平成25年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	岡 田　浩 平
卒業研究題目	原文の表層形を利用した「法令のあらまし」の統計的機械翻訳の改良	

近年、社会や経済のグローバル化に伴い、対日投資の促進や在留外国人への対応、海外での法整備支援などのため、日本の法令を外国語へ翻訳することへの要求が高まっている。それに対して、法務省では「日本法令外国語訳データベースシステム」を公開しており、人手による英訳法令を多数提供している。しかし、人手による法令翻訳では、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い。そのため、2014年2月現在、現行法令の約4%しか英訳を公開できておらず、また、翻訳整備計画から遅れている。

このような背景を基に、法令そのものではなく、「法令のあらまし」を翻訳対象とする統計的機械翻訳システムの研究が進められている。法令のあらましは公布された法令を要約した公的な文書であり、日本法令に関する情報の概要を理解するために有用である。また、法令のあらましは、元の法令より短く、簡潔な文で構成されるため、機械翻訳に適しているといえる。しかし、この翻訳システムでは、法令文の対訳コーパスから学習を行った翻訳モデルによって法令のあらましを翻訳するため、法令のあらまし特有の表現や言い回しに対して適切な翻訳ができない例が多数存在した。

そこで、本研究では、適切な翻訳ができなかった原因である法令のあらましの表層形に着目し、三つの提案手法によって既存の翻訳システムを改良する。

一つ目は、法令のあらましの対訳コーパスと法令文の対訳コーパス両方から学習する翻訳モデルを導入することである。法令のあらましの対訳コーパスは、法令文の対訳コーパスと比較すると少量であるが、法令のあらまし特有の表現を学習できる可能性がある。

二つ目は、漢数字対訳コーパスを追加した翻訳モデルの学習である。法令と法令のあらましでは漢数字の表記が異なる。法令では「十」「百」「千」の位取りを行っているが、法令のあらましでは位取りは行わない。そのため、法令文で学習を行った既存のシステムでは漢数字の翻訳が適切ではなかった。そこで、位取りを行わない漢数字とその英訳のペアをあらかじめ生成し、翻訳モデルの学習データに追加することにより、漢数字に対する翻訳精度の向上を図る。

三つ目は、括弧内の単語列の並び替えに対する制限である。統計的機械翻訳では、原文をいくつかの単語列に分けてそれぞれ翻訳した後、並び替えを行う。このとき、提案手法では括弧内の単語列の翻訳結果がひとまとまりになるように並び替えを制限し、翻訳文における語順改善を目指す。

以上の三つの提案手法をそれぞれ適用した統計的機械翻訳システムを構築し、平成23年公布の法令のあらまし1,371文に対して翻訳実験を行った。翻訳自動評価尺度RIBESによって各システムの出力文に対するスコアを計算したところ、既存のシステムのスコアが66.13であるのに対して、一つ目の提案手法のスコアは68.99、二つ目の提案手法のスコアは66.94となり、いずれも有意にスコアが上昇した。三つ目の提案手法のスコアは66.39となり、既存のシステムからスコアの上昇はあったものの、有意差は認められなかった。