

論文表題の日英機械翻訳システム

長尾 真, 辻井 潤一, 伊吹 潤⁰, 上田 良寛 (京大工)

1. はじめに

近年, 科学文献や商業文に対する翻訳の需要の高まりとともに, 機械翻訳システムの研究が盛んに行なわれるようになり, 近い将来, 実用になるであろうシステムの発表もいくつかされている。しかし実用化のために解決すべき技術的問題は非常に多い。しにがって, 現時点での科学技術情報の国際流通の緊急性を考えると, 論文の表題だけでも翻訳でき, それを用いてデータベースアクセスなどが可能となれば, その効果は非常に大きいと考えられる。ここで述べる翻訳システムは, 対象を論文表題の日英翻訳に限定し, オンラインでの文献検索に利用することを想定して設計した。これは論文表題の文法的構造が比較的単純であること, 文献検索に対して必要な翻訳の質があまり高くないであろうこと¹⁾を考えると, 現時点の技術でも実用になると考えたものである。

システムで用いた辞書, ルールは上記の理由のため, できる限り簡潔なものとした。又, 本研究での翻訳の対象文献は, 情報処理学会誌「情報処理」(Vol.1 No.1 ~ Vol.2/ No.12)に掲載された約7,000の論文の表題である。

2. 処理の方針

論文表題の言語的な特徴として次のものがあげられる。

- (1) 体言止め, あるいは体言+連用修飾助詞で終わっている場合が多い。
- (2) 複雑な係り受けが少ない。連体修飾の場合は, 直後の名詞に, 連用修飾の場合は, 最も近い動詞に係ると考えてよい。

これらの特徴を活用し, 処理の簡略化を図るため, 次のように設計の方針を決定した。

- (1) 原則的に処理を決定論的に進めること。これは特に係り受けの関係を一意に定めることによるあいまいさの解消をわらっている。
- (2) 深い解析をさけること。例えば, 図1では「 N_1 に対する N_2 の N_3 」という名詞句に対し, 2通りの解析結果が得られることがわかるが両方の解析結果に対し, 「 N_3 of N_2 to N_1 」という同じ翻訳が得られることになる。本システムでは, 解析の際に生じるあいまいさをできる限り翻訳結果に残すこととし, 深い解析をさけるようにしている。

係り受けの範囲の決定は次のようにして行なう。図2のような構造の名詞句が入力されたとする。(2)の範囲内の連用修飾句(名詞+連用修飾助詞等)は必ず V_1 に係ることとし, V_2 あるいはさらに外側の動詞句に係ることはないとする。したがって V_2 の埋め込み構造の処理の際には, (2)の部分を無視することができ, よって埋め込み構造の処理を

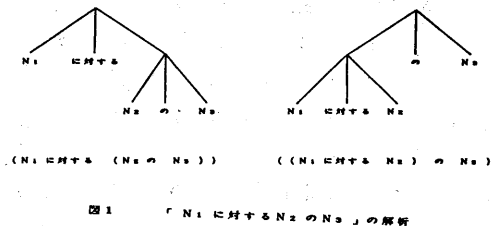


図1 「 N_1 に対する N_2 の N_3 」の解析

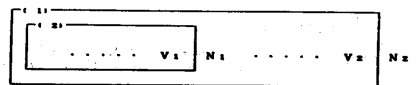


図2 係り受けの範囲 V_1, V_2 : 動詞 N_1, N_2 : 名詞

一段ごとに決定論的に進めることができる。

以上のことから、処理の手順を次のように定めた。

- (1) 最初に、動詞句を含まない名詞句について解析し、部分訳を各々の名詞句につけておく。これによって図2の構造がつくられる。
- (2) 次に、動詞の埋め込み構造を、先頭のものから順に見てゆき、埋め込みの一段ごとに動詞の修飾関係を解析し、部分的に翻訳してゆく。

3. システムの基本構成、辞書、ルールについて

システムの基本構成を図3に示す。

本システムは入カテキストの単語への分割を行なう単語分割システム⁽²⁾と入カ文の解析、英文の表題の生成を行なう翻訳システムから構成されている。

2つのシステムは、共通の辞書から変換されるシステムに用いる辞書によって文法的処理のためのデータを得ている。

辞書における品詞の分類を図4に示す。連体修飾助詞の例でわかる通り、一般には、さらに細分され、解析されることの多い一連の単語でも、文中で果たす機能を考え、一語として扱っている。例えば「に対して」は多くの場合、「連体修飾助詞+動詞の連用形+接続助詞」として扱われるが、本システムの辞書においては連体修飾助詞として扱い、処理の簡略化を図っている。

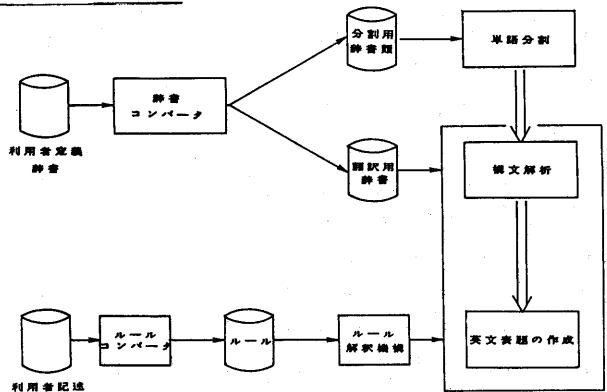


図3 翻訳システムの基本構成

シンボル	品詞名	一般の分類
N	名詞	名詞
D	助詞	助詞
JODO	助動詞	助動詞
RENDAI	連体詞	連体詞、形容詞、形容動詞の連体形
RENVO	連用詞	動詞、形容詞、形容動詞の連用形
TAIJO	連体修飾助詞	の、からの、に対して、等
YOJO	連用修飾助詞	が、から、に対して、等
HEIJO	並立助詞	と、及び、または、等
SETSUJO	接続助詞	接続助詞

図4 本システムにおける品詞分類

動詞

(かいしゃく・する、解釈・する、
(KAISHAKUSURU

= INTERPRET
= INTERPRETING
= INTERPRETED
= OBJ

形容詞

(はっけんてき・だ、発見的・だ、
(HAKKENTEKIDA

= HEURISTIC
= HEURISTICALLY

名詞

(こころみ、試み、
(KOKOROMI

= ATTEMPT
= 1
= AT

(アルゴリズム

()

N = (ALGORITHM)

VQ = 3

QUALIST = (((学習 識別) LEARNING RECOGNITION ALGORITHM)

((数式 処理) ALGORITHM FOR FORMULA

MANIPULATION)

((非決定性) NONDETERMINISTIC AUTOMATON)

((最小 誤り 訂正) LEAST ERROR CORRECTION

ALGORITHM)

))

図5 辞書記述の例

利用者記述の例を図5に示す。動詞については必須格、各活用形に対する記述が書かれる。形容詞に関しては連用形、連体形に対する記述が記述される。名詞に対しては対応する記述の他に名詞連続熟語に関する QUALIST と呼ばれる記述、0-KAKU 属性などが書かれる(これらに対しては第5章で詳細に述べる)。

利用者記述ルールの例を図6に示す。このルールは「～について」のように全体が連用修飾句となる場合の処理を記述したものである。各々のルールは、ルールユニットのいく

```

00100 INITIAL SEQUENTIAL
00200   ( ( (APPLY *PREPROCESS #1) ) )
00300
00400   ( ( (JPOS=8ENV0) *-(1)JPOS = 8H1JPOS = 8Y0J0) )
00500     =1
00600   (ADV #11PREP #3) (INM *NNPJTOTAL #2 #3) )
00700
00800
00900   ( *-(1)JPOS=8H1JPOS=8Y0J0) )
01000     =1
01100   (IPREP #3) (INM *NNPJTOTAL #1 #2) ) )
01200
01300   ( (APPLY *NNPJTOTAL #1) ) )

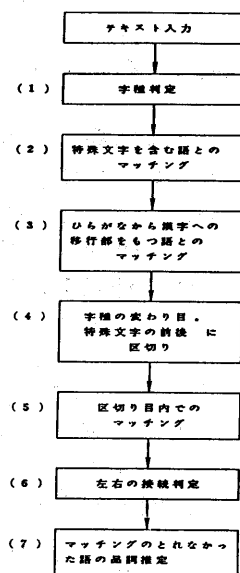
```

れは、品詞列の最後尾に「名詞」+「連用修飾助詞」という品詞があった場合に、
 連用修飾助詞に対応する前置詞を最初にあき、残りの部分に %NP#TOTAL ϕ という
 ルールを適用することを意味している。

主種の分種

カタカナ
ひらがな
漢字
数字（アラビア数字のみ）
英文字
特殊文字（'・:・°・°・°・°など）
左括弧類
右括弧類

処理の流れを図8に示す。(2)では、例えば「A-D変換器」のように特殊文字(‘-’)を含む語に対するマッペンクを行なう。(3)では、「せんし」など、ひらがなではじまる自立語とのマッペンクを行なう。その後、上述の字種の変わり目による区切り目の設定を行ない、1つの区切り目内(文節)でのマッペンクを行なう。この処理における用言の取扱について述べる。用言の各活用形に対するマッペンクは規則活用の場合、各活用型に対する表を用意し、活用 α 型を各用語の辞書に記述することにより各活用形についてのマッペンクを行なうようにしている。「イ音便」「撥音便」を伴う場合については、各活用形をそのまま別々に辞書項目として登録しておく必要がある。



(6)における接続判定は左右の語の品詞を見て、実際にあり得る品詞のなりびであるかを判定するものである。この場合にもシステムに登録された品詞に対しては、接続判定の表が用ゐられているため、品詞名を書くだけで接続判定が表によって行なわれる。ただし、特殊な品詞を利用者が定義する場合は、接続判定の表に対し登録された品詞との接続の情報を書かなければならない。最後の(7)のスタッフで辞書記述とのマッチングがとれなかった文字列について単語とその品詞の推定を行なう。ここでは、以下の処理を行なっている。

```

graph TD
    A["(6) 左右の接続判定"] --> B["(7) マッチングのとれなかった語の品詞推定"]
  
```

図 8 単語分割システムの処理の流れ

- (21)

えはPL/I)があった場合、一語にまとめて名詞とし、英訳としてそれ自体を与える。

d. 他の未知語については、名詞とし、英訳は日本語のままにしておく。

5. 翻訳システム

翻訳システムでは、単語分割システムの出カした単語の並びを解析し、英文表題を生成する。その処理のアルゴリズムを図9に示す。これらの処理は前述したルールによって記述される。

(1)特殊文字、かっこ類の処理(ここで言う特殊文字、かっこ類は単語分割システムの定義によるものである。)

論文表題には、かなりの数の特殊文字、かっこ類が使用されている。かっこ類は先行する名詞の言い換え、内容の説明、又は単に強意のために用いられる。特殊文字は主に表題文の大きな切れ目を指し示すが、コンマ、中黒「・」のように、連続する名詞中に現われる場合もある。これらの処理を最初に行なう。具体的には次のようになる。

a. 表題の左端にかっこ類がある場合、表題をそこで分割し別々に処理する。

(例において「→」は前置の日本語の訳が後置の英文に翻訳されることを示す)

ex. 日本経済のシミュレーション実験(1),(2)

→ simulation experiment of Japanese economics (1), (2)

b. 括弧類の内部が形容詞、又は副詞の一単語である場合、その括弧をとり除き、単語の訳語を括弧で囲んだものに変換する。

ex. "unateji" 論理関数 → "unate" logic function

c. 括弧類の内部の品詞が他のものを含む場合は、名詞句とし、内部の解析・翻訳を行なう。その後、全体を一つの名詞として扱う。

ex. 「プログラムによる differential analyzer」に対する討論

→ discussion to "differential analyzer by program"

d. 特殊文字の内コンマ、中黒については名詞として扱う。他の特殊文字が表題中に現われた場合、表題を特殊文字の前後に分割し別々に処理する。

ex. ALGOL-部分集合について → ALGOL-on subset

ex. 電子計算機を用いた文献情報の蓄積、検索

→ accumulation, retrieval of bibliographic information using computer

(2)等位接続の処理

第2章において、多くの場合係りうけのあいまいさは翻訳の際に保存されると述べたが、等位接続表現が文中にある場合に生じるあいまいさは保存できない。

例えば図10において等位接続の範囲の認定の違いによって2通りの解釈が得られるが、この場合の翻訳結果はそれぞれ「N₂ of N₁ and N₂」、「N₁ and N₂ of N₂」となる。したがって等位接続の範囲を決定する必要がある。現在名詞句の処理の前に論文表題に關するこ

