第２回機械翻訳シンポジウム　電子協主催　1992.10

機械翻訳技術：利用の現状と将来

田中穂積
東京工業大学工学部情報工学科

［はじめに］

機械翻訳システムが商品化されてから既に７年以上が経過した。当時、機械翻訳システムは、ジャーナリズムにもしばしば登場し、それと共にさまざまな反響があった。これは、多くの人にとって、翻訳は苦手な作業であり、それを代行する機械の出現により、これまで「夢のまた夢」と思われていたシステムが現実のものになったという（？マーク付きであったかもしれないが）驚き・喜びがあったからだろう。近い将来、翻訳家は職業として成り立たなくなる、外国語を勉強する必要がなくなる、などといった予測、反響もあった。ともすれば、こうした過大ともいえる期待を担って機械翻訳システムは世にデビューしたといってよい。しかし、翻訳しようとする文を機械翻訳システムに入力すると、翻訳結果が直ちに出力されるといった、多くの人の思い描く機械翻訳システムのイメージと、実際の機械翻訳システムとの間に、大きなギャップが存在していた。

ここで、外国語を学ぶことが、どんなに難しいことであったか、またそれを実際に使いこなすことがどんなに困難なことであったかを思い出してほしい。海外旅行の経験者なら、はじめて海外の地を踏んだ時のことを、思い出してほしい。母国語以外の言語を理解し、使いこなすことがどんなに困難で大変なことか、思い知らされた人が多かったに違いない。翻訳は、人間の知能をもってしても、容易な作業とは到底言えない。著名な翻訳家にしても、誤訳を免れることはない。翻訳とはそもそもそうした困難な作業なのである。

［機械翻訳システムの限界］

容易に想像できることであるが、現在の機械翻訳システムは、あらゆるジャンルの文章の翻訳が可能であるとはいえない。たとえば、詩歌、比喩等を含む文を機械で翻訳することはできない。現在機械翻訳システムが対象とする文は、科学技術関連の文献、マニュアル文などで、分野を限定する必要がある。こうした分野限定型の最近の機械翻訳システムの性能についても、視点の相違により、二つの考え方があるように思われる。一つは、「使う気にならない」というものであり、もう一つは、「意外に使える」というものである。前者の反応は、機械翻訳システムに過大な期待を抱いていた人達に多い。後者の反応は、機械の現在の能力・限界についてある程度の知識を持っている人達に多い。発売当初の機械翻訳システムと比べて、最近のものに対する、筆者の反応は幾らか甘めかも知れないが、分野を狭く限った場合、後者に属する。

後者の反応をした人達にしても、機械翻訳結果に一切手直不要ということではなく、現在の機械翻訳システムの性能で足りないところを、如何にして人間が補っていくかという問題が、将来の機械翻訳システムで重要であるという認識は持ったことであろう。

ヒューマンマシンインタフェースの問題である。これは、使用経験を積むにつれて、次第に改善されようが、システム開発者の側が、使用者からの要求を注意深く聴く姿勢がなければならない。さらに、システム開発者が、自己の開発したシステムを徹底的に使い込む必要がある。それにより、システムの使い勝手を良くするためには、何をなすべきか、どのようなヒューマンマシンインタフェースが必要かを、開発者自ら知り、考える必要がある。

次に、機械翻訳システムそれ自体の性能向上をはかるため、何をなすべきかを考えてみたい。

[辞書の問題]

辞書の充実が機械翻訳システムの性能を左右する大きな要因であることは、これまで多くの人に指摘されてきたところである。知識の問題である。最近、通産省の主導で、基盤技術促進センターの補助を受けた電子化辞書研究所が、大規模な電子化辞書を開発している。ここで開発された電子化辞書が公開されれば、これまで辞書の問題のために、機械翻訳の研究のアクティビティが落ちていた大学や研究所で、再び機械翻訳の研究が活発化すると思われる。それにより、技術の底辺が上がることが期待できる。辞書に関連して、次のことも十分考慮する必要がある。一般に、完全な辞書というものはありえない。時代の変化につれ使用する言葉も変わる。新しい言葉も生まれる。したがって、辞書は絶えず更新されねばならない。辞書の更新が容易なシステムは、機械翻訳システムの運用上、拡張性の面からも極めて重要であるということを、忘れてはならない。辞書の更新は、機械翻訳システムの開発者ではなく、システムの使用者が容易に更新可能でなければならない。これは、前述したヒューマンマシンインタフェースの問題とも関係する。

