

Mu-プロジェクトの英日機械翻訳 システムにおける多品詞語の解消

山本利文(日本・データゼネラル), 辻井潤一, 長尾真(京大・工)

1. はじめに

機械翻訳においては、原言語の持つAmbiguityの解消が大きな問題になるが、特に英語解析においては、統語解析の開始段階ですでに多くの多品詞語があり、これをどのような手順で解消(disambiguate)するかは、一つの大きな問題になっている。また英語においては、多品詞語でなくても動詞のed形やing形のように、本来動詞のものが形容詞的、あるいは名詞的に使われる用法もある。

本稿では、動詞ed形、ing形の形容詞的、名詞的の用法も含め、Muプロジェクトの英語解析文法における多品詞語の解消について述べる。

2. 基本戦略

多品詞語の中には、文中におけるその語の前後の局所的な情報のみで、ほぼ確実に唯一に決定できるものもあるが、文全体の構文的、意味的關係を考慮に入れないと最終的には断定できないものも多い。逆に文全体の解析は、多品詞語がうまく解消しないと、完了しない。つまり、多品詞語の解消と文全体の解析とは、相互に依存しあうものである。このため多品詞語の解消を、どの段階でどのように行うかは、難しい問題となっている。

我々の英語解析文法では、この問題を次のような方針で行っている。

- 1) 解析の初期の段階において、局所的に利用可能なだけの種々の情報を用いて、最も確からしい品詞を一つ選択する。
- 2) それをもとに解析を進める。ある程度構造が決定されてきた段階で、大域的な構造からのチェックを行ない、先に選択した品詞の解釈に矛盾が見つければ、解釈を切り換え、それまでの中間的な解析結果を修復する。

この方針によって、必要な処理量を最少限に抑えつつ、ほぼ決定的(deterministic)に英語の統語解析を行っている。

3. 多品詞語の種類

本稿でいう多品詞語には、以下のものを含む。

- 1) 名詞と動詞など異なった品詞の語と考えられるが、表面上同形のもの。現在の我々の英語解析辞書(18000語)中、約1000語がこのような多品詞語であり、品詞の組みあわせとして約40のパターンがある。
- 2) 動詞のing, en形のように、名詞的、形容詞的にも使われるもの
例 the custom designed LSI/VLSI circuits
この例のように、動詞のed形は局所的には主動詞解釈が可能なものでも、長い専門用語の一部として前修飾の形容詞的の用法と解釈すべきものも非常に多く、通常の解析手法では、局所的な解析パスが増大する。
- 3) do, have, beは主動詞としての用法と助動詞としての用法がある
- 4) thatには、限定詞、代名詞、接続詞、関係代名詞、補文標識の用法がある。このほか、if, when(従属接続詞、間接疑問詞)なども多品詞語となる。(このような特殊な語の品詞決定は単語個別の多品詞語解消規則で処理される)
- 5) 多品詞ではないが、形容詞、序数詞の名詞的用法、年号などのように、品詞の切り換えなどが必要なもの
例 The first gives a general background
～.
In 1972 the industry comprised about 20 companies ～.

多品詞語の解消規則はこれらすべてに対応しているが、本稿ではこのうち1)と2)の一部、名詞と動詞、前置詞と接続詞、動詞のed形について述べる。

4. 多品詞語解消規則概要

4.1 英語解析文法の中での位置

英語解析文法の全体的な流れを以下に示す。

1) 複合動詞のまとめ
2) 複合語のまとめ
3) 引用符で囲まれた部分、括弧などの処理
4) 多品詞語解消規則
5) 並列句の中で前置詞や副詞の並列のように単語単位の単純なものをまとめる
6) 形容詞の直前にあるIntensifierとしての副詞を形容詞にまとめる
7) 名詞句をまとめる。関係節や前置詞句、分詞による後方からの修飾は、この時点ではまだまとめない
8) 現在分詞形、過去分詞形の動詞の用法決定
9) 格を支配しない形容詞で名詞を後方から修飾するものを名詞句に取りこむ
10) 文副詞、接続的副詞の処理
11) 動詞句のまとめ
12) 多品詞語の修復規則 その1
13) 文の一時的な構造化。不定詞句、分詞句、THAT、WH語、形容詞句などを認識し、文単位(その中に唯一の述部をもつ)ごとの処理のスコープを決定する。各種並列文法もここから呼ばれる。この中にも多品詞語修復規則がある