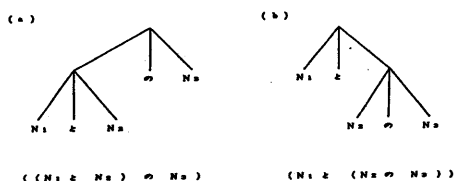


図 10 「N₁ と N₂ の N₂」の解釈

リストアップを用いた次の処理を行っている。

a. 等位接続詞の後に連体詞「その」が続く場合、等位接続詞の前の名詞句全体と等位接続詞の後の名詞句全体が等位接続されているとする。

ex. 語の中の bit を扱う Operation とその回路

→ operation dealing bit in word and its circuit

b. 他の場合には等位接続詞の前後の名詞連続同志が等位接続されているとする。

ex. 機械による文法組立と構文解析のプログラム

→ program of grammar construction and syntax analysis by machine

(3) 名詞連続の処理

等位接続の処理の終了後まず名詞連続の処理を行う。(名詞連続は名詞のみの品詞列を意味する。) ここで重要なのは一連の名詞連続 $N_1, N_2, \dots, N_m$ に対して個々の $N_1, N_2, \dots, N_m$ の訳語をそのまま連結したものとは違う訳を与える場合である。(これを名詞連続熟語と呼ぶ。)

ex. 数式 → formula ; 処理 → processing

数式処理 → formula manipulation

この名詞連続熟語は名詞連続の最後尾の名詞(上記の例の場合「処理」)に対し QUALIST 属性の値として記述している。

これとは別に N_1, N_2 の品詞列に対し N_1 preposition N_2 と訳した方が適切な場合については N_1 に対し ϕ -KAKU 属性がつけられ、その値として preposition (前置詞) を記述している。

処理の流れを以下順に説明する。

a. 全体の名詞句から名詞連続の部分を抜き出す。

b. 末尾の名詞に対し QUALIST 属性がつけられている場合は、前の名詞連続を走査して名詞連続熟語の有無を調べる。存在した場合は、名詞連続熟語の部分について訳を付与し、残りの部分については名詞連続の処理を最初から行う。

c. 名詞連続熟語がないか、た場合は末尾の名詞の ϕ -KAKU 属性の有無を調べ、存在した場合は末尾の名詞と他の名詞連続の位置を反転させ間に前述の前置詞を置く。他の場合は順序はそのままとし末尾の名詞に対して訳を与える。

d. 未処理の名詞連続に対し、b 以後の処理を再び試みる。

e. 全体の処理が終了後、全体を一名詞として扱う。

なお、これらの処理はルールではなく、プログラムにより行われている。

(4) 連体修飾助詞、連体詞を含む名詞句の処理

ここでは名詞間の連体修飾助詞(図4参照)による係り受けの関係の解析を行う。(したがって処理の対象としては「名詞+連体修飾助詞+(連体詞)+名詞+...+名詞」といった品詞列を想定している。連体修飾助詞には「 $N_1 + T_1$ (連体修飾助詞) + N_2 」を「 $N_2 + \text{preposition} + N_1$ 」と訳す際に対応する前置詞が文法ルールにより記述されているが、特定の名詞が N_2 の位置に置かれた場合ルールで指定された以外の前置詞が必要な場合がある。その場合 N_1 の辞書記述に対しあらかじめ連体修飾助詞と前置詞の対として記述しておく。例えば図5の「試みる」のように「NO(の)」に対し「AT」を与えている。

処理は次のようにして行われる。

- 「連体詞+名詞」の品詞列に対し訳を与える。
- 末尾の名詞に対し連体修飾助詞の訳の指定がある場合、直前の連体修飾助詞との一致をとり、一致した場合指定された前置詞を用いて末尾の「連体修飾助詞+名詞」に対する訳をつける。
- 他の部分について同様の処理を未処理の部分がなくなるまで続ける。
- 全体をひとつの名詞とする。

ex. 分類の 一方法 → one method for sorting

(5) 動詞を含む名詞句の処理

ここでは動詞を含む名詞句の処理が終わり、次後第2章で述べたように動詞の埋め込み構造の解析を行なう。処理の対象は、「名詞+連用修飾助詞+...+動詞+(助動詞)+名詞」の形をしている。以下のようにまず動詞の埋め込み構造を解析し、次に図11~13に示す英文生成のバターンのひとつを選択し翻訳を行なっている。

