平成30年度 情報工学コース卒業研究報告要旨

武田 (浩) 研究室	氏　　名	中 原　拓 哉
卒業研究題目	教師なし機械翻訳手法を用いた文体変換	

文体変換とは，ある文を，意味を変化させることなく別の文体で言い換えるタスクである．一般的に，同じ意味内容を持つ文章であっても，文体が異なると，読み手の印象も異なる．例えば,「～だけど,」という表現が「～だが,」となるだけで，少し固い文章の印象をもたらす．したがって，文体を変換することができれば，同じ意味内容を伝える場合でも読み手の持つ印象を変えることができる．

文体変換の代表的な手法には，まず，対象の文体ペアに関する変換規則を作成する手法がある．しかし，無数に種類のある文体ごとに変換規則を人手で作成するのはコストが大きい．また別の手法として，文体変換前後の文を一文ずつ対応させたデータを対訳データとみなし，教師あり機械翻訳手法を適用する手法がある．しかし，一般的に，文体変換の対訳データは存在しないため，対訳データをどのように獲得するかが大きな問題となる．一方，近年，対訳関係にない大規模な単言語コーパスのペアを用いて翻訳モデルを学習する取り組みが行われている．このような手法は教師なし機械翻訳と呼ばれ，英語とフランス語のような比較的類似性の高い言語間の翻訳に関して成果を出している．文体変換のタスクにおいて，それぞれの文体のデータは大量に入手できることから，本研究では，この教師なし機械翻訳手法を用いて，文体変換に取り組む．

本研究においては，Twitterと毎日新聞の文体を対象とし，Twitterでの文体から毎日新聞のような文体にすることを目標とする．まず初めに，教師なし文体変換モデル，すなわち教師なし機械翻訳手法を用いた文体変換モデルを生成した．教師なし文体変換モデルにより文体変換を行った結果,「思ってる」が「思っている」に変換されるような『い抜き言葉』の「い」の補完や,「風呂入る」が「風呂に入る」に変換されるような助詞の補完などの望ましい変換が確認できた．しかし，その一方で,「動画の編集」が「編集の動画」に変換されるような不適切な単語の入れ替わりや,「ありがとう」が「問い合わせ」に変換されるような全く意味内容の異なる文への変換が，変換全体の約30％を占めていた．そこで，二つ目の手法として，教師なし文体変換モデルを，文体変換の学習データの自動生成器として扱うことを考えた．具体的には，教師なし文体変換モデルによる文体変換結果から，編集距離等を手がかりとして，対訳関係にあると考えられる文体ペアを抽出し，教師あり機械翻訳手法に基づく文体変換の学習データとして使用した．その結果,『い抜き言葉』の「い」の補完や助詞の追加などの望ましい変換例を残しつつ，単語の入れ替わりや，全く意味内容の異なる文への変換を大幅に減少させることができた．