自然言語処理 32 － 2
（1982. 8. 2）

パターンマッチングによる英日機械翻訳の一手法

榊 博史　橋本 和夫
（KDD 研究所）

1．まえがき

機械翻訳の方法はその処理過程が意味ネットワークを経由するものとパージング木を経由するものに大別される。本文は後者に属する一つの方法を提案するものである。後者の方法の代表的なものとしては発生したパーズ木から深層的木を生成した後目的言語への変換を行なう方式[1]，パーズ木より節点間の通信を用いて意味関係を抽出する方式[2][4]，パージングと目的言語への変換をほぼ同時に行なう方式[3]等がある。ここで述べる方法はパージングを行なった後パーズ木の最上位部分より下方に向って木の小部分毎に順次目的言語木に変換する方法であり上記小部分の発見にパターンマッチング，変換にプロダクションルールを用いるものである。

2．例を用いた原理の説明

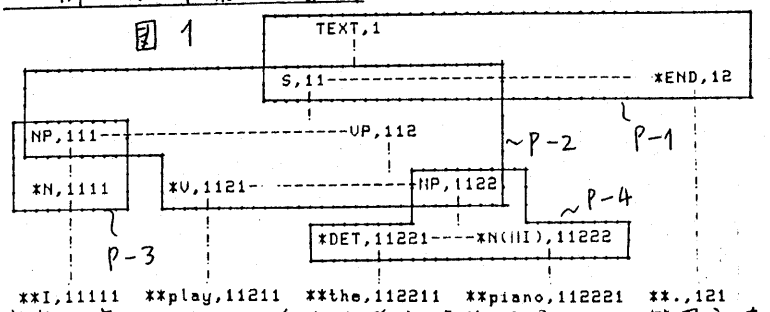


図 1

まず簡単な例を用いて本法の原理を説明する。図1は，

I play the piano. (1)

なる英文のパージング結果である。図中わくで囲んだ部分は、後述する部分である。パージングは適当な適用制限付の自由文脈文法的なパージング用文法を用いたパージング方式、例えば拡張ＬＩＮＧＯＬ等[5][6]を用いて行うことが出来る。図1中**印が付加された節点は終端節点すなわち各単語名に対応する節点であり、＊印が付加された節点は単語のカテゴリーを表わす節点でありここではカテゴリー節点と称することにする。何の印も付加されていない節点は非終端節点である。各節点はその節点の種類を示すラベルとその節点の固有名詞に相当する節点番号を持ち、この順にコンマで区切って示す。節点番号は図に於けるように下位の節点は直上位の節点番号に末尾の数を加える方法で与える。非終端節点ラベルに於て、TEXT，S，NP，VP はそれぞれ文，センテンス，名詞句，動詞句をあらわす。カテゴリー節点のラベルは単語カテゴリーを表わす部分とその右方括弧にかこまれた単語サブカテゴリーを表わす部分から成る。後者は必要に応じ付与される。単語カテゴリーに於てN，DET，Vそして END はそれぞれ名詞，限量詞，終了記号をあらわし、又単語サブカテゴリーに於て（MI）は楽器名（*musical instrument*）を示す。図1に於ける各節点のラベルのうち終端節点のそれは単語名そのものでありパーズ以前に与えられる。カテゴリー節点のそれも単語カテゴリー，単語サブカテゴリー共パーズに先立つ品詞識別時に与えられる。又非終端節点のそれはパージング用文法適用時に自動的に得られる。パージング用文法適用に際し単語サブカテゴリーは無視するものとする。図1に示したわくは原言語パーズ木をパターン文法に対応する領域に分割する方法を示すためのものである。ここには、終端節点に関するものを除いて分割法の全貌が一度に示されているが実際には上位のものよ

