中国語解析システムにおけるヒューリスティックな知識の利用†

楊 顒明†† 堂下修司†† 西田豊明††

本論文では，ヒューリスティックな知識を用いて中国語入力文の部分統語構造を予測するシステムについて述べる．このシステムは，入力文に完全な統語・意味解析を行う前に働かせ，その処理結果（予測された部分統語構造）により全体の解析を正しい方向へしぼり，曖昧性を抑えることをはかる．したがって，このシステムを前処理システムと呼ぶことにする．ここで利用される知識は，中国語において頻繁に現われ，かつ重要な働きをする，限られた数の特徴語に関するものと，それらに対するもっともらしさに関する量的なものである．本システムでは，入力文中の特徴語を手掛りにして部分構造としてまとまりうる部分（ここで断片と呼ぶ）を抽出する．次に，もっともらしさに関する知識を使って，断片のなかから，競合の生じない最大の部分集合を選択して本処理に引き渡す．前処理で用いる知識は，多くの場合正しいが，つねに正しいという性質のものではないので，本システムでは後戻り処理を行って正しさを保証する．ここで述べた方法を，200程度の特徴語に関する規則（101個の B. ATN で記述したもの，3.2節を参照）を用いて，単語ごとに分ち書きされた入力文120例について机上で調査した．その結果94%の場合第1位選択において，98%の場合第2位選択，100%の場合第3位選択までで正解が得られた．さらに，選択処理については，競合の生じた他の80例について調べた結果，83%の場合第1位において，90%の場合第2位選択，98%の場合第3位選択までで正解が得られることがわかった．本方式は現在一部計算機上に実現されており，有効性が確かめられている．

1. まえがき

中国語を計算機によって解析しようとしたとき，おもな問題点は統語上の手がかりが少ないことである．漢字のみを使う中国語では，英語の語尾や日本語の格助詞などのような形態変化が少なく，変化の規則性も弱い．中国語文に単純な統語解析を行うと，単語の多品詞性および，同じ品詞列に対しても多数のパーザ木が存在することにより，爆発的な数の可能な結果が出てしまう．**図1**は中国語文の統語解析における曖昧性の例を示す．例文「在没有摩擦的理想情況下物体将以恆定的速度運動下去。」においては，16個の単語のなかで8個が複数個の統語カテゴリをもち（図1のa），それらに対して図1のbで示したような統語構造（パーザ木）が多数個存在している（1000通り程度にのぼる）．

以上に示したような曖昧性をいかに抑制するかは中国語解析の本質的な課題である．近年，中国語の計算機解析に関して，句構造文法による統語解析[1),2)]，言語の意味範疇（成分分析法）に基づく構文解析[3)]と理解[4)]などの研究が行われているが，曖昧性の問題を解決するための研究はまだ少なく，有効な方法はほとんど提案されていない．人工知能研究の分野では従来から，曖昧性解決の方法として，意味や知識を利用すればよいということが提案されているが，この方法により言語を解析し，理解するためにはぼう大な知識を必要とするため，現段階ではまだ実際的ではなく，また，人間がつねにそのようにしているとも思われない．

本論文では，中国語の統語上の特徴を細かく調べて，曖昧性解消（たとえば，品詞判断や句構造指定など）に役立ちそうな，表層上の知識を集めて，それらを組み合わせて利用することを試みる．本論文で利用する知識は

① 中国語文の構文上において，機能語のように働く一部の単語（ここで特徴語と呼ぶ）：助詞，前置詞，方位詞，助動詞，補助動詞，接続詞などに関する知識

② 断片の最尤さに関する知識

である．ここで利用される知識は，第一に，つねに正しいという性質のものではなく，多くの場合正しいというヒューリスティックな性質をもつ．第二に，完全なものではなく，すなわち，すべての統語現象をカバーできるものではなく，むしろ典型的な現象あるいは特徴を抽出しやすい現象しか取り扱わないものである．第三に，特定の単語や句に依存する個別的なものである．このような知識は，一般の統語・意味解析規則とはかなり性質の異なるものであるから，われわれはこれを統語・意味解析の前処理サブシステムとして

† Use of Heuristic Knowledge in a Chinese Analysis System by YIMING YANG, SHUJI DOSHITA and TOYOAKI NISHIDA (Department of Information Science, Faculty of Engineering, Kyoto University).

