博士（工　学）　范　　莉　馨

学位論文題名

中国語表層構造の特徴を利用した中日機械翻訳手法に関する研究

学位論文内容の要旨

社会の国際化が急速に進むにつれて、国際間における産業情報・技術情報・文化情報の流通は増大しているのもかかわらず、翻訳者の数は世界的にも不足傾向にあり、この需給ギャップを埋めるものとして機械翻訳に対するニーズが高まってきている。現在、日本全体での翻訳の需要は年間数千億円に達すると言われている。また、この量は年々増加している。また、潜在的な翻訳需要は、翻訳のコストを考慮しなければ、数倍にもなると予想されている。中国でも、改革・開放の急速に進むにつれて、翻訳の量がますます増している。

このようなニーズに応えるため、ここ数年の間に機械翻訳システムの商用化が急速に進められてきている。特に、近年のコンピュータ分野の技術進歩は著しいものがあり、半導体技術に代表されるハードウェア技術とこれを支えるソフトウェア技術の進歩により、大規模かつ高速演算が可能となった。また、処理速度の向上とならんで、人工知能の研究の進展が自然言語の取り扱いを容易にさせている。このような情報処理の技術基盤が強化されるに従って、ある程度の機械翻訳が可能となり、そしてその実用化が進められてきている。

しかしながら、翻訳は本来ただ単に１つの言語で表現された文章を別の言語による表現に置き換えるといった単純な技巧で形式的に行われるものではない。すなわち、翻訳は人類のあらゆる文化的産物を背景にして、人間のもつ知識と知能を駆使して行われる。したがって、機械翻訳システムの究極的な姿は、いわゆる人工知能技術を統合した総合的システムということになる。そのような理想的な機械翻訳システムに近づくには、まだまだ遠い道のりがあり、現在の機械翻訳技術では、構文構造の複雑さ、表層形と意味の対応の複雑さ、原言語と目的言語の表現方法の隔たりなどが原因で、正しい翻訳結果が得られないことが多い。

ところで、日本でも中国でも英語を主要な対象とした機械翻訳の研究開発は多いが、中日両言語間機械翻訳の本格的な研究が開始されたばかりであり、英日、英中言語間の機械翻訳と比較すると、未開拓の部分が極めて多い。

本論文では、中日両言語の特徴を有効に利用した中日機械翻訳手法の研究について述べている。良質の中日機械翻訳を実現するためには、日中両言語の特徴、特に機械翻訳の観点から両言語の表現形態を検討しなければならない。すなわち、中日両言語の特徴を把握し、それに基づいて翻訳システムの構造を検討することが必要である。

中国語から日本語への機械翻訳において、一番困難な部分は中国語の形態素解析

である。中国語表層構造の特徴を十分に把握しなければ、質のよい中日機械翻訳システムは構築できないと考えられるが、中国語の固有の特徴により、いまなお中国語の解析に関する研究は十分ではない。それゆえ、本論文では中国語の形態素解析を中心として、中日機械翻訳手法の研究を展開する。すなわち、本研究においては中国語表層構造の特徴を利用した中日機械翻訳アルゴリズムを開発し、システムを構築した。また、アルゴリズムの有効性を確認するための実験および結果の評価を行った。

本論文は、以上の研究成果をまとめたものであり、６章より構成されている。以下にその概要を示す。

第１章では、本論文の背景およびこれまでの世界各国の機械翻訳に関する研究の現状などを概観し、本論文の目的を明らかにしている。

第２章では、機械翻訳の観点から中日両言語の特徴を分析し、以下各章で使われる中日辞書の構造について述べる。また中日機械翻訳における中国語の分類体系、品詞分類と表記を記述する。

第３章では、中国語文の形態素解析に際し、複合語になりうる文字列を抽出して複合語として扱う、複合語の自動合成手法を提案する。これにより、形態素解析の誤りを避け、構文解析における曖昧性を減少または解消することができる。特に原言語文の解析から目的言語文の合成に至る多段の処理を必要とする機械翻訳において、原言語文の形態素解析における曖昧性が後段の処理に大きく影響するので、本手法は有効であると考えられる。また、大量の教科書、科学技術文献中の複合語の調査に基づいて、複合語合成ルールをまとめ、これを組み込んで中日機械翻訳実験システムを構築し、実験により本章で提案した手法の有効性を確認することができた。

第４章では、語素間の結びつきが弱く、その間に他の成分を挿入できる離合詞および関連情報を教科書など大量の実例文から抽出し、この情報を分析し、離合詞の構造上の特徴および挿入成分の検討を行い、これに基づいて中日機械翻訳における離合詞の処理手法を提案する。さらに離合詞を含む３００文を用いて翻訳実験を行った。その結果により、本手法の有効性が確かめられた。