り逐次求められる。境界節点は両方の領域に含まれる。各領域に含まれるパーズ木の部分は原言語パターンと称することにする。パターン文法という名称はパージング文法からの類推により与えたもので、パターン文法は原言語パターンとそれを処理する目的言語パターンを含むプロダクションルールから構成される。1つのパターン文法は1つの原言語パターンと一対一に対応するので各パターン文法及び原言語パターンには同一の名称を与えることにする。各原言語パターンは図1には示さなかった各単語のためのパターンを含めて独立した形でその名称と共に図2に示されている。表1は各パターン文法の内容を示したものであり表1に於て最左端の列はパターン文法名又は原言語パターン名である。その他の列はプロダクションルールで記述したパターン文法の内容である。入力に属する2列はプロダクションルールの入力を示し、出力に属する2列はプロダクションルールの出力を示す。入力のうち上位よりのインストラクションの列はあるパターン文法を処理を受けているある原言語パーズ木に適用しようとする時にその直前のパターン文法適用の結果得られた判断に基き今回の原言語パターンの最上位節点に加えられる指示を示すものであり、下位からのエキストラクションは処理を受けている原言語パーズ木に於ける現在適用しようとする原言語パターンより下位の木構造に関する情報であり構造情報とその発見方法が組となって示されている。表1に示されたP-2-X1はこの発見方法の一例の名称でありこの構造自体は図4に示されている。これに対する説明は本節中で説明する。図3の各図は目的言語パターンの図である。以下表1に示される各パターンの適用法の詳細を適用順に示す。

パターン名又はパターン文法名	入力		出力	
	上位よりのインストラクション	下位からのエキストラクション	目的言語パターン名	下位へのインストラクション
P-1	—	—	P'-1-1	
P-2	—	[play,MI,P-2-X1]	P'-2-1	V,D/hiki
	—	[play,SP,P-2-X1]	P-2-1	V,D/si
	—	[hiss,—,P-2-X1]	P'-2-2	—
P-3	—	—	P'-3-1	—
P-4	—	—	P'-4-1	—
P-9	—	—	P'-9-1	—
P-play	hiki	—	P'-play-1	—
	si	—	P'-play-2	—
P-the	—	—	P'-the-1	—
P-piano	—	—	P'-piano-1	—
P-period	—	—	P'-period-1	—

表1

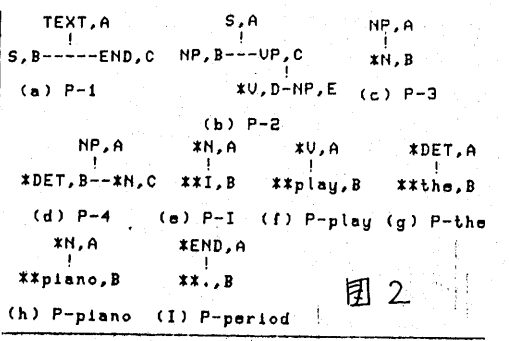


図2

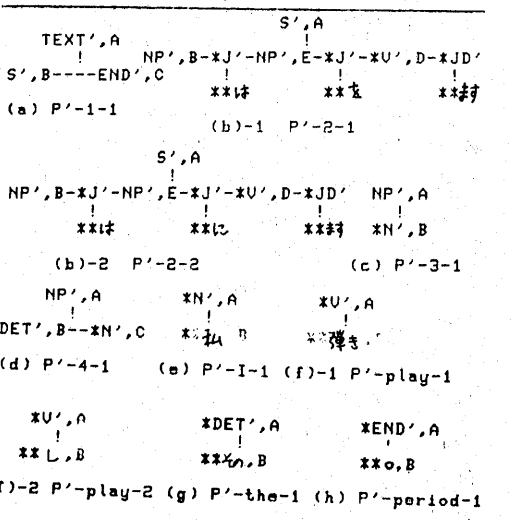


図3

図4

式(1)で与えられる英文入力は図1のパーズ木に解析される。日本語への変換操作に於ては入力として図1のパーズ木が与えられる。これはTEXTをラベル名を持つから、まずTEXTを最上位節点ラベル名として持ち図1のパーズ木の一部分と完全に合致する原言語パターンが表1を含む原言語パターンの表の中からパターンマッチングにより探される。このような形式の一致パターンをここでは上位部分一致パターンと呼ぶことにする。今回は図2(a)に示す原言語パターンP-1が上位部分一致パターンであることがわかり、このパターンに一対一対応する表1中のプロダクションルールP-1に示された操作が行われる。すなわち入力が一印であるので無視し、入力の如何にかかわらず目的言語パターンとしてP'-1-1すなわち図3(a)に示す構造を発生し、下層へのインストラクションとして何も発生しないことを行う。図3に示すように目的言語パターン中の各節点には'(ダッシュ)を付加する。入力に於ける一印は入力を無視することを示し、出力に於ける一印は出力を発生しないことを示す。目的言語パターンはP'の後に適当な文字を付けて示し、後に続く文字の最初のものは対応する原言語パターン名であり、次に続く数字は同一の原言語パターンに対し複数の目的言語パターンがある場合それらを区別するために用いる1より始まる数字である。目的言語パターンが1つしか無い場合でも1を付加することとする。図1と図2(a)との対応に於て図2(a)中のAは1, Bは11, Cは12に対応するという理由により図3(a)のA, B, Cの値をこれらの数字に変えたものが上記目的言語パターンP'-1-1として発生し適当な大モリーに記憶される。次にプロダクションルールP-1に関する最後の操作として図1中の原言語パターンP-1に相当する部分を境界節点を除いて消去する。この操作の後TEXT, 1節点が図1の木より消去され木が2つに分れることになる。

