特集

2-17

2. 各分野における技術の変遷

# 機械翻訳の過去・現在そして未来

田中穂積

田中穂積：正員 東京工業大学大学院情報理工学研究科

Machine Translation : Past, Present and Future. By Hozumi TANAKA, Member (Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152 Japan).

ABSTRACT

機械翻訳のこれまでの研究開発のあらましと製品化の動きを述べ，機械翻訳技術を機械翻訳の方式，自然言語処理技術，人工知能と関連させて説明している．機械翻訳の方式については代表的な二つの方式の利害得失を比較している．自然言語処理技術は機械翻訳技術に直結している．その現状と将来を展望している．人工知能の研究は長期の視点から機械翻訳技術のブレークスルーとなるべき要素を含んでいること，21世紀を展望したとき，高速情報ネットワークと機械翻訳技術との関連が重要になることを指摘している．

キーワード：機械翻訳，自然言語処理，人工知能，情報ネットワーク

## 1. 機械翻訳システムの過去・現在

1980年代半ばから，我が国における機械翻訳の研究と開発が活発化し，相当数の機械翻訳システムが製品化されている．製品化されているといっても，事前に翻訳すべき文を人手により修正する前編集と，翻訳結果を修正する後編集を前提としているものが多い．1985年ごろからコンピュータのメインフレーマが中心となり製品開発を進め，それらの多くは700万円前後と高価で一般のユーザが入手しがたい状況にあった．これに少し先立ち，科学技術庁と京都大学などが科学技術文献の機械翻訳プロジェクトを開始し成果が出始めていた．1980年代の後半には技術移転を目的に，アジア諸言語を対象とした多言語間翻訳プロジェクトが通産省の主導で開始された．1980年代後半から1990年代前半にかけて，ハードウェア技術の進歩に伴うワークステーションの性能向上により，機械翻訳システムはワークステーション上でも動作可能となり，1990年代前半には価格が100万円前後と低下の一途をたどる．このころ，機械翻訳システムの評価に関する基準作りが始まっている．評価基準については電子工業振興協会の報告書があるので参考になる．

1990年代はPC（パソコン）の時代といわれている．ワークステーションからPCへの移行は，PCの性能向上と共に避けられない流れになっている．1990年に入って，PC上で動作する機械翻訳システムの製品が市場に現れ，機械翻訳システムの価格が一挙に30万円以下に低下し，最近では1万円を切る製品も現れている．興味あることは，これらの低価格の機械翻訳システムの購入者層に変化がみられることである．これまで，翻訳を主たる業務とする翻訳会社や対外との関係が深い業務を行う部署が主に購入していたものが，学生など不特定多数のユーザが購入層の主役になっている．

このような低価格化の動きで中心的な役割を担ったのは，機械翻訳システムの先発メーカではなく，中小のソフトウェアハウスに代表される後発メーカであった．このような急激な低価格化の波は，これまで機械翻訳システムの開発に多大の投資をしてきた先発メーカが投資に見

合う資金の回収を困難にすることを意味した．そのため，先発メーカの機械翻訳システムの研究開発にかける熱意は冷めつつあるというのが現状である．

海外に目を転じると，1980年代にヨーロッパ共同体（EC）の主導で多言語間機械翻訳システムの開発を目指すEUROTRA計画が開始されたが，1980年代の後半に試験的な研究を終え，大きな計画に発展することなく現在に至っている．多言語間翻訳が試みられたのは，ECに加盟している国がさまざまな言語を用いているためである．EUROTRA計画では，新しい機械翻訳システムの開発と並行して，既存の機械翻訳システムSYSTRANを導入し，その改良を行っている．このシステムは現在も改良が続き性能向上と共にシステム使用者は漸増しているという．翻訳する言語対の数も増えている．EU（旧EC）では，現在，より広い観点から，言語産業の育成を目指す言語工学プロジェクトを推進している．1980年代後半にはシーメンス社もテキサス大学と共同で英独の機械翻訳システムを開発している．

