

## 論文内容の要旨

博士論文題目 Automatic Construction of Translation Knowledge for Corpus-based Machine Translation

(コーパスベース機械翻訳における翻訳知識自動構築法の研究)

氏名 今村 賢治

(論文内容の要旨)

本論文では、対訳コーパスからの機械翻訳知識自動構築法に焦点を当て、以下の3点の課題について論じる。

1. 対訳コーパスからの知識獲得法の提案
2. 自動獲得された知識の機械翻訳エンジンへの適用と翻訳品質の測定
3. 翻訳品質を低下させる問題点の指摘と解決法の提案

本論文の特徴は、第2、第3の課題に言及したところにある。すなわち、本論文では、対訳コーパスから獲得した知識を用いて実際に機械翻訳を行うことにより、その問題点を指摘し、解決方法を提案する。

まず、第1の課題に対し、本論文では対訳文の階層的句アラインメント方法を提案する。本論文で用いた方式は、句対応スコアと呼ぶ構造類似性評価尺度を用い、アジェンダ中から最良の構文構造と対応関係を同時に決定する。その利点はまず、構文解析の曖昧性を対訳文の類似性を利用して解消することである。もう一つの利点は、構文解析に失敗しても部分木の列を出力することである。このような方式を用いることにより、従来法に比べ、精度を保持したまま約2倍の同等句を抽出することができた。

第2の課題に対しては、階層的句アラインメントを大規模対訳コーパスに適用し、機械翻訳知識を自動構築し、実際の機械翻訳エンジン TDMT に組み込み、翻訳品質を測定する。翻訳知識中に大量に含まれる冗長規則が誤訳や曖昧性増大の原因となることが判明した。冗長規則を排除することにより、自動構築した機械翻訳器は、Hand-coded 機械翻訳 (手作業で作成した知識を用いる機械翻訳) の性能に近づくことができるという結論を得た。

第3の課題に対しては、二つのアプローチを提案する。一つは、対訳コーパス自身に起因する問題に焦点を当て、それを知識獲得の前処理的に解決する方法であり、二つの知識構築法を提案する。一つは、直訳性が高い対訳文だけを用いて知識を構築する対訳コーパスのフィルタリング法である。もう一つは、対訳文を直訳部と意識部に分割し、そこから生成される規則に異なる一般化を適用した、分割構築である。これにより、約8.6%の文について翻訳品質を改善することができた。

第3の課題に対するもう一つのアプローチは、翻訳知識獲得の後処理による。冗長な規則による問題に対して、翻訳品質の自動評価を利用した翻訳規則の取捨選択法(フィードバッククリーニング)を提案する。この方法では、自動評価が出力するスコアを目標関数値と見なし、これを最大化するように規則の取捨選択を行う。自動評価方法にはBLEUを用い、最適化法には山登り法を用いた。訓練コーパスのみでクリーニングを行うために、さらに本論文ではN分割交差クリーニングを提案する。その結果、従来の後処理法に比べ、翻訳品質を大幅に改善することができた。

氏名	今村 賢治
----	-------

(論文審査結果の要旨)

平成16年4月23日に開催した公聴会の結果を参考に、平成16年5月27日に本博士論文の審査を行った。以下のとおり、本博士論文は、提案者が独立した研究者として、研究活動を続けていくための十分な素養を備えていることを示すものと認める。

今村 賢治は、本博士論文において、コーパスベース機械翻訳のための翻訳知識自動獲得方法と獲得された規則のフィルタリングのための手法を提案し、実用的コーパスベース機械翻訳の全体設計を可能にする方法論を確立した。本論文の貢献は、以下のようにまとめることができる。

1. 対訳コーパスの対訳文の統語構造を用いた階層的句アラインメント法という新しい翻訳規則抽出法を提案した。この方法は、完全な統語構造を得ることができなくても部分木による翻訳規則を抽出することができ、従来手法より多くの翻訳規則を得ることができる点が優れている。
2. 抽出された翻訳規則を変換に基づく機械翻訳システムの翻訳規則として組み込み、その翻訳品質を評価することにより、抽出された規則の問題点を客観的に論じた。これによって発見された冗長規則を排除することにより、翻訳性能の向上の可能性を示した。
3. ここまでの結果を受けて、自動抽出された翻訳規則から冗長な規則を排除するため、2つの方法を提案した。一つは、前処理として翻訳規則から冗長なものを取り除く手法であり、規則の直訳性という概念を定義して、その定義に反する規則を取り除くことにより、翻訳性能を向上させた。もう一つの方法は、抽出された翻訳規則を実際の翻訳に実装することによって具体的な例文を翻訳させ、結果を自動評価することにより、翻訳の性能を下げる翻訳規則を発見して取り除く手法である。特に、訓練データのみを用いてこの手法を実現するため、フィードバッククリーニングという新しい手法を提案した。

以上のように、対訳コーパスからの翻訳知識の自動獲得、および、それを実際の機械翻訳システムに組み込む際の問題点を指摘し、冗長な翻訳規則の自動同定の手法を提案することによって実用的なコーパスベース機械翻訳技術の実現を達成した本研究は、独創性が高く、しかも実用的であり、自然言語処理の分野において高い貢献があると評価する。

よって、本論文は、博士(工学)の学位論文として価値あるものと認める。