このようにして発生した2つの木のうち左側のものが次に取り扱われる。複数の木の取り扱いの順序は任意である。これの上位部分一致パターンがやはり表1を含む原言語パターンの表から探され、図2(b)に示すP-2がこれに該当することがわかりこれに対するプロダクションルールP-2が適用される。原言語パターンP-2は目的語1つを含む他動詞の文のパターンでありこれに対応するプロダクション ルール中の入出力の組は表1のP-2の欄に示したる行の他に非常に沢山あるがそれらの多数の行に対応する入力の中からどの入力が発生しているかを調べ発生している入力に対応する出力に関する処置を取ることになる。表1に示すように上位からのインストラクションはP-2に於ては無視される。入力のうち下位からのエキストラクションは抽出(エキストラクト)されたエキストラクション情報とそれが測定された際の測定手段名が組となったものであり大括弧[ ]でかこってその中に1つ以上のエキストラクション情報をエキストラクション測定手段が指定する順序で並べた後にその情報を得る際に用いたエキストラクション測定手段名を並べた形式を取るものとする。表1に於けるプロダクションルールP-2に関するエキストラクション測定手段名は全てP-2-X1である。測定手段名は原言語パターン名の後のXにつづく数字として1より始まる数字を入れて示すことにする。1つしか無い場合でも1を付加する。P-2-X1の構造は図5に示されておりP-2の構造の下位にいくつかの部分が付加されたものとなっている。エキストラクション測定手段はこのように原言語パターンの下位に何らかの測定用の部分が付加された形式を取る。P-2-X1の場合は*V, D節点の下に**Q1節点が、又NP, E節点の下にⒶで示された部分とN(Q2)節点, Ⓐで示された部分がこの順

で左より右へ存在する構成となっている。この測定手段を用いることにより、図4中のカテゴリー節点*V,Dの唯一の下位節点であるQ1に示された位置の節点を見出しこれを大括弧中の最初の値として示し、次にNP,E節点の下に接続する節点のうち名詞Nをカテゴリー名として持つ節点のQ2の部分すなわち括弧内のサブカテゴリー名を見出しこれを大括弧内の2番目の値として発生することが行われる。名詞Nをカテゴリー名として持つ節点はイつに限られることはNP(名詞句)の性質上明らかである。又Ⓐと記した部分は任意の単一節点,木構造であって良く又空であっても良い部分である。上述のようにQ1の添字の順に得られた情報が並べられる。図1の木からTEXT,1節点を除いて得られた木に対して図4のエキストラクション測定手段P-2-X1が適用され結果として[play,MI,P-2-X1]が下位からのエキストラクションとして得られる。このためP-2に関するプロダクションルールに従いまず目的言語パターンとして図3(b)-1に示すP'-2-1が発生し適当なメモリーに記憶される。但しこの際図3(b)-1中のA,B,C,D及びEの値が、図1と図3(b)との節点番号間の関係により、それぞれ11,111,112,1121,1122に変更される。図4(b)-1及び図4(b)-2に於てJは助詞、JD'は助動詞を示すカテゴリー名であり原言語パターンに対応しない節点には節点番号が与えられていない。次に下位へのインストラクション*V,D/hiki により図2(b)に於ける節点*V,D(=対応する節点すなわち図1に於ける節点*V,1121に"hiki"というインストラクションが付加された後P-2に関する節点S,11及びVP,112が消去される。この消去,インストラクション付加によりパターン文法P-2に関する処置が終る。結局今迄の処置により図5に示す原言語の木が未処理の部分として残り、図6に示す木が目的言語の木として適当なメモリーに記憶されることになる。図5中*V,1121,hiki節点の最後の部分は付加されたインストラクションである。