第５章では、長い複文を短い短文に分解することに着目し、中国語文の構造上に重要な連接役割をする要素（関連語）について検討し、この関連語を用いた文の分解に基づく中日機械翻訳手法を提案する。また、関連語の処理を効率的に行うために、多数の中国語教科書および科学技術文献から実データを抽出して訳文関数として整理し、この訳文関数により、入力された中国語文中で関連語が識別されたなら、詳細な文法解析を行わず、訳文関数を用いて直接訳文を生成する手法と、関連語に関係する要素の意味属性を用いて関連語の多義性を解消する手法を提案する。これにより、中国語の長い複文において多数出現しやすい兼用品詞の曖昧性および構文上の多義性がある程度に抑えられると考えられる。さらに、以上の考慮に基づく中日翻訳実験システムを構築し、実験により本手法の有効性を確認した。

第６章では、本研究で得られた成果の総括を行っている。

学位論文審査の要旨

主　査　教　授　栃　内　香　次
副　査　教　授　小　川　吉　彦
副　査　教　授　宮　本　衛　市
副　査　助教授　宮　永　喜　一

学位論文題名

中国語表層構造の特徴を利用した中日機械翻訳手法に関する研究

機械翻訳の研究は近年ますます盛んに行われているが、その多くは英語と他言語間あるいは欧州各国語間の翻訳を対象としており、それ以外の各国語間の機械翻訳に関する研究は未だ少ない。中国語－日本語間機械翻訳に関する研究もその一つで、両国間の交流が深まるのに伴い、その進展が待たれている分野であるが、本格的な研究は開始されたばかりであり、未開拓の部分が極めて多い。

本論文は、このような現況にある中日両言語間機械翻訳について、特に中国語の形態素処理を中心として、中日両言語の特徴を有効に利用し、良質な翻訳結果を得ることのできる翻訳手法を構築することを目的として行った研究の結果をまとめたものであり、以下に要約されるような成果を得ている。

まず第一に、機械翻訳の観点から中国語の特徴を分析し、漢字のみで表記され、語尾変化等の屈折現象のない中国語文を対象とする機械翻訳における最大の問題は形態素解析における誤りと曖昧性の発生であることを明らかにするとともに、それらを解消して高精度な解析結果を得るのに適した中国語の品詞分類体系を求め、それに基づいて翻訳用辞書の構造の検討を行った。

次に、中国語文の形態素解析に際し、複合語になりうる文字列を抽出してあらかじめ複合語化して扱う、複合語の自動合成手法を提案し、これにより形態素解析が容易になり、誤りを減少させることができることを示した。そして、教科書、科学技術文献等の実文書中の複合語の調査に基づいて具体的な複合語合成規則を求めて実験システムを構築し、実文書による実験を行って提案した手法の有効性を確認した。

第三に、単語を構成する各文字間の結合が弱く、その間に他の語を挿入できるという性質を有する中国語離合詞に着目し、教科書等の実文書から、通常使用される離合詞及び関連する情報を抽出して、その構造上の特徴ならびに離合詞の中間に挿入される語の検討を行い、中間挿入語と日本語訳文の関係に基づいて離合詞の処理・翻訳を行う手法を提案した。さらに、離合詞を含む３００文を用いて提案手法に基づく翻訳実験を行い、本手法の有効性を確認した。

第四に、複文の単文への分解法の検討を行い、複文を構成する各文間の連接要素である関連語について検討し、関連語を手掛かりとする文の分解法を考察し、複文中の関連語が識別された際に、詳細な文法解析を行わずに直接訳文を生成する手法を導いた。ついで、関連語に関係する要素の意味属性を用いて関連語の多義性を解消する手法を提案し、これにより、中国語の複文において多数出現する兼用品詞の曖昧性および構文上の多義性を抑制することが可能であることを示した。さらに、以上の手法について、実験により有効性を確認した。

以上の諸点は、漢字でべた書きされるために極めて複雑な処理を必要とする中国語の形態素解析処理においては、欧州系言語の処理に際して通常用いられる方法と異なり、形態素処理の段階で単語の属性を考慮した意味処理を導入する必要があることを示し、また、その中でも主要な要素について具体的な処理手法を提案し、さらに実験によりその有効性を確認したものである。

これを要するに、著者は、中国語文の構造の特徴ならびにその形態素解析手法について、実例

に基づく詳細な検討を行い、中日機械翻訳アルゴリズムの確立に有益な新知見を得たものであり、自然言語処理工学、ことに言語認識・理解工学の発展に寄与するところ大なるものがある。よって著者は、北海道大学博士（工学）の学位を授与される資格あるものと認める。