a. 動詞の後に「時の」、「ための」、「た」(完了の助動詞)などの語句がある場合、図11のその語に従って生成パターンから選択する。

- 1: N to V [...]
- 2: N of Ving [...]
- 3: N for Ving [...]
- 4: N in Ving [...]
- 5: N by Ving [...]
- 6: N Ving ... (N Ving N)
- 7: N Ven ... (N Ven N)

ex. 文字認識を設計するためのアルゴリズム

→ algorithm for designing character recognition logic

b. 「必要性」と特定の名詞が動詞に後置される場合、その名詞に従って図11の生成パターンから選択する。

ex. 微分方程式を標準型式に変える
必要性

→ necessity to change differential equation into canonical form

c. 動詞に受動の助動詞「れる、られる」が付加されている場合、表層格を考慮して生成パターンを図12のように定める。その際主格についての扱いも同時に決定する。

d. 上述の条件に合致しない場合、表層格を考慮して図13により生成パターンを決定する。(動詞のVI, VIIは連用修飾助詞「を」に対応する格がobjectとなるかを決定する。)この場合同時に主格、目的格に対する格の扱いも決定する。

e. 残った格の認定及び前置詞句による訳出を行なう。必須格は動詞辞書に対し連用修飾助詞とそれに対応する前置詞句の前置詞を記述している。動詞の必須格との一致が得られた場合はその辞書情報を使用して格を前置詞句として翻訳する。他の格はすべて自由格として処理する。自由格は次のように、各々の連用修飾助詞に対し対応する前置詞が決める。

から: from へ: to
を用いて: using によつて: by

where N : noun
V : verb, infinitive
Ving : verb, ing-form
Ven : verb, past participle

図 11 英文生成のバターン

VERB TYPE CASE	VT
S	N with S V + n
-	N V + n

WHERE S : SUBJECT

図 12 受動の助動詞「れる、られる」がある場合の生成パターンの決定

verb-type cases	vt	vi
S O	N with S Ving O	—
S -	N Ven by S	N with S Ving
- O	N Ving O	—
- -	N Ven	N Ving

WHERE S : SUBJECT
O : OBJECT

図 13 能動語の文の場合の生成パターンの決定

6. 英文キーワードによる検索システム

この翻訳システムは情報処理学会誌「情報処理」に掲載された論文の表題をデータベースとした検索システムを備えている。これは英語キーワードを入力とし論文表題の翻訳結果を出力とするものであり、日本語をま、よく知らない人でも英語のふで文献検索を行なえる。

検索は以下のように行われる。

- (1) 登録された英語キーワードの表を検索し 入力英単語を含むキーワードを見つける。
- (2) 英語キーワード（一般に複数）を日本語キーワードとの対訳表により、日本語へ翻訳する。
- (3) 日本語キーワードにより、文献を検索し文献の番号を求める。
- (4) 文献数を表示し、他の英語キーワードとの AND 又は OR の操作を利用者に求める。
- (5) 新たな操作の指定がある場合は、集合の AND 又は OR 操作をくり返す。
- (6) 表示の指示がある場合は、各文献について翻訳結果を出力する。

7. システムの評価及びこれからの方角

現在本システムを用いて、実験を論文表題約700題に対して進めている。現時点で正しい解析結果の得られる割合は約8割である。ただし冠詞、数については扱っていない。失敗例のうち特に顕著なものを以下にあげる。

(1) 並列句を含む場合

失敗例の大半は割合を占めている。現在のシステムにおける並列の範囲の認定に失敗しているためである。

ex. 経営の意思決定と計算機

→ decision making and computer of management

(2) 連用中止を含む場合

これはルールに記述していないため失敗している。連用中止法（動詞の連用形＋「て」を含む）は、動詞句の等位接続、あるいは従属節を導き、理由手段などを説明するのに使用されるが、広い意味範囲を同一の表現で表わすため、意味の特定のためには、深い解析が必要となる。

