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	重野 泰和
卒業研究題目	ニューラルモデルと翻訳メモリを併用した機械翻訳	

高度に国際化が進む今日において，文書翻訳の必要性が高まっている。特に，製品マニュアルや公文書などの翻訳では，訳語の正確性，統一性を保持することが求められている。

それに対して，小酒井ら (2015) は，翻訳メモリと統計的機械翻訳を併用することにより翻訳性能を向上させる手法を提案した。翻訳メモリは，既存の原文とその訳文の組を保存するデータベースである。翻訳メモリを利用すると既存の訳文を再利用できるため，翻訳の統一性を保つことができる。しかし，再利用した訳文を翻訳対象文に合わせて編集する必要があるうえ，翻訳対象文と類似する原文を保存していない場合，訳文を提案できない。一方，統計的機械翻訳は，原文を複数のフレーズに分割し，それぞれのフレーズを確率的に翻訳する機械翻訳手法である。統計的機械翻訳を利用すると，原文と一致する文が存在しない場合でも，一致するフレーズが存在すれば，訳文を出力できる。そのため，小酒井らは翻訳メモリから訳文を提案できない場合に，統計的機械翻訳を用いることにより訳文を提案できるようにした。

しかし，統計的機械翻訳には未知のフレーズをうまく翻訳できないという問題がある。これは，統計的機械翻訳が原文の表層的な変化に脆弱な手法であるがために発生すると考えられる。この問題を解決できる手法の一つにニューラルモデルを用いた翻訳がある。ニューラル翻訳は，原文をベクトルに変換し，そのベクトルを用いて訳文を生成する機械翻訳手法である。ベクトル化によって，文意を保持しつつ，文全体を抽象化して扱うことができる。そのため，原文の表層的な変化に対して頑健な翻訳手法である。したがって，統計的機械翻訳とは異なる原理に基づくニューラル翻訳を用いることにより，この問題を解決できると期待できる。

そこで本研究では，ニューラル翻訳と翻訳メモリを併用した翻訳手法を提案する。本手法では，翻訳メモリに保存されている各原文に対して翻訳対象文との類似度を求める。求めた類似度があらかじめ与えられた閾値以上となる原文が翻訳メモリ中に存在する場合には，その原文に対応する訳文を出力する。そうでない場合は，翻訳対象文をニューラル翻訳によって翻訳して訳文を出力する。ただし，翻訳対象文との類似度が閾値を超えるものが複数存在する場合は，類似度が最大のものを出力とする。類似度には Cosine 係数，Dice 係数，Jaccard 係数，および編集距離を使用する。

本手法の有効性を検証するため，英語と日本語の文対応のある法令文 28 件（10,307 文）を用いて実験を行った。実験の評価スコアには，$1 - \mathrm{TER}$ を用いた。TER は，機械翻訳の出力を参照訳と一致させるために必要な最小の編集コストを参照訳の語数で割った値である。

閾値を 1.0，すなわち，翻訳メモリの原文と完全一致した文のみを引用した場合，本手法のスコアの平均は 52.39%（分散は 304.39）となり，ニューラル翻訳単独の出力に対するスコアの平均 44.75%（分散は 64.36）と比較して有意な改善が見られた。次に，本手法と小酒井らの手法を比較すると，類似尺度に Cosine 係数，Dice 係数を用いた場合は閾値を 0.9 としたときに，類似尺度に Jaccard 係数を用いた場合は閾値を 0.8 としたときに，それぞれ有意な改善が見られた。

以上の結果から，本手法は既存の翻訳支援手法よりも翻訳性能が向上し，よりよい翻訳支援が期待できることが確認できた。