特集・機械翻訳の現状と未来

機械翻訳のむずかしさと課題

新田義彦／田中穂積

一　はじめに

機械翻訳（machine translation）という言葉に新奇性を感じる人は、今日では少なくなっているのではないだろうか。市場には様々な形式・規模の翻訳システムが数多く出回るようになってきている。「機械翻訳」が世間に知られるようになっていくという過程は、見方を変えれば、機械（computer）が行う翻訳は人間の行う翻訳とは違う、あるいは、機械はベテラン翻訳者と同じような翻訳文を完全自動で作り出すことは出来ない、という認識が広まっていくことでもある。これは何も皮肉な見方をしているわけではない。「機械による翻訳」の本質を丁寧に調べて、「人間による翻訳」との相違点をなるべく正確に知ることこそが、（現状の）機械翻訳システムを上手に使いこなしたり、より良い（将来の）機械翻訳システムを作り上げていくための基礎であると考えるからである。

このような観点に立って、本稿では左記のような項目ごとに機械翻訳が抱えている課題を見ていきたい。

(1)　人間による翻訳と機械による翻訳――機械の翻訳の仕方は、人間の仕方とどのように違うのであろうか。

(2)　言語のモデル化と機械翻訳のメカニズム――機械はどのようにして言語（原文）をその内部に取り入れ、どのような処理を施して訳文を作り出すのであろうか。

(3) 多義性の解消と言語理解――人間が日常用いる言語（自然言語）には多義性（曖昧性）が潜んでいる。それは人間の言語運用には効率をもたらすが、機械にはトラブルをもたらす。言語を理解できない機械は、それをどのように解消するのであろうか。

二 人間による翻訳と機械による翻訳

「翻訳（translation）」を一応形式的に定義してみると：翻訳とは、ある言語(A)によって表現されている意味内容（メッセージ）を、その実質を最大限度保存しながら、別の言語(B)による表現に移し換えることである（新田、一九七九 p. 735）、

のようになる。このような定義をさらに進展させて、人間による翻訳（HT：Human Translation）のメカニズムを形式化しようとすると、非常な困難に直面する。例えば（1e）を（1j）に日本語訳することは、ほとんど全ての人にとって無意識・瞬時のうちに可能であろう。

（1e） Thank you.

（1j） ありがとう。

ところが、どうして"Thank you."が理解できて、それが日本語の「ありがとう。」に訳せるのかということを、計算機のプログラムの形式で記述することは困難である。［人間の］翻訳者ならば"Thank you."を耳にしたときに、その意味内容として喜びの感情のようなある種のイメージをもつであろう。ところが、このイメージを形式化・記号化する方法がほとんどわかっていないので、機械翻訳では利用できない。

したがって、現時点で機械翻訳システムを構築するには、ベテラン翻訳者のやり方をそのまま真似するわけにはいかない。ところで、我々が外国語を外国人として初めて学び始めた頃のことを思い出してみよう。辞書あるいは単語帳を片手に、次のような「対訳語の置換」と「語順の置換」によって翻訳をした経験があるだろう：

（2e） S（主語）＋V（他動詞）＋O（目的語）

において英単語と日本語の単語の置換をしながら、

（2j） S＋は／が＋O＋を＋V［する］。

を導くのである。このやり方は逐語訳とか直訳（literal translation）と呼ばれることがある。

あえて極論すれば、今日の機械翻訳システムのやり方は、逐語訳・直訳方式を高度に発展させたやり方と言える。機械翻訳（MT：Machine Translation）が行う翻訳とはどのようなものか、感触をつかむために訳例を見てみよう。ただし、左記の訳例は特定のシステムの出力ではない。前述の直訳型のMTの機能を理想化して構成した仮想MTによるものである。本稿の目的は異種MT間の訳例比較ではなく「人間の翻

訳（HT）と機械の翻訳（MT）との比較」にあるからである。以下においては、ダッシュ（′）の付いたインデクスを持つ文が仮想MTの出力であり、その他の無印のインデクスを持つ文は、原文あるいはHTの結果である。

(3j) 国境の長いトンネルを抜けると雪国であった。

(3e′) After passing through the long border tunnel, it was the snow country.