[統計情報の利用]

統計を用いた機械翻訳技術が注目されたことがある［IBMのPeter Brownら］。これは、文の対（S，T）を考え、この対に対して、Pr(T|S)を割り当てる。Pr(T|S)は、ソース言語の文Sが与えられて、翻訳家がターゲット言語の文Tを作り出す確率と考えられる。このような二つの言語の文の対の集合（対訳例文集）のもつ統計的な性質から、ソース言語の文が与えられたとき、もっとも確からしい翻訳結果を作り出す。このとき、音声認識で使われるn-granモデルなどを使う。当初この方法は、単文レベルの実験であるとは言え、研究者の予想を上回る成果を挙げたこともあって、一時、翻訳システムの作成に、言語学は不必要かもしれない、対訳例文集（コーパス）さえあれば、統計的な手法により機械翻訳が可能かもしれない、などといった幻想をいだく研究者もいた。現在では、こうした考え方は否定される傾向にあり、統計的手法のみでは、明らかに限界があると考えられるようになった。しかし、翻訳に、直接、上記した手法を取り入れるのは無理かも知れないが、確率文法の考え方など、言語の持つ統計的性質を、機械翻訳システムの解析結果の優先度に反映させる試みや研究は、これから極めて重要である。人工知能の研究で用いられている確率付き推論の方法なども、機械翻訳の立場から検討されてよい。

［翻訳例の利用］

翻訳例を利用した機械翻訳システム構築が考えられている［京大の長尾、佐藤ら］。これはmemory based MT, example based MT, analogy based MTなどと呼ばれているが、人工知能の分野の、事例ベース推論とも関係が深い。。たとえば，"I take a taxi" という文の日本語訳が「私はタクシーに乗る」であるという翻訳例を用いて、"He takes a bus" を翻訳することを考えてみようる。もし、"he"と"I"が対応し、"taxi"と"bus"とが対応することが分かれば、前者の対はどちらも「人間」であり、後者は「乗り物」であるという共通点があるので、"I take a taxi"の翻訳結果を利用して、「彼は自動車に乗る」という翻訳結果を得ることができる。

このような翻訳は、二つの文の類似性を計算することが必要になり、それのため、概念の上位・下位関係を用いることが多い。こうした方法は、exact matchでない限り、常に、特定の翻訳例を一般化したり特殊化する必要がある。我々は、翻訳を行う場合、もしある特定の翻訳例に対して、常にある特定の一般化を行うことになれば、一般化したパターンを学習して翻訳するだろう。先の例でいえば、「HUMAN take VEHICLE」 は「HUMANがVEHICLEに乗る」と訳すというパターンで覚えておくだろう。したがって、こうした一般化を自動的に行う（学習する）機構が必要になるだろう。一たび一般化が行われたら、これは、通常の翻訳方法で翻訳することになる。最後に、"I"が"he"に対応するなどという対応関係を知るためには、翻訳例文をあらかじめ解析して蓄えておく必要がある。解析結果を含めた対訳例を利用する方法についても研究がなされている［IBMの丸山］。

［定型パターンの利用］

特定の分野では、特殊な言い回しや定型的な言い回しで、簡潔に事実を述べることがある。たとえば、気象情報に関するラジオのニュース文には、定型化された言い回しが多数含まれる。このような文に対して、あらかじめ多少一般化したテンプレートを用意して解析し、このテンプレートに対する翻訳結果を利用して、直ちに翻訳する方法が考えられる。NHK技研では、外電経済文に対してこの方法を適用し、翻訳精度の著しい向上が得られたことを報告している［情報処理学会第４５回全国大会］。テンプレートを用いた解析は、人工知能の研究分野では意味文法と呼ばれていた。この方法は、テンプレートマッチを基本とし、テンプレートマッチに失敗したときのみ、通常の、文法と辞書を用いた解析・文生成を行うので、翻訳速度も向上する。意味文法の考え方と、翻訳例文を利用する方法とを組み合わせた方法であるが、これは、特殊な分野での特殊な言い回しの翻訳を行う場合の、速効的な翻訳精度向上の方法であると結論することができる。しかし、一般の文を翻訳する技術としてこの方法を応用するには問題があるだろう。

