位置換えによる機械翻訳の一方式

2B-1

沢村孝至　宮永喜一　栃内香次

北海道大学

1　はじめに

本稿では形態素解析と位置換えにより英文を日本語に翻訳する方法を提案する[1)2)]。本方式はトランスファ方式とダイレクト方式の中間に位置する方式である。トランスファ方式の構文解析においては「形態素をいかに結び付けるか」が重要な問題であるが、本方式では、「いかに分割するか」「分割されたものをいかに順序づけるか」がテーマである。

この方式はトランスファ方式[3)4)]と比較してアルゴリズムが単純なため、システムを簡単に構築でき、処理も速い。また辞書の構造も簡単である。

2　処理方式

この方式は最終的に英単語列を逐語訳したときにその日本語単語列が正しい日本語文と一致するようにあらかじめ英単語列を並べ換えるものである。

英文において特定の形態素Cはそれより前に位置する単語列Aおよび後ろに位置する単語列Bとの間で次のような性質をもつ。すなわち、

パターン1：英文におけるA／C／Bという語順を訳文においてはA／B／Cという語順にする
（例：Cは助動詞・動詞）

パターン2：英文におけるA／C／Bという語順を訳文においてはB／C／Aという語順にする
（例：Cは前置詞・分詞・従属接続詞など）

パターン3：英文におけるA／C／Bという語順を訳文においてもA／C／Bという語順にする
（例：Cは等位接続詞）

これらの関係を満たす単語列{A／C／B}をひとまとめにセグメントと呼ぶことにする。図1に各パターンの具体例を示す。

以下、Cとなる形態素をセグメントにおけるコアと呼ぶことにする。セグメントの長さ（AおよびBの長さ）と位置換え処理はコアごとに決まり、より長いセグメントにあるコアほど処理の優先順位が高く、先に処理される。たとえば従属接続詞をコアとするセグメントは複文全体であり、AおよびBはそれぞれ主文・従属文になる。動詞のセグメントは単文全体であり、Aは動詞より前の句（多くは名詞句である）・Bは動詞より後ろの句（名詞句または副詞句）である。したがって従属接続詞は動詞より先に処理される。つまり、文全体でみるとセグメントは入れ子状になる。この例を図2に示す。

このセグメントの考え方によって、翻訳処理を行う。

全体の処理は大きく分けて形態素解析・分割・位置換え・逐語訳の四段階に分けられる。形態素解析ではおもに品詞の隣接情報をもとに形態素を決定する。分割処理の過程では文を位置換えの単位となる各要素に分割する。そして位置換え処理において分割された各要素の位置関係を変える。この結果となる単語列を逐語訳すると日本語訳が得られる。

以下に各段階の詳細を述べる。

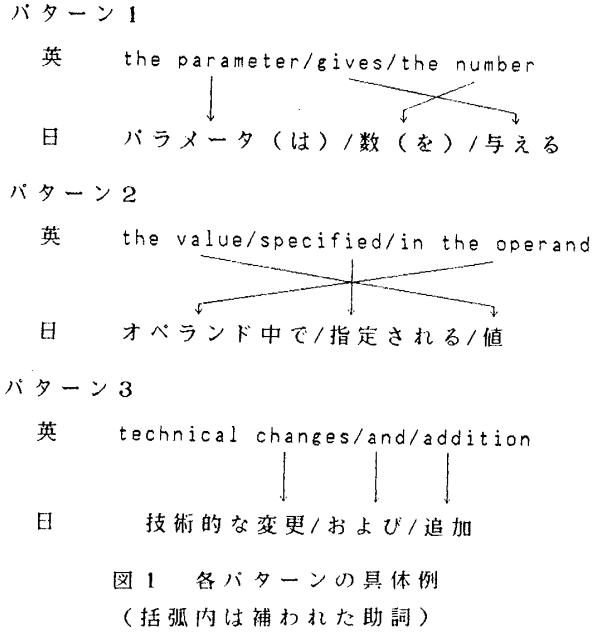

図1　各パターンの具体例
（括弧内は補われた助詞）

2．1　形態素解析

ここでは従来の形態素解析方法[3)]に加えて、品詞の隣接情報をもちいて各々の単語の品詞を決定する。隣接情報とは、例えば冠詞の直後には形容詞か名詞が来やすい、といったような2つの品詞の並びの組合せの頻度情報である。

本方式では必ずしも全ての単語の品詞を決定する必要はなく、語とその位置によっては曖昧なままで構わない（例:冠詞と名詞の間の

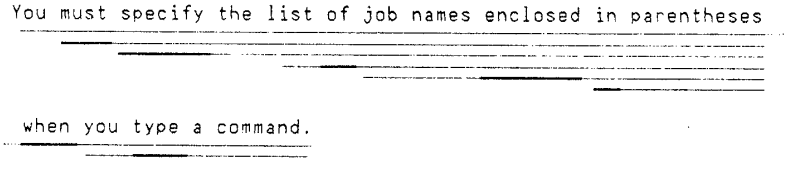

図2　セグメントの入れ子
（細線がセグメント・太線がコアを表す）

MACHINE TRANSRATION USING WORDS-INTERCHANGE METHOD

Takashi SAWAMURA, Yoshikazu MIYANAGA, Koji TOCHINAI

Hokkaido University

語など）。多品詞語はこの時点で品詞と訳語を決定する。多義語も前後の品詞からその訳語が決定できるのなら決定する。また、イディオムは一つの形態素と見なし、分割しない。

