次世代の機械翻訳技術について

第１回　従来技術の到達点

はじめに

日英言語間の翻訳を中心に、現在多くの機械翻訳システムが開発され、翻訳現場への適用例も増えてきました。ここ何年間かの研究開発の結果、日本の機械翻訳技術は世界の最先端に躍りでた観があります。しかし、日本語と英語の違いは、英語、仏語などのインドヨーロッパ系の言語間の違いに比べて大きいこともあって、その翻訳は決して容易ではありません。言語はそれを使用する集団の文化的産物ですから、ものの見方や捉え方の違いが、言語の構造や表現法にも反映しています。このような観点から、第二世代と言われる現在の翻訳技術の問題点と、最近の新しい試みを中心に、次世代の機械翻訳技術の展望について述べたいと思います。

第二世代技術の特徴

機械翻訳を考えると、誰もがまず、単語対単語に置き換えて翻訳することを考えます。第一世代は、文や表現の形式を手がかりに、原言語と目的言語の間で、対応する単語の関係を決めれば翻訳できると考えられていた時代で、構文解析の技術を背景に、投資をすれば機械翻訳システムは実現できると考えられていました。しかし、一つの単語にも数々の意味があるため、それが文中でどのような意味で使われているか知らなければ、目的言語に対応させることが出来ません。表現の構造を解釈す

るにも、単語の意味が影響します。そこで、第二世代では、構文解析技術に加えて、単語の意味を決めるための種々の技術が研究されてきました。動詞と名詞との関係で単語の意味を扱う格文法や、単語の意味を意味素に分解する方法など、単語の意味を処理する方法により、第二世代では、かなりの程度単語対単語型の翻訳が出来るようになったわけです。

前処理の内容

機械翻訳の前に人手で行なわれる前処理の内容を見ますと、①一つの単語は一つの意味でしか使用しないこと、②修飾句は修飾するものの直前に置いたり、埋め込み関係を複雑にしないことなどにより、なるべく構文として複数の解釈を生じないようにすること、③主語や目的語を補って、英語にするとき不足する要素が内容にすること、④諺や熟語などの慣用的な表現は使用しないこと、等を基準に日本文を書き換えています。これらの基準を見ますと、いずれも直訳しにくい表現を直訳できるように編集していることに気がつきます。

直訳法の前提

単語対単語置き換え型の翻訳、いわゆる直訳の方法は、技術的には要素合成法とも言われ、大きな前提（フレーゲの定理）があります。すなわち、言語表現の意味は単語の意味に分解できること、また、分解して処理した結果を組み合わせれば元の意味が復元できると言う前提です。言語は慣用表現の集まりと思っても良いと言われるくらいですから、自然言語では原理的には、この前提の成り立たないことが明かです。そして、言語の差が大きいほど、この前提は成り立ちにくくなります。

現在、前編集をすればある程度訳せると言うことは、逆に言えば、日本語と英語は元来、違いの大きい言語であるにもかかわらず、明治以降の欧米の文化の輸入に伴う日本語表現の変化もあって、要素合成法の成り立つ表現の領域がかなりあり、そこを実用化で狙ってきたことになります。

従来方式の反省

第二世代の翻訳方式は意味処理を考えることに特徴があると言いましたが、そこで言う意味処理の考え方は、計算言語学におけるチョムスキーの生成文法の意味論に基づいています。すなわち、意味は実際の表現（表層構造）とは別のところ（深層構造）にあるもので、言語に共通のものだとされてきました。中間言語を意味的表現とし、それを介して翻訳する現在の機械翻訳の方式は、このような意味論に支えられています。しかし、言語は、まず「対象」があって、それを見る話者の「認識」が生まれ、さらにその認識が「表現」に表されると言う過程を持つことを考えれば、言語に共通なのは「対象」だけであり、認識（見方、捉え方）以降は言語によって異なることになります。

現在までの結果を見れば、言語共通の意味としての「深層構造」は、要素合成法の扱う範囲の意味と等価で、直訳型の翻訳の実現によって、その役割を終えたと言えるかも知れません。

