審 査 の 結 果 の 要 旨

氏　名　　呉　先超

本論文は、「STATISTICAL MACHINE TRANSLATION USING LARGE-SCALE LEXICON AND DEEP SYNTACTIC STRUCTURES（大規模辞書及び深い文法構造を用いた統計的機械翻訳）」と題し、８章より構成される。

統計的機械翻訳は、対訳データから対訳辞書および翻訳規則を自動獲得して、対象となる言語間の翻訳を実現する技術である。本論文では、実用レベルの翻訳システムの構築を目標として、専門用語を含む新語にも柔軟に対応可能な大規模対訳辞書と、文の意味的な構造に基づく翻訳規則の獲得および適用手法を新たに提案し、評価実験を通して有効性を示すとともに、今後の展開の方向性について論じている。

第１章は「Introduction（序論）」である。数多く存在する言語間での自動翻訳の重要性を指摘するとともに、統計的機械翻訳における対訳辞書および翻訳規則の役割を解説している。また、既存の翻訳システムの問題点を例示して本論文における具体的な目標設定を行い、本論文の構成を記している。

第２章は「Background（背景）」と題し、統計的機械翻訳の歴史を概観し、最近の翻訳技術を提案手法との関係を明らかにしながら紹介している。また、括弧表現を利用した対訳抽出に関する過去の研究例について述べている。

第３章は「Semi-supervised Approach for Bilingual Lexicon Mining （半教師学習に基づく対訳辞書獲得手法」と題し、大規模な対訳辞書をウェブデータから獲得するための手法について論じている。具体的には、中国語テキストにおいて新たな翻訳語が使用される際に、対応する英語表現を括弧内に入れて併記する場合が多いことに注目し、これを利用した中英対訳辞書の自動抽出手法を提案している。論文中では、代表的な３つの対訳関係に対応して、英語略記に対する中国語表現の抽出、英語から中国語への翻字、語彙レベルで意味的に対応づけられた翻訳、という３つの抽出法を提案し、さらに、これらを多段階に組み合わせることによって、精度高い辞書の抽出が実現できることを示している。また、提案手法を実際に大規模なウェブページおよび技術論文コーパスに適用して百万語規模の対訳辞書を作成した上で、サンプリングおよびwikipedia見出し語の対応を正解とした比較評価を試みている。これにより、先行研究に対する優位性が示され、提案手法が大規模の対訳辞書の自動獲得に有効であると結論づけている。実験結果の誤り解析によれば、高い精度を実現するためには、英語表現に対応する中国語表現の範囲を正確に同定することが必要である。この問題に関して論文中では、形態素解析結果の利用により性能が改善できることを示している。

第４章は、「Syntax-based Translation Using HPSG（主辞駆動句構造文法による構文解析結果に基づく翻訳法）」と題し、HPSG解析器が生成する構文解析木を利用した翻訳規則の抽出法について論じている。まず、文脈自由文法に対して線形時間で翻訳規則を抽出する手法であるGHKMを、HPSG解析器が出力する構文解析木に適用するための拡張について議論している。また、構文解析木上で述語項構造を含む最小の被覆木を求め、

これに基づき複合翻訳規則を抽出するための線形時間手法を提案している。論文中ではさらに、上記で獲得した翻訳規則に基づき、以下に述べる２つの翻訳モデルの枠組みを与えている。１つは英語を入力言語とする木構造から文字列への翻訳モデルである。HPSG解析器により得られる構文解析木をボトムアップに走査して最適の翻訳文を見つける。もう１つは英語を出力言語とする文字列から木構造への翻訳モデルである。入力文の解析と出力文の木構造の解析を同時に行いながら、木構造から文字列への翻訳規則を逆方向に適用し、最終的な翻訳確率を比較して最適の翻訳文を求める。

第５章は、「Experiments on Tree-to-String Translation（木構造から文字列への翻訳に関する実験）」と題し、論文抄録から得られた百万文の英日対訳文ペアを用いた統計的機械翻訳の評価について述べている。英語から日本語への翻訳タスクにおいて、提案手法を用いることにより、従来手法に対してBLUEスコアによる評価値が大きく改善されることを示すとともに、獲得された翻訳規則の有効性や翻訳誤りについて詳細な解析を行っている。

第６章は、「Experiments on String-to-Tree Translation（文字列から木構造への翻訳に関する実験）」と題し、約二万文の中英対訳文ペアを用いた統計的機械翻訳の評価について述べている。述語項構造に関する翻訳規則を適用するための変換法を定めた上で、中国語から英語への翻訳タスクに提案手法を適用してBLUEスコアにより評価を行った結果、従来手法に対して大きな改善が得られることを示している。また、翻訳結果に対する誤り解析を行うとともに、第３章で抽出した大規模対訳辞書の利用による性能改善についても報告している。

第７章は、「Conclusion（結章）」であり、本論文の成果をまとめている。

第８章は、「Future Directions（今後の展開）」と題し、本論文における検討を通して明らかになった課題や今後の展開の方向について論じている。

以上を要するに、本論文では、大規模対訳辞書およびコンパクトな翻訳規則を用いて、実用的な統計的機械翻訳システムを構築するための手法について論じている。本論文の手法を用いると、大規模なコーパス資源から効率的に有効な翻訳知識を抽出することが可能である。本論文では、実際に大規模なデータセットを用いた評価により有効性を示すとともに、提案する枠組みの中に最先端の抽出アルゴリズムを取り込む際の課題を詳細に検討し解決法を提示しており、統計的機械翻訳分野への貢献が期待される。

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる。