4W-07

# 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†‡

## 1. はじめに

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．

## 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[1]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ると示した．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．Bolukbasi ら[1]は部分空間の射影を用いた Hard Debias を示した． 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[2]によって日本語の単語埋め込みにも応用可能であると示されている．しかし，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[3].

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[3]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．この研究は，既存のデバイアス方法[1]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，Hard Debias[1]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り，デバイアスによるモデルへの影響が考慮されていない．

Visualizing Gender Bias in Word Embedding
Arisa Sugino† Takayuki Itoh‡
Dept. of Information Sciences, Ochanomizu University†‡

## 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 の日本語版を使用した．MeCab と Unicode で英単語や記号を除外した．

### 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．「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[4]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用することで，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．エルボー法によって最適なクラスタ数を特定した．

### 3.4. デバイアス

Hard Debias[1]を応用する．まず，性別を表す単語群[1]をもとに，主成分分析で性別の基準となる軸を計算し，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．次に，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．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

### 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(左)である．図 1(左)より，クラスタ数を

3 に決定した．

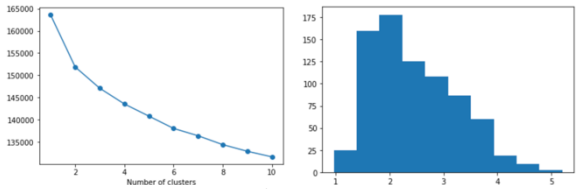


図 1: エルボー法(左) とデバイアス前後の L2-ノルムの差(右).

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: 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[2]は，日本語の単語埋め込みにおいてひらがなは女性的，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[2]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 1(右)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わかる．

### 4. 3. Tensor Board による可視化

図 2(左)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 2(右)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 2 は「本名」と類似度が高い単語群を表示したものである．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」を，「母親」は「女」を意味する単語とする．図 2(左)より，「本名」と「父親」のコサイン距離は 0.497 である．しかし「母親」は図 2(左)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 2(右)より，デバイアス後の「本名」と「父親」のコサイン距離は 0.507 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 2: デバイアス前(左)とデバイアス後(右)

## 参考文献

[1] 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.

[2] 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.

[3] 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

[4] 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.