平成29年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	石 飛 篤 志
卒業研究題目	法令文のニューラル機械翻訳に関する研究	

社会のグローバル化が進む近年において，国際取引の円滑化や対日投資の促進，海外での法整備支援などを図るために，日本法令を国際的に発信することが求められている．そうした需要に基づき，法務省は日本語外国語訳データベースシステム (JLT) を開設し，日本法令の英訳を公開している．しかし，法令を人手で翻訳するには，言語知識に加えて法令に関する深い知識も必要となるため，翻訳作業にかかる時間は大きく，翻訳整備計画から遅れている．

こうした背景から，人手による法令翻訳を支援することを目的として，法令文の機械翻訳に関する研究が行われている．法令文の機械翻訳については，従来，法令文の統計的機械翻訳 (SMT) が研究されてきたが，SMT では，長文に対する翻訳精度に課題があった．新たな翻訳手法として，ニューラル機械翻訳 (NMT) が提案されている．NMT は，入力文の各単語を分散表現に変換し，これらを合成して入力文全体を表す文の分散表現を生成したのち訳文を出力するものである．従来の SMT では，入力文を単語に分割する際に形態素解析器が用いられてきた．しかし，NMT に対しては形態素解析器よりも，Sentencepiece と呼ばれる単語分割手法を用いたほうが，一般文においては翻訳精度が向上することが報告されている．

そこで本研究では，法令文翻訳においても同様の傾向が見られるか調査するために，NMT の前処理として形態素解析器と Sentencepiece が与える影響を調査した．調査対象として以下の翻訳モデルを作成した．

日本語の単語分割が NMT へ与える影響を調査するために，MeCab+IPADIC，MeCab+NEologd，MeCab+UniDic,KyTea，Sentencepiece(4k)，Sentencepiece(8k) 及び unigram 分割を用いて 7 種類の翻訳モデルを作成した．対訳コーパスとして，JLT で公開されている対訳文 563 法令 266,560 文を利用し，上記の 7 つの単語分割手法それぞれを用いた学習コーパスを作成しそれぞれ翻訳モデルを作成した．これらの学習データ 7 種類それぞれに対して翻訳モデルを作成した．翻訳モデルの作成には，オープンソースである OpenNMT を用いた．

作成した 7 種類の翻訳モデルを用いてテストデータを翻訳することにより，比較実験を行った．本実験ではオープンテストとして JLT で公開されている対訳文 29 法令 12,455 文を用いた．また，テストデータは，学習データに改正前が含まれているもの，含まれていないものの 2 グループに分けた．翻訳結果の評価には自動評価尺度である BLEU 及び RIBES を用いた．

実験の結果，法令文の NMT では Sentencepiece による翻訳精度向上が確認できなかった．これは，今回用いた法令文のコーパスサイズが大きくないため，Sentencepiece の効果が出なかったと考えられる．