単語埋め込み空間の異方性が社会的バイアス測定に与える影響:WEAT の妥当性に関する予備的検討

小野 晟太郎　京都大学大学院人間・環境学学研究科
齋木 潤　京都大学大学院人間・環境学学研究科

近年，単語埋め込みモデルに内在する社会的バイアスを測定する手法として，Word Embedding Association Test (WEAT)が注目されている．WEAT は，IAT の原理を応用し，単語間の意味的類似度からバイアスを定量化する手法である．WEAT を用いたこれまでの研究は，多くの言語モデルに社会的バイアスが内在することを報告してきた．しかし，単語埋め込み空間にはベクトルが特定の狭い領域に偏って分布する「異方性」と呼ばれる現象が知られており，この空間的歪みがバイアス測定に影響を与える可能性がある．本研究では，提案する白色化変換による空間の異方性の補正前後で WEAT スコアを比較した．結果として，異方性補正後に人種バイアスのスコアが大幅に減少した．この知見は，WEAT で検出されるバイアスの一部が意味的連合ではなく空間の幾何学的アーティファクトを反映している可能性を示唆する．

Keywords: word embedding, WEAT, implicit social bias

問題・目的

単語埋め込みは，単語を高次元の連続ベクトル空間に写像することで，単語間の意味的関係を数値的に表現する自然言語処理の基盤技術である．GloVe（Pennington et al., 2014）に代表される単語埋め込みモデルは，大規模テキストコーパスにおける単語の共起パターンを学習することで，意味的に類似した単語が空間内で近傍に配置されるような表現を学習する．

近年，このような大規模言語コーパスで学習された単語埋め込みが，人間と類似した社会的バイアスを獲得することが報告されている．Caliskan et al.（2017）は，社会心理学におけるImplicit Association Test (IAT)の原理を単語埋め込み空間に適用したWord Embedding Association Test（WEAT）を開発した．WEATは，ターゲット概念（例：花，昆虫）を表す語と属性概念（例：快，不快）を表す語の間の埋め込みベクトルの類似度の差異から，バイアスの強度を定量化する手法である．彼らは，GloVeのような代表的な単語埋め込みモデルにおいて，ジェンダー，人種などの社会的バイアスが存在することを示した．この発見は，言語モデルが学習データに含まれる社会的偏見を内在化することを示す重要な知見として広く注目を集めた．WEATは現在，言語モデルの公平性評価のみならず実社会に存在する潜在的偏見の予測指標としても活用されている（Caliskan et al., 2022）．

一方で，単語埋め込み空間にはベクトルが特定の狭い領域に偏って分布する「異方性」と呼ばれる現象が報告されている（Mu & Viswanath, 2018）．異方性の高い空間では，意味的に無関連な単語ペア間でもコサイン類似度が高くなる傾向があり（Ethayarajh, 2019），この空間的歪みがWEATによるバイアス測定に影響を与える可能性がある．

本研究では，白色化変換による異方性の補正がWEATスコアに及ぼす影響を検討し，検出されるバイアスの妥当性を評価することを目的とする．

方法

WEAT　WEATは，2組のターゲット語セット（X, Y）と2組の属性語セット（A, B）間の相対的な連合強度を測定する．まず，単語 w と属性セット A, B との連合強度 $s(w, A, B)$ は以下のように定義される：

$$s(w, A, B) = mean_{a \in A} \cos(w, a) - mean_{b \in B} \cos(w, b) \quad (1)$$

ここで$\cos(w, a)$は単語 wとaの埋め込みベクトル間のコサイン類似度を表す．次に，ターゲット語セット X と Y の連合強度の差異を表す検定統計量 $S(X, Y, A, B)$を以下のように定義する：

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (2)$$

S(X, Y, A, B) が正の値をとる場合，X は Y と比較して属性セット A とより強く連合していることを示す．統計的優位性は permutation test によって評価する．また WEAT では，バイアスの強度を表す指標として効果量 d を使用する．効果量 d は，両ターゲット集合における平均連合強度の差を，全ターゲット語の連合強度の標準偏差で標準化した値として以下のように定義される：

$$d = \frac{mean_{x \in X}\, \mathrm{s}(x, A, B) - mean_{y \in Y}\, \mathrm{s}(y, A, B)}{std_{w \in X \cup Y} s(w, A, B)} \quad (3)$$

この効果量の解釈は Cohen (1988) の基準に従い，$|d| \approx 0.2$は小さな効果，$|d| \approx 0.5$は中程度の効果，$|d| > 0.8$は大きな効果とみなされる．WEAT では，この効果量 d をターゲット概念と属性概念間の連合の非対称性，すなわちバイアスの強度を表す指標として使用する．