機械翻訳システムの開発は米国で始められた．この試みが1960年代半ばに挫折したことはよく知られている．しかし，EUが改良を重ねて現在も使用しているSYSTRANシステムの原型はこのころ開発されたものである．1980年代の後半になると，カーネギーメロン大学，ニューメキシコ州立大学で機械翻訳の研究が始まった．製品化についてはメインフレーマではなく中小のソフトウェアハウスが行っている．米国で機械翻訳の研究を遂行している大学や研究機関の数は意外に少ない．

音声と機械翻訳の結合は，我が国の郵政省の主導で設立したATRで，1980年代後半から研究が開始されている．最終的には同時通訳電話を目指しており，そのための基礎研究が進められている．これは最近では，日本，米国，ドイツの三国共同研究に発展している．

## 2. 機械翻訳の方式

機械翻訳の方式は，トランスファ方式と中間言語方式の二つに大別することができる．中間言語方式は，ソース言語とターゲット言語の双方に共通な，中間言語とよばれるレベルを設定する（日英翻訳の場合には，日本語がソース言語で英語がターゲット言語になる）．このレベルはソース文の意味を理解したレベルであると考えてもよい．多言語間翻訳を目指す場合には，中間言語という言語によらない普遍的なレベルが設定できれば，ソース文を解析して中間言語のレベルの（意味）構造を抽出し，その構造からさまざまなターゲット文を生成することができる（図1参照）．こうしてソース言語とターゲット言語の対を陽に考慮することなく翻訳システムを構築することができるので，多言語間翻訳に都合の良い方式であるとされている．中間言語方式の問題は，中間言語の設計が難しいこと，たとえそれが可能であるにしても，ソース文を解析して中間言語のレベルの構造を抽出する技術が未熟であることである．中間言語方式は，意味理解に深くふみ込んだ理想に近い方式であり，CICC(用語)，EU，カーネギーメロン大学での試みのほかに，NECや富士通の試みもあるが，なお一層の研究が必要である．

中間言語方式ではなく現実的な方式としてトランスファ方式がある（図2）．この方式は，ソース言語の解析結果を，ソース言語に近い表現にとどめ，それをターゲット言語に近い表現にト

### 用語解説

CICC　財団法人国際情報化協力センター（通産省の外郭団体）の英語名の略称．

LFG, GPSG, HPSG　1980年代から現在にかけて発展した言語理論（文法理論）でLFGはLexical Functional Grammar，GPSGはGeneralized Phrase Structure Grammar，HPSGはHead-Driven Phrase Structure Grammarの略．これまで主流であった変形文法理論と異なり，変形という操作を排除し，辞書項目記述を重要視している点に特徴がある．各理論の基本的な考え方に興味のある読者は，P. Sells（著），郡司，ほか（訳），"現代文法理論，" 産業図書，を参照されるとよい．

