平成26年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	小酒井 款雄
卒業研究題目	統計的モデルと翻訳メモリを併用した機械翻訳	

あらゆる文書のデータ化・国際化の進む今日において、文書翻訳のニーズが高まってきている。商品マニュアルや公文書などの翻訳では、統一性・正確性を保持するための支援が求められており、翻訳メモリや機械翻訳などが利用されている。

翻訳メモリを用いた支援では、まず、原文と訳文が組となったデータベースを作成する。その中から翻訳対象文に類似する文を探し出し、その訳を提示することにより翻訳のヒントを与える。一方、既に存在する翻訳を参照できるため、翻訳の統一性・正確性の保持を図ることができる。機械翻訳を用いた支援では、計算機による翻訳結果を提示することにより、翻訳のヒントを与える。機械翻訳には複数のアプローチがあるが、現在の主流は統計的機械翻訳である。この統計的機械翻訳では、文対応のついた対訳コーパスから翻訳規則を学習するため、そのコーパスの特徴を反映した翻訳ができるという利点がある。

しかし、翻訳メモリや統計的機械翻訳を利用した翻訳支援には、それぞれ問題がある。翻訳メモリを利用した支援では、翻訳メモリに類似文が存在しない場合に翻訳のヒントを提示できない。また、統計的機械翻訳を利用した支援では、ある程度意味のわかる訳を提示できるものの、文法的に正しくない場合もあり、その精度は人手による翻訳に遠く及ばないという問題がある。

上記の問題を解決する試みとして、翻訳業界では、翻訳メモリを利用した従来の翻訳支援に、機械翻訳を追加したシステムが既に考案されている。このシステムでは、翻訳メモリからの類似文の提示を基本としているが、文中の指定した部分を統計的機械翻訳によって翻訳することもできる。そのため、翻訳メモリに類似文が存在しない場合でも、翻訳のヒントを提示できる。また、統計的機械翻訳に翻訳メモリからの引用機能を追加したという観点で見ると、人手による翻訳を利用することになるため、翻訳精度の向上が期待できる。

しかし、上記のようなシステムについては、学術的な観点からの性能評価は行われていない。そこで、本研究では、翻訳メモリと統計的機械翻訳を併用した翻訳手法 (ハイブリッド翻訳) について述べ、その性能を評価する。

本手法では、翻訳メモリ中の各原文に対して翻訳対象文との類似度を算出し、閾値以上だった場合はその対訳を、そうでない場合は統計的機械翻訳による訳を出力する。ただし、翻訳メモリ中に閾値以上の文が複数存在する場合は、類似度が最も高いものを出力とする。また、類似度としては cosine 係数、Jaccard 係数、Dice 係数を使用した。本手法の有効性を検証するため、法令 3,817 文を用いて実験を行った。閾値が 1.0、つまり、完全に一致する文のみを翻訳メモリから引用した場合、本手法における BLEU、RIBES、WER のスコアはそれぞれ 0.4213、0.6723、0.6064 となり、統計的機械翻訳の各スコア 0.4082、0.6652、0.6280 に比べて有意な改善が見られた。また、本手法では翻訳メモリからの引用を用いており、その分、統計的機械翻訳よりも自然な文が多く出力できている。翻訳メモリからの引用文は、翻訳対象文との部分的な相違を含む場合があるものの、統計的機械翻訳とは異なり、文法的に正しい文であるため、間違いの修正がしやすい。以上の結果から、統計的モデルと翻訳メモリを併用することにより、システムによる翻訳の質が向上し、さらに効率のよい翻訳作業が期待できることを確認した。