自然言語処理 48－7
（1985. 3. 27）

日英機械翻訳システムにおけるプリエディットについて

有田 英一　福島 正俊　進藤 静一
（三菱電機（株）中央研究所）

1．はじめに

機械翻訳の目標は(1)対象を制限せずに，(2)人間の介入なしに，(3)高品質の翻訳文を作り出す事である。（他に多言語間翻訳の目標もあるが，本報告では扱わない。）

現在発表されている機械翻訳システム〔1〕は，対象を制限することにより，人間の介入なしに高品質の訳文を得て，実用的なシステムを作ろうとしている。これは大量の同質の文書を高速に翻訳する業務向きのシステムである。このアプローチの研究は，次の段階では辞書を大きくし，文法ルールを拡充して受理可能な入力文を増す方向に進むと思われる。言語現象は対象を限定しても種々の様相が出現するので，このアプローチでは，文法ルールの増大と構築されるシステムの柔軟性の不足をどう解決していくかが問題となる。

我々はこれらの問題に対応するために，対話型の翻訳システムを作成している。即ち，人間の介入を許して対象の制限を前提とせずに高品質の訳文を作り出すことにより，実用的なシステムを作ろうとしている。これは対話型であるので大量，高速に翻訳することはできないが，対象の制限を前提としないので一般個人向きのシステムであるといえる。このアプローチの研究は今後，人間とのインターラクションを減らす方向に進む。

本報告では2．でシステムの特徴について述べる。3．4．5．でシステムの特徴を実現するための処理について説明する。また6．に現在得られている実験結果を示す。なお，本システムの入力日本文はローマ字分かち書きであり，各単語はその属性の1つとして漢字かな表記を持っている。

2．システムの特徴

我々は次の5つの点を考慮して対話型の日英機械翻訳システムの開発を進めている。

(1)ガイダンス機能

対象の制限を前提としない場合，入力文は非常に多種多様な表層表現となる。それらをすべて受理できるような文法ルール及び辞書を予め用意することは難しい。従ってシステムの枠組から外れる表現をシステムの枠組に引き込むメカニズムが必要となる。例えば，未登録語があった時，それが未登録表現であるから処理できないとするのではなく，入力表現とほぼ同じ意味を持ち，かつシステムが受理できる表現をシステムがユーザーに提示する枠組が必要である。

(2)アクティブメニュー方式

ユーザーとの対話まで自然言語で行うと，ユーザーの意図がわからない場合があるので，対話はメニュー方式がよい。メニュー方式では，明らかに関係のないことまでメニューに含めると，表示時間が無駄であり，ユーザーの負担も大きいので，適切な処理を行い，真にあいまいなものだけをユーザーに問い合わせるようにしなければならない。

(3)デフォールト

解釈が複数個存在しても，その内の1つの頻度が非常に大きいものはデフォールトとしてユーザーに問い合わせることなく処理を進める。

(4)モニタリング

システムがどういう動きを内部でしているのかわからなければユーザーは不安であるので，システムの判断を監視する機能が必要である。

(5)インタラプト

システムの判断の誤りにユーザーはいつ気付くかわからず，またシステムの処理のフェイズとは関係なく判断の誤りに気付くので，そのようなインタラプトをうまく処理する必要がある。

3．未登録語の処理

未登録語の処理に関しては〔文献2〕がある。〔2〕では，名詞の処理を中心に考察しているが，我々は用言及び関係表現について考察した。

3．1　未登録語の検出の方法

未登録語の検出の方法としては次の3つの場合を想定している。

(1)入力単語がシステムの辞書のどの単語ともマッチングしない場合

この場合はシステムが未登録語を検出し，ユーザーに問い合わせる。

(2)システム内の同品詞同音異義語に入力単語がマッチングした場合

システム内の複数の同音異義語にマッチングした場合は，システムがユーザーにそのどれであるか問い合わせてくるので，ユーザーがどれでもないと指定することにより検出できる。システム内の1つの同音異義語にマッチングした場合，システムはマッチングした単語を漢字で表示するので，ユーザーがそれを見ることにより検出できる。

(3)システム内の他品詞同音異義語に入力単語がマッチングした場合

システム内の複数の同音異義語にマッチングした場合は，システムがユーザーにそのどれであるか問い合わせてくるのでユーザーがどれでもないと指定することにより検出できる。システム内の1つの同音異義語にマッチングした場合，5．で述べる品詞の色分け表示をユーザーが見ることにより検出できる。

