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	1911113	提出日	令和 3年 1月 25日
学生氏名	佐藤　義貴		
論文題目	Pseudo-data generation for grammatical error correction considering learner's first language 英語学習者の母語を考慮した文法誤り訂正のための擬似データ生成		

要旨

文法誤り訂正(Grammatical Error Correction: GEC)は，第二言語学習者の書いた文法的に誤りを含む文を入力し，その誤りを訂正した文を出力するタスクである．

本研究では特定の言語を母語にもつ英語学習者が書く文法誤りに頑健な，GECモデルの実現を目指す．GECの研究分野においては，モデルの訓練に必要な学習者コーパスの不足を対処するために，擬似誤り文を生成し活用する研究が活発に行われている．従来法によって多種多様な擬似誤りを生成することができるが，学習者の母語によって誤りの傾向が異なることが考慮されているとは言えない．

そこで，本研究では以下の2つの手法によって母語の影響を考慮した擬似誤り文を生成することを目指す．
提案手法I. 訓練済みの逆翻訳モデルを特定の言語を母語にもつ学習者によって書かれた学習者コーパスでfine-tuneし，母語の影響を考慮した擬似誤り文を生成する．
提案手法II. 誤りタイプを明示するトークンを付与した学習者コーパスで逆翻訳モデルを訓練し，生成時には特定の言語を母語にもつ学習者によって書かれた学習者コーパスの誤りタイプの傾向に応じて，生成する誤りタイプの比率をコントロールする．
これら2つの手法によって母語の影響を考慮した擬似誤り文を生成し，学習者の母語を考慮した文法誤り訂正モデルの性能向上を目指す．その結果，提案手法Iで生成した擬似誤り文を用いることで，従来の逆翻訳モデルで生成した擬似誤り文を用いた場合よりも，GECモデルの性能が向上することが確認できた．提案手法IIでは，50種類以上の誤りタイプに対し，開発セットで76.8%の確率で特定の誤りタイプの疑似誤り文を生成できた[1]．提案手法IIは最終的なGECモデルの性能向上には寄与できなかったが，今後の発展が期待できる．

1 従来の逆翻訳モデルは開発セットにおいて19.3%の確率で特定の誤りタイプを狙って生成できた．