†† 京都大学工学部情報工学教室

1　2　3　4　5　6　7　8　9 10 11 12 13　14　15　16

例文：在　没有　摩擦　的　理想　情況　下　物体　将　以　恆定　的　速度　運動　下　去。

において，ない，摩擦，の，理想，情況，下，物体，(未然形)，で，定常，な，速度，運動，していく。

訳文：摩擦のない理想情況の下において，物体は定常な速度で運動していく．

a．複数個の統語カテゴリを持つ単語（例文において□で示されたもの）：

在 (1)：動詞 (V)：いる／前置詞 (P)：において
没有 (2)：動詞 (V)：持っていない／副詞 (AD)：なかった
摩擦 (3)：動詞 (V)：摩擦する／名詞 (N)：摩擦
理想 (5)：形容詞 (A)：理想的／名詞 (N)：理想
下 (7,15)：動詞 (V)：おりる／補助動詞 (VAUX)：下へ／動量詞：回，度／方位詞：…の下
運動 (14)：動詞 (V)：運動する／名詞 (N)：運動
去 (16)：動詞 (V)：行く／補助動詞 (VAUX)：していく

b．例文に対応する可能なパーザ木

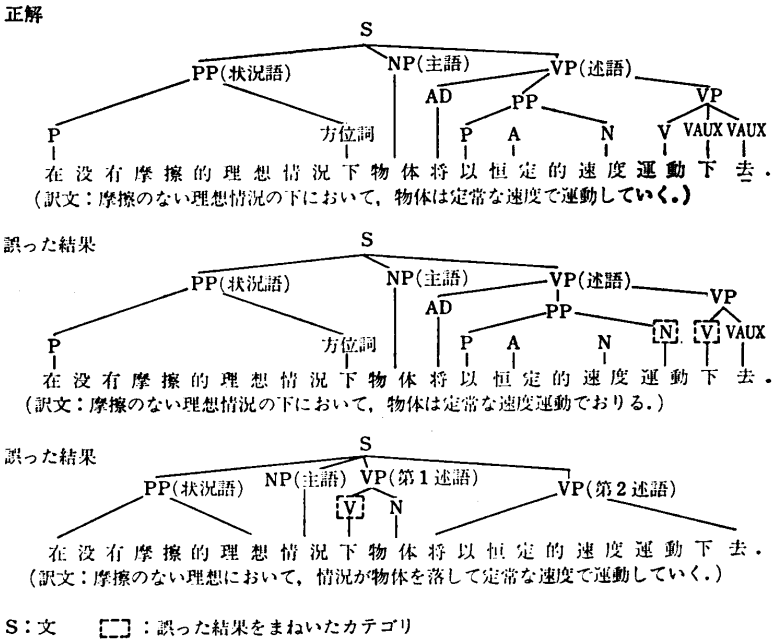


図 1　例文「在没有摩擦的理想情況下物体将以恆定的速度運動下去。」の統語解析における曖昧性

Fig. 1　The ambiguity in the syntactic analysis of a Chinese sentence.

