バイアス除去がもたらすNLPモデルの性能劣化

東京工業大学 小林一樹 kobayashi.k.bk@m.titech.ac.jp

自然言語処理(NLP)のバイアス除去(デバイアス)の研究は、NLPタスクを解く能力への影響をほとんど考慮していない。本研究ではこの問題に着目し、デバイアスされた単語ベクトルのNLPタスクを解く能力の劣化を確認し、その原因が単語ベクトルが曖昧な表現になっているためであることを特定した。

NLPのバイアス

computer programmer - man + woman = homemaker [Bolukbasi+16]

社会に存在する格差や差別を助長する可能性があるので除去する必要がある。

NLPのデバイアスの課題

[Bolukbasi+16,Zhao+18]: 単語ベクトルのデバイアス

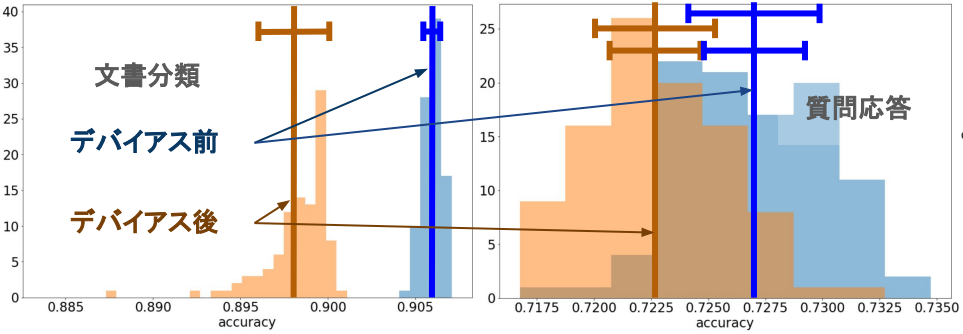
・NLPタスクを解く能力への影響は考慮されていない。

[Utama+20]: 自然言語理解モデルのデバイアス

・タスクを解く能力への影響も考慮している。

デバイアスによる単語ベクトルの劣化

[Bolukbasi+16]によってデバイアスされた単語ベクトルは、
文書分類タスク、質問応答タスクを解く能力が劣化していた。



劣化の理由の考察

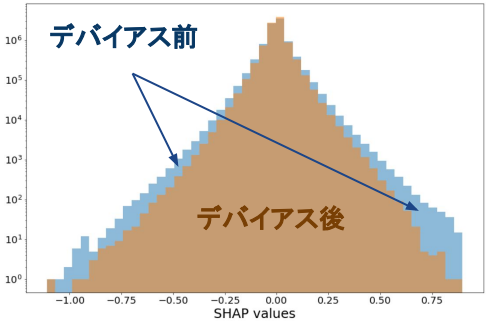
デバイアスによって曖昧な表現になってしまった。

SHAP値の分布(右図)

SHAP値が減少傾向にあるので、
分類に使われていない入力が多いと考えられる。

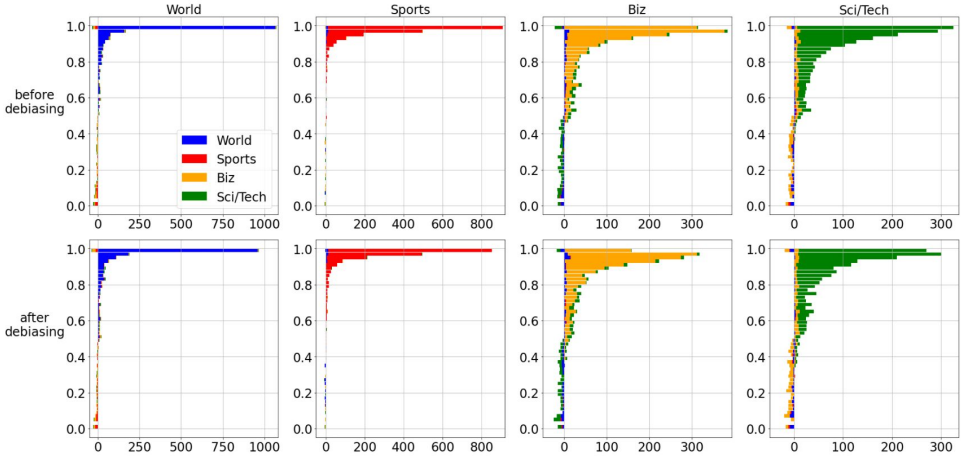
📝SHAP値 [Lundeberg+17]

・入力の出力に対する寄与率を表す



分類確率の分布(下図)

正解クラスへの分類確率が減少傾向にあるので、各クラスへの分類確率の差が小さくなっている。従って、デバイアスされた単語ベクトルはどのクラスに分類されるかわからない曖昧な表現になっていると考えられる。



参考文献、ポスターはQRコードを参照