# 単語埋め込みのバイアス評価における RNSB と WinoBias との相関関係の分析

加藤大晴　宮尾祐介
東京大学大学院情報理工学系研究科コンピュータ科学専攻
{kato_taisei, yusuke}@is.s.u-tokyo.ac.jp

## 概要

本研究では単語埋め込みの直接バイアス評価手法である RNSB と、間接バイアス評価の WinoBias との相関関係を調査する。単語埋め込みは膨大なテキストコーパスから単語の意味を実数ベクトルの形で学習することができるが、同時にバイアスも学習してしまう。バイアスの評価手法としては直接評価と間接評価があるが、両者の関係を調査した研究は少ない。そこで本研究では既存研究に倣って様々な度合いのバイアスを持つ単語埋め込みを作成し、RNSB、WinoBias 両方でそのバイアスを測定し、両者の相関係数を計算した。結果として RNSB は WinoBias と負の相関を持つことが分かった。

## 1　はじめに

自然言語処理技術が進歩していく中で、これらの技術が我々の日常生活に与える影響も拡大している。一方でこれらの技術が男女差別や人種差別といった偏見、バイアスを助長しているという問題点にも注目が集まっている [1, 2]。本研究ではこういったバイアスに対処する一歩として単語埋め込みのバイアス評価手法に焦点を当てる。

単語埋め込みは膨大なテキストコーパスを用いて単語の分布的意味を事前に学習することで、様々な後段の自然言語処理タスクにおいて性能の向上を図ることができる技術である。有用性が高い反面、テキストコーパスに含まれているバイアスを学習し、それが後段の多くの自然言語処理タスクに影響を与える危険性を孕んでいる。それゆえ単語埋め込みのバイアスを正確に測定し、自然言語処理技術がバイアスなく動作することを担保することは重要な課題となる。

単語埋め込みのバイアス評価には大きく分けて直接評価 (Intrinsic bias metric) と間接評価 (Extrinsic bias metric) の二種類がある。直接評価では単語埋め込み表現そのものに着目し、ある単語とある単語の距離が適当かどうかなどを以ってしてバイアスを評価する。それに対し、間接評価では評価対象の単語埋め込みを用いてある自然言語処理システムを学習させ、そのシステムがバイアスのかかった挙動を示すかどうかを以ってして元の単語埋め込みのバイアスの有無を判断する。間接評価では単語埋め込みが実際に自然言語処理システム内で用いられたときの振る舞いを見るという、より現実に即したシナリオでバイアスの分析を行なっており、バイアスが社会に与える影響が分かりやすいという利点がある。一方、直接評価では調べたいバイアスを単語集合を指定することで決められるので、比較的自由に様々な種類のバイアスを調べることができる。直接評価と間接評価、それぞれに長所があるため両方の視点を以ってバイアスの問題に取り組む必要がある。

しかし、両者の関係性はよく分かっているとは言い難い。Goldfarb-Tarrant et al. (2021) [3] は広く使われる直接バイアス評価手法の WEAT [2] と間接バイアス評価が多くのケースで相関しないことを発見した。ただ、その他の直接バイアス評価も同じ問題を持つのかはまだ分かっていない。そこで本研究では別の直接バイアス評価手法である RNSB [4] と間接バイアス評価との相関関係を調査する。RNSB は直接バイアス評価の枠組みの中にありながら、間接バイアス評価と同様に機械学習モデルを学習させてバイアスを計測するプロービング手法を取っている。このことから間接バイアス評価との相関が高いことが予想される。

実験方法としては Goldfarb-Tarrant et al. (2021) に倣い、様々な度合いのバイアスを持つ単語埋め込みを作成し、WEAT、RNSB、間接バイアス評価でバイアスを測定する。その結果 RNSB は間接バイアス評価と負の相関を持つことが示された。一方で相関の強