14) 動詞の格パターンや自由格を決定する一般規則等によって、前置詞句THAT節などを支配する述部の決定や、名詞句の係受け関係を決定する。また、このような統語的關係を決定すると同時に、その意味的な関係を決定する。

15) 決定された係先に従って、句構造を作る

16) 深層テンスやアスペクトを決定し、句構造を次のトランスファ過程で想定している依存構造に変える

図 1

多品詞語解消の規則は、上記の処理ステップの中で、4), 8), 12), 13)にある。4)は中心となる解消規則で、比較的局所的なさまざまな情報にもとづき、最も確からしい品詞を一つだけ選択する。8), 12), 13)は、誤った選択を修復する処理である。

4.2 多品詞語解消規則の概要

図1の4)が中心となる解消規則で、これは以下の手順をとっている。8), 12), 13)の誤りの修復規則には名詞と動詞、前置詞と接続詞、that、動詞ed形などに関するものがあるが、これについては具体例を次章以下で述べる。

- (i) ほぼ確実な規則 (規則数 4)
前後の語を見るだけで、ほとんど確実に品詞が決定できるもの。例えば、冠詞の直後の名詞あるいは形容詞など
- (ii) 単語個別の規則 (規則数 9)
moreやenoughなど、一般の規則によらず、その語固有の規則を設けたほうが良いと思われるものを集めている
- (iii) thatの品詞決定、動詞のing, en形の名詞的あるいは形容詞的用法の判断、be, have, doの本動詞か助動詞かの判断に関する規則 (規則数 39)

(iv) 並列や前置詞と接続詞の多品詞語など広い範囲を見る必要のある規則(規則数25)

(v) 多品詞語の連続で可能性の高いものを選択する規則(規則数6)

多品詞語が連続する場合、一個一個の品詞を解消する規則では、決め手を欠く場合がある。このため確実度の高いものについては、連続した多品詞語を同時に解消する規則を設ける

例 This increases the output ~.

代名詞+動詞(数の一致などもみる)

(* thisには限定詞、increaseには名詞の解釈もある)

(vi) 前後数語により最も可能性の高いものを選択する規則(規則数98)

これらの規則は、直前の語の品詞、直後の語の品詞および多品詞語が、文頭あるいは文末にあるかによって分類されている。たとえば直前に前置詞があることを一つの手掛りとする規則には、名詞と動詞の多品詞語を名詞とする規則、限定詞を優先する規則などいくつかある。これらをつまみとめ、ある多品詞語の直前に前置詞があればその規則群だけの適用を試みる。これによって、ある一つが多品詞語について、98個全部の規則を試みる必要がなくなり効率的に実行できるようになっている。

一般に、一文中にはいくつもの多品詞語があり、ある語の品詞決定には、その前後の語の品詞がわからないと、困難なものも多い。このため、多品詞語が何個も並んでいるような場合、規則の適用順序によっては、うまく解消されたり、解消されず多品詞語のまま残ったりすることがある。しかし、多品詞語の実際の文中での出現の組みあわせは、膨大なものであり、その組みあわせに合わせて、規則の適用順序を制御することは、現実的に不可能に近い。そのため、ここでは、上記の解消規則全体を、最大3回まで繰返し適用するようにしている。これによって、段階的に多品詞語の解消を図っている。

個々の規則には、2つのタイプがある。1つは可能性のない品詞を排除していくタイプの規則であり、も

う1つは、最も可能性の高いものを抜き出してくるタイプの規則である。解消規則は、この2つのタイプの規則をできるだけ効率的に組みあわせて構成している。特に(i)は、排除していくタイプの規則が多く、最初に可能性のない品詞を排除して、Ambiguityを減らしておいてから、以降の規則を適用する戦略になっている。

例 ~ universally-oriented tasks were most effectively linked to external areas by direct project member communication.