別のモジュールに集めた．一般の言語理論では，構文解析は形態素解析一統語解析一意味解析のように分けられている．しかし，中国語解析にとってはこのような分け方はむずかしい．ここでの解決方式では，形態素解析と統語解析をはっきり分けることをせずに，一括して扱うものであり，前処理は統語解析の一部であると考えられる．

現在の前処理サブシステムの入力は，中国語式ローマ字（漢語拼音）によって単語ごとに分ち書きされた中国語文である．システムは，ヒューリスティックな知識によって入力文の中の句のまとまり（断片）を抽出して，もっともらしい順に統語・意味解析サブシステム（本処理）に引き渡す．断片は，単語の統語カテゴリや句構造などを局所的に指定または予測することによって得られた入力文の部分構造であり，曖昧性を抑え，本処理の統語・意味解析での無駄な探索を減らす役目も果たす．本処理では，統語規則（正しく，完全であり，CFG で記述されたもの）を用いて前処理で指定されなかった部分の統語解析を完成しながら，意味解析を行う．ただし，本処理が失敗した（予測された統語構造がありえない）場合，システムは前処理のその出力を拒否して，バックトラッキングを行う．

以下では，本方式について詳細を述べ，この方式がどの程度有効に機能するかについて検討する．

2. ヒューリスティックな知識を利用した曖昧性解消

2.1 断片リンク構造

前処理を説明するために，まず断片構造を導入する．前処理で得られた結果を断片リンクと呼ぶデータ構造を用いて記述する．**図2**に本論文で用いる断片リンクの種類を示す．図2において，各リンク構造の意味は次のとおりである：入力系列 s 番目の単語から e 番目の単語までに関して

リンク①：断片の構造が発見された．

リンク②：断片の構造が予測された（実際の構造はまだ生成されていない）．

リンク③：断片の構造が一部発見された（他の部分は生成されていない，すなわちリンク③の中にリンク②を含んでいる）．

リンク④：ある構造がありえない．

ただし，リンクのないところは，「統語構造がない」ということではなく，「わからない」という意味である．

2.2 特徴語の利用

中国語では，虚詞（前置詞，助詞，接続詞，副詞など）といわれるものと，助動詞，補助動詞，方位詞などの語は，数が少なく，使用頻度が高いため，解析における曖昧性を減らすための有力な手掛りになる．本文式では**表1**に示したような特徴語を利用して，次のように処理を行う．

（1） 特徴語の前後の品詞決定

特徴語の前後に曖昧性のある語が来たとき，その特徴語を利用して**図3**のように品詞を決定することができる．

例1で示したように，助詞「了」，「過」，「着」，「得」の直前の｛〈動詞〉，〈名詞〉｝（動詞でもあり名詞でもある）の語は動詞であると判断する．

例2で示したように，補助動詞（たとえば「起来」）が｛〈動詞〉，〈名詞〉｝の語，あるいは｛〈動詞〉，〈前置詞〉｝の語の直後にあると，前のほうは動詞，後のほうは補助動詞であると判断する．

例3で示したように，単語「要」，「会」などは助動詞でもあり動詞でもある．それが｛〈動詞〉，〈名詞〉｝の語の直前にあると，前者は動詞でなく助動詞であり，後者は名詞でなく動詞であると判断する．