さは WEAT よりは高く、改良次第で信頼に値する直接バイアス評価になりうることも示唆された。

## 2 関連研究

### 2.1 バイアス評価指標

ここでは本研究で使用した直接バイアス評価・間接バイアス評価指標について概説する。

#### 2.1.1 WEAT

WEAT は Caliskan et al. (2017) [2] によって設計された、広く使われる単語埋め込みの直接バイアス評価指標である。WEAT はバイアスの対象となる概念 2 つを表す、等しい大きさの単語集合 (ターゲット語集合) $X, Y$ ($|X| = |Y|$) と、何を基準にそのバイアスを調べるのかを定める 2 つの単語集合 (アトリビュート語集合) $A, B$ を必要とする。例えば、理系科目・文系科目におけるバイアスをジェンダーの目線から調べたいのであれば、理系科目にまつわる単語集合 $X = \{\text{science}, \text{math}, ...\}$、文系科目にまつわる単語集合 $Y = \{\text{art}, \text{history}, ...\}$ をターゲット語集合に、男性にまつわる単語集合 $A = \{\text{man}, \text{he}, \text{brother}, ...\}$ と女性にまつわる単語集合 $B = \{\text{woman}, \text{she}, \text{sister}, ...\}$ をアトリビュート語集合として指定する。このように $X, Y, A, B$ が与えられたとき、WEAT におけるバイアスの評価値は $c(X, A) + c(Y, B) - c(X, B) - c(Y, A)$ と定義される。ここで $c(T, A) = \sum_{t \in T} \frac{1}{|A|} \sum_{a \in A} \cos(t, a)$ であり、cos はコサイン類似度を表している。$c(T, A)$ はターゲット語集合 $T$ とアトリビュート語集合 $A$ との類似度を表している。前に挙げた例では、もし理系科目にまつわる単語と男性にまつわる単語の類似度、文系科目にまつわる単語と女性にまつわる単語の類似度が、他の組み合わせの類似度よりも高ければ、WEAT の値も大きくなり、単語埋め込みがバイアスを孕んでいると言える。逆に文系にまつわる単語と男性にまつわる単語が、理系にまつわる単語と女性にまつわる単語の類似度の方が高ければ、WEAT は負の値を取る。これもバイアスを孕んでいると言える。よって WEAT の値が 0 に近ければ近い程バイアスが少ない単語埋め込みと言える。

#### 2.1.2 RNSB

RNSB [4] はプロービングの手法を用いる直接バイアス評価指標である。RNSB では 2 つのターゲット語集合 $X, Y$ と任意個のアトリビュート語集合 $A_1, ..., A_n$ を取る。本来の RNSB ではアトリビュート語集合の大きさはどれも 1 だが、後に Badilla et al. (2020) [5] によって任意の大きさのアトリビュート語集合を取れるように拡張された。この拡張版について記す。

まずロジスティック回帰モデル $M$ を、$X$ を負例に、$Y$ を正例として学習させる。学習後、$M$ に単語 $w \in A := A_1 \cup ... \cup A_n$ を与え、それが正例である確率 $p_w$ を計算させる。本来であればアトリビュート語集合とターゲット語集合の間には関係はないはずなので、$p_w$ はどれも近い値となるはずである。この $p_w$ の分布と一様分布との差が RNSB でのバイアス評価値となる。具体的には $\sum_{w \in A} p_w = 1$ となるように $p_w$ を正規化して $p_w$ が $A$ における確率分布と見做せるようにし、それと一様分布 $U$ との KL ダイバージェンス $D_{\mathrm{KL}}(p, U)$ をバイアスの評価値と定義する。

#### 2.1.3 WinoBias

WinoBias [6] は共参照解析システムのバイアス評価のためのデータセットであり、本研究では間接バイアス評価を行うために用いられた。WinoBias は Type 1 と Type 2 とに分離されており、各 Type がさらにステレオタイプセット、反ステレオタイプセットに分離されている。例えば、Type 1 のステレオタイプセットには “The physician hired the secretary because he was overwhelmed with clients.” のような文が含まれており、ここでは “The physician” と “he” が共参照関係にある。これは「医者は男性である」というステレオタイプを反映している。一方反ステレオタイプセットには “he” を “she” で置き換えた文が含まれており、“she” と “The physician” が共参照関係にある。男女のバイアスを持たない共参照解析システムであれば、ステレオタイプセット・反ステレオタイプセットでの性能の差は生じないはずなので、その差がバイアスの評価値となる。Type 2 には “The secretary called the physician and told him about a new patient.” のように文構造も共参照解析の手がかりとなる文が含まれている。