この例では、directに、動詞、形容詞、副詞があるが、前置詞byの直後であるため、動詞の解釈は排除される。その後、形容詞と副詞の一般の規則を適用して、多品詞を解消する。こうすることによって、さまざまな品詞の組み合わせごとに、別の規則を準備する必要性が少なくなり、規則の数を減らすというメリットもある。

5. 具体的な多品詞語解消規則

本節では具体的な多品詞語解消規則として、動詞と名詞、前置詞と接続詞、動詞のed形について述べる。

5.1 動詞(V)と名詞(N)

VとNの多品詞語は、多品詞語の中で最も多く、それに応じて解消規則もいろいろな場面に対応しなければならない。下記 a) に中心となる解消規則に用いている具体的な手掛りと、その適用例をあげる。ただし、ここでの解消規則は必ずしも常に正しい結果を与えるに限らない発見的な規則となっており、選択に誤りがある場合には、b) に述べるような規則により、誤りの修復を行う。

尚、以下の手掛りは、ごく簡単に記述しており、実際の規則にはここに記していない細かな制限が付加されている。

a) 解消規則具体例 (V,N)

a-1) Nと判断する規則

・形容詞の直後

The book is intended for general use as an introduction ~.

- Vと解釈すると他動詞なのに目的語がこない (能動態のときのみ)
~ to completely eliminate this cause of electrical failures.
- Vと解釈すると他動詞であるが、文末にある Elements of front panel design.
- to以外の前置詞の直後
~ the performance of project groups with ~.
- 冠詞+動詞ing, en形に続くもの
~ has led to an alarming increase in electrical hazards at work.
- 他動詞の直後
~ a daily monitoring system to verify work in process.
- 関係代名詞の先行詞の位置にあるもの
~ induction heating and melting which are established ~.
- 複数の数詞の直後の複数形のN
~ through 1979, 405 reports of electrical accidents ~.
- 特別な名詞の直前に置かれたing形: system, technology など、全体として複合名詞を作ることが多い
Progress in manufacturing technology for electronics is discussed.
- ofの直前の動詞のing形
- 名詞所有格の直後

a-2) 動詞と解釈する規則

- 直後に冠詞など名詞句の始まりを示す語がある
The following article presents the aims of the available training courses, ~.
- 助動詞の直後
Note that a failure to function most often does not present a hazard.
- 複数形のNの直後
Four major factors affect R&D planning ~.
- 主格代名詞の直後
They face complex problems ~.
- 文頭に名詞句が連続することはない
Present a profile of the engineer ~.

- 前に主語と思われるNがあり、後ろにも冠詞など名詞句の始まりを示す語がある。かつ数が一致する。

The second part covers circuit analysis techniques

- 主語と動詞あるいは限定詞と名詞の数の一致から名詞の解釈がとれない場合
This presentation attempts to answer the question
- WH語 (whoseを除く) の直後
~ describes presently used techniques and devices which attempt to ~ .
- 前に二重目的語をとる動詞がない限り、名詞句は2個連続しにくい
~ has a checking time ~ and causes the system ~.
- 動詞の並列
~ who apply, design, and maintain digital switching systems.

b) 誤りの修復規則 (VとN)

解消規則では多品詞解消の決め手のなかったVとNの多品詞語は、とりあえずVとし、そして以降の解析を進める。また前述した多品詞解消の規則も多くの場合に正しい結果を与える発見的な規則であり、常に正しいとは限らない。したがって、解析が進み、文の全体的な構造が見えてきた段階で、あらためて、先の品詞決定を見なおし、矛盾があれば切り換える。

1) 名詞句、前置詞句がまとまった直後

(図1の流れ図では、8)の内部)

主語と主動詞の数の一致を検査し、矛盾があれば切り換える

例 Fatal accidents of electrical maintenance works in factories increase in the area ~.

