単語埋め込みのジェンダーバイアスの可視化

杉野　有咲†　　伊藤　貴之†

†お茶の水女子大学 理学部情報科学科 〒112-8610 東京都文京区大塚 2-1-1

E-mail:　†{g1920521, itot}@is.ocha.ac.jp

あらまし　本稿では，日本語の Wikipedia から学習した単語埋め込みから，ジェンダーバイアスの分布を可視化し，可視化結果からバイアス緩和を支援する手法を提案する．本稿では「辞書では性別を意味に含んでいないが，単語埋め込みが「男」または「女」の一方に偏っている単語」を「ジェンダーバイアスが生じている」と定義する．バイアス緩和のためのデバイアス処理がもたらすモデル全体への影響を目で確認しながらデバイアス処理のパラメータを操作すべきという観点から，本研究では可視化を導入する．モデルへの影響を抑えたデバイアスが可能で，かつ多様な言語モデルに適用可能なシステムの開発を本研究の最終目標としている．

キーワード　可視化，データ分類，ジェンダーバイアス

1. はじめに

近年の AI 技術の発展により，IoT 家電や AI による画像及び文章生成など，AI を活用したサービスが増加し，AI が我々の生活のあらゆる場面で活躍している．中でも，自然言語処理は機械翻訳や AI アシスタントなどに活用される重要な技術である．

単語埋め込みは，自然言語処理において広く利用されている．しかし，単語埋め込みには，言語モデルを作成する際の事前学習に用いたデータセットにバイアスが内包されていた場合，そのバイアスの影響が生じる欠点がある．Farkas ら[1]は，英語と性別に中立な代名詞を持つハンガリー語の職業名を，Google 翻訳を用いて比較し，Google 翻訳にジェンダーバイアスが生じていることを示した．単語埋め込みのバイアス及びデバイアスは研究途上分野である．特に日本語の単語埋め込みは先行研究が少なく，そのほとんどの研究対象が英語の単語埋め込みである．

また SDGs に「ジェンダー平等の実現」が項目に含まれているように，近年世界的にジェンダー問題への注目が集まっている．世界経済フォーラム(WEF)が発表した 2022 年の日本のジェンダーギャップ指数[1]は 0.650，順位は 146 カ国中 116 位であり，先進国の中で最低水準である．日本では特に工学分野において，ジェンダーギャップの解消への研究が遅れている[2].

本稿では，日本語の学習済み Wikipedia2Vec のデータセットから，ジェンダーバイアスを検出し，それらをデバイアスした結果を可視化することで，バイアス緩和を支援する手法について検討する．

[1]https://www3.weforum.org/docs/WEF_GGGR_2022.pdf

2. 関連研究

3 種類に分けて関連研究を紹介する．1 つ目は単語埋め込みの類推方法の研究である．Bolukbasi ら[3]は，「女王=王様-男+女」のように，単語埋め込みはベクトルの足し引きで単語の類推を表現することが可能であり，単語埋め込みはジェンダーバイアスを内包していると示した．例えば，「Computer Programmer」が男性，「Homemaker」が女性の名詞と類推できてしまう．本研究では，足し引きによる単語埋め込みの類推方法を参考にして，本来「男」または「女」から中立であるべき単語が，「王様」と「女王」のように共起している単語をジェンダーバイアスが生じていると定義する．

2 つ目はデバイアスに関する研究である．デバイアス手法は多数存在している．Bolukbasi ら[3]は部分空間の射影を用いた Hard Debias を示した．Hard Debias は，竹下ら[4]や Wang ら[5]や Liang ら[6]など，広く応用されているデバイアス方法である．Hard Debias は主に英語の単語埋め込みに対して使用されているが，竹下ら[4]によって日本語の単語埋込にも応用可能であると示されている．Lu ら[7]は，学習済みデータセットに性別を反転させたデータを追加し，CDA (Counterfactual Data Augmentation)と呼ばれる，文法的な矛盾が生じないジェンダーバイアスの緩和方法を示した．しかし，以上の単語埋め込みのデバイアスにはモデルの性能が劣化する問題点がある．今現在もモデルの性能劣化とデバイアスの関係性は研究対象になっている[8]．Sun ら[9]は，デバイアス手法は，再学習でデバイアスする方法と，既存のモデルにデバイアスプログラムを適応させて出力を調整しデバイアスする推論法の 2 種類に分類できると示した．現段階では，本研究は推論法で議論を進める．これらの研究では，デ