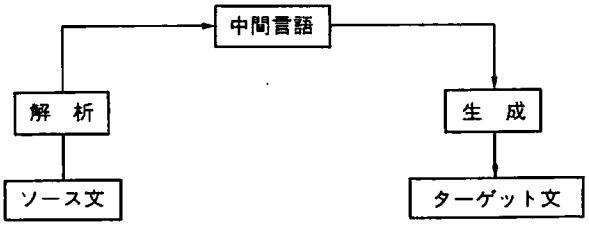


図1　中間言語方式　中間言語はソース文とターゲット文によらない普遍的な表現

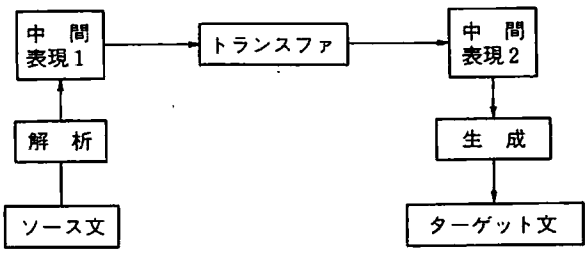


図2　トランスファ方式　中間表現1と中間表現2は，それぞれソース文とターゲット文に近い表現

ランスファ（変換；移行）する．そしてそこから最終的なターゲット言語の文を生成する．この方式はソース言語とターゲット言語の対を考慮するレベルが設定されていることが特徴である．解析レベルをもっと深いレベルに設定すれば中間言語方式に近くなり，解析を行わずに単語単位の翻訳を行うだけにすれば直接翻訳方式になる．

現在では直接翻訳方式の機械翻訳システムは皆無といってよいから，中間言語方式とトランスファ方式とを比較することになる．トランスファ方式の擁護者は，人間の翻訳ではソースとターゲット言語の対を絶えず考えながら翻訳していると主張する．理想的な意味での中間言語方式はまだ確立されていないので，商品化された機械翻訳システムは，解析レベルの差はあるとはいえ，トランスファ方式を採用しているといってよいだろう．但し中間言語方式を標ぼうとするシステムでは，トランスファ方式と比べて，より深い意味解析を行おうとする姿勢が強い．

## 3.　自然言語処理技術と機械翻訳技術

機械翻訳システムの製品化動向，開発動向をこれまで簡単にみてきたが，技術の観点から過去を振り返りつつ，将来を展望してみたいと思う．機械翻訳は複数の自然言語を処理する総合的なシステムである．自然言語処理技術の進歩は機械翻訳技術の進歩に直結している．

自然言語処理技術は，文中の単語を認定し，辞書引きを行い，品詞をみて，文中の品詞がどのような並びをしているかを文法を用いて調べる．文中の単語を認定する仕事を形態素解析とよぶ．品詞の並びが文法にかなっているかどうかを調べる仕事を構文解析とよぶ．自然言語の解析では，同音異義語のあいまい性解消や，係り受けに関するあいまい性解消などを行う必要がある．これらの仕事を意味解析とよぶが，意味解析にあたり前後の文脈を参照することも必要になる．文脈を参照する仕事には省略語の補強，代名詞の参照先の決定などが含まれるが，これらを文脈解析とよぶ．

形態素解析については，英語などの言語では単語と単語の間に空白があるので，大きな問題にはならない．日本語，韓国語，タイ語などにはその空白がなく，形態素解析は大きな問題になる．表層レベルの形態素解析については成熟した技術がある．構文解析についても高速な解析技術が開発されている．しかし，意味解析・文脈解析技術については問題が多い．現在の機械翻訳システムの翻訳精度が不足する最大の要因は，意味解析，文脈解析技術の未熟さにある．

意味解析についていえば，比喩的な表現が多用される文章の機械翻訳は，現在の技術では困難である．個別に対処する以外に方法がない．比喩的な文は，部分的な意味解析の結果を合成して全体の意味を計算するこれまでの意味解析の方法が適用できないからである．そこで，このような表現の少ない科学技術文献やマニュアル文の機械翻訳がこれまで試みられてきた．

意味解析，文脈解析が困難な理由をもう少し考えてみたい．結論をいえば計算機がもっている知識の量が十分でないこと，知識をどう表現するかという技術が十分でないことが挙げられる．更に文脈をどのような形式で記憶しておき

それをどう検索するかなど，未解決の問題が山積している．知識を中心に据えた機械翻訳の方式は知識ベース機械翻訳ともよび，計算機による深い意味理解を目指すものである．この方式のもつ困難な問題を避けるために，最近，具体的な翻訳例を利用した例文ベース翻訳とよぶ機械翻訳の方式が提案されている．これは，翻訳すべき文と似た例文を検索し，その翻訳結果を利用した翻訳を行うものであり，意味解析や文脈解析をバイパスすることができる方式として京都大学の長尾らが提案し注目された．