次回は、このような反省に立った最近の研究を紹介したいと思います。

第２回　新しい翻訳方式の考え方

新しい文法の発見

最近、生成文法の意味論の限界を超える考え方と翻訳の方法についての研究が行なわれるようになってきました。今回は、その例として、言語過程説とそれをベースにした翻訳方式の考え方について紹介します。言語過程説の考え方は、生成文法の意味論になれた人には珍しく見えるかも知れませんが、最も常識的な考え方と言えるもので、伝統的な国語学の中にあるものです。

本居宣長は日本語は「詞」と「辞」によって話者の捉え方の大きな違いが表現されることを指摘しましたが、この流れを汲む時枝誠記は、言語が「対象」、「認識」、「表現」からなる過程的な構造を持つことに着目し、言語過程説（時枝文法）を提唱しました。生成文法よりも３０年も前のことです。

言語過程説に従えば、言語の過程的構造、すなわち「対象」と著者の「認識」と「表現」の関係から文法や意味の考え方を見直さなければなりません。文法は認識を表現に結び付ける際の約束ですから、通常の辞書に書いてある単語の意味は、本当はその語の使い方の約束を示すと言う点で、厳密には文法の一部だと言うことになります。また、意味とは対象と認識と表現の対応の関係であり、この関係は単に単語と概念のレベルの対応だけでなく、対象の構造と認識の構造と表現の構造との対応としても見なければならないことになります。従って、表現の構造を変えれば意味も変わってしまいますから、意味を変えない変形は原則としてあり得ないことになります。

新しい翻訳方式の発想

このように考えたとき、言語過程説から学ぶべき大切な点が二つあります。一つは、表現には話者の感情や意志を直接表現する主体的表現と、話者が対象を概念化して捉えて表現する客体的表現が結合していること、もう一つは、表現の構造は意味の一部であるから、バラバラにして翻訳し、組み立てようとする（要素合成法）と、元の意味が失われると言う点です。前者は主体的表現と客体的表現を見分けて言語間に対応づけることが考えられますが、二番目の点を重視すると、原言語の考えられるすべての表現に対して、対応する目的言語の表現をあらかじめ用意すればよいことになりますが、それは工学的に成り立ちません。

構造の段階的抽象化

多段翻訳方式では、この問題を表現構造の段階的な抽象化の方法によって近似的に解決しようとしています。客体的表現が対象を概念化して表現するものであり、対象の持つ構造が直接表現に反映されるため、異なる言語間でも比較的共通性が高いのに対して、主体的表現は話者や言語による見方、捉え方の差が強く現れますが、言語間での構造的対応は少ないと言えます。

そこで、多段翻訳方式では、主体的表現と客体的表現を分け、主体的表現の示す意味を目的言語に対応させる仕組みを持つ一方で、客体的表現に対して、以下の３段階の構造の抽象化を行なって目的言語に対応させています。抽象化の第１段階では単語と単語の関係が固定的でその組み合わせが変われば意味の変わるもの、第２の段階では単語自身に付随する表現の構造、第

３の段階では品詞や意味カテゴリーのレベルで構造をルールとして表現できるもの、を集めそれぞれに必要な日英変換のルールを設けています。ルール適用の順序は１～３の順で、個別性の高いルールが適用できる表現ほど良い翻訳が出来ることになります。個別性の高いルールに重点がありますから、従来、例外とされていたものが本論となり、本論とされていたものが例外の扱いとなるなど、発想が逆転していると言えます。

このようなルール構成とするとき、第３のレベルは抽象性が高いため、ルール数は比較的少なくてすみますが、第１、２のレベルは個別性が高く、万の単位のルールが必要になります。しかし、言語は元来個別性の高いもので、一語一語の微妙な違いも使い分ける人間の言語能力を考えれば、この程度の数のルールは、まだ少ないかも知れません。また逆に、個別性の高いルールであれば、この程度の数のルールをシステムに組み込んで選択的に使用することは、さほど困難ではありませし、追加削除も容易ですから、工学的には問題ないと言えます。

構造と意味を一体化した変換のためのルールは日本語の解析においても、大変役に立ちます。構文解析で生じる多義は文の構造上の多義ですが、文の構造の解釈が違うということは、意味の解釈も違うと言うことです。第１、２のレベルのルールを調べると、意味のまとまる単位、すなわち構造のまとまる単位が捉え易くなりますから、構文上の多義が減少し、解釈が容易になります。

