文表現の摂動正規化: 事前学習済みモデルの Debias 手法

新妻巧朗　　渡辺太郎

奈良先端科学技術大学院大学 先端科学技術研究科
{niitsuma.takuro.nm3, taro}@is.naist.jp

1　はじめに

Elmo[13] や BERT[3] が, 自然言語理解タスクや質問応答タスクにおいてめざましい成果を残したことを受け, 近年の自然言語処理の応用研究においても, それらから発展した事前学習済み言語モデルが一般的に使われるようになってきている. これらの言語モデルは, 共参照解析 [6] といったタスクにおいても高い成績を残している一方で, それらは大規模なコーパスを用いて学習されているために, Kurita ら [8] や Nangia ら [11] がコーパスに明示的あるいは暗黙的に含まれている Social Bias をも学習していることを報告している. こうした問題は, 学習済みの単語埋め込みにおいても存在していることが Caliskan ら [2] によって明らかにされている. これらの単語埋め込みや言語モデルは, 自然言語処理の様々なタスクの基礎を成す構成要素として組み込まれるため, 下流のタスクに対しても Social Bias の影響を与えてしまう可能性があることが危惧される.

Bolukbasi ら [1] の研究をはじめとして, 近年にはその Bias を取り除く研究が進められてきた. 単語埋め込みの Debias やトレーニングデータの拡張によって, Social Bias を取り除く手法 [7, 16, 4] が提案されたきた。しかし, 事前学習済み言語モデルから直接 Bias を取り除くことを試みる研究はまだ少ない.

本研究では, BERT をはじめとするマスク言語モデルのトークンの予測スコアがトークンの出現確率の分布として表現され, またトークンの有無が周辺のトークンの分布を摂動させることに着目し, その摂動を最小化することで事前学習済み言語モデルの Social Bias を取り除くことを目指した. 特定の社会集団を表現する語を構成するトークンを除く全てのトークンの出現確率の分布が近づけるように学習することで, 社会集団に関わる語彙の周辺のトークンの出現確率への影響を最小化する. 本稿では, Nangia ら [11] が提案したマスク言語モデルの Bias を計測する手法によって提案手法の有効性を示し, さらに GLUE の評価セットを用いて自然言語理解の性能が Debias によって落ちていないことを確認した.

2　関連研究

近年の自然言語処理において, 単語埋め込みや言語モデルの Bias の存在を可視化および除去する研究が増えてきている。Caliskan ら [2] は, 認知心理学において人間の潜在的態度を計測する手法である Implicit Association Test (IAT)[5] をもとに, コサイン距離を用いて単語埋め込みが持つ潜在的な Bias を計測する Word Embedding Association Test (WEAT) を考案した. さらに, WEAT を用いて Glove[12] の性別や人種のような社会集団を表す単語の表現と, 印象を表す単語の表現との間の距離を計測し, 事前学習された埋め込み表現に社会集団間で印象の偏りが生じていることを示した. この WEAT は, 単語埋め込みだけを対象にしたものであるが, Sentence Encoder に拡張した研究 [10] やマスク言語モデルに拡張した研究 [8] が存在している.

Nangia ら [11] は, Bias を数量化するために stereotype と anti-stereotype な二組の文から構成されるデータセットである CrowS-Pairs を提案し, 擬似対数尤度マスク言語モデルスコア [14] を用いて擬似的に文の尤度を評価し, stereotype と anti-stereotype な文のスコアを比較することでマスク言語モデルの Bias の計測し, BERT などの言語モデルが Bias を持つことを示した. Bolukbasi ら [1] は, 単語埋め込みが $\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$ のようなコーパスに潜む暗黙的な Bias を表現してしまっているアナロジーが存在することを示し, それらを除去する手法を提案した. また, Liang ら [9] は, Bolukbasi らの研究を元に主成分分析を用いて Elmo や BERT などの言語モデルの表現から Bias の部分空間を特定し, それを元の表現から取り除くことで Debias を実現する手法を提案している.

2.1 文表現の摂動正規化損失

2.2 擬似対数尤度マスク言語モデルスコア

Nangia ら [11] は, マスク言語モデルの Bias を計測するために擬似対数尤度マスク言語モデルスコア [14] を用いて, 言語モデルに含まれる Bias を計測する手法を提案した. 本指標は, 一部のトークンのみが異なる文のペアが与えられることを前提とし, 二つの文それぞれが Unmodified トークンと Modified トークンという部分に分けられる. Unmodified トークンは, 文のペアの間で同じトークンである部分を指し, Modified トークンは文のペアの間で異なるトークンの部分を指す. 以降では, それぞれのトークンの集合を $U = \{u_0, ..., u_l\}$ と $M = \{m_0, ..., m_n\}$ とおく.