3．2　未登録語の処理

未登録語は，普通名詞以外はシステムに登録されている代替表現で言いなおす。名詞は一単語が一概念を表しているので，その単語がわからなければ処理ができないが，動詞などは少し表現を変えれば既登録語が存在する場合が多い。また，動詞は解析のための多くの情報を持っているので，未登録動詞を予めわかっている既登録の類語におきなおすのは有効なことである。

未登録語があった場合，システムは入力単語の表層の漢字を媒介として既登録の類語を検索し，ユーザーに表示する。

代替表現の検索のために次のテーブルを用いる。

(1)漢字－用言対応表（表1）

各漢字には，その漢字の字義の単語と登録語の中でその漢字を表層に含むものとが類語として対応づけられている。

国語辞典を使い，当用漢字1850字のうち，動詞，形容詞，副詞の字義を持つ漢字 992字を抽出し，その字義の単語の集合（約1300語）を基本語いとしてシステムに登録している。大部分の動詞は当用漢字を使って表記されるので，基本語いは日常使う動詞の概念を網羅していると考えられる。従って，日常使うほとんどの動詞に対応する類語がシステム内に存在すると考え

（表1）漢字－用言対応表

漢　字	用言
愛	慈しむ，可愛がる，愛する，好む，
哀	悲しむ，哀れむ，哀願する，
悪	悪い，下手である，苦しい，忌む，
⋮	⋮

（表2）漢字グループ表

No．	漢字
1	重，崇，尊，貴
2	制，定，締，期，契，盟，約
⋮	⋮

られる。

(2)漢字グループ表（表２）

似た字義を持つ漢字をまとめた表である。

(3)漢字－関係表現対応表（表３）

各漢字には，その漢字を表層に含む関係表現が対応づけられている。

(4)漢字－概念対応表（表４）

各漢字から連想される概念項目が対応づけられている。

(5)概念－関係表現対応表（表５）

概念項目とそれを表す関係表現の対応表である。現在，原因・理由，目的，否定・除外，比較，限定，話題・関連など28項目に分類している。関係表現は〔文献３〕をベースに集めている。

