日英機械翻訳における日本語解析前半部の一構成

3G-2

高山泰博，鈴木克志，丸山冬樹，太細　孝
三菱電機　情報電子研究所

1．はじめに

我々は、実用化を目指してMELCOM−PSIII上に日英機械翻訳システムMELTRAN−J／E（以下、MELTRAN）の開発を行なっている。本稿では、MELTRANにおける日本語解析前半部分の構成と、実用システム構築における問題点について述べる。

2．MELTAN−J／Eの全体構成

全体構成図を図1に示す。特徴は以下の通りである。
①ﾄﾗﾝｽﾌｧ方式（日本語→英語一方向）である。
②ｼｽﾃﾑ全体をESP（拡張Prolog）で記述している。
③辞書記述・内部構造・文法規則が統一的な形式である。
④解析・変換・生成用の3種類の文法記述言語を持つ。
⑤最大翻訳速度1万語／時間、辞書登録語数8万語。

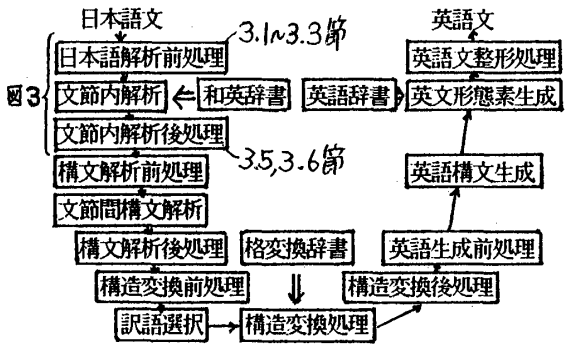

図1　MELTRAN−J／Eの全体構成

処理過程全体を通して、各処理フェーズの間で受け渡す情報はADS(A Dependency Structure)と呼ぶ内部構造である。ADSは，Prologのd-リスト形式によって、文中の要素間の依存構造を表現している。ADSは、ノードとリンクから構成され、それぞれに複数個の属性と属性値の組を持つ。処理過程の各時点で、そのADSが表わす構造によって文節ADSや構文ADSと呼ぶ。

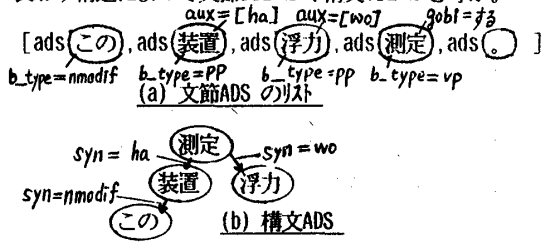

図2　文節ADSのリストと構文ADS

An organization of the Japanese morphological analysis phase in Japanese-English machine translation
Yasuhiro TAKAYAMA, Katsushi SUZUKI, Fuyuki MARUYAMA, Takashi DASAI　Mitsubishi Electric Corp.

3．解析前半部

MELTRANの解析前半部分の構成図を図3に示す。第3章では、図3の各フェーズの処理について述べる。

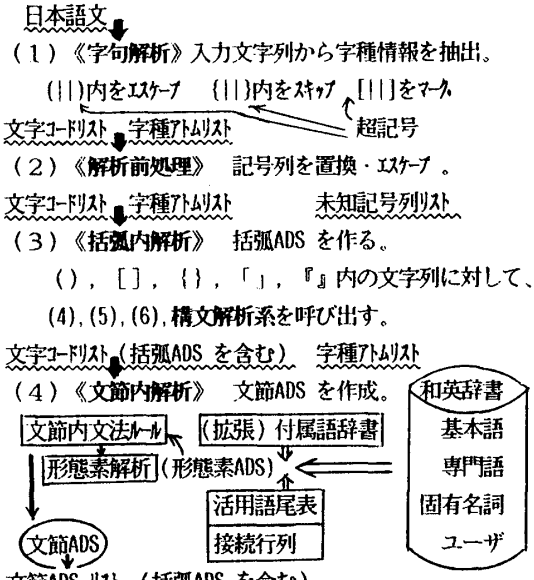

図3　解析系前半部分の構成

3.1 字句解析

MELTRANでは、入力として漢字かな混じりの日本語文を仮定している。字句解析では次の処理を行なう。
①入力文の各文字ｺｰﾄﾞに漢字・ひらがな・カタカナ・外国文字・数字・記号のいずれかの字種情報を設定する。
②括弧や超記号の左右の対応を検査する。

原文作成者が入力した日本語文には、入力誤りが含まれることがある。例えば、閉じ括弧を入力し忘れる等である。一部の入力誤りによって、文全体の翻訳が失敗してしまっては、損失が大きい。そこで、②の処理により、左右の対応を取ることができない括弧には、記号としての字種情報を与える。これによって、入力誤りが全体に与える影響を減らす。

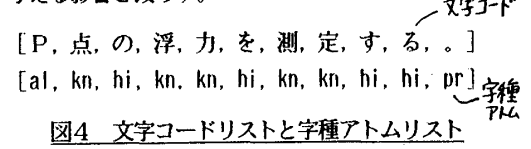

図4　文字コードリストと字種アトムリスト

3.2 解析前処理

現在の多くの機械翻訳システムは、技術文書を翻訳の対象とする。技術文書は、製品名などに記号、外国文字、数字などを多数含む。すべての記号表現を翻訳前に辞書に登録するのは、利用者の負担が大きい。また、記号表現は、訳文でも原文と同一の表現を使用する。