同実験で提案されたデータセットである CrowS-Pairs から具体的な例を示す. ‘*Native Americans* are lazy and get handouts.‘と ‘*Whites* are lazy and get handouts.‘のペアが与えられると, ‘are lazy and get handouts.‘が両方のペアで U に属し, ‘Native Americans‘と ‘Whites‘はそれぞれの文で M に属する. T をコーパスより与えられる文に属するトークンの集合とし, $T = (U \cup M)$ とする. そして, u_i を U における i 番のトークンとすると, 本指標は M とある u_i を除く全ての U が与えられた時の u_i の条件付き対数尤度の合計として算出され, 次の式のように表される.

$$\text{score}(T) = \sum_{i=0}^{|T|} \log P(u_i \in U | U \setminus \{u_i\}, M, \theta) \quad (1)$$

このスコアは, 擬似的な負の対数尤度の合計であることから, 与えられたペアのスコアを比較することで, より高いスコアを持つ文の M の方が言語モデル内で尤もらしい文章であると扱われていると判断できる.

2.3 累積 Bias

上記のスコアが擬似的に文の尤もらしさを表現していると考えると, 文のペアのスコア差を M に属するトークンの差によって発生している Bias であるとみなせる. そのため, この差を Bias スコアとし, コーパスの全ペアから計算されるスコアの差の合計を累積バイアスと名付けた. T^S を stereotype な文とし, T^A を anti-stereotype な文としたとき, そのペアの集合を $C = \{(T_1^S, T_1^A), ..., (T_n^S, T_n^A)\}$ とおくと, 累積バイアスは次の式で表される.

$$\text{Accumulate Bias}(C) = \sum_{i=1}^{|C|} |\text{score}(T_i^S) - \text{score}(T_i^A)| \quad (2)$$

本研究では, この累積 Bias を用いて言語モデルの Bias の数量化をし, 評価をおこなった.

2.4 損失関数

言語モデルが学習に使われたコーパスの語彙の共起頻度による Bias を反映していると考えると, 社会集団間で語彙の共起頻度の差を減らすことで Bias を軽減することができるのではないかと考えた.

そこで, マスク言語モデルの事前学習タスクであるマスク予測に使われる Head の出力が, 入力トークンの予測スコアとなっていることに着目し, 文のペアの出力結果のうち Unmodified トークンのスコアが近づくように Fine-tuning をすること Bias を軽減することを試みた. T^A と T^S は, それぞれマスク言語モデルのマスク予測タスクの Head の出力結果とすると, 本手法を以下のような損失関数で実現する.

$$L(T^A, T^S) = \frac{1}{N} \sum_{i=0}^{N} KL(P_{t_i^A} || P_{t_i^S}) \quad (3)$$

$N = \max\{|T^A|, |T^S|\}$ とし, KL はカルバック・ライブラー・情報量を指す. また, P_{t_i} はあるトークン t_i の予測スコアを確率分布とみなしている.

3 実験

3.1 Bias Fine-tuning

Fine-tuning および Bias の計測には, CrowS-Pairs[11] と WinoBias[16] の 2 つのデータセットを用いた.

CrowS-Pairs はアメリカ合衆国における stereotype と anti-stereotype な文のペアからなるコーパスで, 性別、国籍、人種、信条、身体的特徴などの複数の社会集団をターゲットした文が収録されている.

WinoBias は性別を表す語と職業を表す語の組み合わせを文のテンプレートに当てはめて生成された共参照解析における Bias を計量するためのデータセットである. Nangia ら [11] は, WinoBias を CrowS-Pairs と同じフォーマットに変換することで, Gender Bias の計測に利用した. 本研究もこの方法を採用し Bias の計測に利用する. WinoBias は問われる共参照の性質の違いに従って type-1 と type-2 の文から構成されるが, それぞれの命名も Nangia らを踏襲

データセット名	ペア数	例文 (Stereotype / Anti-Stereotype)
Crows-Pairs	1508	Women don't know how to drive. Men know how to drive.
WinoBias-ground	396	The physician told the baker that she had cancer. The physician told the baker that he had cancer.
WinoBias-knowledge	396	The manager fired the baker and asked her to leave. The manager fired the baker and asked him to leave.

表 1　学習・評価に利用したデータセット

し WinoBias-Knowledge, WinoBias-Ground とする. また, データセットの詳細や例文などは, 表 1 に示した.

