機械翻訳における訳語選択

田中穂積

1 はじめに [Hutchins 86]

翻訳の問題を一般化して言えば、ソース言語の単語の列をどのようなターゲット言語の単語の列に変換するかということになるだろう。以下では、機械翻訳における訳語選択を次のような意味で使うことにする。機械翻訳というときには、ソース言語の単語の列の長さは少なくとも1文を越える。一方、訳語選択というときには、変換すべきソース言語の単語の列の長さは1文を越えない。ソース言語の文の一部をターゲット言語の文の一部に変換することを問題にしている。

以上のことから、訳語選択は比較的局所的な処理であるように見える。文全体ではなく、ソース言語の一部を対象にすれば良いと思えるからである。訳語選択にあたり、ソース言語の単語がターゲット言語の単語と1対1対応していれば、単語レベルでの変換で済むことも考えられる。多くの人は訳語選択という言葉から、このような単語レベルの変換を連想するのではなかろうか。確かにそのような変換も少なからずある。また、現在の機械翻訳システムの多くが、それを期待してシステムを作成してきたという経緯もある。機械翻訳システムが、科学技術文献や機械の操作マニュアル文を対象にしてきたのは、そこに含まれる科学技術用語には、ソース言語とターゲット言語との間に1対1の対応関係があるものが多い、ということと無関係ではない。機械翻訳で翻訳対象分野を特定化しようとするのも、上述したことと関係がある。しかし、容易に予想できるように、訳語選択は文の部分にだけ着目した局所的な処理とはいえない。この問題は2章で論

じる。3章では2章にあげた諸問題を解決するための基礎として、ソーラスの重要性を指摘する。

2　訳語選択の諸問題

前章で訳語選択は、ソース言語の単語をターゲット言語の単語に変換するといった単語のレベルに代表される局所的な変換ではない、ということを指摘した。訳語選択は、ときには1文を越えた大局的な見地が必要になることがあり、そこに訳語選択の難しさがある。

訳語選択の問題は、見方を変えると、曖昧さの解消問題としてとらえることができる。曖昧さの解消は自然言語の解析で最も重要な問題であり、曖昧さの解消にあたって様々な知識が必要になることが多い。本章ではこの問題をもう少し検討してみよう。

2・1　単語レベルの訳語選択

ソース言語の単語が同音異義語である場合には、それをどの様なターゲット言語の単語に翻訳すべきかが問題になる。この場合には、ソース言語の単語の側に既に曖昧さが含まれており、ソース言語の単語とターゲット言語の単語との間に1対1対応がとれないからである。同音異義語のうち多品詞語に関する曖昧さの解消は、統語解析により部分的に解決できる。統語解析によりソース言語の文に含まれる各単語の品詞を決定することができるからである。品詞の決定は、品詞に関する曖昧さを解消してくれるからである。品詞に関する曖昧さが解消されない限り、訳語選択が行えないことは明らかだろう。例えば、

[a]　I can give him a book.

という文に含まれるcanは、少なくとも二つの品詞、助動詞と名詞とを持つ。それぞれの品詞に応じて「(蓋っきの)缶」、「できる」等と訳し分けなければならない。また

[b]　He said that it was cold.

[c]　He saw that dog that was able to swim.

という文に含まれるthatは、指示代名詞、接続詞、関係代名詞の少なくとも三つの品詞を持つ。それぞれの品詞に応じて日本語の訳語が異なることは明らかだろう。英語を勉強した人なら、上記の各文のcanとthatが、いずれの品詞の語であるかを知っている。それは、我々は無意識の内に言語的な解析を頭脳で行っているからだと考えられる。

コンピュータでは、canとthatの品詞をどのようにして決めたら良いだろうか。最終的に意味解析や文脈解析が必要になることはいうまでもないが、先ず最初に行わなければな

らない解析は統語解析である。統語解析は、文を構成する単語の品詞の並びが文法にかなっているかどうかを、文法に照らし合わせることをいう。したがって統語解析により、文中の各単語の品詞が何であるかを解析結果として知ることができる。幸いにして統語解析は最も研究が進んでおり、解析手法についてはほぼ確立しているといってよい。我々は、与えられた文の統語解析を行い、can と that の品詞を決定することができる。文法を用いるということは、ある単語に着目したとき、その単語の前後の品詞がどのようなものかを調べていることに他ならないから、すでに統語解析は、単語にだけ注目した局所的な処理であるとはいえない。

同音異義語の曖昧さのうち、同一品詞で様々な意義(sense)を持つ単語の場合には、統語解析を行っても曖昧さが解消できないので、訳語選択にとって難しい問題となる。

[d] 1 second later, he got there.
[e] Today is the second of April.
[f] Put a few of seals to your letter.
[g] We saw many seals in a zoo.
[h] We bought a seal yesterday.

