知識工学と人工知能 43－4パ5
（1985・11・8）

機械翻訳と知識工学

辻井 潤一
（京都大学・工学部）

1 はじめに

科学技術庁の機械翻訳プロジェクト（Muプロジェクト－以下では，Mu）が開始し、すでに3年と6カ月が経過した。自然言語処理技術とソフトウエア技術の現時点での到達点をできるだけ活用して、機械翻訳という『夢』の実現にどこまで迫れるかを明らかにしたい、というのが、プロジェクトを開始するときの意気込みであった。

このプロジェクトを行ってみて実感したことは、機械翻訳システムは、

(1)人間の言語活動が、人間の持つ実世界知識に支えられたものであること、
(2)翻訳を行うためには、両言語の文法的、語彙的知識という(普通言われる)実世界知識とは異なる知識を、大量に集積しなければならないこと、

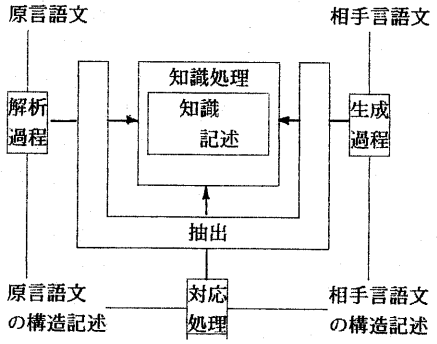
という2重の意味で、知識情報処理の典型である、ということであった。本稿では、この2点について、議論する。

2．現実世界に関する知識と翻訳

人工知能的なアプローチからの自然言語理解から機械翻訳へと研究対象を変えて、最も興味を持ったのは、『単語と概念』の関係であった。『概念』を、仮に、『人間からの自然言語による要求に答えるために推論機構や問題解決機構が使用する内部的な記号系での基本語彙』というように、計算機処理的な立場から定義するとすると、データ・ベースの自然言語によるアクセスや初期の自然言語理解が対象としたToy-Worldでは、『単語と概念』との相互関係は比較的直接的であった。Toy-Worldやある特定分野でのデータ・ベース・アクセスのように『視点』の固定した世界では、内部処理で使われる記号系、すなわち、概念の規定、が明確であり、かつ、有限個の離散的な『概念』だけを対象とすれば良い。したがって、このような分野での単語と概念の対応は、比較的直接的である。しかし、分野を少し広げると、この対応はそれほど直接的ではなくなる。自然言語は、有限の語彙で連続的に無限な外界世界を分節化しており、我々人間の認識能力はある意味で言語による分節化よりもはるかに微細なものまで区別できるために、単語と概念の対応はそれほど直接的ではなくなってくる。

2つの言語を対象とする機械翻訳では、2つの言語の分節化の差を調整する必要が生じるために、この問題は特に深刻になる。比較的単純な科学技術分野の翻訳においても、このような2言語間の分節化の差が生じる。従来の自然言語理解が問題にした自然言語のAmbiguityよりも、これまであまり深刻に議論されてこなかった自然言語のVaguenessの問題が、より重要な問題となる。例えば、この問題は、有限の語彙で無限の世界を記述する自然言語の宿命である『言語の比喩的使用の問題』とも密接に関係していよう。現在の機械翻訳は、この問題を解決しているとは思えない。

すなわち、言語理解と知識処理の関係について、機械翻訳は、これまでの自然言語理解の研究とは明らかに異なる側面を持つ。事実の因果関係だけを推論するためには、『太郎は花子を殴った』も『花子は太郎に殴られた』も、同じである。従って、事実の因果関係やその上での推論という観点からは、この2つの文を共通の意味構造に縮退することが、望ましい。しかし、翻訳の立場からすると、この2つの文には、話者の視点など明らかに相違が見られる。2つの文を、共通の『意味表現』で捉える立場は、文の『意味』を、文の表現する事実との関係で捉えていることになる。同じことは、『太郎が首相を殺した』と『太郎が首相を暗殺した』という2つの文にも言える。



現実世界の同一事象は、実際には、話者というフィルタを経て、様々な形式で言語表現される。機械翻訳と知識処理とを結び付けるためには、翻訳の対象となる文の全体としての記述と、推論などの知識処理を行うために文から抽出された『意味』とを分離して持つ必要がある。このためには、図1のように、文から『知識処理』のための記述を抽出する部分を設けて、言語処理と知識処理とを結合する必要があろう。

3　機械翻訳における知識獲得の問題

我々のシステムでは、言語的知識の集積としての辞書の重要性が強調されている。特に、単語個別の知識を取り扱うために、単語ごとの個別処理を強調してきたが、実際に、単語個別処理を指定することができるということと、その個別処理の機構をどのように使用するかは、別問題である。

少数の語彙を対象に機械翻訳の研究をしている場合には、個別処理が必要な単語がその１語だけであっても、大量語彙を対象にした場合には、かなり多くの語彙が同じような個別処理を必要とすることも多い。この場合には、それら一連の単語に共通な言語的性質を発見し、どのような基準とテストを使ってその種の単語を見付けるかを整理しなければならない。単語ごとの辞書規則を定義できる我々のシステムの機能は、処理の枠組を用意しただけである。これを乱用することは、語彙的知識の言語学的な整理段階をバイパスすることになり、大量語彙に対して辞書記述する場合には、だれがどのような基準で語彙的規則を指定できるかという問題が生じる。同じような事情は、辞書記述の中に手続きを付加してゆく手法やオブジェクト指向型的な手法を辞書記述に取り入れる場合や、『知識』を概念階層で整理する枠組を提供する場合も同様である。

