中国語表層構造の特徴を利用した中日機械翻訳手法

6Q－3

范　莉馨　＊任　福継　宮永　喜一　栃内　香次
北海道大学電子工学研究科　＊（株）ＣＳＫ技開本

１．はじめに

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，前述のように，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，処理速度の工場とならんで，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したがって，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ところである。従って，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明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。従って，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給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る。

機械翻訳の現状では，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。また，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・開発が多いが，中日両言語間の機械翻訳に関する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る。中日機械翻訳を実現するために，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。

中国語から日本語に訳す機械翻訳なので，一番難しいのは勿論，中国語の解析にある。特に，中国語の固有の特点により，いまなお中国語の解析に関する研究が十分ではない。中国語表層構造の特徴を十分に把握できなければ，質のよい中日機械翻訳システムが構築できないと考えられる。それゆえ，本論文では中国語の解析を中心として，中日機械翻訳手法の研究を展開する。即ち，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，システムを構築する。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う。

２．中国語複合語の自動合成

中国語文ではいくつかの漢字が結合して複合語となっている場合が多く，また，特に科学技術文献では次々と新しい複合語が生まれている。一方，中国語文の解析という立場からは，なるべく複合語を単位とする方が曖昧性が減少し，都合がよい。それで，中国語文の解析に際し，すでに明らかになっている複合語だけでなく，複合語になりうる文字列を抽出し，複合語として扱うことが必要だと考えられる。この処理を複合語の自動合成という。

本章では，以下の２点に着目して中国語複合語の自動合成手法を提案する。

２．１　構文解析の曖昧性の解消

例えば，例①”機器翻譯是自然語言処理的一部分”に対し，形態素解析の結果は”機器／翻譯／是／自然／語言／処理／的／一／部分”になる。この中で”翻譯”，”是”，”処理”の３語は動詞の属性をもつ品詞なので，いずれもこの文の述語になりうる。従って，このままでは構文解析の結果に多大な曖昧性が発生することになる。もし複合語自動合成手法を導入すれば，形態素列は次のとおりになり，上述の構文解析の曖昧性を解消した上，日本語訳文も容易に生成される。

例文①：機器翻譯／是／自然語言処理／的／一部分。

↓　↓　↓

S　V　C

訳文①：機械翻訳は／自然言処理の／一部分／である。

２．２　複合語合成による形態素解析の誤りを回避

中国語の特徴により，全ての複合語を固定化辞書に登録することができない。例えば，例②”他用機器翻譯技術文献／彼は機械で技術文献を翻訳する”に対し，もし”機器翻譯”，”技術文献”を複合語として登録，あるいは例①のように合成すれば，形態素解析列は”他／用／機器翻訳／技術文献”になり，これは誤りである。即ち，本例文では”翻譯”は動詞で，文の述語になっている。このような誤りを避けるため，動詞の属性をもつ品詞からなる複合語を固定化辞書に登録せずその都合によって合成する。

上述のような観点から，中日機械翻訳のための中国語文の複合語自動合成手法を提案している．本手法の特徴として，①形態素解析の誤りを避けるため，動詞性の兼用品詞からなる複合語は辞書に登録せず，その都度合成すること，②構文解析の曖昧性を減少または解消するため，できるだけ早い段階で複合語を合成すること，③複合語合成ルールを用意して複合語合成を容易に実現し，処理時間を短縮すること，などがあげられる。

本章では中日機械翻訳システムの構築のため，これらの中国語文の特徴を利用した中国語複合語の自動合成の手法を提案した．更にこの手法に基づく実験システムを構築し，科学技術に関連する400文を対象とした実験を行った．その結果，複合語合成の正解率は95％であった．これにより，本手法の有効性を確認できた．

３．離合詞の処理

中国語の単語には２文字以上からなるものが多い．そのうち語素間の結びつきが弱く，その間に他の成分（挿入成分と呼ぶ）を挿入できる単語がある．このように語素を連結してもよく，離してもよい使い方ができる単語を離合詞とよぶ．例えば，”投稿（投稿する）”は離合詞であり，時には文の中に”投了稿（投稿した）”，”投過稿（投稿したことがある）”，”投了三回稿（三回投稿した）”，”投得了稿（投稿することができる）”.....などの形態で出現する。自然言語理解，特に機械翻訳では，上述の離合詞を正しく認識できなければ，中国語文の形態素解析，さらに構文解析に対しては大きな障害になると考えられる．

A Method of Chinese-Japanese Machine Translation Using Chinese Surface Structural Characteristics
Fan/Lixin,*Ren/Fuji,Miyanaga/Yoshikazu,Tochinai/Koji
Faculty of Engineering, Hokkaido University;
*Machine Translation Development Dept.CSK

本章では，教科書など大量の実例文から離合詞および関連情報を抽出し，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。本手法の要点としては，①離合詞の構造特徴を検討し，離合詞の分類を行ったこと，②実例を用いて文の意味的表現に重要な役割をもつ離合詞の中間挿入成分の類別をまとめたこと，③中日機械翻訳においては訳文関数を用いて離合詞を含む中国語表現を日本語表現に変換すること，との３点が挙げられる。

本章では離合詞を含む300文を用いて翻訳実験を行った。その結果，離合詞を含む表現を正しく処理したのは 283文で，正解率が約93％であった。

４．文の分解に基づく中日機械翻訳

一般に中国語文の実際上の使用度から言えば，単文は極めて少なく，複文の方が多いと考えられる。従って機械翻訳システムの実用という見地から言えば，複文の方が極めて重要である。一方，中国語は単語間の切れ目のない「べた書き」形式で表記され，特に日本語文と異なり，助詞，助動詞などの機能語がすべて漢字で書かれていることと，また動詞，助動詞，形容詞などの語尾変化，すなわち活用現象もないことなどのため，複文の構文解析，翻訳処理をする際に曖昧性が大きく，翻訳の質に大きな影響を与える。