次回は、これらの新方式を支える技術の現状と今後の課題について述べる予定です。

第３回　新技術の現状と今後の課題

多段翻訳方式を支える技術

多段翻訳方式の基本的な考え方は、前回述べた通り表現構造の段階的な抽象化による意味と構造の一体化にありますが、これを支える技術としては、精密な意味カテゴリーがあり、また、応用した技術としては、日本文の自動書き換えや文脈的要素補完の技術があります。

従来、意味カテゴリーとしては名詞の意味を通常数十通り、最大で２００通りぐらいに分類する方法が実現されていましたが、上記の第１、２段階で万単位のルールを記述するには粗すぎますので、一般名詞２，８００カテゴリー、固有名詞２００カテゴリーまで精密に分類しています。このカテゴリーは動詞や名詞の訳し分けのほか、名詞句の構造を解析するのにも役に立ちますし、複合語の構造を解析して訳すのにも有効です。日本語は膠着言語の特徴として、造語が容易で、新しい複合語が限りなく生み出されるのに対して、従来は、まとめて辞書に登録するしか良い方法がありませんでしたが、このカテゴリーを使用することによって、単語に分割して扱うことが出来ます。要素合成法を超えるための準備が、ここでは要素合成法を助けるという効果も生でいます。

応用技術

意味のまとまりが分かれば、これを日本語の段階で翻訳し易い表現に変えて訳すこともできます。直訳しにくい日本語を直訳し易い日本語に自動的に変えて翻訳する日本文自動書換の技

術でも、書き換えて良いかどうかの判断は、やはり構造的な意味のまとまりを手がかりに決められます。また、日本語は相手に分かっていることはなるべく書かないと言う特徴がありますが、省略された主語や目的語など、英語に訳すとき不足する情報を補って翻訳する文脈要素補完の技術も、広い意味では、日本文書換の技術に入ります。

新方式の効果と課題

多段翻訳方式ではこれらの方法によって、名詞、動詞の意味による訳し分けや、複合語、同格等の名詞句、慣用表現等の翻訳、主語、目的語等の文脈補完等の基本技術について見通しが得られるようになりました。しかし、残された課題も沢山あります。意味カテゴリーを精密化するだけでは訳し分けられないものも多いこと、複数の用言の関係を見て適切な英文構造を決める技術、冠詞と数を決める処理、など当面急ぐべき課題です。

また、言語過程説の立場からは、従来の品詞体系の見直しや日本語の認識構造をシミュレートした入れ子型構文解析技術等も必要です。

期待されるサービス

マニアルや技術文などの記述文であれば、前編集すれば翻訳できると言うのが現在の機械翻訳であるのに対して、前編集しなくても意味の分かる翻訳がきると言うのが次世代の機械翻訳の目標となります。即時性という機械翻訳本来のメリットが前編集作業によって損なわれていますが、前編集が不要となれば、速報性の要求されるニュースの配信や国際データベースサービス、国際メールサービス、テキスト翻訳通信、等が可能となり

ます。

長期的課題

しかし、会話文の翻訳や英語として品質の良い優れた翻訳を実現するのは、まだまだ容易ではありません。

会話翻訳では、相手に分かっていることはなるべく言わないと言う原則が記述文以上に働きます。この事情は英語でも同様です。このため、文脈や背景に関する知識がいっそう重要になってきて、内容が理解できなければ訳せないような場面が多くなります。このような翻訳に接近するには、対象世界の知識を背景に、表現の意味、すなわち話者が表現に結び付けた「対象」と「認識」の対応関係を解析する方法を確立する必要があります。

また、記述文でも、論文の執筆では単に分かりやすいだけでなくある程度の格調も要求されますし、小説などの文学的な文章では直接文字に表現されたもの以上の意味を表現していることがあります。このような文の翻訳は永遠の課題と言うべきかも知れませんが、現在、心の社会のモデル化など人間の精神的活動に関する研究も行なわれています。これらの研究が、話者の視点の構造や学習メカニズムの解明に結びつき、言語表現から話者の認識を追体験する仕組みや対象の捉えなおしの仕組みなど、人間の認識構造に着目した知的言語活動のモデル化が進むことを期待したいと思います。