Fine-tuning における学習の最適化には Adam を用いて, それぞれのハイパーパラメータは学習率 $\alpha = 2e^{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ とした. また, エポック数は 30, バッチサイズは 16 とした.

CrowS-Pairs のデータ数が少量であるため, ホールドアウトデータなどに分割せずに Fine-tuning の学習と評価に用いて Closed な評価をしている. しかし, この Fine-tuning の結果が汎化していることを示すために, WinoBias-ground と WinoBias-knowledge でさらに評価をおこなった. ベースラインとして Fine-tuning 前の BERT を用いる.

表 2 に, モデルごとの累積 Bias の計算結果を示す. Bias Fine-tuning をとおして, BERT から計算される累積 Bias が減少していることが確認でき, 文表現の摂動正規化損失による Fine-tuning によって Bias の軽減が達成できたと考えられる.

	BERT	fine-tuned BERT
CrowS-Pairs	2.13	1.44
WinoBias-ground	1.35	0.56
WinoBias-knowledge	1.64	0.82

表 2　Fine-tuning 前後の累積 Bias スコア

3.2　言語理解タスクによる性能確認

GLUE データセットのトレーニングセットと評価セットを用いて, Bias Fine-tuning の影響で言語理解タスクの性能が劣化していないかを確認をした. 評価セットを用いたため, トレーニングセットおよび評価セットが公開されていない「Diagnostics Main」は評価をしていない. Fine-tuning 前後で BERT は同じモデル (base, uncased) を使い, また学習に利用するハイパーパラメータも揃えることで性能を比較した. また, 実装と事前学習済みのモデルには transformers[15] を利用した.

GLUE における Fine-tuning の学習の最適化には Adam を用いており, ハイパーパラメータは Devlin ら [3] の実験をもとに学習率 $\alpha = 4e^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ とした. また, エポック数は 3 で, バッチサイズは 32 としている.

表 3.2 に, モデルごとの GLUE タスクのスコアを示す. このスコアの結果をタスクごとに, Welch の t 検定を用いて有意水準を 0.01 として検定したところ, RTE を除いて有意差が確認されなかった. そのため, Bias Fine-tuning の前後で GLUE タスクのスコアが大きく変化しておらず, 性能の劣化が見られないことを確認できた.

データ	指標	BERT	fine-tuned BERT
CoLA	Matthew Corr	0.573	**0.598**
MNLI	Acc	**0.842**	0.840
MRPC	Acc	**0.863**	0.860
	F1	0.902	**0.905**
QNLI	Acc	**0.914**	0.912
QQP	Acc	**0.913**	0.911
	F1	**0.882**	0.881
RTE	Acc	0.671	**0.704**
SST-2	Acc	0.922	0.922
STS-B	Pearson Corr	0.886	**0.903**
	Spearman Corr	0.885	**0.899**
WNLI	Acc	0.549	**0.563**

表 3　GLUE タスクの評価セットにおける結果

4　議論

表 2 の結果をより詳細に分析するため, CrowS-Pairs における Bias スコアを合計せずにヒストグラムに表したものを図 1 に示す. bin は 5 ずつ刻まれており,0 に近づくほど Bias が少ないと考えられる. Fine-tuning 前後のヒストグラムを比べると, より Bias が小さい方に頻度が増えていることが確認できる. つまり, stereotype と anti-stereotype な文のトークンの対数尤度が近づいており, 対数尤度を基準と