（表３）漢字－関係表現対応表

漢　字	関係表現
内	の内で，以内に，の内側で
為	のため，のために
度	のたびに，程度
⋮	⋮

（表４）漢字－概念対応表

漢　字	概念
添	添加
逆	逆接，場所
⋮	⋮

（表５）概念－関係表現対応表

場所	場所	に，で，の場所で，において
	内部	の中に，で，に，の内側で
	⋮	⋮
時	着点	まで，へ，に，への，
	起点	から，より
	⋮	⋮

処理のながれを図１に示す。まず，未登録語の品詞をユーザーが指定する。

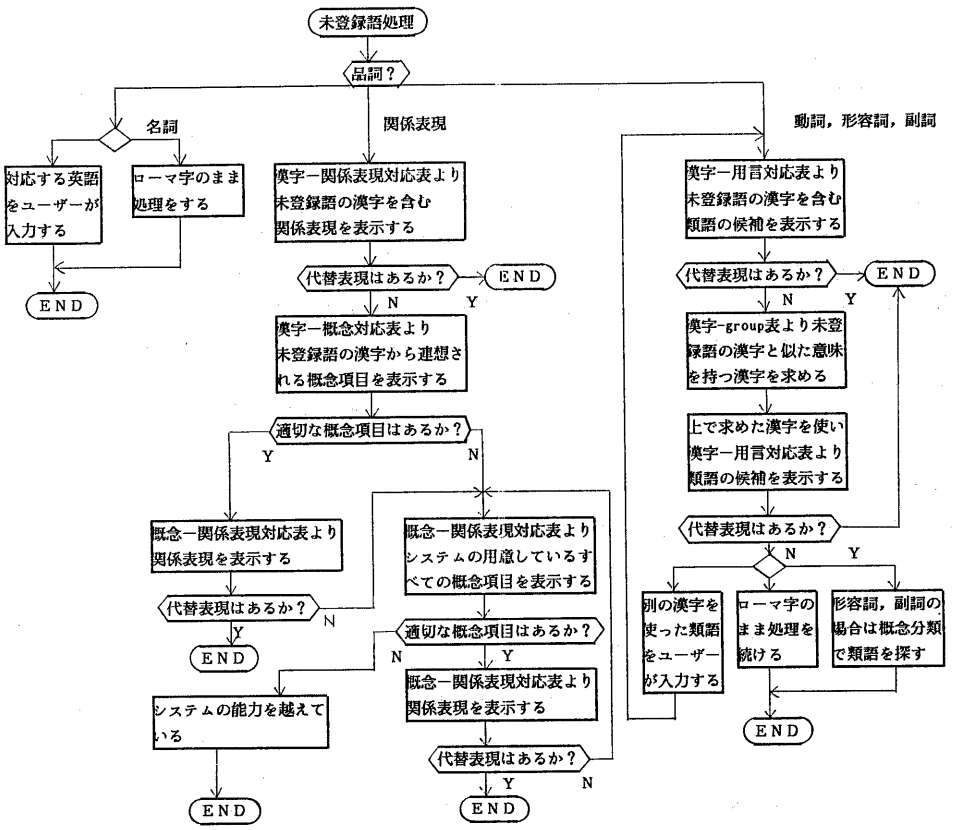

図１　未登録表現の処理

(1)未登録語が名詞の場合

対応する英語をユーザーが入力する。わからない場合はローマ字のまま処理を進める。

(2)関係表現の場合

まず，漢字－関係表現対応表より未登録語の漢字を含む関係表現をシステムは表示する。その中に代替表現があればユーザーはそれを採用する。例えば，「～を基盤として」が未登録表現であっても，「～に基づいて」という既登録表現が得られる。代替表現がなければ，未登録語の表層の漢字から連想される概念項目をシステムは表示する。ユーザーは適切な概念項目があればその概念を表す関係表現を選ぶ。無ければシステムは用意しているすべての概念項目を表示するのでユーザーはその中から選ぶ。

(3)動詞，形容詞，副詞の場合

まず，漢字－用言対応表より類語の候補をシステムが表示する。その中に代替表現があればそれをユーザーは採用する。代替表現がなければ，システムは漢字グループ表より類義語をもつ可能性のある漢字を求め，それらの漢字から再び漢字－用言対応表を使って類語の候補を表示する。その中にも代替表現がない場合は，ユーザーに別の漢字を使った類語を入力してもらい，その漢字を手がかりにして，システムは類語の候補を検索する。形容詞，副詞の場合は，予め設定された概念分類に従って検索することもできる。形容詞の概念分類は〔文献　5，6〕をベースにしている。

4．あいまいさの処理

4．1　付属語の意味のあいまいさの処理

付属語の意味に多義がある場合は，次の5つの方法により多義の可能性を絞る。1つに絞りきれない場合はユーザーに問い合わせて解決する。

本報告と同じ考えのもとに，ユーザーに付属語の意味を問い合わせて解析するものに〔文献4〕がある。〔4〕では機械翻訳のみを前提としていないので，多義のある付属語の意味は代替表現のあるものまでしか絞らない。例えば，「で」の用法のうち「主体，場所，原因・結果」などは「が，において，の結果」と言換えができるので区別するが，「材料，構成要素，方式，観点」は区別しない。本システムではユーザーに問い合わせる際，言換え表現だけでなく，内部の意味ラベル（材料，構成要素など）も表示し，意味を1つに絞る。また〔4〕では多義の中のどの意味か問い合わせる際にその付属語の意味をすべて表示しているが，本システムでは用法の簡単なチェックをかけて，選択枝の数をできるだけ減らしている。

(1)局所的なsyntaxのみから決まる

例．　ため ┌目的　　たため→理由
　　　　　└理由

　　Pから ┌原因・理由　　シテから→動作（後）
　　　　　├完了
　　　　　└動作（後）　　シタから ┌原因・理由
　　　　　　　　　　　　　　　└完了

(2)大域的なsyntaxから決まる

例．P1がP2（接助） ┌逆接確定　P1とP2が同じ用言であり，
　　　　　　　└順接確定　P1に「ない」があり，P2にかかる用言に「ない」がない場合
　　　　　　　　　P1に「ない」がなく，P2にかかる用言に「ない」がある場合
　　　　　　　　　逆接確定 ←

(3)簡単な意味で決める

例．　Nのために｛目的／原因・理由／利益｝　Nが「人」の時→｛原因・理由／利益｝

(4)表層の共起で決める

例．　Nの間に｛場所／時｝　Nが「不在，出張…」の時→時

(5)デフォールトとする

例．　Vたら｛順接仮定／時（後）／順接確定｝
Vに「もし，もしも，明日…」がかかっている時は順接仮定
Vが「終える，終わる…」の時は時（後）
それ以外の時は順接確定とする

