氏　　名	金子（かねこ）　正弘（まさひろ）
所　　属	システムデザイン研究科 システムデザイン専攻
学 位 の 種 類	博士（情報科学）
学 位 記 番 号	シス博　第 135 号
学位授与の日付	令和 3 年 3 月 25 日
課程・論文の別	学位規則第 4 条第 1 項該当
学位論文題名	Grammatical and Semantic Biases in Representation Learning from Raw Datasets （生データを用いた表現学習における文法・意味バイアス）
論文審査委員	主査　准教授　小町　守 委員　教　授　山口　亨 委員　教　授　高間　康史 委員　教　授　岡崎　直観（東京工業大学）

【論文の内容の要旨】

自然言語処理における表現学習は，単語や文の文法や意味などに関する情報をデータからベクトルやモデルとして学習することである．生データは人手で情報を付与していないデータであり，容易に大規模なデータを収集できるため，表現学習の学習データとして用いられることが多い．深層学習が主流である昨今の自然言語処理において表現学習は基盤的な役割を果たしており盛んに研究されている．そして，単語の情報を表現することを単語分散表現，文の情報を表現することを言語表現と呼ぶ．既存の表現学習の多くは単語や文の共起から文法や意味に関する有益な情報を学習している．一方で，生データは単語や文の共起が偏っており，このバイアスが文法や意味の表現学習に悪影響を与えることが知られている．そのため，本研究では表現学習における文法と意味に関するバイアスの問題に取り組む．

(1) 文法バイアス : 表現学習に使われているコーパスのほとんどはネイティブにより書かれており，文法的に誤ったテキストはほとんど含まれていない．そのため，コーパスで学習した単語分散表現や言語表現は文法的に誤った表現を考慮することができない．これは文法的に誤ったテキストを扱うタスクにおいて最適な単語分散表現や言語表現になっていないため問題である．文法誤りを含むテキストを扱うタスクとしては，入力文の文法的に誤った箇所を検出する文法誤り検出や入力文の文法的に誤った箇所を文法的に正しく書き換える文法誤り訂正などがある．これらは言語教育や言語学習の支援を行うことができる．

(2) 意味バイアス：表現学習では偏った単語の共起から，差別的な意味表現を学習してしまうことが知られている．例えば，コーパスでは単語doctorは単語sheより単語heと共起し，単語nurseは単語heより単語sheと共起する．共起を基に学習した単語分散表現では，共起する単語同士は類似度が高くなるように学習される．そのため，単語doctorは単語heと類似度が高く，単語nurseは単語sheと類似度が高くなる．このように単語分散表現は意味バイアスから差別的な単語の関連性を学習することがある．そして，性差別的な情報を含んだ単語分散表現を用いることで，下流タスクにも性差別的な影響を与えることが知られている．これは公平性や倫理的な観点から問題である．

本研究では，単語分散表現と言語表現の文法バイアスを，表現学習に文法誤りを考慮することで解消する．まず，単語分散表現の学習に擬似的な文法誤りと文法的に正しいテキストの両方を学習に用いる手法を提案する．そして，文法誤り検出により言語表現に事前学習された情報を保持しながら文法誤りを考慮する手法を提案する．さらに，言語表現の文法情報を効果的に活用するために，言語表現の各層に対してアテンションを計算する手法を提案する．アテンションとはどのベクトルを重要視するかを学習する仕組みのことである．意味バイアスに関しては，事前学習された単語分散表現の有益な性別情報を保持しながら性差別バイアスを除去する手法を提案する．この手法はバイアス除去に自己符号化器と回帰モデルを用いている．

文法誤り検出と文法誤り訂正の結果から，提案手法は単語分散表現と言語表現に文法誤りを効果的に考慮できることを示した．さらに，提案手法を用いることで文法誤り検出と文法誤り訂正の2つのタスクにおいて世界最高精度を達成した．そして，性別単語とステレオタイプ単語の単語分散表現間の類似度，事前学習された単語分散表現を用いた共参照解析モデルの性別単語とステレオタイプ単語を含む文の予測の均一さ，類似性やアナロジーに関するベンチマークデータセットの結果から，性差別情報を除去する提案手法は差別的ではない性別情報を保持しながら，既存手法と比較して最も性差別バイアスを除去できていることを示した．これらの実験から，本研究は表現学習における文法バイアスと意味バイアスの2つを解消することに成功した．

本稿では，まず第1章で文法バイアスと意味バイアスの背景と関連タスクについて述べる．そして，第2章で単語分散表現の文法バイアスを低減する手法について説明する．第3章で言語表現の文法バイアスを減少させる手法について述べる．さらに，第4章で文法バイアスを含む言語表現の文法情報を効果的に活用する手法について解説する．そして，第5章で単語分散表現の意味バイアスを除去する手法について説明する．最後に，第6章で本稿の総括とこの研究の今後の展望について述べる．