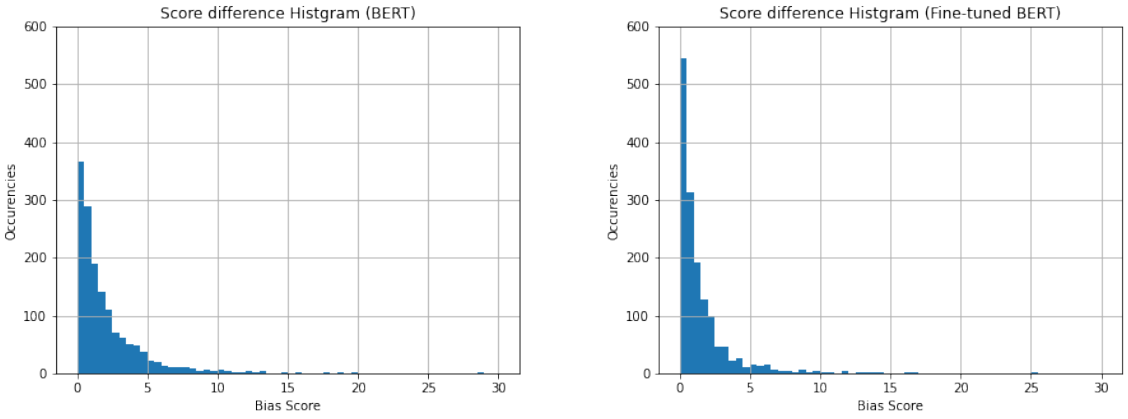

図 1　Bias スコアの分布

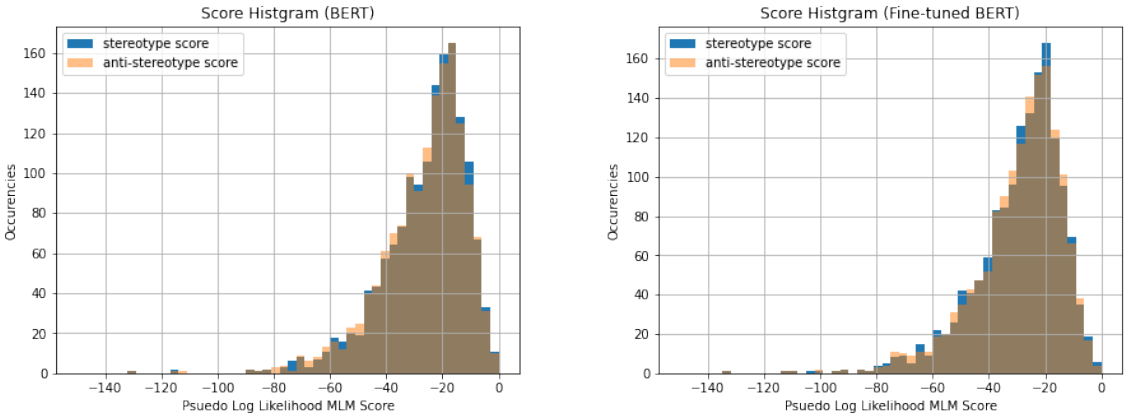

図 2　擬似対数尤度スコアの分布

した Bias を軽減できてると考えられる. また, 個々の stereotype および anti-stereotype の文それぞれの擬似対数尤度マスク言語モデルスコアをヒストグラムにしたものが図 2 である。このヒストグラムの形が対数正規分布に近かったことから, 負の対数尤度から絶対値を取ることでスコアを正の数に変えて対数変換をしたところ、正規分布のような形となった. そこで, それぞれのモデルごとに前述の変換後の stereotype と anti-stereotype のスコアが同じ分布であるという仮説のもと, 対応のある t 検定で確認した. その結果が表 4 である. 有意水準を 0.1 として, この結果を読み取ると Fine-tuning をする前の BERT では, stereotype と anti-stereotype のスコアが有意に異なる分布をしていると考えられ, 一方で Bias Fine-tuning 後では有意差がないため, それぞれの文のスコアが似た分布に近づいていると考えられる.

	BERT	Fine-tuned BERT
p-value	$3.72e^{-07}$	0.214

表 4　擬似対数尤度スコアの検定結果

5　おわりに

本研究では, stereotype と anti-stereotype な文のペア間のトークンの予測スコアを近づけるように学習する損失関数と, 事前学習に使われるコーパスの頻度によって引き起こされている Bias を計量するのに効果的な指標を提案した. 実験により文表現の摂動正規化損失は言語モデル内の文の尤度の視点でバイアスを軽減させることに成功していることが確認された.

しかし, この損失関数によって Fine-tuning された表現が後続のタスクの Bias に対してどのような影響を及ぼしているかの評価や本研究で用いた尤度ベースのスコアとは異なる視点から Bias を計測する指標による評価はまだおこなっていない. そのため, 引き続き議論を深めて言語モデルの Bias を軽減する手法の検証をしていく必要がある.

参考文献

[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, June 2016.

[2] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *American Association for the Advancement of Science*, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. 2018.

[5] Anthony G. Greenwald, Debbie E. McGhee, and et al. Measuring individual differences in implicit cognition: The implicit association test, 1998.

[6] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *arXiv preprint arXiv:1907.10529*, 2019.

[7] Masahiro Kaneko and Danushka Bollegala. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy, July 2019. Association for Computational Linguistics.

[8] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. *arXiv preprint arXiv:1906.07337*, 2019.

[9] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics.

[10] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November 2020. Association for Computational Linguistics.

[12] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[13] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. of NAACL*, 2018.

[14] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July 2020. Association for Computational Linguistics.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[16] 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)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.