バイアスによるモデルへの影響が考慮されていない．

3 つ目はデバイアスによるモデルへの影響を評価する研究である．小林ら[8]は，単語埋め込みの「表現の曖昧さの増大」によりモデルの性能が劣化する可能性を示している．デバイアス前後の単語埋め込みのL^2-ノルム及びそれを用いて作成した分類モデルの SHAP 値の差を比較した．この研究は，Bolukbasi ら[3]の既存のデバイアス方法を用いているが，デバイアス方法の評価および改善は実施されていない．本研究では，Bolukbasi ら[3]のデバイアス方法のパラメータを操作することで，モデルへの影響が少なくなるようにデバイアスを改善することを目標とする．

以上の研究はいずれも単語埋め込みモデルを一括で扱っており，特定の単語に注目した可視化方法やデバイアス方法は見当たらない．また，全て英語の単語埋め込みが対象である．

3. 提案手法

3.1. 使用したデータ

本研究では，デバイアス対象として Wikipedia2Vec の日本語版[2]を使用した．まず MeCab を用いて固有名詞のみを抽出する．次に Unicode で CJK またはカタカナまたはひらがな以外の名詞を除外した．英単語や記号の除外が目的である．

3.2. バイアス検出

バイアス検出のために，まずデータの単語全てに対して「-男+女」及び「+男-女」を計算し，コサイン類似度上位 10 位をそれぞれ出力する．出力結果を比較して，単語が重複かつコサイン類似度が一定値離れているペアを，バイアスが生じているとみなして抽出する．

例として「本名」という単語で説明する．「本名-男+女」「本名+男-女」の出力結果には「旧姓」が共通している．「本名-男+女」と「旧姓」のコサイン類似度は 0.7679801，「本名+男-女」と「旧姓」のコサイン類似度は 0.6627757 である．この場合「本名」が「男」寄り，「旧姓」が「女」寄りにバイアスが生じているので，「本名」と「旧姓」を抽出する．また，「兄」や「女優」など性別を意味に含む単語はバイアス検出対象から除外した．除外する単語リストの作成は，Lu ら[7]が公開していた「Gender Pairs」を参考にした．

3.3. バイアスクラスタリング

バイアスが生じている単語群を，k-means 法でクラスタリングする．クラスタ別に最適なデバイアスを適応して，デバイアス後のモデルの性能の劣化を抑えるためである．最適なクラスタ数は，エルボー法で検出した．

3.4. デバイアス

Bolukbasi ら[3]の Hard Debias を応用する．「男」または「女」の一方に偏った単語が，「男」「女」両方から等距離の位置に近づくほど，その単語のジェンダーバイアスが緩和される．

処理手順は以下の通りである．

1. 性別を表す単語群をもとに，主成分分析で性別の基準となる軸を計算する．今回は{男，女，男性，女性，彼，彼女，父，母，息子，娘，少年，少女}を用いて軸を計算した．これらの単語は，Bolukbasi ら[3]の性別を含む単語群を日本語に訳したものである．
2. バイアスを除去する単語から，性別の基準となる軸成分を減算する．
3. 性別を含む単語群を，軸から等距離になるように移動する．

減算する軸成分の割合を調整することで，バイアスの緩和度を操作できる．

3.5. デバイアス前後のモデルの評価と Tensor Board

デバイアス前後の単語ベクトルの大きさの差で評価する．デバイアス後の単語ベクトルの大きさが小さい場合，デバイアスによって情報を損失したとみなすことができる．本研究では，デバイアス前後の単語同士の関係の変化の確認を目的として，Tensor Board を用いてモデル全体を可視化する．

4. 実行結果・考察

4.1. デバイアス前

バイアスが生じている単語群をクラスタリングする，クラスタ数を判断するためのエルボー法が図 1 である．図 1 より，クラスタ数を 3 に決定した．

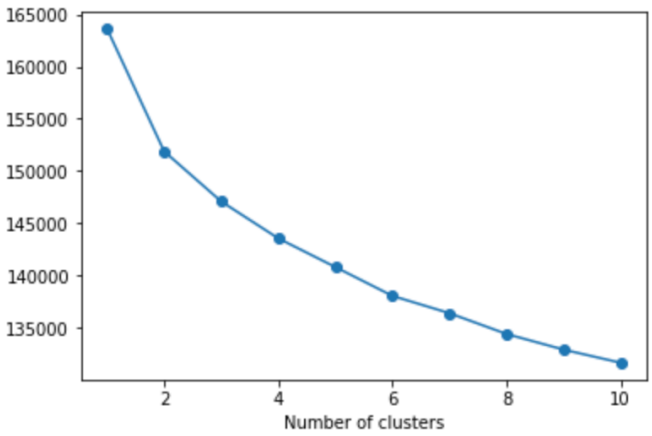

図 1: エルボー法．

[2]http://wikipedia2vec.s3.amazonaws.com/models/ja/2018-04-20

抽出した単語は，人名等の固有名詞を除くと以下表 1 のように分類された．表 1 の「例」の項目は，それぞれ「男に偏った単語」と「女に偏った単語」の順である．

表 1: バイアス分類

種類	例
想定	村長と村議，本名と旧姓，弔辞と祝辞
	出っ歯と雀斑，船長と船医
類義語	ストーリーとシナリオ，別名と別称
	境目と境界，時勢と世情，策略と姦計
対義語	右寄りと左寄り，先頭と後尾，後手と先手
	マジョリティとマイノリティ，硬球と軟球
1 字違い	アイディアとアイデア
	サーモグラフィーとサーモグラフィ
同音語	ケジメとけじめ，シャボンとしゃぼん
	ハサミと鋏，囀りとさえずり