上記した知識を用いるだけでなく，最近は統計的な知識を用いてあいまい性を解消する研究が盛んに行われている．例えば“crane”という単語には「起重機」という意味と「鶴」という意味がある．それぞれの意味ごとに，“crane”という語を含む文を大量に集め（これをコーパスとよぶ），“crane”という語の前後にどのような語が現れやすいかに関する統計データをとり，この統計データを用いて多義語のあいまい性を解消する．大量のコーパスを集め，そこから抽出した統計的な知識を利用した自然言語処理技術は，これ以外にもさまざまな技法が提案されている．この技法は，コーパスを集めさえすればよいので，大規模な知識ベースを設計する必要がない．自然言語の深い意味解析の安定した技術が開発されるまでのつなぎとしてコーパスベースによる自然言語処理は，速効性があり有効な方法であるといえる．ただ，大量のコーパスの収集が我が国では個別に行われているのが問題で，こうした現状を今後改める必要がある．

文を生成する技術については，一文ごとの翻訳で逐語的な翻訳結果を得たいのであれば，現在の技術で間に合うことも多い．しかし，省略を含む文や適当な代名詞化を行い，こなれた訳文を生成する技術の確立は今後の課題である．

機械翻訳は言語理論の発展に支えられている面もある．言語理論は必ずしも計算機で処理可能なレベルにまで詳細化されていないため，言語理論を駆使した機械翻訳システムはまだ存在しない．しかし，LFG，GPSG，HPSG[用語]などの最新の言語理論は形式化が進み，計算機で扱うことも意識されている．これらの言語理論の今後の発展に注目する必要がある．いわゆる計算言語学との結びつきが強い言語理論であるといえる．そのほかに，翻訳と密着した対照言語学の進展も望まれる．

## 4. 機械翻訳と人工知能

機械翻訳のレベルを今一段向上させるためには長期の視点も必要だろう．人工知能の研究成果を機械翻訳システムに組み込むことは，長期の視点から注目しなければならない．その一端を見てみたい．

言語の翻訳は人間の知的な活動に支えられている．人工知能の研究と機械翻訳の研究とは極めて密接な関係にある．翻訳家が翻訳のエキスパートであることを考えると，機械翻訳システムをエキスパートシステムとしてみなすこともできる．このエキスパートシステムは，翻訳用の知識をもち，この知識を用いて推論し，ソース言語とターゲット言語間の翻訳を行う．翻訳用の知識として，言語的な知識と非言語的な知識がある．言語的な知識の中には，辞書的な知識と文法的な知識がある．非言語的な知識としては日常生活で使う常識などが含まれる．

前者については言語学者や辞書記述者による研究の蓄積がある．辞典はその典型例である．機械翻訳システムで用いる辞書は電子化され，計算機が理解可能な構造をしていなければならない．既存の人間用の辞書から計算機が理解可能な構造の辞書をどのようにして（自動的に）構築するかは，人工知能の分野では知識獲得，学習の問題とよばれている．

機械翻訳システムでは，翻訳用の知識の質だけでなく量的な問題も解決しなければならない．文法規則については，相当広範囲の日本語の文をカバーしうる文法規則の構築に興味を示す言語学者の数が少ないのが問題である．現状では翻訳システム開発者が文法規則を作成しなければならない．そこで，係り受け関係を示す

括弧付きの例文を集め，それから文法規則を自動的に獲得しようとする試みがある．これも，人工知能の研究と大いに関係する．

常識のうち概念間の関係については，「人間は液体を飲む」という一般化した常識を用いて，「花子は酒を飲む」，「太郎は水を飲む」など，人間や液体の下位概念を用いた無数の文の意味的妥当性を判断することができる．最近数万から数十万の数の概念間の関係を体系化したものとして，国立国語研究所の分類語彙表，電子化辞書研究所の開発したもの，プリンストン大学の開発したWordNetが利用可能になってきた．大学での機械翻訳の研究が必ずしも十分でなかった背景には，こうした大規模な知識ソースが手に入らなかったことも一因である．こうした状況は次第に改善されつつある．