［解析・生成技術］

解析技術は、形態素解析、構文解析、意味解析、文脈解析に分けて論じられることが多い。日本語などのように、語と語の間に空白をおかない言語の文の形態素解析が問題になるが、この場合、品詞より細かい文法的カテゴリー相互の接続関係を表（接続表）にしたものを用いて、きめ細かな形態素解析を行う方法が良く用いられる。しかし、このレベルでの解析と同様なことが、構文解析でも重複して行われることがある。部分的

に得られる形態素解析結果を、構文解析側に送りながら、両者を並行的に動作させて解析する方法は考えられるが、両者を今少し融合した方法も考えられている［田中ら］。これは一般化ＬＲ法（ＧＬＲ法）を用いて、両者の融合化をはかるものである。ＧＬＲ法では先読み語とＬＲ表とを用いる。そもそもＬＲ表は、先読み語と現在解析中の語との接続関係を表にしたものである（follow関数を考えよ）。したがって、接続表に書かれた情報をＬＲ表に反映させることが可能である。接続表の情報が反映されたＬＲ表を用いて構文解析を行うと、その過程に自然に形態素解析が含まれる。我々は現在この方法の有効性を検討しているところである。

意味解析・文脈解析についても問題が多い。特に、種々のレベル（語彙的曖昧性、構造的曖昧性など）の曖昧性解消の問題、省略語の補強や、代名詞、指示詞の指すものを決める問題などに、今後真剣に取り組む必要がある。言語学者からの知見で利用可能なものがあるかどうかも、徹底的に調査する必要がある。生成技術についても、適切な省略や代名詞化を行った、こなれた翻訳文を出すために、今後より一層の研究が必要になるだろう。

［中間言語方式］

先に述べた、［統計情報の利用］、［翻訳例の利用］、［定型パターンの利用］は、意味解析、文脈解析には本質的に困難で未解決の問題が多数含まれているため、それらをサイドステップした速効的な方法として最近注目されている技術である。しかし、翻訳例を利用する場合にも、翻訳対象文に "I take it" などのように指示代名詞（"it"）を含む場合には、"it"が何を指しているかを知る必要がある。省略語の補強や、代名詞、指示詞の指すものを決めるためには、知識を用いた深い推論（解析）が必要になるが、前項で述べたように問題は多い。深い推論は、言語理解の問題とも関係し、中間言語方式の目指す方向と一致する。困難ではあっても、この問題の研究が更に進むことを期待したい。多くの国が参加したＣＩＣＣの多言語間機械翻訳システム構築の経験も、現在テクニカルレポートとして取りまとめ中と聞く。早期に公表されることを期待したい。

［終わりに］

現存する機械翻訳システムを評価する方法の検討を進める必要がある。これは、使用者側からの評価基準、開発技術の側からの評価基準の二つに分けて検討する必要があるだろう。電子協の委員会での検討結果の詳細が、本シンポジウムで報告される。詳細はそれに譲りたい。ハードウェア技術については、これから相当数の処理装置が利用可能になると予測される。機械翻訳システムと（超）並列処理の関係についても研究する必要がある。私見によれば、構文解析については、十分高速なアルゴリズムがすでに開発されている。構文解析を並列に行う必要は現在のところあまり感じられない。並列処理は、むしろ多数の例文検索や、意味解析、文脈解析の場面で必要[illegible]は、電子化辞書を含む大規模知識ベースの存在が前提になる[illegible]は否定できないが、現実には、解析技術の進歩が、（分野を狭く限定し[illegible]は別として）機械翻訳システムの性能向上に最も大きく貢献すると思わ[illegible]