たとえば [d] から [h] に含まれる second の品詞には、形容詞と動詞と名詞の三通りある。seal の品詞にも、名詞、他動詞の二通りある。おそらく統語解析を行えば、[d] から [h] に含まれる second と seal の品詞は名詞であると決めることができよう。問題は、名詞である second と seal が、それぞれ複数個の意義を持っていることである。複数個の意義に応じて日本語の訳語は相当異なる。second については「第二」と「秒」という日本語に対応した意義があり、seal については「印章」、「封印」の他に「おっとせい」という日本語に対応した意義がある。

以上の事実は、機械翻訳における訳語選択の問題が、統語解析を行っても解決できないことを示している。second と seal とを適切に訳すためには、意味解析が必要になるだけでなく、我々の保有している常識のようなものを用いた解析が必要になることがあるのである。たとえば [e] の second については、月の二番目であるという意味から、二日と訳さなければならないだろう。これはおそらく(数の数え方に様々な呼び方があるという)日本の文化に関係する事柄かも知れない。さらに [h] の seal については、意味解析によっても曖昧さが解消できない。なぜなら [h] の seal の日本語訳として、先にあげた「印章」、「封印」、「あざらし」のいずれも適切であると考えられるからである。したがってそのいずれが適切であるかは、周囲の文脈がどのようなものであるかによって決まる。周囲の文脈がどの様なものであるかに応じて、seal の訳語を決めなければならない。この問題についてはまた後で

ふれる。

［h］については、文脈解析により「おっとせい」かそれ以外のもの（「印章」、「封印」）であるかを決めることは、さほど困難ではないと思われる。なぜなら前者は生物であり後者は無生物であると判断できるからである。したがって訳語選択にあたり困難だと思われるのは、無生物である「印章」と「封印」のうちいずれを選択すべきかという問題であろう。

この例は、ソース言語のレベルでみると、語源が同じであるという理由で一つの辞書項目に納められているものが、ターゲット言語では意義が異なるとして複数個の訳語が対応している場合である。問題を明確にするために更に幾つかの例を考えてみよう。動詞 take を考えてみよう。

［i］ I take a plane.
［j］ I take his blood pressure.
［k］ I take a plane to Moscow.
［l］ I take a plane with a gun.
［m］ I send two characters.
［n］ I hate his character.

我々は意味解析による［i］から［l］の take と、［m］から［n］の character とを適切な訳語に翻訳するシステムを既に開発している［田中 86］。［i］から［j］では、take の目的語がどのようなものかを意味解析することにより、take の訳語を「乗る」、「計る」のように選択できる。これらは take と目的語との間の意味的共起関係（二項関係）を調べれば良く、さほど困難とはいえない。しかし［k］から［l］については、目的語の他に前置詞句をも考慮しなければならないので、計算コストの観点からは若干問題である。多くの場合、前置詞句の意味が動詞の訳語選択に影響することはないから、常に前置詞句を考慮することは効率的とはいえない。

［i］から［l］は、目的語や前置詞句の意味により動詞の訳語が決まる例であった。ところが訳語の選択には常にこの様な方向性があるとはいえない。次の［m］から［n］の例では、動詞の意味により目的語 character の訳語がきまる。

単語レベルの訳語選択で最も困難だと思われるものは、ターゲット言語の側で、様々なニュアンスをもつ多数の訳語が存在する場合であろう。現在の機械翻訳システムの多くは、このようなニュアンスの違いまでを考慮した翻訳は行っていない。

文脈の影響については、すでに［h］で見てきた。最後にもう一つ例をあげておく。

［o］ Press the suit.
（その訴訟に圧力をかけよ：そのスーツをプレスせよ。）

[o]のsuitは、文の前後の状況を知らない限り、適切な訳語を選択することができない。ここで、名詞suitの訳語だけでなく、動詞のpressの訳語が同時に適切なものに決められていることを、また目的語のsuitに付加する助詞も異なることにも注意。このような文脈に依存した訳語選択の方式として、見るべき技術はまだないといってよい。これからの研究課題である。なお[o]については、「その訴訟をつぶせ」と意訳すべきかもしれない。意訳も今後の研究課題である。

2・2 構造的な曖昧さと訳語選択

2・1では単語レベルの訳語選択において、統語解析により多品詞語の(品詞レベルでの)曖昧さが解消可能なことを述べた。また意味解析により意味的な曖昧さを解消して適切な訳語が選択できる場合もあるということを述べた。意味解析については技術的にほぼ成熟のところがあるが、統語解析については技術的に未成熟のところがあるが、統語解析については技術的にほぼ成熟しており問題は少ないと述べておいた。これは若干誤解を招く言い方であったかも知れない。それは統語解析に問題が少ないということの意味が、統語解析結果としていつも意味的に正しいものが一つ得られるというように解釈される可能性があるからである。

たとえば次の[p]を考えてみよう。

[p] Time flies.