（2） グループ化

特徴語を利用して，どこからどこまでが一つの断片として集められるかを決定する．

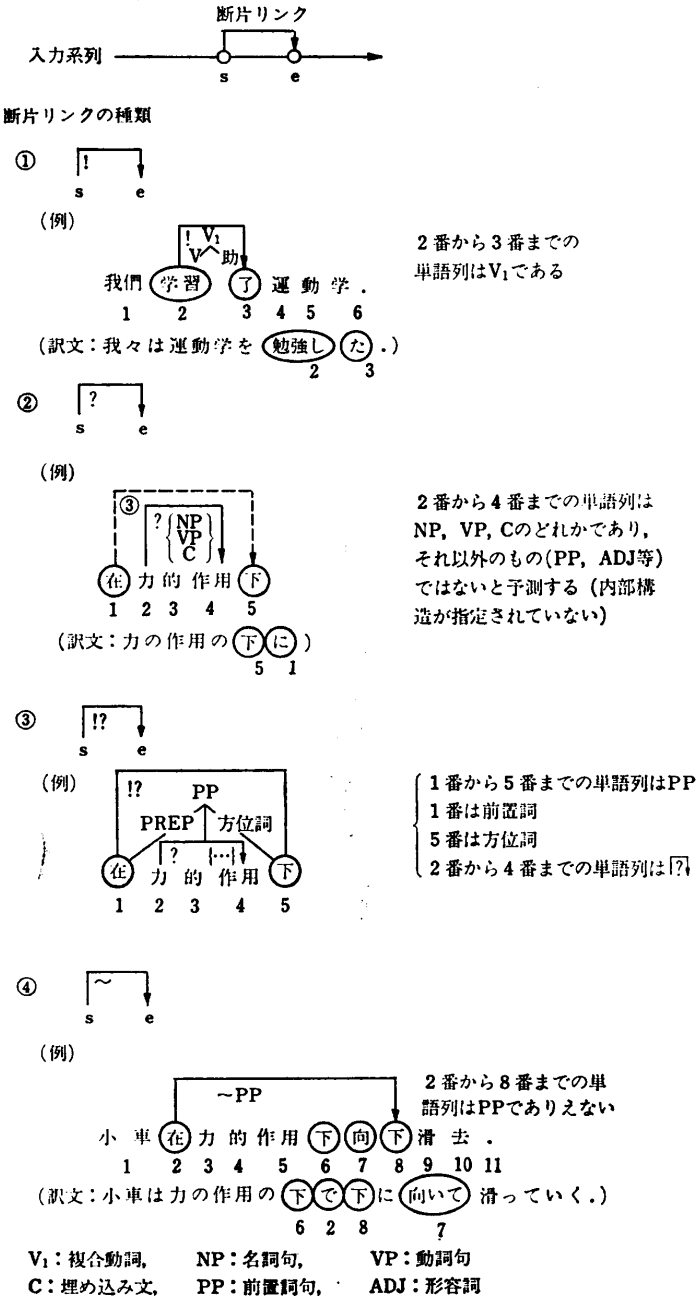


図2 断片リンク構造の種類
Fig. 2 Kinds of fragment link.

表 1 本システムで取り扱う特徴語
Table 1 The characteristic words used by the system.

助　詞	7個	了，過，着，的，地，得，所
補助動詞	17個	上，下，来，去，到，起，起来，過，進，出，回，開，完，成，掉，作，為
助動詞	10個	能，可能，能够，会，要，可以，必須，应，应當，应該
接続詞	5個	跟，和，与，而，因
前置詞	13個	在，到，向，当，从，於，沿，随，用，根据，由，对，依
方位詞等（方向，位置時間などを表す特徴名詞）	95個	上，下，前，後，…
特殊動詞	8個	是，来，去，進行，開始，持続，保持，発生
慣用型	13個	也就是説（即ち），对…来説（…にとって），是…的（…のである），…
綴り	5個	_性，_化，_状態，_状況，_情況
「標点符号」（区切り記号）	2個	，。
専門用語（現在30個）		匀速直線運動，平抛運動，…
etc.		将，…

たとえば，前置詞の場合，他の語（方位名詞など）との呼応関係を利用してグループ化ができる．他の例では，文中にある種の特徴語があれば，その語から文の終りまで一つの動詞句であると判断できる．

図4の例1において「在」から「下」までが一つの前置詞句（PP）としてグループ化され，例2においては「是」から文の終りまでが動詞句としてグループ化される．

（3） 単語の役割（role）の決定

特徴語を利用して，その前後の語が主文の述語かどうかというような役割を決定する．ここで用いる規則は次のとおりである．

規則1：特徴語「的」と「是」の前後の単語は，主文の述語にならない．

規則2：特徴語「所」，あるいは前置詞の後の語は主文の述語にならない．

規則3：形容詞は，もし入力文の最後の単語でなければ，主文の述語にならない．

規則4：すでに他の断片の一部に組み込まれたものは主文の述語にならない．

これらの規則を図5の例文に適用すると，初め7個の動詞，形容詞が主文の述語になる可能性があったが，このうち5個は主文の述語になれないことがわかる．この結果，述語動詞句のスコープは二つの可能性に限定される．