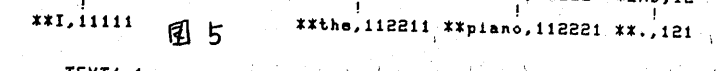


図5

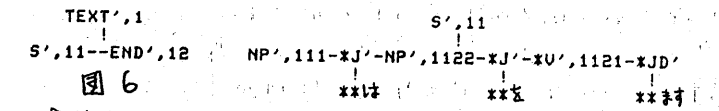
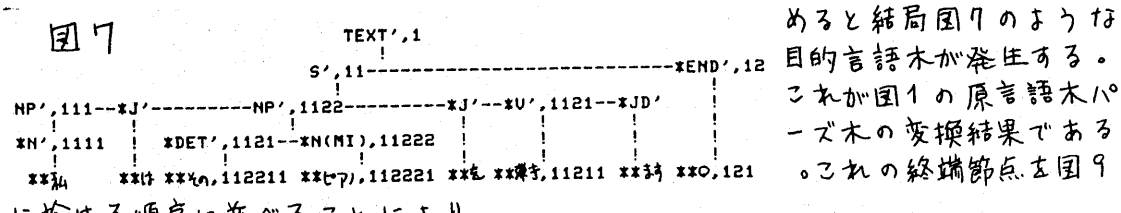


図6

今後処理すべき木は図5に示されている。複数の木がある場合どの木を先に処理しても良いのであるがここではまず左から2番目のplay,1121を含む木を先に処理することにする。この場合もこれの上位部分一致パターンがやはり表1を含む原言語パターンの表の中から探され図2(f)に示すp-playが該当することがわかりこれに対する表1のp-playの項に示されたプロダクションルールが適用される。p-playのプロダクションルールに記された2つの行に対応する入力のうち上位よりのインストラクションが"hiki"である上の行が今回の入力であると判定され上の行の出力に対応する処置が取られる。結果として図3(f)-1に示すp'-play-1が目的言語パターンとして発生し図5の4つの木のうち左から2番目のものが消去されp-playに関する処置が終る。

このように順次上位部分一致パターンの発見とプロダクションルールの適用を続けて行くと最後に処理されるべき原言語木が空となり、目的言語を記憶するメモリーに図3に示す目的言語パターンのうちp'-1-1, p'-2-1, p'-play-1, p'-4-1, p'-the-1, p'-piano-1, p'-period-1が記憶された状態となる。但しこれらの節点番号は図6の中間段階に於けるように対応する図1にあらわれる番号に変えられる。これらの目的言語の複数の木を同一節点で融合することにより1つにまと

図7



めると結局図7のような目的言語木が発生する。これが図1の原言語木パーズ木の変換結果である。この終端節点を図9に於ける順序に並べることにより、

私はそのピアノを弾きます。　(2)

という目的言語である日本語への翻訳結果が得られる。なおもし、

I play ping-pong　(3)

なる文が入力した場合は、ping-pong の単語カテゴリーがN、単語サブカテゴリーがスポーツ名に相当するSP(Sports)でありN(SP)というラベルのうち[play, SP, P-2-X1]が下位からのエキストラクションである行の操作が行われこれが、playの翻訳として"し"を発生するので結果として、

私はピンポンをします。　(4)

という日本語訳を発生する。playをkiss に変え目的語を元に戻すと下位からのエキストラクションが[kiss, ー, P-2-X1]である行に対する操作が行われ目的言語パターンP-2-2が発生することにより、

私はピアノにキスします。　(5)

という文が発生する。これら木の操作は計算機上ではS式を用いて行われる。[2]

3. パーズ木変換に関する基本的検討

前節の例により明らかであるので本文の方法による一般的な動作サイクルの記述は行わない。本文の方法のように原言語木、目的言語木の対応により変換を行う方法では考えられる原言語木の全てに対し対応する目的言語木を用意するのが理想的であるが、この場合木の種類が膨大となるので実用的ではない。本文の方法は処理中のパターンの上下位の情報をもとにパターン毎の変換を行い上記木全体の変換と等価なことを行おうとする方式であるが、完全に等価とするためには各プロダクションルールの上位よりのインストラクションの項数は全ての直上位に来ることが可能な全原言語パターンの全出力の数だけ必要となる。すなわち、ある原言語パターンを取り扱う場合、直上位パターンの名称とその出力を知らなければ正しい変換は行えない。但し、さらに上位のパターンに関する情報は、直上位のパターンに織り込まれているので考慮する必要はない。各原言語パターンに於ては下位からのエキストラクションを適切に用いることにより下位部分の必要な全情報を得ることが出来るので下位からのエキストラクションに関してはあまり問題がない。この方式は木全体の変換を行う方法と等価であるがほぼ同様の困難を伴う。この困難性を除くためには、上述のインストラクションの選定方とは逆に上位のパターンが下位に合わせる形で下位に接続されると予想されるパターンのプロダクションルールが発生しなければならない出力を選択発生するためのインストラクションを選定するという選定法が選択肢の数が前述の出力の数よりはるかに少ないから実際的である。この場合、上位パターンは下位パターンを文字通りインストラクト(指導)することになる。この場合、下位に来るパターンが複数個あれば必要なインストラクションの種類はそれら複数個のパターン全体に必要なインストラクションの種類の和となる。これに対応して下位のパターンに於て上位よりのインストラクションの一部を無視する機能が必要となると思われ