竹下ら[4]は，日本語の単語埋め込みは文字種によってジェンダーバイアスの特徴が異なると仮定し，ひらがなは女性的，カタカナは男性的と考察した．表 1 の同言語の「ケジメ」と「けじめ」のように，カタカナは「男」，ひらがなは「女」に偏る同音語のペアが複数確認できた．また竹下ら[4]は，漢字は中性的と考察していた．表 1 より同音語において，漢字は「カタカナの単語」と「漢字の単語」，または「漢字の単語」と「ひらがなの単語」という形で出現した．

4.2. デバイアス後

図 2(左)は，バイアスを含む単語埋め込みを主成分分析で次元削減してクラスタリング(クラスタ数 3)した結果である．図 2(右)は，モデル全体に同一のデバイアスを実行した単語埋め込みを，同様にクラスタリングした結果である．デバイアス後は，デバイアス前よりもベクトル空間が小さくなっていることが分かる．

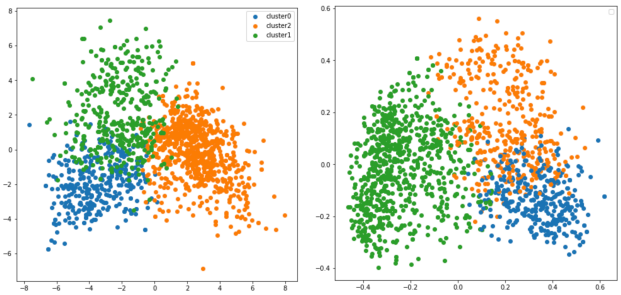

図 2: デバイアス前(左)とデバイアス後(右).

図 3 はデバイアス前後の単語ベクトルの大きさの差(デバイアス前-デバイアス後)のヒストグラムである．横軸が全て正であり，デバイアスにより単語ベクトルが全体的に縮小していることが分かる．

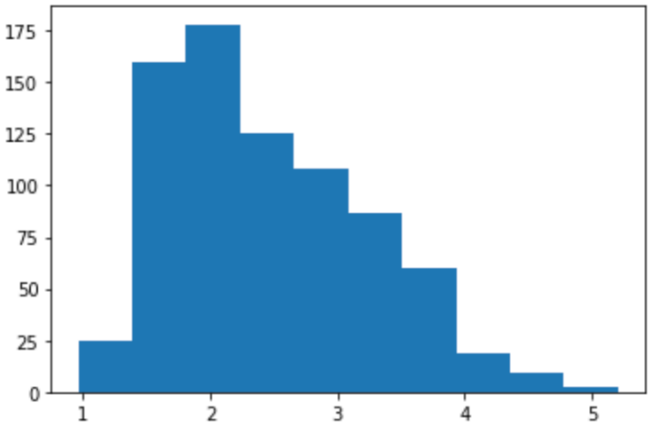

図 3: デバイアス前後のL^2-ノルムの差.

4.3. Tensor Board による可視化

図 4 はデバイアス前の単語埋め込みを Tensor Board で可視化した図である．図 5 はデバイアス後の単語埋め込みを同様に可視化した図である．単語のラベルを一つ指定すると，その単語に近い単語が画面右側にリストアップされる．

図 4 及び図 5 は「本名」と類似度が高い単語群を表示したものである．ラベルの「M:」及び「W:」は，それぞれその単語が「男」「女」に偏っていることを意味する．また，性別を意味に含む単語は，本研究ではデバイアスの対象外だが，デバイアス前後のバイアス量変化の基準として可視化対象のデータセットに追加している．

デバイアス前の「本名」は「男」寄りに偏っている．以下「父親」は「男」，「母親」は「女」を意味する単語とする．図 5 より，「本名」と「父親」のユークリッド距離は 0.997 であり，3 番目に類似度が高いことが分かる．しかし「母親」は図 4 では確認できない．デバイアス前の「本名」と「母親」は類似度が低い．

図 5 より，デバイアス後の「本名」と「父親」のユークリッド距離は 1.007 であり，デバイアス前より類似度が低下していることがわかる．一方，「母親」は 7 番目に類似度が高く，デバイアス前よりも「本名」と「母親」の距離が近くなっていることがわかる．以上より，Tensor Board を用いることで，バイアスを緩和できていることがわかると同時に，デバイアス前後の単語の関係の変化も確認することができた．