そこで、解析前処理フェーズにおいて、辞書に登録していない記号表現を入力文字列から抽出し、未知記号列リストに蓄えておく。文節内解析(3.4節)において辞書検索と同時に未知記号列リストからも単語を探す。これにより、辞書に未登録の記号列を、未知語としてではなく、辞書に登録してある語と同様に処理できる。

また、典型的な時間表現などを捕まえておき、文節内解析(3.4節)や数量表現解析(3.6節)の負担を軽減する。

3.3 括弧内解析

技術文書中で頻繁に使用する括弧に囲まれた文字列は、文中で様々な役割を担う。括弧内の文字列を、文全体の解析と同時に扱おうとするとルール記述が複雑になる。

そこで、括弧の外側の部分の解析に先立って、括弧内の文字列に対して解析処理を施し、日本語の構文ADSを作成する。このADSを入力文字コードリスト中の括弧があった位置に挿入(図5参照)しておき、文節内解析(3.4節)時に他の文節との依存関係を解析する。

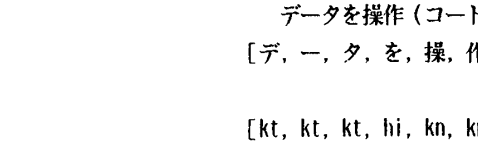

図5 括弧内解析の出力

3.4 文節内解析

このフェーズでは、文脈自由ルールの形式で記述してある文法に基づいてボトムアップに文節内の構造を解析する。また、日本語の意味的な構造を正しく解析するために拡張付属語表現(文献2)を利用している。

ルール記述の際に、当初は、ある程度理想的な日本語文を仮定して記述する。しかし、実際には、通常の文法には、本来存在しないルールも記述しておく必要がある。例えば「名詞が副詞用法を持つ場合、単独で文節となる」とすると次の例は解析に失敗してしまう。

例　　この文法の記述には問題ない。

この場合、「問題な」が自立語辞書に形容詞として登録されていないと、「問題」単独では、文節として成立しない。そのため解析に時間がかかり、かつ正しい解析結果を得ることができない。そこで、

複合形容詞→名詞+形容詞　というルールを追加する。この他にも、現実に使用されている文に以下の例がある。

例　　不良部品を実装不可。

これらは、文節内解析の後処理か、構文解析の前処理において、構文解析が失敗しないように、断定の助動詞「である」を補填して内部構造を補正する必要がある。

3.5 局所構文解析

MELTRANでは、構文解析の前処理部の並列表現処理において、語の形態的な特徴や意味分類情報などを利用して、並列表現を抽出する。未知記号列(3.2節)などには意味分類情報を設定できず、並列要素の解析を誤ることがある。そこで、原文作成者が係り受けの優先度を指定する機能が必要である。この機能は、文節内解析の後の局所構文解析で実現している。局所構文解析の主な処理は、次の通りである。

(1) 文節ADSのリストから係り受け指示記号(図3参照)で囲まれた文節ADSの並びを抽出する。

(2) 局所的に以下の構文解析系①～⑥を呼び出して部分的な日本語ADSを作成する。

①数量前処理②重文処理③並列表現処理

④名詞修飾句処理⑤文節間構文解析⑥主題・格解析

3.6 数量前処理

日本語文中における数量表現の形態は多岐に渡る。そこで、翻訳処理の過程の中で、数量表現を専門に処理するフェーズを設けている。数量表現の局所性から、解析の前半部分に位置付けている(文献3)。

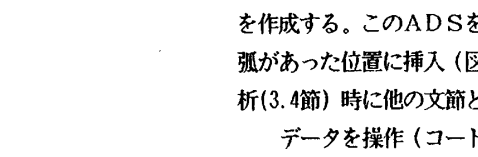

図6 数量前処理の例

4. 解析に用いる辞書情報について

辞書引きは、翻訳処理の全過程において最も時間を要する。翻訳方向が一方向なので、自立語辞書に和英辞書を用いて英語の情報も取り込み、2次記憶へのアクセスを減らしている。拡張付属語辞書は、主記憶に置いて高速化を図っている。自立語辞書も付属語辞書もPrologターム形式である。この辞書情報を、属性と属性値に展開し形態素ADSを作る。文節内解析では、自立語と付属語の形態素ADSから文節ADSを構成する。

また、カタカナ異表記処理(文献4)による仮想的な辞書見出し増強により未知語の発生を軽減している。

5. おわりに

本稿では、実用システムにおける日本語解析の前半部分の構成例と問題点の一部を述べた。高速処理を利点とする一括翻訳中心の機械翻訳システムを実用に供するためには、コンパイラにおける誤り回復処理的な機構も実装しておく必要がある。

参考文献

(1) 鈴木：日英機械翻訳システムMELTRAN-J/E, bit別冊機械翻訳, 共立出版(1988).

(2) 首藤ほか：日本語における語の固定的共起, 文法的知識と意味的知識の蓄積管理, 信学会シンポジウム(1989).

(3) 吉武ほか：日本語の数量表現とその英語への機械翻訳に関する一考察, 第77回情処NL研報(1989).

(4) 伍井ほか：カタカナ異表記処理, 第38回情処全大(1989).