアスの影響を同程度に抑えた本手法によるデバイアスと，既存手法の Hard Debias のモデルへの影響の差の可視化結果を示しており，縦軸は正解ラベル，横軸には予測ラベルである．対角線成分が全て正の値を示していることから，本手法により全てのカテゴリにおいてカテゴリ分類精度が向上していることが確認できる．

表 1: デバイアス条件別の各カテゴリの θ.

	モデル性能重視	両立	デバイアス重視
politics	0.0, 0.6	0.7	0.8〜1.0
science	0.5〜0.7	0.8	0.9〜1.0
business	0.6	0.7	0.8〜1.0
sports	0.6	0.9	1.0
entertainment	0.7〜0.8	0.9	1.0

4 まとめ

本手法を用いて各カテゴリに対するデバイアスの度合いを，可視化結果にもとづいて調整することで，モデル性能の劣化を抑止したデバイアスを実現できた．さらに，バイアス緩和の度合いとモデル性能の劣化をデバイアスの度合いごとに可視化・比較することで，デバイアス重視・性能劣化抑止重視・両者を等しく重視するパターンのいずれに対しても柔軟に対応できることがわかった．分類結果を参照しながらスライダーでデバイアスの度合いを調整できる UI の作成を試みたが，実用的な応答時間を得るのは難しかった．今後は，単語の特徴ごとにユーザがデバイアスの度合いを自由に調整できるシステムの実現を目指すとともに，精度と速度を両立させたプログラムの開発に取り組む．

参考文献

[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama and Adam Kalai, Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4356–4364, 2016.