### 2.2 WEAT と間接バイアス評価の関係

直接バイアス評価と間接バイアス評価の関係性を調べた研究としては Goldfarb-Tarrant et al. (2021) [3] が挙げられ、本研究でもその手法を踏襲する。Goldfarb-Tarrant et al. (2021) は様々な度合いのバイア

スを持つ単語埋め込みを作成し、WEAT と間接バイアスの値を計測し、その散布図から両者の関係性を調査した。

バイアスの度合いを調節するためにはデータセットバランシング [7] と ATTRACT-REPEL アルゴリズム [8] が用いられた。データセットバランシングは機械学習モデルの学習に使うデータセットに含まれているアンバランスさを解消することで、よりバイアスの少ないモデルを入手する方法である。ここではステレオタイプな文を sub-sampling することでバイアス削減が、反ステレオタイプな文を削除することでバイアスの強化が行われた。ある文がステレオタイプ、反ステレオタイプかどうかは WEAT でも用いられたターゲット語集合 $X, Y$、アトリビュート語集合 $A, B$ を用いて判断する。ある文が $X, A$ 両方の単語、もしくは $Y, B$ 両方の単語を含んでいればステレオタイプであり、逆に $X, B$ の単語、$Y, A$ の単語を含んでいれば逆ステレオタイプである。ATTRACT-REPEL アルゴリズムは元々は単語埋め込みを強化するために考案されたアルゴリズムである。類義語集合と反義語集合を受け取り、元々の単語埋め込みの形を保ちつつ、類義語同士の埋め込みを近づけ、反義語同士の埋め込みを遠ざける。バイアス削減にはステレオタイプな単語の組み合わせを反義語集合に、反ステレオタイプな単語の組み合わせを類義語集合としてアルゴリズムが実行された (バイアス強化の際はその逆を行う)。ここでも同様にターゲット語集合 $X, Y$、アトリビュート語集合 $A, B$ が必要となる。ステレオタイプな単語の組み合わせの集合は $X \times A \cup Y \times B$ であり、反ステレオタイプな単語の組み合わせの集合は $X \times B \cup Y \times A$ となる。

実験の結果、WAET は多くの場合で間接バイアス評価と相関しないことが示された。

## 3 実験手法

基本的には Goldfarb-Tarrant et al. (2021) [3] の手法を踏襲して実験を行う。様々な度合いのバイアスを持つ単語埋め込みをバイアス調整手法を用いて生成し、WEAT、RNSB でバイアスを評価する。その単語埋め込みを元に共参照解析モデルを学習させ、WinoBias でバイアスを評価し間接バイアス評価とする。そして直接・間接バイアス評価の相関係数を計算する。単語埋め込みの学習用コーパスには Wikipedia [9] を使用する。 Wikipedia データは

**表 1** WEAT 6, 7, 8 [2] の詳細

| 名称 | ターゲット語集合 | アトリビュート語集合 |
|---|---|---|
| WEAT 6 | (男性, 女性) | (仕事, 家庭) |
| WEAT 7 | (数学, 芸術) | (男性, 女性) |
| WEAT 8 | (科学, 芸術) | (男性, 女性) |

NLTK [10] によってトークン分割し、頻度が 10 未満のトークンは特別なトークン ⟨unk⟩ で置き換える。最終的に 3,121,412,445 個のトークンのコーパスが得られた。そこから Word2Vec [11] を用いて次元数 300 の単語埋め込みを生成する。共参照解析の学習用コーパスには OntoNotes [12] を用いる。直接バイアス評価のためのターゲット、アトリビュート語集合には WEAT の元論文 [2] で使われたジェンダーにまつわるバイアスを測定するための単語のセットである WEAT 6, 7, 8 (表 1) を使用する。元々の WEAT 6 の男女を表す単語には人名が用いられているが、WinoBias のデータセットには人名は現れないため、WEAT 7 で使用されているのと同じ一般名詞や代名詞で置き換えたものを使用する。バイアス調整のために使用するターゲット、アトリビュート語集合としては WEAT 6, 7, 8 に加えて、それらの単語の和集合をとり Lauscher et al. (2020) [13] の方法を用いて拡張したものを用いる。具体的には spaCy [14] の事前学習済み単語埋め込みを用いて、WEAT の各単語について spaCy 上で最も近い 100 個の単語を追加する。ただし、その単語が元々の WEAT に含まれていた場合は無視する。これら 4 個のターゲット、アトリビュート語集合のペアを用い、バイアス調整手法は 2 通り、各手法でバイアス削減・強化を行うので、生成される単語埋め込みの数は、バイアス調整を受けない単語埋め込みを合わせて 17 個となる。