ex. 事項検索と推論して質問に答える情報検索システム

ex. 算術計算における誤りを検出し訂正する2進符号

(3) 完全な文となっている場合

これは現在のシステムが名詞句あるいは名詞句＋連用修飾助詞という入力を想定しているため、文法ルールをもたないためである。

ex. アドレス、何ですか？

ex. 計算機は1000 Mc動作に向かっている。

(4) 形容詞の格認定

形容詞の格認定の際、特に問題となるのが「の」で表わされる格である。

これは名詞、連体修飾助詞と表層上は、よく同じ場所（名詞の後、形容詞の前にあらわれる。現在「の」はすべて連体修飾助詞として扱っているため失敗例が出ている。

ex. 速さの異なる記憶装置 → different memory device of velocity

前述の様々な問題点（特に(1), (4)）に対して 名詞に意味マーカを付与し、深
解析に役立てようとの試みを行っている。

・並列句に対しては意味マーカによる接続範囲の限定、解析結果の木構造の類
似性を考えた処理を進めている。

・形容詞の格認定に対しても意味マーカによる解析をする予定である。

今後の研究の方向としては前述の問題点の解決を図り、翻訳対象を拡大すると
ともに生成される英文の質の向上があげられる。冒詞の処理を含めた語形変化の
処理の精密化、又英文を句単位の構成とするこゝにより、表現の自由度を増大さ
せる試みを現在進めている。

本システムによる翻訳例を付録として示しておく。なお本システムは京都大学
大型計算機セクタ FACOM M382 上の PL/I を用いて開発された。

謝辞

数々の有益な助言をいただいた、中村順一助手、言語データの整理などを手伝
っていただいた、久米雅子嬢他、研究室の方々に深く感謝します。

参考文献

- (1) 建部周二、技術論文表題の英和自動翻訳、京都大学工学部修士論文(1979)
- (2) 中村敏明、日本語文を単語に分割するプログラムの作成、
京都大学工学部特別研究報告書(1982)
- (3) 上田良寛、論文表題の日英機械翻訳、京都大学工学部修士論文(1982)

付録 翻訳結果

TITLE NUMBER = #000693

OBJECT TITLE IS ...

与入力パラメترونによる4変数以下の論理関数の実現法

ITS TRANSLATION IS ...

REALIZATION METHOD OF LOGICAL FUNCTION UNDER 4 VARIABLES BY 5 FAN INS
PARAMETRON

TITLE NUMBER = #000694

OBJECT TITLE IS ...

転力の計算制御とプログラム処理の問題点

ITS TRANSLATION IS ...

COMPUTER CONTROL OF ROTARY KILN AND PROBLEM OF PROGRAM PROCESSING

TITLE NUMBER = #000695

OBJECT TITLE IS ...

FACOM 222 FAST (CDT 402 型)

ITS TRANSLATION IS ...

FACOM 222 FAST (CDT 402 MODEL)

TITLE NUMBER = #000696

OBJECT TITLE IS ...

常微分方程式の数値解法における不安現象に対する一対策

ITS TRANSLATION IS ...

ONE COUNTERPLAN TO INSTABILITY PHENOMENON IN NUMERICAL SOLUTION OF
ORDINARY DIFFERENTIAL EQUATION

TITLE NUMBER = #000705

OBJECT TITLE IS ...

記憶装置の多重構成

ITS TRANSLATION IS ...

MULTI STRUCTURE OF STORAGE DEVICE

TITLE NUMBER = #000702

OBJECT TITLE IS ...

記憶装置の将来

ITS TRANSLATION IS ...

FUTURE OF STORAGE DEVICE

TITLE NUMBER = #000703

TITLE NUMBER = #000704

OBJECT TITLE IS ...

計算機の方式と記憶装置

ITS TRANSLATION IS ...

METHOD OF COMPUTER AND STORAGE DEVICE

TITLE NUMBER = #000699

OBJECT TITLE IS ...

記憶装置の概要

ITS TRANSLATION IS ...

OUTLINE OF STORAGE DEVICE

TITLE NUMBER = #000700

OBJECT TITLE IS ...

内部記憶装置の最近の進歩

ITS TRANSLATION IS ...

RECENT PROGRESS OF INTERNAL STORAGE

TITLE NUMBER = #000701

OBJECT TITLE IS ...

外部記憶装置の最近における進歩

ITS TRANSLATION IS ...

RECENT PROGRESS OF EXTERNAL STORAGE