知識情報処理システムの場合、システム中のどこかにオープン・エンドな部分を設けて、この部分を活用することによって、システム設計時には考えていなかった現象に対処するのが普通である。多くの場合、このオープン・エンドな部分は、個々の知識記述の中に設けられる。例えば、この部分に手続き付加的な機能を持つことで『何でもできる』ように作るのが普通であり、ある特定の具体的な現象や推論を見せられると、この枠組を使ってそれはそれなりに対処できるのが普通である。このような潜在的なシステムの拡張可能性と実際にその枠組を使ってどれだけの知識が系統的にコーディングできるかの議論は、明確に区別する必要がある。処理の枠組の汎用性とそこにどのような知識を記述するかの議論を混同してはならない。

例えば、機械翻訳のための知識ベースである辞書を作成する場合においても、それを系統的に行うことを考えると、次のような種々の問題があった。

[単語の表記]日本語では、送り仮名やカナ表記での長音表記、異文字種による表記など、複数の表記を持つ単語が多い。英語の大文字、小文字の区別やハイフンも同様である(on-line, on line, online)。

[単語の単位]複数の語からなる表現を辞書に登録する場合、どの範囲までを登録するかが問題となる。

[品詞の設定]『安全』を名詞とするか、形容動詞とするか、あるいは２義性があるとするか、名詞と形容動詞を区別しない品詞体系を採用するか、などの問題がある。また、『従来』、『将来』のように名詞と副詞の判断に迷う場合もある。

[意味用法の認定]『プログラムを走らせる』の『走る』と『汽車が走る』の『走る』とは別の用法であるとするのか、あるいは、同じ用法とするか。また、『合図を送る』と『手紙を送る』の『送る』はどうか、など多分に作業者の主観による。また、名詞にも多くの多義性(AmbiguityとVagueness)があり、通常の概念階層を作ることがそれほど簡単でないことが判る。

[格関係の認定] システムごとに格の設定に相違があるように、格の概念は安定したものではない。個々の格の認定基準を明確にしておかないと、作業者の迷いや辞書記述に一貫性がなくなる。

[意味分類の付与基準] 意味マーカの付与は、最も基準設定が難しく、作業コストの大きい部分である。豊富な実例とテストの手段がないと作業できなくなる。

[対象分野への依存の程度] あまりに一般的・網羅的な辞書は逆に文法処理への負担が大きくなる。対象分野での使用頻度の低い用法は入っていない方が良い。どの範囲までの用法を辞書に入れるかは、結局個々の単語ごとに作業者が判断せざるを得ない。このことによる作業能率の低下、辞書記述の網羅性の低下などは、システムを大規模化してたり、他分野へ移行したりする際の障害になる。

以上のような問題点は、いずれも作成された知識ベースとしての辞書の網羅性や一貫性に大きく影響する問題であり、大規模知識ベースを作成する場合には、どの分野を対象としようとも、現われてくる問題であろう。

4．知識ベース作成とソフトウエアの支援

３で述べた辞書記述の基準の問題は、言語的な知識の整理という長期的な研究課題と密接に関係しており、一朝一夕で解決するものでわない。工学的に実際のシステムを作成する上では、現時点で可能なかぎり知識を整理するが、足らないところはシステム的な工夫によってできる限り援助することになる。作成された文法規則と辞書記述の整合性や文法規則の不備を検出するために様々なソフトウエア・ツールを準備して

おくことは、機械翻訳のような知識情報処理システムにとっては不可欠である。知識というものは、文法規則であれ、個々の単語ごとの語彙的知識であれ、記述された段階で完全で、なんらの誤りも含まないという理想的な状態は期待できない。知識情報処理システムの研究者の多くが、ソフトウエア支援の重要性を強調するが、これは偶然ではない。この種のシステムにおいては、初期の不完全な知識をシステム運用を通じて完全なものにしてゆく必要があり、開発、運用、保守が分ちがたく結び付いている。

我々のシステムにおいても同様であって、プロジェクト１年目で文法を記述するためのコアのシステムが完成したが、文法規則のデバッガ等はそれより半年間程度完成があとになった。辞書管理用のツールも同様であり、入力データのチェック・プログラムやパンチ入力されたデータの形式変換ツール、辞書用のエディタなど、機械翻訳のコア・システムよりもむしろ先行して作成しておくべきであった。この種のソフトウエア・ツールが完備してからの開発の速度と、それまでの速度とはかなりの差が見られた。

この他、翻訳実験を能率良く行うためには、適切なテスト・サンプルを選択しておくこと、それ以前の翻訳結果と新しい翻訳結果との比較を容易にするためのソフトウエア・ツールなどを準備しておくことも重要である。

５．終りに

機械翻訳あるいは自然言語理解の研究には、本稿では議論できなかった数多くの研究課題が残されている。特に、機械翻訳の実用化においては、人間との協力を含めた運用形態の問題、また、それを支援するためのソフトウエアの問題、ユーザ辞書やシステムの分野依存性の問題などは、実用化段階に達しつつある他の知識工学システムにも共通するものであろう。プロトタイプのシステムから実用システムへ至るためには、広い意味での知識獲得の問題が、今後真剣に議論されるべきであろう、と思われる。