この例では、workを動詞とする解釈は、主語の主名詞がaccidentsであることがわかった段階で排除される。

2) 文の構造が判ってきた段階

(図1の流れ図では、13)の内部)

一つのSの中に、主動詞が一個しか存在しえないと

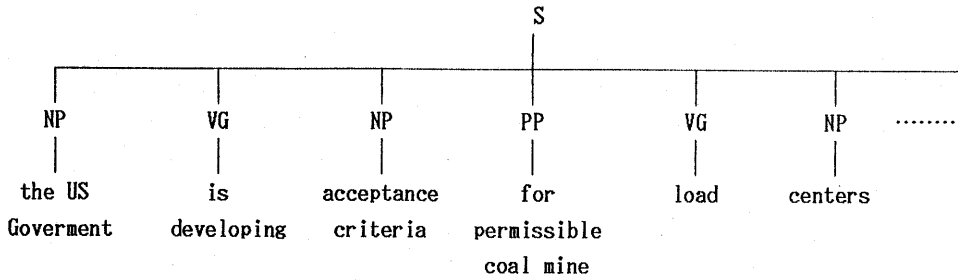


図 2

いう条件を用いる。一つのSの中にVが複数個あった場合、Nに切り換えられるものがあれば、切り換える。

例 The US Government is developing acceptance criteria for permissible coal mine load centers (PLC's) up to 15000 V and 2000 kVA.

この検査を行う直前までは、loadはVとみなされており、一つのSの中に、is developing と loadの2個の動詞が存在することから、loadはNに切り換えられる。(図 2)

5. 2 前置詞 (PREP) と接続詞 (CONJ)

before, as, forなどの語は、後ろに名詞句をとって、PREPとして、あるいは文をとってCONJとして使われる。しかしながら以下に文が続く場合であっても、その文の主語となる名詞句が続いていると、局所的な手掛りだけではPREPかCONJかの判断はできない。したがって、このタイプの多品詞語は、それらの語に続く構造をまず解析してみて、その結果にもとづいて品詞決定を行う。

a) 簡単な解消規則 (PREPとCONJ)

まず、次の手掛りがある場合には、通常が多品詞解消規則と同様に、局所的な手掛りだけで品詞が決定できる。

- ・以降に動詞がなければ、PREP

Remarkably, the printed board can be

executed as a one-sided or a double-sided unit.

- ・直後に名詞句が2つあれば、PREP

For safety purpose a site is considered as ~.

- ・for ~ to の形で使われ、to不定詞句の主語を与える前置詞

Many oppotunities occur for contractors to obtain electrical maintenance work in factories.

b) 上のような規則で品詞が決定できなかったものについては、次のような手順によって品詞を決定する(図1の流れ図の中の12)。PREP, CONJの多品詞語から後ろを切り取り、そこだけに対して解析文法全体を最後までかけてみる。その結果、文にまとまれば、その多品詞語はCONJと判断し、名詞句にまとまれば、PREPと判断する。

例 The solderbility of reflowed tin and tin/40 percent load coated copper has been examined after thermal aging designed to induce extensive copper-tin intermetallic compound growth.

この例では、after以降が名詞句にまとまるので、afterはPREPと判定される。また、次の例ではas以下が文にまとまるので、CONJと解釈される。

例 As power density and power dissipation in VLSI logic chips increase, designing the package for good thermal performance

becomes critical.

次の例では、for以下を文として解釈しようとしても、主語と動詞の数が一致しないので、forはPREPと解釈される。

例 An approach to discover suitable media for silicon epoxy copolymers is outlined.