白色化変換による異方性補正　異方性を補正するため，共分散行列の固有値分解に基づくZCA（Zero-phase Component Analysis）白色化を単語埋め込み空間に適用した．埋め込みベクトル集合の共分散行列を Σ とし，その固有値分解を $\Sigma = V\Lambda V^{\top}$ とする．ここで V は固有ベクトルを列に持つ直交行列，Λ は固有値を対角成分に持つ対角行列である．ZCA白色化変換行列 W は以下のように定義される：

$$W = V\Lambda^{-\frac{1}{2}}V^{T} \quad (4)$$

各埋め込みベクトル x は $x' = Wx$ へ変換される．この操作により，変換後のベクトル集合の共分散行列は単位行列となり，次元間の無相関化と各次元の分散の均一化が達成される．結果として，特定の支配的主成分がコサイン類似度に及ぼす影響が抑制され，より等方的な空間が得られる．ZCA白色化は，変換後ベクトルと元ベクトル間のユークリッド距離を最小化する性質を持ち，元の座標系における幾何学的構造を保持する点で，意味的情報の保存が求められる単語埋め込みの処理に適している．

手続き　単語埋め込みモデルとして，先行研究（Caliskan et al., 2017）との比較可能性を考慮し，300次元のGloVeベクトルを採用した．共分散行列と白色化行列の推定には，WikiText-103データセットに含まれる語彙のうちGloVeに収録されている5万語の埋め込みベクトルを使用した．推定された白色化行列をWEATの刺激語の埋め込みベクトルに適用し，白色化された空間におけるバイアスを測定した．測定対象のバイアスは，IATおよびWEATの先行研究で標準的に用いられている以下の3種類とした：①花−昆虫 vs. 快−不快，②楽器−兵器 vs. 快−不快，③ヨーロッパ系アメリカ人名−アフリカ系アメリカ人名 vs. 快−不快．①②は文化を超えて普遍的に観察される感情価に基づくバイアスであり，WEATの妥当性検証のためのベースライン条件として機能する．③は人種バイアスを評価する際に用いられ，本研究の主要な検討対象である．

結果

Table 1に白色化前後のWEATスコアを示す．白色化前の結果は先行研究（Caliskan et al., 2017）をほぼ再現し，3条件すべてにおいて有意なバイアスが検出された．花−昆虫バイアス（$d = 1.434$, $p < .001$）および楽器−兵器バイアス（$d = 1.699$, $p < .001$）は大きな効果量を示し，人種バイアス（$d = 1.068$, $p < .001$）も大きな効果量が観察された．

白色化後においても，花−昆虫バイアス（$d = 1.456$, $p < .001$）および楽器−兵器バイアス（$d = 1.419$, $p < .001$）は有意であり，効果量も大きいままであった．楽器−兵器バイアスでは効果量が約16%減少したものの，バイアスの検出という点では白色化前と同様の結果が得られた．一方，人種バイアスについては，白色化後に効果量が大幅に減少し（$d = 0.244$），統計的有意性も消失した（$p > .05$）．

考察

本研究では，単語埋め込み空間の異方性がWEATによるバイアス測定に及ぼす影響を検討した．白色化による異方性補正後，楽器−兵器バイアスの効果量が減少し，人種バイアスが消失したことは，従来のWEATで測定されたバイアスの一部が異方性に由来する幾何学的アーティファクトを含むこと，またそれによってバイアスが過大評価されている可能性を示唆する．

また白色化前後で花−昆虫および楽器−兵器バイアスは維持された一方で，人種バイアスが消失した点は注目に値する．この知見は，人間で観測される人種バイアスが，単語埋め込みモデルの学習だけでは捉えきれない，純粋な意味的類似性を超えた社会的文脈に依存して生起している可能性を示唆している．今後は，人間の行動データとの対応分析を通じて，埋め込み空間で測定されるバイアスと人間の潜在的態度との関係を検討する必要がある．

Table 1
WEAT effect sizes (d) and p-values before (Raw) and after (White) whitening

Targets / Attributes	Raw (d)	Raw (p)	White (d)	White (p)
Flowers -Insects / Pleasant-Unpleasant	1.434	0.0001	1.456	0.0001
Instruments-Weapons / Pleasant-Unpleasant	1.699	0.0001	1.419	0.0001
European-African / Pleasant-Unpleasant	1.068	0.0002	0.244	0.3311

引用文献

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186.

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis and reflections. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2109–2120.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 55–65.

Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. International Conference on Learning Representations.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. Proceedings of the 2014 Co nference on Empirical Methods in Natural Language Processing, 1532-1543