付属語の意味は次のように問い合わせる。

私の姉はこの小説をよむと必ず涙をながす

1 ：順接仮定 （ ならば ）
2 ：順接確定 （ 時 ）
3 ：内容

ここで2を選ぶと，次からは表層表現を「と」からあいまいさのない「時」に変えて表示する。

私の姉はこの小説をよむ時必ず涙をながす

これはシステム内部の制限言語をユーザーにteachingすることにもなる。次回からはユーザーが「と」の代わりに「時」という表現を入力すれば，「時」には意味にあいまいさがないので，システムからの問い合わせが1回減る。

4．2　構文のあいまいさの処理

かかりうけは基本的には連用か連体かによって判断している。しかしそのままでは可能性が多くなるので，ヒューリスティックな条件でかかりうけの可能性を減らしている。例えば，

形容詞（連体形）は「は」を越えてうしろの名詞にかからない。

格助詞「を」のすぐうしろが動詞の時はその動詞にかかる。

などである。現在このような制約条件を21個用いている。もちろんこのようなヒューリスティックの条件にあわない文もあるが，その場合は，5．で述べるインタラプトの機能でユーザーが訂正する。

制約条件を使ってもかかりうけにあいまいさが残る時は，次のようにしてユーザーに問い合わせる。

1 ：私の姉はこの小説をよむと必ず涙をながす
2 ：私の姉はこの小説をよむと必ず涙をながす

番号を入れて下さい ：

「姉は」が「よむ」にかかるのか「ながす」にかかるのかを尋ねている。この時，複数指定してもよい。

5．システムの判断のモニタリングとインタラプト

マン・マシンシステムでは簡単なことまでシステムが人間に判断を求めると，インタラクションの回数が多くなりすぎるので，ある程度わかりきったことはシステムが判断する必要がある。しかし，その判断が必ずしも正しいとは限らず，それが正しいかどうかいちいち人間に確認を求めることはできない。一方，

ユーザーはシステムがどう判断したのかわからなければ不安である。従って，システムが自動的に処理したことに対してはユーザーに報告するだけで判断を求めない機能と，ユーザーがその報告を見て，誤りを発見した時，任意の時点で訂正できる割込みの機能が必要となる。

5．1　モニタリング

本システムでは次のようにしてシステムの内部状態のモニタリングを行っている。

(1)単語の品詞

システムは品詞を色分けして表示する。

私の姉はこの小説をよむと必ず涙をながす
名詞　関係表現　形容詞　副詞　動詞　限定詞

(2)付属語の意味

システムは多義のある付属語の意味を判断するとそれと同じ意味を含む別の表現におきかえる。例えば，「病気のために」であれば「ため」の意味を原因とシステムは判断し，「病気が原因で」と表層を書き換える。

かのじょはびょうきのためにかいしゃをやすんだ
彼女は病気が原因で会社をやすんだ

(3)かかりうけ

キャラクタディスプレイでかかりうけ関係を表現するのでforeground colorとbackground colorを使って表示する。２つの文節にかかりうけ関係があれば，かかる文節のforeground colorと受けの文節のbackground colorを同じ色にして表示する。

私の姉はこの小説をよむと必ず涙をながす

(foreground colorをアンダーラインで，background colorをオーバーラインで示している)

5．2　インタラプト

ユーザーがシステムの判断の誤りを発見した時は，システムの入力待ちの時はいつでも割込みをかけてその誤りを訂正することができる。インタラプトをした時はその理由を次のメニューで聞く。

リジェクトした理由は何ですか？
1：品詞の誤り
2：同音異義語の誤り
3：係り受けの誤り

(1)は「まばたき」のように動詞の連用形から名詞と成ったもので，その動詞形と名詞形のうちどちらか一方だけが登録されている時，生じる誤りである。(2)は同音異義語の一部だけが登録されていて，登録されていない単語がひらがな入力された時に生じる誤りである。(3)はシステムがかかりうけ関係をヒューリスティックを使って決める時に生じる誤りである。システムはインタラプトされると，始めに戻って処理をやりなおす。この時，１度ユーザーに問い合わせたことは覚えているので，同じことを２回聞きなおすことはない。

6．実験結果

現在得られている結果を下に示す。(a)は入力されたローマ字分かち書きの文，(b)は読みやすいように入力文をひらがな表記したもの，(c)はあいまいさが少なくなるようにシステムが書き換えた文，(d)は英訳である。(c)を入力文とすると，(1)(2)(3)のインターラクションの回数はそれぞれ3回，3回，2回となる。

(1)