5.3 動詞のed形

動詞の過去、過去分詞形が同形のものには、

- a) 主動詞
- b) 名詞を後方から修飾する、後位修飾用法
- c) 名詞を前方から修飾する、前位修飾用法

の3つの可能性があり、それぞれの解釈ごとに大きく異なった構文解析結果が得られる。多品詞語解消文法では、b)、c)のうち、単独で名詞を修飾する用法と思われるものを、形容詞(ADJ)に切り換えている。

a) 解消規則 (動詞のed形)

中心となる解消規則では、次のような手掛りをもとに、解消を行っている。

- ・冠詞、前置詞などの直後
the fourth edition presents a unified treatment ~ .

- ・直後が動詞や助動詞のとき
Reasons for the steps taken and the low fire risk liquid selected are indicated
- ・冠詞、形容詞などと名詞にはさまれたもの
The best solution is to realize a technically well conceived construction ~ .

- ・be+過去分詞+NPの形になっているが、二重目的語をとる動詞ではないので、受動態と解釈できないもの
~ melting which are established industrial processes ~ .
- ・形容詞用法の動詞の並列
~ a trained and motivated workforce.
- ・there + be に続くもの

b) 修復規則 (動詞のed形)

ed形の動詞の直後に名詞がきている場合、二重目的語をとる動詞でない限り、後位修飾の用法である可能性は少ない。考えられるのは、主動詞としての用法か、あるいは前位修飾の形容詞的用法である。

このような場合には、解消規則では、とりあえず主動詞の解釈をとる。解析が進んで、文や節のまとまりの範囲がわかるようになった段階で、矛盾がないかを検査する。一つのレベルに主動詞が複数個存在することはあり得ないので、そのような場合には、ed形の動詞で前位修飾の解釈が可能なのものを、形容詞に切り換える。この規則は、流れ図(図1の13)に組み込んでいる。

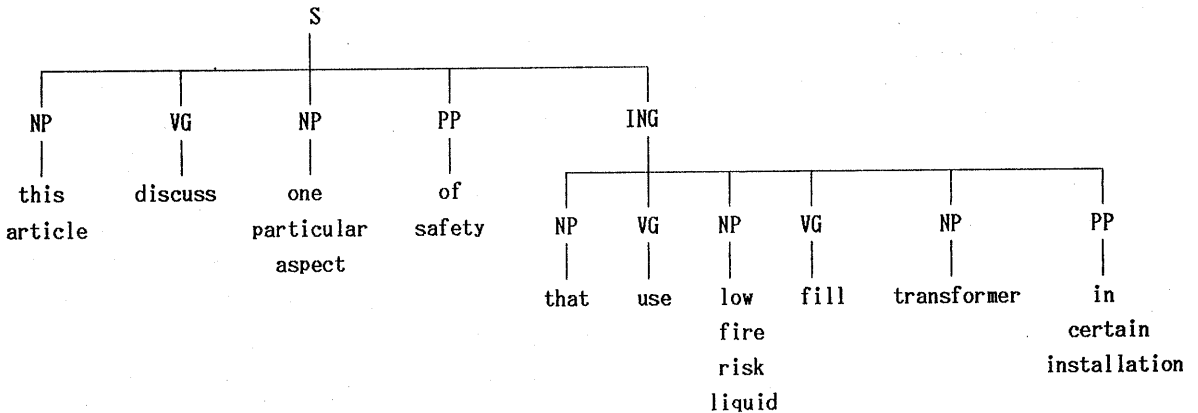


図 3

次の例によって、その動作を具体的に述べる。

This article discusses one particular aspect of safety using low fire risk liquid filled transformers in certain installations.