学習した単語埋め込みに ATTRACT-REPEL アルゴリズム [8] を適用してバイアス削減・強化を行う。また Wikipedia にデータセットバランシング [7] を適用し、バイアス削減・強化されたテキストコーパスを作成し、単語埋め込みを学習する。こうして得られた複数の単語埋め込みについて WEAT、RNSB の計測を行う。

また、これらの単語埋め込みを用いて Lee et al. (2017) [15] の共参照解析モデルを学習させ、WinoBias でバイアスを計測する。間接バイアス評価には WinoBias のステレオタイプセットの適合率、再現率から反ステレオタイプセットの適合率、再現率を引いたものを用いる。適合率、再現率の算出には MUC [16], B-CUBED [17], CEAF [18] での尺度の

**表 2** WEAT、RNSB と WinoBias との相関係数。行は直接バイアス評価を、列は間接バイアス評価を表す。第一列はどのターゲット、アトリビュート語集合のペアを用いたのかを表す。

| | | Type 1 | | Type 2 | |
|---|---|---|---|---|---|
| | | Pre. diff | Rec. diff | Pre. diff | Rec. diff |
| WEAT 6 | WEAT | 0.25 | 0.22 | 0.16 | 0.051 |
| | RNSB | − 0.37 | − 0.35 | − 0.49 | − 0.64 |
| WEAT 7 | WEAT | 0.33 | 0.31 | 0.29 | 0.19 |
| | RNSB | − 0.32 | − 0.31 | − 0.32 | − 0.50 |
| WEAT 8 | WEAT | 0.34 | 0.30 | 0.29 | 0.18 |
| | RNSB | − 0.36 | − 0.36 | − 0.31 | − 0.49 |

**表 3** WEAT と RNSB の相関係数

| | 相関係数 |
|---|---|
| WEAT 6 | $9.8 \times 10^{-3}$ |
| WEAT 7 | –0.015 |
| WEAT 8 | $8.5 \times 10^{-3}$ |

平均を取ったものを用いる。

RNSB の計測には WEFE [5] というバイアス評価フレームワークを使用する。共参照解析モデルの実装には AllenNLP [19] が用いられる。実験は Wisteria/BDEC-01 スーパーコンピュータシステム上で行う。

## 4 結果と考察

実験の結果を表 2 に示す。WEAT と WinoBias との相関係数はどれも正なのに対し、RNSB と WinoBias との相関係数は負の値を取った。一方で相関の強さ (絶対値) に着目してみると、WEAT 7-Type 1 の場合を除いて RNSB の方が大きく、WEAT 7-Type 1 の場合でも同程度の相関の強さがある。

また、直接バイアス評価同士の相関関係を表 3 に示す。直接バイアス評価同士の相関係数は、WinoBias との相関係数と比較しても分かるように非常に小さな値となった。

