日英機械翻訳システムにおける長文の解析

1F－1

藤井洋一　鈴木克志　丸山冬樹　太細孝
三菱電機株式会社　情報電子研究所

１．はじめに

機械翻訳システムで長文を処理する場合の問題点は、長文を構成する「文」と「文」の係り受け関係の認識と処理時間の増加にある。このうち係り受けの問題を処理する手段として、文脈処理等の高度な技術研究が活発に行なわれている。しかし、文脈処理は十分確立した技術とは言い難く、現在の機械翻訳システムで導入できるような状況とはいえない。したがって、現在の翻訳システムでは前編集の指針として、[1] のように長文を短文に分割する事などを推奨している。しかし、長文を短文に分割する方法は、ある種類の文では可能であるが、すべての種類の文に対してうまくいくとは限らない。そこで、本稿では長文を分割することなく正しい係り受けを解析することを目標として、我々が翻訳対象としている技術文書（輸出用パソコンマニュアル）を分析し、係り受けの決定に、構文情報を基にしたヒューリスティックスを用いることを考察した。

２．検討の対象

長文の定義は曖昧であるが、今回は以下に定義する「要素文」の数が３つ以上のものとして検討を進める。

＜定義＞

一般に文章は単文・複文・重文に分類される。そのうち、述部と述部の関係を現す表現（「連用中止」，「接続助詞」等）に注目した。具体的には等位接続詞で結ばれる重文と、副詞節で結ばれる複文がそれに当たる。ここで、述部と述部で区切られるそれぞれの文を「要素文」と呼ぶ。

～し、～した場合、～である。
要素文　要素文　要素文

３．日本語表現の抽出

今回用いた例文は、当社の輸出用パソコンのマニュアル文であり、約１８００文からなる。そのうち、長文（３つ以上の「要素文」からなる）は１３０文ほどであり、それらの文について検討を行なった。注目した日本語の表現としては以下のようなものがある。

(a)　～し（連用中止），～して（連用形＋接続助詞『て』）
(b)　～するが（連用形＋接続助詞『が』）
(c)　～したり（連用形＋並列助詞『たり』）
(d)　～するよう，～するように
(e)　～するとき（時），～するとき（時）は，～するとき（時）には，～する場合，～する場合は，～する場合には
(f)　～しながら
(g)　～する前に，～た後で，～してから，～するまで（迄）
(h)　～するため（為），～するため（為）に
(i)　～すると
(j)　～のに
(k)　～なら，～ならば

４．抽出文の考察

３．で述べた１３０文に対して考察をおこなうと次のようなことがわかる。なお、以後、(a) ～(k) の記号はそれぞれ３．で分類した要素文を表すこととする。

・まず、要素文の数について

(1) 多くは、要素文の数が３つ、または４つからなっていた。また、構造は、極端に複雑な構造になることは少なく、「～する前に～する場合は、～する。」といった分かりやすいものが多く見られた。これは、マニュアル文の性質として、物事をわかりやすく、明確に表現しなければならないことが理由といえる。

(2) 要素文が多い場合は、主に次の２つのことが原因であった。

(a) の要素文が連続

(例) システム・ユニットからカバーを傾け持ち上げて取りはずし、脇へ置きます。

「～ならば～し、～ならば～する。」といった一定のパターンが連続。

(例) サイクルを停止したい場合はＹを入力し、そうでないときはＮを入力します。

・次に、係り受けの特徴について

(3) (a) の要素文の係り先は基本的に近い。遠くに係る場合は、(2) で示した「～ならば～し、～ならば～する。」のような一定のパターンが多い。

(例) 電源が入ると、赤色のランプが点滅し、しばらくしてビープ音が聞こえます。

Analysis of Long Sentences in Japanese-English Machine Translation System
Youichi FUJII, Katsushi SUZUKI, Fuyuki MARUYAMA, Takashi DASAI　MITSUBISHI Elec.Co.Ltd.

(4) (a) の要素文の連続に対しては最後のモーダルが伝わる。具体的には、マニュアル文で「～し、～してください。」といった表現は翻訳された場合には、ともに命令文として処理されるべきである。

(例) 前面パネルの左下に付いている調節ツマミを回して、明るさとコントラストを調整してください。

5．係り受けヒューリスティックス

4．で考察したことをヒューリスティックスとしてまとめると、次のようになる。それぞれについて、典型的なパターンを示した。

述部の並列があれば、優先的に処理する。

(1) (c) のある場合、後続する(C) があればそれに係り、なければ近い要素文に係る。

～したり、～たりするために、～する。

パターンに合ったものを優先的に処理する。

(2) (a) を含む一定のパターンの場合、(a) が遠くに係るように優先的に扱う。

～した場合、～し　～した時は、～する。

以下順に処理する。

(3) (b) の要素文は遠くに係る。

～するが、～し、～である。

(4) (d),(f),(g),(h),(j),(k) は直後に読点がなく、直前の要素文に読点がある場合は、直前の要素文が係らない。

～し、～するよう～する。

(5)(a)は、近い要素文に係る。

～すると、～し、～する。

(6)(e),(i)は(a) に係らず、それ以降の節に係る。

～する時、～し、～する。

最後にモーダルを伝える。

(7) 最後が命令文であれば、その情報を(a) に伝える。

～する時は、～し、～して下さい。
(命令)

このように、要素文間の係り受けの性質を分類し、処理することができる。

6．ヒューリスティックスの評価

マニュアル文は、物事をわかりやすく、明確に表現しなければならないので、入力文が極端に複雑な構造になることは少なく、定型的な表現が多い。この事から、5．のヒューリステックスは有効に働くと考えられる。また、ひとつの前編集の指針にもなると考える。

7．おわりに

係り先のパターンは、5．で述べたヒューリスティックスだけではうまくいかない。それはどうしても意味情報を考慮しなければうまく係り先を決定できない場合が生まれてくるからである。今後は、これらのヒューリスティックスを実装してみるとともに、うまくいかない場合に、どの様な意味情報まで考慮すれば、よりうまく係り先を決定できるかの考察が課題である。

（参考文献）

[1] 奥田他：「仕様記述用制限日本語のための長文分割」，
情報処理学会第33回全国大会予稿集(1986).

[2] 信国　：「自然言語における長文分割方式」，
情報処理学会第39回全国大会予稿集(1989).