文の一時的な構造化の処理では、まず文や節の切目を認識し、図3のような右下りの解析木を作る。

この木をもとに、いろいろな係先決定文法を適用して、正しい解析木に変形していくわけである。その過程で、それまでの多品詞語解消規則の誤りを修復する

規則が適用される。

この例においては、現在分詞系の存在を示すINGのノードの下に、主動詞が2個(useとfill)が存在しており、許容できない構造となっている。これを修復するには、2つの可能性がある。一つはfill以下を一段上のレベルに上げることである。もう一つは、fill以下を前位修飾の形容詞用法と判断し、品詞を切り換えてしまう方法である。しかし、前者の方法でfill以下を1レベル上に上げると、今度はその上げたレベルにdiscussとfillの2個の主動詞が存在するようになってしまう。したがって、この場合には、形容詞に切り

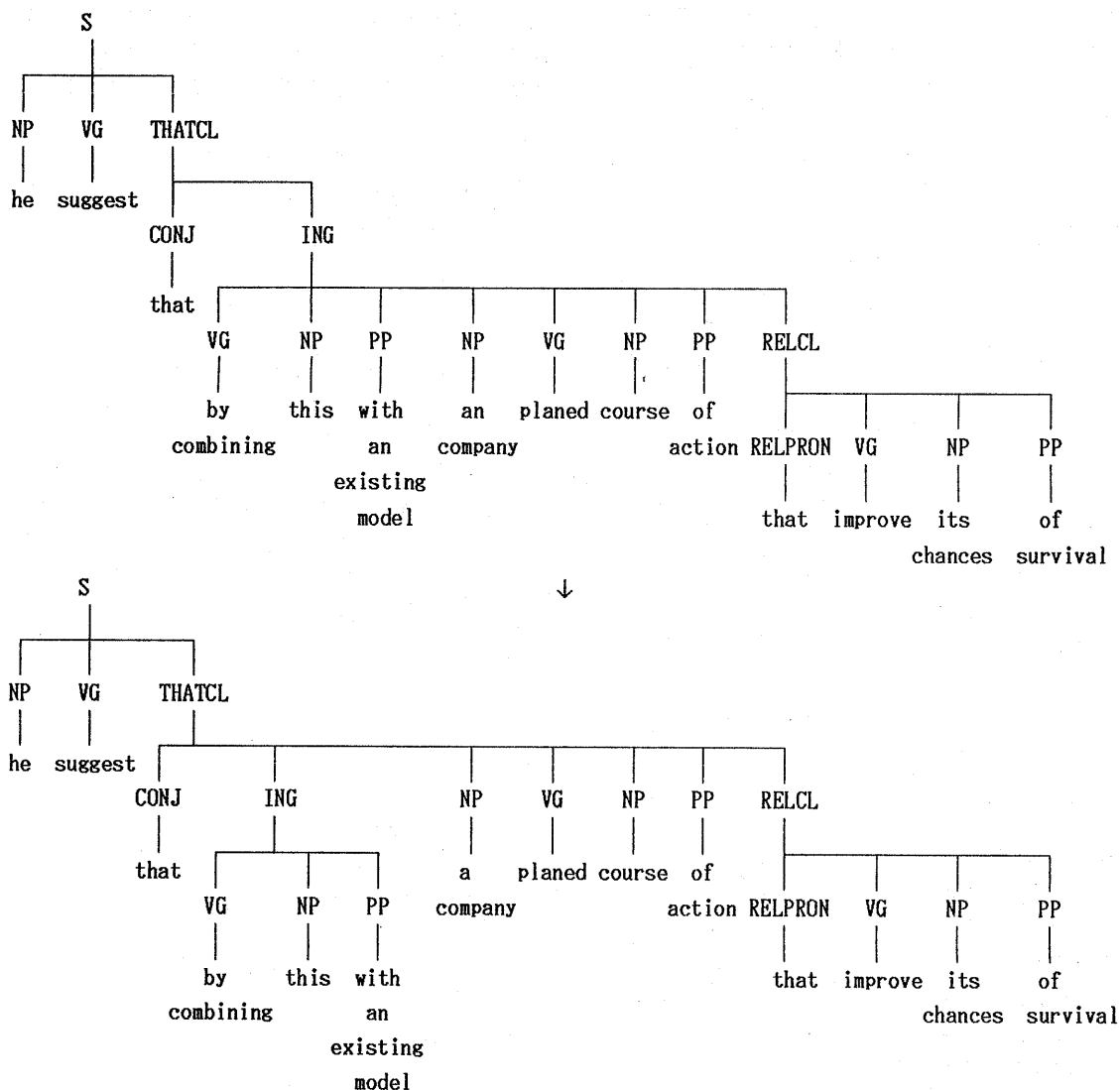


図 4

換えるのを最も妥当な解として採用する。

次の例は、局所的にみると動詞のed形が前後の名詞に挟まれており、先の例と同様な形であるが、これは形容詞には切り換えず動詞の解釈を採用する。

He suggested that by combining this with an existing model a company planned courses of action that would improve its chances of survival.

この例ではINGのすぐ下のレベルにはcombiningとplanの2つのVGがあるが、その上のTHATCLのレベルにはVGがないので、図4の下の方のように動詞planned以下(この例ではplannedの主語であるa companyを含めて)を一段上のレベルに持ち上げる。

6. まとめ

一般に、英語の文には多品詞語がいくつも存在する場合が多いので、解析開始時点ではきわめて多くのあいまいさがあることになる。これらの多くは構文解析が終了した時点では、文全体の解釈に結び付かず、その文中の品詞としては棄却される。したがって、構文解析を最後まで行くと、品詞レベルのあいまいさは解消できる。しかしながら、これを実際に実行するとすると、すべての単語のすべての品詞解釈について、組み合わせ的に解析を行わなければならないこと、また最終的な解析結果に至らない品詞解釈でも局所的には、それを含む可能な解釈が存在することなどから、実行に要する時間、必要記憶量とも爆発的に増大してしまうことになる。