(3j) は川端康成の『雪国』の冒頭の一文である。(3e′) が不適当であることは、サイデンステッカー氏の名訳 (3e) と対比するとわかりやすい。(3j′) は (3e) の仮想MT訳である。

(3e) The train came out of the long tunnel into the snow country.

(3j′) 列車は長いトンネルを抜け出て雪国にはいった。

(3e′) が不適当である理由は、「何か（it）が国境の長いトンネルを抜けた後で、その何か（it）が雪国になってしまった。」という風に読まれてしまうからである。

少しばかり理屈っぽい分析をしてみる。(3j) は二つの述語、「抜ける」と「であった」を含んでいるが、その対訳は 'passing through' と 'was' として (3e′) に埋め込まれている。さらに目的語「トンネル」、補語「雪国」、および接続詞相当語「～すると」なども (3e′) においては、それぞれ 'tunnel'、'snow country'、および 'after' として適正に訳されている。ここまではMTも成功している。問題は (3j) の二つの述語の主語が、二つとも省略されていることに起因している。(3e′) では「は」という一つの語だけで省略主語を復元したために誤訳となったのである。

HTの結果である (3e) では、この問題は見事に解決されている。何故だろうか。(3j) の分析的な解釈は「〈主人公が乗っている列車〉が国境の長いトンネルを抜け出たときに、〈列車の窓を通して主人公の目に飛び込んで来たもの〉は、雪国であった」というようなものであろう。人間（HT）は「主人公の視点」という言外の状況を察知して、文の意味――不正確な言い方ではあるが、文が表現しているイメージ――を理解することができる。理解した結果から相手言語の文法に従って、文を再び生成したものが翻訳文となっている。HTは本質的に「意訳」、つまり「意味理解形の翻訳」と言える。

これに対して、今日の機械（MT）は、既に述べたように「構造束縛形の翻訳」しか出来ない。つまり、原文の構造を形式的に解析し次に変形しながら、対応する語や句をはめ込んでいくというやり方である。意味に対する配慮はもちろん精一杯行うが、語句の並びや接続の意味的整合性をチェックする程度が限界である。文あるいは文章全体の意味を理解することは、到底無理である。複雑で精緻な意味表現図式を展開したり、外界世界の常識をデータベース化する工夫などをし

て、意味理解形のMTを実現しようとする努力が行われている。しかし、現状は語彙と話題に非常に強い制限を持つ研究モデルの域を出ていない。

構造束縛形のMTの出力としては、(3e) の出来映え (performance) はほぼ満点である。あとの修正——例えば 'it was' を 'the train came into' に置換することなど——は人間の仕事（校正）としても仕方がないといえよう。今日の「構造束縛形の」機械翻訳システムにとっては、適正な訳文を作り出すのは非常に大変な仕事なのである。

三　言語のモデル化と機械翻訳のメカニズム

MTは、直訳あるいは逐語訳をどのようにして作り出していくのか、ということを見るのが本節の目的である。MTのメカニズムを知ることはまた、MTの構築の難しさ、MTの改良の苦労、MTに言語理解機能を持たせることの難しさ、などを知ることでもある。これらの知見は、将来の言語理解形MTを考える、ヒントを与えてくれるかもしれない。

原文の意味を理解できないMTが、訳文を作り出すための唯一の拠り所は、原文の構造（構文＝syntax および語彙項目＝lexical items）そのものである。文字あるいは語の連鎖としての表層文から、その内部構造を抽出するために、あらかじめ用意しておく枠組のようなものを「言語モデル」という。したがって「言語のモデル化」とは、自然言語表現の多様性を縮退させ、計算機による形式的処理が可能となるように変形することと言えよう。