表 2 では WEAT と RNSB で全く異なる相関関係が得られたが、原因として両者が取りうる値の範囲が考えられる。今ターゲット語集合として $X, Y$ が、アトリビュート語集合として $A, B$ が与えられているとして、単語埋め込みに $X$ を $A$ と関連づけ、$Y$ を $B$ と関連づけるようなバイアスがあれば、WEAT でも RNSB でもバイアス評価値は正となる。ここでバイアス削減の結果、この関係が反転し $X$ が $B$ に、$Y$ が $A$ に結びつけられるようになると WEAT では負の値を取るのに対し、RNSB では KL ダイバージェンスを用いるため依然として正の値を取ることになる。つまり WEAT ではステレオタイプなバイアスは正の値として、反ステレオタイプなバイアスは負の値として計測されるが、RNSB では両者を区別することができない。WinoBias の方ではバイアスの度合いをステレオタイプセットの性能から反ステレオタイプセットの性能を引いて求めているため、WEAT と同様に二種類のバイアスを見分けることができる。そのため RNSB とは上手く相関しなかったのではないかと考えられる。実際に単語埋め込みのバイアス評価値を観察すると WinoBias で負の値が確認されたのは 1 回だけだったが、WEAT においては負の値を取っているものが一定数見られた。RNSB でも WEAT のようにステレオタイプなバイアスと反ステレオタイプなバイアスを区別できるようになれば正の相関を持つ可能性がある。

表 3 では直接バイアス評価同士で相関がないことが示されたが、これは Badilla et al. (2020) [5] での結果に反する。Badilla et al. (2020) ではジェンダーバイアスの評価については WEAT や RNSB などのバイアス評価手法で相関があることを発見した。この齟齬は相関関係の計測の仕方によるものと推測される。本研究では相関関係の計測には単純に相関係数を使用したのに対し、Badilla et al. (2020) では各バイアス評価手法で単語埋め込みを順位づけし、その順位間の相関を調査した。WEAT は単語埋め込みのコサイン類似度について線形に変化するが、RNSB ではロジスティック回帰モデルや KL ダイバージェンスが関係するためより複雑に変化する。この両者を線形な関係を前提とする相関係数で計測しようとしたために齟齬が生じた可能性がある。

## 5 おわりに

本研究では Goldfarb-Tarrant et al. (2021) [3] に倣って RNSB と WinoBias との相関関係を調査した。結果としては負の相関が観測され、RNSB が信頼に値する直接バイアス評価手法となるような証拠は得られなかった。一方で、相関の強さのみをみると WEAT よりも大きな相関を示しており、改良によっては WEAT を上回る直接バイアス評価ともなり得る可能性も示唆された。

# 参考文献


[1] Latanya Sweeney. Discrimination in online ad delivery. **Communications of the ACM**, Vol. 56, No. 5, pp. 44–54, 2013.

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. **Science**, Vol. 356, No. 6334, pp. 183–186, 2017. Publisher: American Association for the Advancement of Science.

[3] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In **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. 1926–1940. Association for Computational Linguistics, 2021.

[4] Chris Sweeney and Maryam Najafian. A transparent framework for evaluating unintended demographic bias in word embeddings. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1662–1667. Association for Computational Linguistics, 2019.

[5] Pablo Badilla, Felipe Bravo-Marquez, and Jorge Pérez. WEFE: The word embeddings fairness evaluation framework. In **Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence**, pp. 430–436. International Joint Conferences on Artificial Intelligence Organization, 2020.

[6] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In **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. Association for Computational Linguistics, 2018.

[7] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In **Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society**, AIES '18, pp. 67–73. Association for Computing Machinery, 2018.

[8] Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 309–324, 2017.

[9] Wikipedia. https://www.wikipedia.org/.

[10] Steven Bird, Edward Loper, and Ewan Klein. **Natural Language Processing with Python**. O' Reilly Media Inc., 2009.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[12] Ralph M. Weischedel, Eduard H. Hovy, Mitchell P. Marcus, and Martha Palmer. Ontonotes : A large training corpus for enhanced processing. 2017.

[13] Anne Lauscher, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. A general framework for implicit and explicit debiasing of distributional word vector spaces. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 34, No. 5, pp. 8131–8138, 2020. Number: 05.

[14] spaCy · industrial-strength natural language processing in python. https://spacy.io/.

[15] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**, pp. 188–197. Association for Computational Linguistics, 2017.

[16] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In **Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995**, 1995.

[17] Amit Bagga and Breck Baldwin. Entity-based cross-document coreferencing using the vector space model. In **Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**, Vol. 1, pp. 79–85. Association for Computational Linguistics, 1998.

[18] Xiaoqiang Luo. On coreference resolution performance metrics. In **Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05**, pp. 25–32. Association for Computational Linguistics, 2005.

[19] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson H. S. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. A deep semantic natural language processing platform. 2017.