我々のシステムでは、このような品詞解釈のあいまいさに伴う処理量の爆発を、次のような手順で解決している。

1. 解析初期の段階で、多くの場合に正しい解釈を与える多品詞解消規則によって、品詞を暫定的に判定する。
2. 英語解析の主要部分は、これをもとに統語解析を行う。1.の品詞判定のために、組み合わせ的な処理の爆発は避けられる。
3. 1.の品詞判定は誤ることもあるが、これは解析を行った後、大域的な構造を参照することによって、修復される。

また、ある単語の品詞解釈が、人間にとって正しい解釈から見ると誤っている場合でも、文全体として統語的には正しい解釈を作り出すことがある。たとえば次の例

Many opportunities occur for contractors to obtain electrical maintenance work in factories.

では、forを接続詞、workを動詞としてcontractor以下を文とする解釈もありうる。我々のシステムでは、最初に多くの場合に正しい品詞解釈を与える規則が適用され、その解釈が矛盾に導かない限り、その解釈を採用することになるために、このような(思ってもない)解釈を排除することができる。上の例では、5.2節で述べたような、for ~ toという強い手掛りがあるために、forを前置詞とする解釈が優先して採用されることになる。

完全な文脈処理、意味処理が期待できない機械翻訳においては、このような表面上の手掛りを利用した解釈の優先度づけは不可避であると考えられる。

7. おわりに

Muプロジェクトの英日機械翻訳システムにおける多品詞語解消について述べた。

英日機械翻訳システムは、現在INSPEC抄録文の中から抽出した約1000文を対象に実験を行っている。これまでに書かれた解消規則では、約9割の正答率(単語数で勘定した場合)を得ているが、現在なお改良中である。

[参考文献]

1. 久米 他; Muプロジェクトにおける英語解析過程
情報処理第31回全国大会 pp.1333-1334
2. 辻井 他; Muプロジェクトにおける英日翻訳システムの構成
NL研50-3, 1985.7.