

論文内容の要旨

博士論文題目 Translation Knowledge Acquisition for Pattern-based Machine Translation
(パターンベース機械翻訳のための翻訳知識獲得)

氏名 北村美穂子

(論文内容の要旨)

機械翻訳の性能向上は、その機械翻訳が有する翻訳知識の質及び量が決め手となる。したがって、性能の良い機械翻訳を構築するためには、従来のような手作業での翻訳辞書の登録でなく、機械的又は半機械的に翻訳知識を構築できる技術及びその構築された知識を容易に追加・拡張できる機械翻訳技術が重要となる。

本論文は、上記2点の技術課題について、対訳コーパスからの翻訳知識の獲得技術及びパターンベース機械翻訳技術を用いた解決方法を論じる。

第一の提案は、素性の扱いが柔軟なパターンベース機械翻訳システムの提案である。従来のパターンベース機械翻訳は、翻訳規則の記述に多くの制限があり人間の手が入れにくく、翻訳結果のチューニングが難しいという課題があった。我々が提案するパターンベース機械翻訳は、そこで扱う全ての翻訳知識を人間が理解し易いパターンの形式で表す。パターンは人間自ら作成することもできる。また、対訳コーパスを用いて機械的に作成することもできる。

第二は、上記のパターンベース機械翻訳で利用される翻訳パターンの自動抽出方法の提案である。文対応済みの対訳コーパスにおいて、共起する原言語と目的言語の単語列を対応付けることにより、翻訳パターンを自動的に抽出する。統計的な手法によって単語列組の対応関係の強さを計算し、対応関係が強いと認められた単語列組から順に抽出する。出現回数による閾値を設定し、その値を徐々に下げながら計算対象を増やしていく方法を採用することにより、単語列の組み合わせ爆発を抑えることができ、高精度で翻訳パターンを抽出することができる。

第三の提案では、第二の提案手法をベースにして、対訳辞書、及び、対訳コーパス中の文の構文情報を利用して、翻訳パターンを獲得する。本方法では、実用化に配慮して、人間が効率良く抽出結果を確認する手法も導入する。第三の手法は、第二の提案手法に比べて、より高精度、高いカバレッジで翻訳パターンを抽出することができる。

上記の二つの手法を用いることにより、連続単語列からなる専門用語や慣用表現に関する翻訳パターンを獲得することができる。しかし、より複雑な翻訳パターン、例えば、連続単語列内に変化部が存在する翻訳パターンや、訳語選択のための翻訳規則等を獲得することはできない。そこで最後に、統計的手法と構造照合手法を利用した翻訳規則の自動獲得方法を提案する。構造照合手法とは、対訳文間を依存構造レベルで対応付けるための手法である。

上記の四つの技術の融合により、翻訳知識獲得機能を有する実用的な機械翻訳システムの構築が可能となる。

氏 名	北村美穂子
-----	-------

(論文審査結果の要旨)

平成16年10月26日に開催した公聴会の結果を参考に平成16年11月26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が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。

北村美穂子は、本博士論文において、対訳テキストからパターンに基づく機械翻訳のための種々の手法を提案し、その有効性を示した。提案手法は次のようにまとめることができる。

1. パターンに基づく機械翻訳の一つとして、階層化された翻訳知識をもつモジュール性の高く、かつ、利用者にとって理解しやすく、翻訳知識の追加が容易な機械翻訳システムを提案した。
2. 文対応済みの対訳コーパスにおいて、共起する原言語と目的言語の単語列を対応付けることにより、任意の大きさの翻訳パターンを自動的に抽出する手法を提案した。また、実験により高い精度で翻訳パターンを抽出できることを示した。
3. 対訳コーパス中の文の構文情報を利用して、長さには制限があるが構文的に均質な翻訳パターンを獲得する手法を提案した。また、抽出結果を最小限の人手によって確認するステップを翻訳規則獲得フェーズに入れることにより、高い精度とカバレッジで翻訳パターンを抽出できることを示した。
4. 対訳コーパスの両言語文を統語解析し、単語係り受け結果の構造照合を行うことにより、統語構造をもった翻訳規則を抽出方法を提案した。また、提案した種々の方法を統合的に用いることにより、パターンに基づく機械翻訳システムのための翻訳規則が柔軟に自動獲得できることを示した。

対訳コーパスを利用して様々な翻訳パターンを自動獲得する手法を提案した本研究は、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。