機械翻訳の原理的メカニズムは、言語モデル上に投影された入力文（原文）の構造を、色々と変形操作しながら出力文（訳文）を組み上げることである。この変形操作を実行する際に、コンピュータが参照する代表的な知識ベースが、「辞書」と「文法」である。この他に、「世界常識」を表現している形式的知識を利用して翻訳精度を向上する工夫も行われている。ここで理解していただきたいポイントは、機械翻訳のプロセスは原文の構造の強い束縛の中で進行しているという点である。これは、原文の意味を理解し、頭の中にイメージを浮かべてから、それを基にして流暢な訳文を作り出すといったベテラン翻訳者（HT）のやり方（例えば中村、一九七三を参照されたい）とは程遠いものである。

辞書の主な役割は、原文の語や句の文法的カテゴリー分類——例えば、名詞、動詞、などの品詞コードとか、具象、行為、状態などの意味コード——を与えたり、対訳語句を与えたりすることであり、文法の役割は、可能な変形のパターンを示したり、受理してよい入力文のパターンや、生成すべき出力文のパターンを示すことなどである。

ここで注意すべき点は、日常の言語学習で用いられる辞書

や文法と、MTで用いられるそれらとの違いである。MT用の辞書は、対訳の列挙による二言語間の橋渡しだけではなく、システム内での処理（解析・変換・生成）の手掛かりとなる文法的情報を、語彙ごとに与えているという点である。MT用辞書は、通常その使用目的別に、解析用辞書／変換用辞書／生成用辞書といったように、複数個用意される場合が多い。当然、一つの語句には二つ以上の対訳語が対応したり、複数の文法カテゴリーを兼備する、といった多義性が存在する。多義性に対しては、出来る限り正確な制約条件（例えば、依存関係を結び得る語句の意味コードなど）を併記しておく必要がある。受理してよい原文の構造や、生成すべき訳文の構造なども考慮に入れながら、詳細な分析的記述項目を持つ辞書を作らなければならないという点が、MT構築の難しさの一つと言えよう。

MT構築のもう一つの——というより最も特徴的な——難しさは、MT用文法の構築にある。言語学における多くの文法は、適正な（あるいは自然な）文の生成（構成）や不適格な文の抑制を、説明しようとする規則体系あるいは理論体系である。これに対して、MT用文法は、入力文の解析（構造を抽出してその結果をモデル上に投影すること）、様々な多義性の解消、構造の変換（異言語間における構造的差異の解消・橋渡し）、出力文の生成（モデル表現から、文字列・単語列としての訳文を組み上げること）、などを行うための知識体系である。この体系には言語現象の外にある知識（extra-linguistic knowledge）（例えば、世界常識や専門知識など）も含まなければならないから、なおさら厄介である。

MT用文法の構築にあたっては、計算機プログラムによる処理がしやすい形式、人間（文法作成者）が判読・記述・改訂しやすい形式などの、工夫が必須である。そのために、テーブル（表）のような書式／エキスパート・システムで用いられるような「IF（条件）＋THEN（操作）形のルール」のような書式／木（tree）や網（network）のような図式とその書式／これら相互間の変換プログラム（文法コンパイラ・エディタ・コンバータ）など、様々な記述方法の工夫が行われている。いずれにせよ、機械翻訳のメカニズムの本質部分は、「文法」によって記述されるという点が肝心である。

MT用文法の構築は通常、入力言語と出力言語の構造を比較して、両者間の構造的平行性（structural parallelism）を最大限度利用するようにして行われる。前節の（2e）と（2j）の関係は、非常に単純化された素朴な平行性の例である。しかし、異言語間には通常、発想の差異による大きな構造的ギャップ（Nitta, 1986a&b）があり、単純な平行性は見い出しずらい場合が多い。MT用文法は、いきおいアドホックなヒューリスティック・ルール（百%正しいという形式的厳密性

は保証できないけれども、経験的にみて大抵はうまく作用するルールのこと）の集合となりがちである。

MT用文法作りの難しさ、したがって機械翻訳の難しさは、まさにこの点――異言語表現間における発想のギャップを、形式的規則によって強引に解消せねばならぬ点――にあると言えよう。