多段翻訳方式を構成する技術

構造と意味の一体化による多義の解消＝意味を失わない翻訳

・語彙の拡大に耐える多義解析法

（固有名詞辞書等６０万語）

・名詞、動詞の多義（語義）

・意味カテゴリーの精密化

（一般名詞２８００、固有名詞２００カテゴリー）

・書き換え技術

・構文辞書（１５、０００パターン）

・慣用表現の扱い

・複合語の扱い

・名詞句、同格の処理

・文脈による要素補完

現在の到達レベル

・名詞の訳し分け

・動詞の訳し分け

・慣用表現の翻訳

・省略表現の翻訳（格要素）

第４回　今後の課題と展望

当面の課題

・意味カテゴリーの二次元化またはシソーラスの併用

・大域的な英文構造の生成

・用言の意味的関係の解析

・冠詞と数の処理（認識法からの接近）

長期的課題と展望

・話者の視点の導入（観念的に自己分裂する話者）

・言語理解の方法論の確立

・学習型言語知識の獲得

・人間の知的言語活動のモデル化

・人間でも不可能なことはやはり不可能

＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝＝

これらの前処理を不要とするには、・大量の語彙を対象にそれが使われた意味を調べて訳す技術、・慣用的な表現とそうでない表現を区別して訳す技術、・英語にするとき不足する要素を補って翻訳する技術、等が必要となります。

人手翻訳のプロセス

さて、それでは今後の機械翻訳のあり方について考えてみたいと思います。日英機械翻訳で見れば、日本語らしい日本語、直訳しにくい日本語の翻訳に挑戦することが今後の課題です。

人間の翻訳家が翻訳するプロセスを見ますと、翻訳家は、まず表現を見て、原言語で決められた約束（言語規範：文法）を手がかりに、その表現に結び付けられた著者の認識（見方、捉え方）とその背景にある対象の姿を追体験し、自分のものにします。その後、目的言語の枠組みに合わせて捉えなおしを行ない、目的言語の言語規範に従って表現に結び付けます。この過程を計算機でシミュレート出来れば良いのですが、人間の追体験と捉え直しの能力は高度に知的なものであるため、まだその仕組みが分かっておりません。しかし、認識と表現を対応づけるのは言語の約束であること、また、認識は表現の部分にだけでなく、表現の全体に対応付けられることを考えれば、原言語で対応づけられた関係を目的言語の枠組みに合わせて、どう対応づければ良いかについて考えれば良いことになります。

日英言語の文化的背景

かって日本は中国文化を受け入れるに当たって、日本語の中に漢字を取り入れ、音読みや訓読みなど日本語としての表現法を成長させてきましたが、それと同様、明治以降は欧米の文化を取り入れるため、新しい語の概念に留まらず、欧米のものの見方や考え方を表現する枠組みを成長させてきました。漢文に返り点を打って日本語読みする方法に習って、単語と単語を対応させ、翻訳する直訳の方法は、日本語の中に欧米のものの見方、捉え方と共通する表現の仕組みが成長してきたことに支えられています。これに対して、日本文化が欧米の文化に与えた影響はまだまだ限られた範囲にあると思われますから、日本人のものの見方、捉え方を表現する枠組みが欧米の言語の中に育ってきたとは考えられません。これらの事情によって、日本語

を英語に翻訳するには、英語を日本語に翻訳する場合とは異なった難しさがあると言えます。

以上の背景を見ると、現在の機械翻訳の対象が技術文中心となっていること、また、技術文と言っても、テクニカルライトされていることや前編集を前提にしていることは止むを得ないと言えるでしょう。

ＮＴＴの取り組み

ＮＴＴが日英機械翻訳の方法として提案した多段翻訳方式では、この問題を正面から取り上げ、表現構造の段階的な抽象化と言う方法でこの矛盾を調和させ、近似的に解決しています。この方式では、従来、例外と考えられていたものが本論となり、本論で扱われていたものがむしろ例外として扱われるなど、発想の反転があります。