

先端科学技術研究科 修士論文要旨

所属研究室 (主指導教員)	自然言語処理学 (渡辺 太郎 (教授))		
学籍番号	1911404	提出日	令和 3年 7月 21日
学生氏名	新妻 巧朗		
論文題目	Studies on debiasing social biases from pre-trained language models 事前学習済み言語モデルから社会的バイアスを取り除く手法の研究		
要旨			
近年の自然言語処理研究において事前学習済み言語モデルも活用が盛んであるが、そのようなモデルはコーパスに潜む偏りが反映されてしまうことで差別的な分類をしてしまう危険が存在すると指摘されている。 そこで、本稿は事前学習済み言語モデルから社会的バイアスを取り除く手法を研究した。 次の二つの研究から構成されている。 (1) 言語モデルが単語の共起頻度をモデリングしていることから同じ構造を持つ文章であれば類似のアイデンティティを入れ替えても予測確率を表すスコアが変化しないように学習する損失関数を設計し、社会的バイアスを除去できるかを実験した。 (2) 小規模なデータセットでの学習において汎化性能の高い頑健な表現を獲得するための学習手法が、社会的バイアスを引き起こすラベルの偏りに寄与するのではないかと考え、Supervised Contrastive LossをとMixoutをBERTのFine-tuningに組み込んで社会的バイアスの学習を抑制できるかを実験した。 二つの手法はそれぞれで目的とする社会的バイアスを計測する指標においてベースラインを超えて改善できていることが確認でき、モデルから社会的バイアスを取り除くことができたと考えられる。			