四　多義性の解消と言語理解

前節では、文表現の発想の差異に起因する機械翻訳の難しさについて述べた。実は、もっと身近で到る所に発生する難しさがある。それは、自然言語の「多義性」とか「曖昧性」と言われるものである。多義性の背景には、人間が言語を運用する際には、なるべく簡単な形式で、やかましい論理性には気をつかわずに済ませようといった、「節約原理」のようなものが存在する。よって、多義性を悪者扱いして、MTが受理できる入力文から追い出してしまうわけにはいかない。

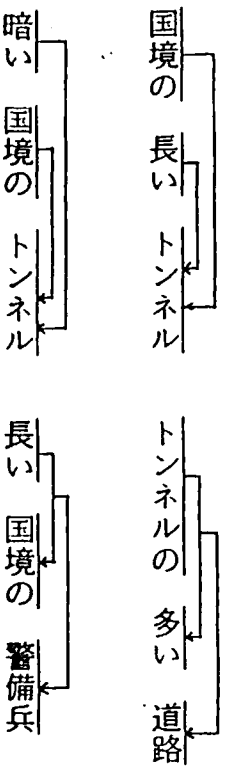
多義性は、単一言語の範囲内で［人間が］言語運用をしている際には気付かずに済ませられるが、翻訳、つまり他の言語に言い換えようとすると、とたんに顕在化するものが多い。例えば、「良い」という日本語を英訳しようとすれば、'good', 'smart', 'fine', 'beautiful', 'well', 'gentle', 'efficient' などのうちの一つを多義性の解消により選択しなければならなくなる。選択は、主語や被修飾語の意味的な性質（例えば、判断、思考、天候、外観、健康、人格、効率）に依存して行われる。

第二節では、機械翻訳システムは意味を理解せずに、原文の構造のみを拠点として言語表現の変形をするだけと極論したが、MTが上記のような多義性を解消するためには、単純な構造変換と語彙置換だけでは済まされない。何らかの形で言語理解を近似する必要がある。最も広く行われる言語理解の近似法は、連続する語句間における意味素性（semantic feature）の両立性（consistency）を判定する方法である（意味素性を判定する具体的なロジックやアルゴリズムについては、例えば田中・新田、一九六を参照されたい）。例えば、'take' という英単語は、その使用環境に応じて、乗る（take a bus）、乗っ取る（take a plane with a gun）、持って行く（take him the book）、のように訳し分ける必要がある。そのためには、主語（subject）と目的語（object）を占める名詞の意味素性ごとに、異なる訳語を記述しておく必要がある（長尾・田中、一九五五 pp. 51-55）。この記述は通常、辞書の 'take' という見出しの下に与えられる。MT用辞書作りの難しさの典型例とも言えよう。

多義性は単語レベルに止まるわけではない。もっと手強いのは、構文レベル、特に文の係り受け構造に関わるものである。第二節で取り上げた例文、(3j) と (3e) を再び見直して

みよう。(3e)においては、「国境の長いトンネル」という日本語の名詞句を、'the long border tunnel'というように、MTは〔紛れ当りかもしれないが〕冠詞の選択や形容詞の位置を誤まらずに見事に翻訳している。

名詞＋の＋形容詞＋名詞／形容詞＋名詞＋の＋名詞という典型的な日本語の名詞句の構造には、概略次に示すような係り受け(依存)構造の多義性が潜んでいる(新田、一九八七 p. 888)。さらにイディオム的な意味の多様性にも対処する必要がある。したがって、〔近似的〕意味理解処理機能の装備はMTにとって必須と言える。