また，中国語と日本語は異なる語族に属し，文の構造および意味表示手段などでは大きなギャップが存在しているので，従来の手法をそのまま利用すると，不都合なところが多い。特に，中国語の兼用品詞は形態上の区別がないので，長い複文の解析はとても難しいと考えられる。本論文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（関連語と呼ぶ）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。即ち，まず入力された中国語文（複文）をいくつかの基本文（単文）と関連語に分け，基本文に対しては構文解析，変換処理などを行い，関連語に対しては解析を行わず変換処理を直接行う。そしてこれらを日本語の構文規則によって合成処理を行い，最終の訳文を生成する。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。さらに関連語の多義性を解消するため，関連語に関係する要素の意味属性を用いて行う手法も提案する。これにより，中国語長い複文における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。

以上の考慮に基づく中日翻訳実験システムを構築した。その後，実験用文献及び教科書から関連語 150個を抽出し，手作業で関連語ﾃｰﾌﾞﾙ及び訳語条件ﾃｰﾌﾞﾙを作成した。また関連語の含む複文200文に対して翻訳実験を行った。その結果，全文の正翻訳率は74％，関連語だけの正翻訳率は91％となった。

５．中国語表層構造の特徴を利用した中日機械翻訳の流れ

例文③：機械翻訳是自然語言処理的一部分，所以自然語言処理的範疇更廣。

①　最長距離マッチング法により形態素解析を行う→表１

②　複合語合成ルールにより中間処理を行う→表２

上表中，４つの品詞は文の述語になりうる。従ってこのままでは構文解析の結果に多大な曖昧性が発生することになる。ここで複合語合成ルールにより中間処理を行う必要がある。

③　関連語の抽出および入力文の分解を行う

上表中，単語「所以」の属性記号は”ｃ３”（ｃは関連語の属性記号，ｃ１は一語性関連語，ｃ２は多語性関連語，ｃ３は両方とも属する関連語の記号である）なので，スタックにこの「所以」を入れる。そして，いまスタック中に要素があるので，これを取り出して関連語テーブルの項目と照合する。勿論，照合が成功した。その後，複文分解が確認され，この関連語を抽出し，単文に分解する。

単文Ａ：機械翻訳是自然語言処理的一部分。

単文Ｂ：自然語言処理的範疇更廣。

関連語：所以

上述の単文に対して翻訳処理を行うと次の訳文が得られた。

訳文Ａ’：機械翻訳は自然言語処理の一部分である。

訳文Ｂ’：自然言語処理の範疇はさらに広い。

引続き関連語の変換処理を行うと，関連語「所以」は多義性がないので，直接に訳文関数を取る。

Ａ，所以Ｂ。　→　Ａ’，それでＢ’。　　.....<7>

最後に，<7>式の訳文関数と単文の翻訳結果を用いて最終の日本語訳文を合成する。本例では相応する日本語動詞の形態的な変化がないので，訳文は次のとおりである。

訳文③：機械翻訳は自然言語処理の一部分である，それで自然言語処理の範疇はさらに広い。

表１　形態素解析の結果

No	単語	属性記号	No	単語	属性記号	No	単語	属性記号
1	機器	ｎ４	7	的	ｕ１	13	処理	ｖｎ
2	翻訳	ｖｎ	8	一	ｍ１	14	的	ｕ１
3	是	ｖ７	9	部分	ｎ０	15	範疇	ｎ０
4	自然	ａ２	10	所以	ｃ３	16	更	ｄ３
5	語言	ｎ４	11	自然	ａ２	17	廣	ａ１
6	処理	ｖｎ	12	語言	ｎ４			

表２　中間処理の結果

No	単　語	属性記号	No	単　語	属性記号
1	機器翻訳	ｎ９	7	自然語言処理	ｎ９
2	是	ｖ７	8	的	ｕ１
3	自然語言処理	ｎ９	9	範疇	ｎ０
4	的	ｕ１	10	更	ｄ３
5	一部分	ｎ９	11	廣	ａ１
6	所以	ｃ３			

６．終わりに

社会の多方面で国際化が進む今日，機械翻訳の必要性はますます高まっている。しかしながら，約半個世紀にわたる機械翻訳に関する研究・開発により，いくつかの機械翻訳システムが実用化されているにもかかわらず，現在の機械翻訳システムはごく限られた分野で実用されているにすぎない。特に，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，良好の翻訳を妨げる多くの問題が残されている。

また，日中両国でも，英語を主要な対象とした機械翻訳の研究開発が多いが，中日両言語間の機械翻訳に関しては本格的研究が開始されたばかりであり，特に中国語表層特徴の分析および中日翻訳を目的とした構文解析がまだ十分ではなく，未開拓の部分が極めて多い。

本論文では中国語表層の構造特徴を利用して，文の分解に基づく中日機械翻訳手法を提案した。また中国語の構文解析の曖昧さを減少または解消，形態素解析の誤りを回避するため，中国語複合語の自動合成アルゴリズムを研究・開発し，中国語離合詞の処理手法を提案した。更に実験システムを構築し，検討を行った。

［参考文献］范莉馨：”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”，北海道大学電子工学研究科博士論文（1993.12）。