コアが等位接続詞の場合、並列のレベル（接続する単語列が何かということ、すなわち単語か句か節か）によって分割処理の優先度が異なってくる。したがって、分割よりも前に等位接続詞の前後の単語列を比較し、どんな品詞が含まれているかによって並列のレベルを判断する。等位接続詞と共に用いられる並列の区切りのカンマも同様の処理を行う。

2．2 分割

セグメントの処理においては、語順の転換（位置換え）に先だってまずコアごとに、コアの直前および直後で形態素を分割する。分割する場所には区切り記号"／"を入れるがそれらは２つずつコアに対応している。処理は再帰的に行われる。区切り記号から区切り記号までの間がセグメントである。（１文全体は区切りで囲われていると考える。つまり分割処理の最初は文全体をセグメントとして開始する。）その中で処理の優先度の高い形態素からコアとして、セグメントの先頭からコアCの直前までの単語列をA・およびCの直後からセグメントの最後までの単語列をBとし、A／C／Bと分割する。次のステップでは今のAやBが新たなセグメントとなる。セグメントを次第に短くしていき、コアとなる単語がなくなるまで分割する。図３に分割処理の例を示す。

```
1-a.The personal computer is invaluable as a personal filing system.
                                C
1-b.The personal computer/is/invaluable as a personal filing system.
                 A          C                    B
2-a.                                invaluable as a personal filing system
                                              C
2-b.                                invaluable/as/a personal filing system
                                         A      C            B
3.  The personal computer/is/invaluable/as/a personal filing system.
```

図３　分割処理の例　（3.が最終結果）

2．3 位置換え

優先度の高いセグメントから順に、そのコアの品詞と区切りに従ってセグメント内の位置換えを実行する。これもやはりトップダウンに、入れ子状のセグメント全てについて長いセグメントから短いものへと行う。例を図４に示す。

```
1-b.The personal computer/is/invaluable as a personal filing system.
                 A          C                    B
1-c.The personal computer/invaluable as a personal filing system/is.
                 A                                B ,                C
2-b.                                invaluable/as/a personal filing system
                                         A      C            B
2-c.                                a personal filing system/as/invaluable
                                              B            C      A
4.  The personal computer/a personal filing system/as/invaluable/is.
```

図４　位置換え処理の例

（1-bと2-bは図３に対応している。4.が最終結果）

2．4 逐語訳

全ての位置換え処理が終ったとき、その単語列の各単語について逐語訳を行い、適当に助詞をおぎなえば和訳が得られたことになる。例を図５に示す。

得られたテキストの日本語の整形（動詞の活用語尾など）は日本語文書処理で自動に行えるが必要ならば形態素解析の結果を用いてもよい。

```
4.  The personal computer/a personal filing system/as/invaluable/is.
5.  パソコン（は）/個人用ファイリングシステム/として/重要/である。
```

図５　逐語訳の例

（4.は図４に対応している。括弧内は補われた助詞。5.が最終結果）

3　辞書と文法

意味解析を行わないため、単語辞書そのものは単純でよい。形態素情報とそれに対応する訳語がそろっていれば足りる。

文法辞書は持たない。文法知識はアルゴリズムと一体化しているため、文法が複雑化すれば、プログラムも複雑化する。しかし、対象とする分野を限ると辞書及び文法は単純化できる。

4　試作システム

以上のアルゴリズムをもとにプロトタイプの翻訳プログラムを試作した。これはprolog-KABA上で２００行程度であり、上に述べた全ての機能を組み入れてはいない。

なおサンプルとして用いたテキストはコンピュータマニュアルの英文である。

このプログラムによって、アルゴリズムの基本的な正当性は確認されている。

5　終わりに

出力された日本語テキストの後処理など、まだ実現していない機能があるのでそれらを実現することが急務である。

形態素解析において一意的に決定できない単語が出現した場合どうするかという問題もある。（分野を限った場合、その出現頻度はかなり低いと思われる。）それについては、統計的な重みをつけていくつか候補を絞ることも考えている。

参考文献

1)沢村・宮永・栃内:「品詞間の結びつきを用いた機械翻訳方式の提案」,昭和６２年電気関係学会北海道支部連合大会講演論文集,283.

2)沢村・宮永・栃内:「単純な位置換えによる機械翻訳方式の提案」,昭和６３年電子情報通信学会春期全国大会講演論文集,D-271.

3)情報処理,「機械翻訳」特集,Vol.26,No.10 (1985).

4)電子情報通信学会誌,「自然言語処理」小特集,Vol.70,No.9 (1987).