意味を用いない統語解析では、[p]から少なくとも次の二通りの解析結果が得られる。

(1) [s [np Time] [vp [v flies]]]

(2) [s [vp [v Time] [np flies]]]

ここでの問題は、(1)と(2)の構造的な相違が多品詞語のtimeとflyの訳語選択に影響を及ぼすことである。もちろん我々は、(2)の統語解析結果からは、妥当な翻訳結果が得られないことは知っている。それは我々が意味解析や常識的な判断を無意識の内に行っているからだろう。しかしコンピュータで行う統語解析の場合には、一般に解析結果が多数得られることが多いが、そのいずれが妥当であるかを決めることは統語解析の役割ではないとされている。訳語選択では、このような構造的な曖昧さの問題を考慮しなければならない。このことは統語解析だけでは充分でなく、意味解析や文脈解析を考える必要がある、ということを意味している。

3 シソーラスと訳語選択

ところで意味解析や文脈解析で必要になるものは知識である。すでに前節のtakeの訳し分けのところで、目的語の意味に応じてtakeを「乗る」や「計る」と訳すということを

述べておいた。文［i］の場合には、takeの目的語が「乗り物」であればtakeを「乗る」と訳すということになるだろう。そのためにはplaneが「乗り物」であるという知識が必要になる。これはシソーラスとよばれるものの一部を成すものである。

これまで、意味解析を行う場合に、意味マーカの考え方が良く用いられてきた［長尾 85］。筆者は訳語選択には意味マーカは充分ではないと考えている。訳し分けに当たって、意味マーカ以上に微細な知識が必要になることが多いからである［田中 86］。

この問題を理解するためには、次の［q］、［r］に含まれるtakeの訳し分けを考えれば良いだろう。

［q］ I take a cup of water.

［p］ I take medicine.

文［q］の場合には、意味マーカを使って、目的語が+liquidという意味マーカを持てば、takeを「飲む」と訳す、という知識を記述することは可能だろう。しかし文［r］については、目的語に貼るべき意味マーカが用意されていないとみるのが妥当であろう。そのため"take medicine"を熟語として登録することも考えられよう。このようにすれば"take a pill"をも熟語と見なさなければならなくなる。このとき"take a bitter pill"を正しく訳すために、なんらかの仕掛を用意しておく必要がある［上脇 85］。形容詞bitterがaとpillの間に挿入されることを考慮した解析を行わなければならない。さらにいえば、takeの目的語として薬の製品名がきたらどうなるかという問題もある。これら全てを熟語として登録しておくことは好ましいこととはいえない。

以上の問題を解決するためには、意味マーカによる粗い意味分類ではなく、もう少し微細な知識の分類体系が必要になる。上位・下位にだけ限った概念階層を狭義の意味でのシソーラスとよぶことにすれば、このようなシソーラスを用意することは訳語選択に極めて有効である。文［q］、［r］の例では、medicineやpillは「薬」であるという記述がシソーラスにあるものとして（この程度に詳しい概念階層知識を用意する必要がある）、それを用いて、takeの目的語が「薬」であればtakeを「飲む」と訳す、という知識の記述が可能になる。そしておそらく、このようなシソーラスの階層の上位には、意味マーカの体系が位置することになるだろう。訳語選択の立場からシソーラスの重要性を強調しておきたい。

4 おわりに

本稿では、訳語選択における種々の問題を論じてきた。そして問題を主として解析の観点から分析し、解析結果が多数得られる場合の問題点を検討してきた。そして訳し分けの立

場から、概念の上位・下位関係に限定した（大規模な）シソーラス（知識）の重要性を指摘した。

最後に訳文を作り出す文生成過程で、どのような文体の翻訳文を生成するかという問題があることを指摘しておきたい。これは一般に、文の生成結果が一つに決まらないという問題である。この問題については、充分に研究が行われているとはいえない。文体については、文脈的な状況を考慮して、省略とか代名詞化を行う問題もある。いずれも研究が緒についた段階にある。今後の研究に期待したい。

現段階で、訳語選択の立場から、文解析過程と文生成過程とを比較して、どちらが重要であるかを考えてみる。少なくとも次のことが言えるように思われる。なんらかの理由により解析過程で曖昧さの解消に失敗し、誤った解析結果を得てしまうと、生成過程が如何に優れていても、翻訳結果に誤りが含まれることになる。したがって訳語選択でまず考慮すべきことは、解析における曖昧さの解消であると思われる。

参考文献

[Hutchins 86] Hutchins, W. J. : Machine Translation, ellis Horwood (1986)。

[上脇 85] 上脇正他：辞書の TRIE 構造化と熟語処理、Proc. of Logic Programming Conference'85, ICOT, 329—340 (1985) .

[長尾 85] 長尾真他：科学技術庁機械翻訳プロジェクトの概要、情報処理学会論文誌、Vol. 26, 329—340 (1985)

[田中 86] 田中穂積他：科学技術における曖昧さの構造の計算機による検証、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 (1986)。

（たなか・ほづみ 東京工業大学教授）