構文レベルの多義性――複数の語句の連鎖に関する多義性――の解消には、「イディオム登録」の方法が特効薬としてしばしば利用されている。つまり、規則によるめんどうな形式的変換などはやらずに、入力言語の語句と出力言語の語句をダイレクト・カップリングする方法である。このカップリング記述(語句の直接対応表示)もまた、辞書の中に登録する必要がある。この特効薬の副作用は、メモリーの爆発的浪費(combinational explosion)である。文法的サブカテゴリーなどを導入して、複数のイディオムを一つにまとめる工夫をしたとしても、なお膨大な組合せが存在する。さらに、イディオムを〔MTに使える形で〕網羅的に、かつ体系的に拾い出すことも大変な仕事である。机上版辞書に記述されているイディオムは、MTという観点からすれば、ほんの一部分の典型例があげられているにすぎない。人間(HT)の場合には、典型例から他の色々な具体例が容易に推論できるからそれで十分なのである。機械(MT)の場合には、類例から推論して適訳を作り出すわけには行かない。

以上見てきたように、現状水準の計算機プログラム(MT)は、人工知能とか機械知能などという名称がついてはいるが、その知能のうちの言語理解に関わる部分は、人間のそれには程遠い物である。この「言語理解能力の不十分さ」こそが、「機械翻訳の構築の難しさ」の実体であり、MT用文法やMT用辞書作りの難しさの源である。

五　おわりに

機械翻訳のむずかしさは、自然言語表現の多義性とそれを

解消するのに必要な「近似的」言語理解能力の実現という対比から捕えられることを示した。したがって、機械翻訳の今後の研究課題は、計算機処理への応用という目的のもとに、言語理解について多面的に研究すること、と言えよう。

この課題は言語理解プロセスの認知科学的分析とか計算機論理による近似といったような狭義に捕えるべきではない。機械処理がしやすく、かつ人間にとっても読みやすい文(章)の構造の研究、入力文を書き換えたり、不正確な出力文を校正したりする方法や援助システムの研究など、ヒューマン・インターフェイスとの関わり(新田、1986b)にも目を向けるべきである。また異言語間の対比分析(comparative study)を膨大な文例について系統的・網羅的に実行し、翻訳用文法や翻訳用辞書(特に、イディオムの収集)に結実させることも、重要な課題である。このためには、言語学、計算機科学、情報処理、心理学、認知科学、などの異分野の研究者の実質的な協力が肝要である。

いずれにせよ、高性能なコンピュータがふんだんに利用できる今日において、機械翻訳は、科学・工学・技術の多方面にわたって、包括的で歯応えのある、魅力的な課題を数多く提供していると言えよう。機械翻訳がこれまでに成した一つの貢献は、「言語学」というややもすれば専門領域に押し込まれがちであった学問分野を、コンピュータ科学・技術といった、一般人と係わりの深い領域に引っ張り出してきて、数学や物理などと同様の基礎部門にしつつあること、と言ったら言い過ぎであろうか。

【参考文献】

長尾真、田中穂積(一九八五)、『科学技術文における曖昧さの構造の計算機による検証』昭和五七、五八、五九年科学研究費補助金(言語の機械処理における標準化)研究成果報告書、東京工業大学工学部

中村保男(一九七三)、『翻訳の技術』中公新書 no. 345、中央公論社

Nitta, Y. (1986a), "Idiosyncratic Gap: A Tough Problem to Machine Translation." *Proceedings of COLING'86* (The 11th International Conference on Computational Linguistics), Bonn, West Germany pp. 107-111

Nitta, Y. (1986b), "Problems of Machine Translation Systems: Effect of Cultural Differences on Sentence Structure." *FGCS* (Future Generation Computer Systems), vol. 2, no. 2, North-Holland, Amsterdam, The Netherlands pp. 101-115

新田義彦(一九八七a)、「機械翻訳」矢田光治(監修)、『AI総覧』フジ・テクノシステム/エス・ディ・シー pp. 731-950

新田義彦(一九八七b)、「ヒューマン・インタフェースからみた機械翻訳」『エレクトロニクス』vol. 32, no. 4 pp. 66-74

田中穂積、新田義彦(一九八六)、「ロジック・プログラミングと計算言語学」『情報処理』vol. 27, no. 8 pp. 940-946

(にったよしひこ・計算言語学)

(たなかほづみ・計算言語学)