(a) (ATARASHII &YU-ZA-KONPYU-TASOHUTOUEA& HA BETSUNO HOUHOU DE KEISANKI NI DOUSA NO SHIJI WO ATAERUKOTOGADEKIRU GA MATA OUYOU&PUROGURAMU& KAN NO &INTA-RAKUSHON& WO YORI ISSOU SUSUMERUKOTOMODEKIRU)

(b) あたらしいユーザーコンピュータソフトウエアはべつのほうほうでけいさんきにどうさのししをあたえることができるがまたおうようプログラムかんのインターラクションをよりいっそうすすめることもできる

(c) あたらしいユーザーコンピュータソフトウエアはべつの方法でコンピュータに動作の指示を与てそしてまた応用プログラムかんのインターラクションをよりいっそうすすめることもできる

(d) A new user-computer software can provide a computer with an indication of an action by an other way and can promote an interaction among application programs still more

使用した構文制約条件 : (KAISEKI1 KAISEKI5 KAISEKI3 KAISEKI18 KAISEKI7)
インターラクションの回数 : 4

25168 msec used. (total time)
6262 msec used. (not including disk access time)

(2)

(a) (TATOEBA &WINDO-& WO TSUKAEBA 1 DAI NO &KONPYU-TA& DE DOUJI NI DOUSASHITEIRU IKUTSUKANO CHIGATTA &PUROGURAMU& NO UGOKI WO &GURAPHIKARU&NI HYOUJISURUKOTOGADEKIRU)

(b) たとえばウィンドーをつかえば1だいのコンピュータでどうじにどうさしているいくつかのちがったプログラムのうごきをグラフィカルにひょうじすることができる

(c) たとえばウィンドーをつかえば1台のコンピュータの中でどうじにどうさしているいくつかのちがったプログラムの動きをグラフィカルにひょうじすることができる

(d) ??? can graphically display an activity of several different programs which is simultaneously operating in one computer by using a window, for example,.

使用した構文制約条件 : (STRUCT1 KAISEKI13 KAISEKI18 KAISEKI13 KAISEKI18)
インターラクションの回数 : 4

23688 msec used. (total time)
7007 msec used. (not including disk access time)

(3)

(a) (&YU-ZA-& HA &SUPUREDDOSHI-TO_PUROGURAMU& WO &RAN&SASETE SONO KEKKA WO &GURAPHIKKU_PAKKE-JI& NI NYUURYOKUSUREBA &REPO-TO& YOU NO ZUHYOU WO TSUKURUKOTOGADEKIRU)

(b) ユーザーはスプレッドシート プログラムをランさせてそのけっかをグラフィック パッケージににゅうりょくすればレポートようのずひょうをつくることができる

(c) ユーザーはスプレッドシートプログラムをランさせてそしてその結果をグラフイックパッケージに入力するならばレポートようの図表をつくることができる

(d) If a user runs a spread-seat program and enters the result into a graphic package, he can produce an illustration for a report.

使用した構文制約条件 : (KAISEKI7 KAISEKI7 KAISEKI7)
インターラクションの回数 : 5

19751 msec used. (total time)
6293 msec used. (not including disk access time)

７．まとめ

対話型の日英機械翻訳システムの前処理について述べた。

(1)漢字を使って未登録表現からシステムに登録されている類語表現を導出する方法について述べた。これにより，小さな辞書で，広範囲の文を処理できる。

(2)入力文のあいまいさをユーザーに問い合わせる時，真にあいまいなものだけを選択枝の中に入れる，動的なメニューの手法について述べた。これによりユーザーの負担を軽くできる。

(3)システムの内部状態をモニタする手法とインタラプトの機能について述べた。これによりシステムの判断の誤りを処理の途中で訂正できる。

参考文献

〔１〕長尾　真：「科技庁機械翻訳プロジェクトの概要」　情報処理学会ＮＬ研究会資料３８－２

〔２〕増山　他：「ＡＴＬＡＳ／Ⅱにおける未知語処理」　情報処理学会第29回全国大会 p.1239

〔３〕吉田　将：「日本語の規格化に関する基礎的研究」　昭和58年度科学研究費成果報告書

〔４〕長尾　他：「制限文法にもとづく文章作成援助システム」　情報処理学会ＮＬ研究会資料４４－５

〔５〕岡田直之：「自然言語および図形理解のための形容詞の概念の分類－単純概念の場合」
情報処理学会ＮＬ研究会資料３８－１

〔６〕岡田直之：「自然言語および図形理解のための形容詞の概念の分類－非単純概念の場合」
情報処理学会ＮＬ研究会資料３９－２