概念間の関係についての常識のほかに，社会的な規約，法律，物理法則などの常識がある．これらをどう計算機の内部に表現し，どう利用するかは，人工知能の研究分野では知識表現の問題として研究が進められている．人間の翻訳に，これらの常識が反映されていることは疑いようもない．理解に基づく翻訳が理想であるとすれば，人工知能の研究成果を機械翻訳システムにどう取り込むかは，今後新たなブレークスルーを目指す上で重要なことといえよう．

## 5．情報ネットワークと機械翻訳

これからの情報処理技術は通信技術と結合し技術の幅と奥行を増していくに違いない．コンピュータを中心とした将来の情報処理技術におけるキーワードをあげれば，次の三つに集約できる．

（1）　超並列と超信頼性コンピュータ
（2）　大規模分散知識ベースと知的インタフェース
（3）　超高速情報ネットワークインフラストラクチャ

機械翻訳は（2）と（3）に関係する．インタネットに接続されている端末数が急激に増大している．21世紀初頭には，超高速情報ネットワークが世界的な規模に成長することは間違いない．音声・映像メディアだけでなく，さまざまな言語で書かれた大規模な言語メディア知識が，ネットワーク上の世界各所に分散して蓄積していく．言語の差異によるギャップを解消する機械翻訳技術は，分散した多種多様な言語メディアを検索する技術としてこれから重要な役割を果たすだろう．具体例を挙げれば，既にインタネット上でのモザイク検索や，電子メールの翻訳に機械翻訳技術を応用する試みが始まっている．電子図書館的なものも世界各国に建設されるだろう．世界に分散した電子図書館にアクセスし必要な情報を検索し取り出す場面では，機械翻訳技術は今後重要な役割を果たすと思われる．そのためにも，機械翻訳に関する一層のブレークスルーが望まれる．

## 6．おわりに

以上みてきたように機械翻訳技術は我が国がリーダシップをとり得る数少ないソフトウェア技術である．しかしそこに含まれる問題は困難で長期を要す性質をもっている．現在商品化されている機械翻訳システムの翻訳性能は到底十分とはいえない．人手によるソース文の修正（前編集）や，ターゲット文（翻訳結果）の修正（後編集）が必要である．我が国では機械翻訳システムの開発意欲が衰え，むしろその反動として機械翻訳技術の進展にとって憂慮すべき現象が起きている．このような現状を打破するためには，今一度初心に帰り，速効性のみを追求した機械翻訳技術の開発ではなく，長期的な視野に立つ基礎的な研究を行うべきであろう．既に述べたように，そのための知識ソースは着々と整備されつつある．これらを利用した機械翻訳システムの研究を大学などでも行うことが可能な環境が整備されつつある．このような基礎的な研究から，機械翻訳の新たなブレークスルーが必ず生まれると筆者は考えている．その意味で，機械翻訳は「あの技術」ではなく「今もホットでこれからの技術」であり，我が国ではそうあらねばならないと考えている．

最後に機械翻訳の研究をこれから行おうとする読者は，"Machine Translation" というジャーナルや，2年に一度開かれる "Theoretical and Methodological Issues on Machine Translation"，"International Conference on Computational Linguistics" などの会議録を参照されるとよい．そこから更に孫引きして必要な文献を探すとよいだろう．最新の製品化動向を探るためには機械翻訳協会の活動を参考にするとよい．

田中 穂積（正員）

昭39東工大・理工・制御卒．昭41同大学院修士課程了．同年電気試験所（現電総研）入所．以来，OS，自然言語処理の研究に従事．現在，東工大大学院情報理工学研究科教授．工博．著書「自然言語解析の基礎」など．