平成28年度 情報工学コース卒業研究報告要旨

外山 研究室	氏　　名	栗 本 雄 太
卒 業 研 究 題 目	法令文の統計的機械翻訳における訳語統一の効果	

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、高いコストがかかる。そのため、すべての現行法令が英訳されているわけではないのが現状である。

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、法令文の統計的機械翻訳 (SMT) に関する研究が行われている。SMT は、統計モデルに基づき、大量の対訳コーパスから機械翻訳システムを自動的に構築するものであり、入力文と同じドメインの対訳コーパスを用いることにより、特定分野への適応が可能となる。先行研究では、大規模な法令文対訳コーパスとして、JLT で公開されている対訳データを利用することにより、法令文用の SMT システムを実現している。しかし、JLT の法令文対訳データには訳語が統一されていないという問題がある。この問題は、外国人による日本法令の解釈に混乱を与えるだけでなく、SMT の学習データとして利用する場合に、翻訳精度に悪影響を及ぼすと考えられる。

そこで本研究では、法令文の統計的機械翻訳における精度向上を目指して、法令文対訳コーパスにおいて訳語を統一する手法を提案し、その効果について示す。本手法は、日本語文が同一であるにも関わらず、その訳文が異なっている対訳ペアを対訳コーパスから抽出し、それらすべての訳文を、同一の日本語文に対して最も頻出する訳文に書き換える。本手法により訳語統一を行った対訳コーパスを用いて SMT システムを構築することにより、翻訳結果において訳語の揺れが小さくなることが期待できる。特に、法令文では特定の表現が多用されるため、訳語統一による効果は大きいと考えられる。

法令文の統計的機械翻訳における訳語統一の効果を検証するために、2種類の SMT システムを構築した。一方のシステム (以下、従来システムと呼ぶ) では、対訳コーパスから取り出した文をそのまま学習データとした。もう一方のシステム (以下、本システムと呼ぶ) では、本手法による訳語統一を対訳コーパス全文に施したものを学習データとした。対訳コーパスとして、JLT で公開されている 758,199 文を利用し、それらすべてを学習データに用いた。SMT のデコーダには Moses を、言語モデルの学習には SRILM を、学習データの単語アライメントには GIZA++を、日本語文の形態素解析には MeCab をそれぞれ使用した。

両システムで同一のテストデータを翻訳することにより、比較実験を行った。本実験ではクローズドテストを行い、テストデータは、対訳コーパスのうち 1,220 文を用いた。テストデータの正解訳には訳語統一が施されたものを使用した。翻訳結果の評価には自動評価尺度である BLEU を用いた。

実験の結果、本システムと従来システムの BLEU スコアはそれぞれ、47.23 と 46.40 となり、訳語統一による翻訳精度の向上を確認した。オープンテストの実施は今後の課題である。