る。この形式のインストラクションは前述の木全体の変換と等価な場合のインストラクションが下位パターンの出力の種類毎に合併したものであると言える。

本方式での変換誤り発生の原因として入力段階に於けるパージング誤りのほかにパターン発見誤り、上述のインストラクションの合併による情報の欠除があげられる。このうちパターン発見誤りの大部分が下位の部分でのパターン発見不能を引き起こし、この時点で取り除かれると思われる。

次にパターン選定法に関して述べる。原言語パターン、目的言語パターン双方共木に於ける二次元的部分を構成する。これらパターンの最上位節点は1つに限る。これは1つの非分離の木のみをパターンとするということと同意義である。この意味でパターンは係飾・被係飾の関係を記述するものであると言える。又通常の文法以外にイディオム的なものの記述も可能でありこの場合のパターンは終端節点をも含む。

なおインストラクション及びエキストラクションの種類は自然言葉の多彩さを反映して多彩な形態を取る。例えば、*I never see her but I want to kiss her.* の例では否定文を肯定文にするインストラクションが必要である。

複数のパターンが適用可能な場合、節点の数の多いものを先に適用することが変換誤りを少なくする上で有効であると思われる。この際終端節点、次いでカテゴリー節点を多く含むものを更に優先させることも必要であろう。パーズ木は通常複数個生じる。更に1つの木に対してもパターンの適用法の違いにより更に複数個の変換結果が生じる可能性がある。パージング用文法は直上下位の関係のみを規定し利用し得る適用条件に関する情報は下位のもののみであるのに対しパターン文法は文中の全ゆる情報を利用し相当緻密な適用条件を書ける。このことを利用して1つのパーズ木に対しては前記パターンの適用優先により変換結果を1つにしぼり、複数の木の間の比較に於ては優先度の高いパターンを含むものを選ぶことにより パーズ木の数より変換結果を少くすることが可能であると思われる。

4. まとめ

パターンマッチングとプロダクションルールを併用した、上位部分より順次部分毎に変換を行う木構造変換法について述べた。この方法に於けるパターンとして単語、文法、イディオム等全ゆる形態の自然言語表現法が表現法の基本的な形態とほぼ同じ形で導入できるので文法が書き易いという特徴がある。筆者5はこの方法をKPP(KDD pattern matching and production rule)法と名付けて筆者らが開発を予定している英日国際会議資料翻訳システムに適用する準備を進めている。なおこの方法は木構造解析できる全ゆる言語に適用できる。今後はパターン文法を導入するためのインストラクション、エキストラクションの設計法、正しい変換結果を選択する方法に関する具体的な検討を行う予定である。

○末筆ながら筆者らは機械翻訳に関する研究の重要性を指摘し、この研究開始のきっかけを作って下さったKDD研究所鍛治所長に深謝し、又適切な御指導を賜っている同研究所寺村副所長、中井次長、川井第一特別研究室長に感謝する。

長尾京大教授、草薙筑波大助教授には本検討の初期に貴重な御討論を頂いた、又電子総研田中氏、東工大市川教授、畝見助手には拡張[5]及びLINGOL[6]に関し種々御教示頂いた。厚くお礼申し上げる。

[文献 1)情報処理システムに関する新技術調査(日本電子工業振興会,昭和55年3月)所載のアルノーアル大関係文献. 2)田中,計算機による自然言語意味処理に関する研究,電子総研報告977号,昭和58年7月, 3)新田,他,英和機械翻訳のための構文解析技法,自然言語処理研資料28-4 1981.11.20 4)田中 他,意味表現言語SRLの機械翻訳への応用,自然言語処理研資料31-5,1982.5.21. 5)電総研推論機構研究室,拡張LINGOL,1978 6)畝見,LINGOLの日進木への拡張,東工大修論,1980]