5. まとめと今後の課題

本稿では，バイアスを含む単語のみを対象にデバイアスして，バイアス前後の単語埋め込みの性質変化を確認する際の可視化の有用性を示した．デバイアスに

より，単語ベクトルとモデル全体が縮小することが確認できたが，それにより失われた情報の重要性がまだ確認できていない．本研究では，モデルに対して一括でデバイアスするのではなく，バイアスを分類し，バイアスの特徴ごとにパラメータを調節することで，モデルの性能劣化を抑えたデバイアスの実現を試みた．本研究では K-means 法によるクラスタリング結果でバイアスを分類した．一方で，バイアスを持った単語は，表 1 のように分類することも可能である．よって K-means 法以外の分類方法も検討する余地がある．

今後の予定として，文書分類タスクによるデバイアス後のモデルの性能比較，クラスタごとの最適なデバイアス度を調整可能なプログラムの実装，単語間の共起関係を確認できる単語埋め込みの可視化プログラムの実装があげられる．最終的には，翻訳アプリや AI 対話システムなどアプリケーションのジェンダーバイアスを緩和する助力となるシステム開発を目標とする．

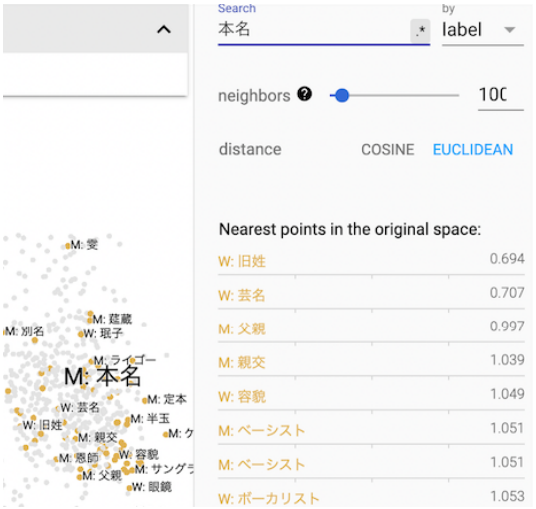

図 4: デバイアス前．

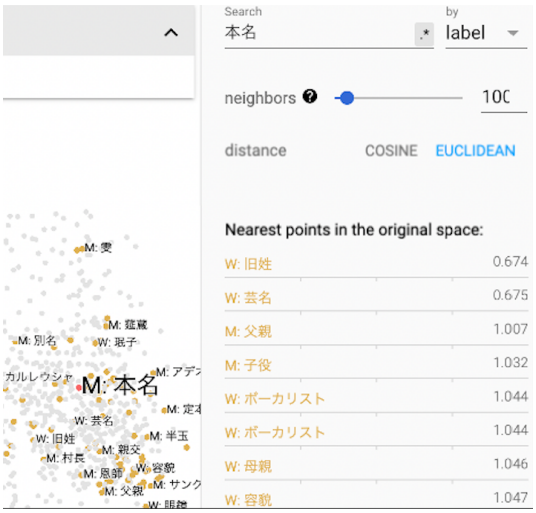

図 5: デバイアス後．

参 考 文 献

[1] Anna Farkas, Renata Nemeth, “How to measure gender bias in machine translation: Real-world oriented machine translators, “multiple reference points”, Social Sciences & Humanities Open 5, 2022

[2] 渡邊智子，北川尚美，田中真美，“日本の工学系女性研究者の現状-八大学工学系連合会に着目して-”，公益社団法人日本工学教育協会 2022 年度工学教育研究講演会，pp.116-117，2022.

[3] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, 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.

[4] 竹下昌志，ジェプカ・ラファウ，荒木健治，“日本語の単語埋め込みにおける文字種による性別バイアスの相違の分析”，電気・情報関係学会北海道支部連合大会，pp. 129-130，2020.

[5] Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez1 Caiming Xiong, “Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation”, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5443-5453, 2020.

[6] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, Louis-Philippe Morency, “Towards Debiasing Sentence Representations”, The 58th Annual Meeting of the Association for Computational Linguistics, pp. 5502-5515, 2020.

[7] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, Anupam Datta, “Gender Bias in Neural Natural Language Processing”, Logic, Language, and Security, pp. 189-202, 2019.

[8] 小林一樹，脇田建，“バイアス除去がもたらす NLP モデルの性能劣化”，The 36th Annual Conference of the Japanese Society for Artificial Intelligence，2022

[9] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, William Yang Wang, “Mitigating Gender Bias in Natural Language Processing”, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1630-1640, 2019.