2.3 量的情報の利用

句の長さに関する情報も，曖昧性解消に有益である．ヒューリスティック規則の競合が生じたとき，最長のものをとるか，最短のものがよいかは場合により異なる．図6では選択処理の各種の場合を示す．

例1のような断片に最長優先の規則を適用する．その結果，いちばん外側の断片が選ばれる．

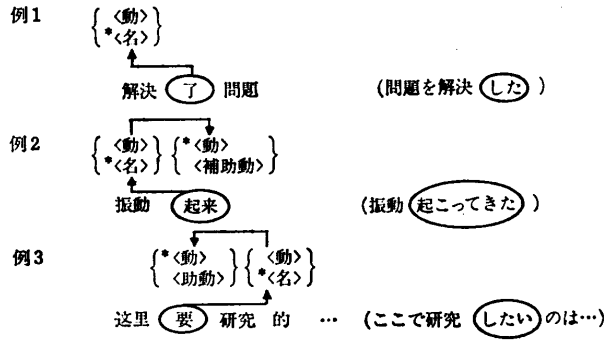


図 3 特徴語と語順関係による品詞判断の例
Fig. 3 Examples of category decision by characteristic words and their position.

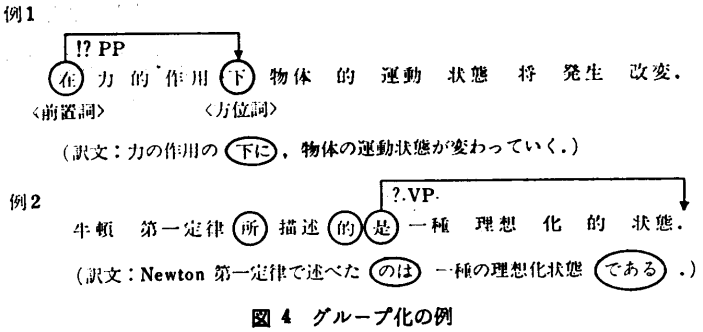


図 4 グループ化の例
Fig. 4 Examples of grouping.

例2には，最短優先規則の適用例を挙げる．

例3の前置詞の場合，四つの前置詞句 $\{f_1, f_2, f_3, f_4\}$ のなかで，正解は最長のほうでもなく最短のほうでもない．その場合，前置詞から呼応語までの距離は最短，しかも呼応語列の長さは最長のほうを優先する規則を用いる．その結果，$f_3 > f_4 > f_2 > f_1$（>：優先）の優先順が得られ，四つの断片から中間のほう（正解の f_3）が選ばれる．

ここでは，以上の選択規則をスコアを用いた優先度処理としてシステムに組み込む．

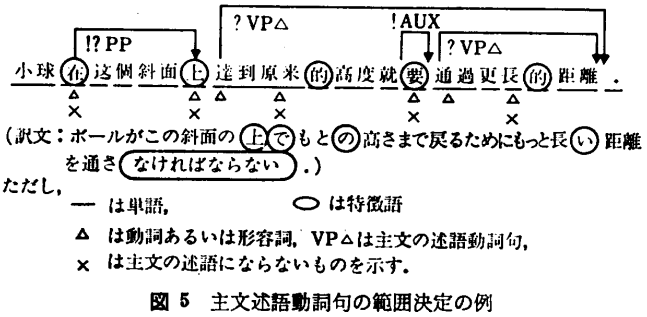


図 5 主文述語動詞句の範囲決定の例

Fig. 5 Examples of deciding the scope of the main sentence predicate.

3. ヒューリスティックな知識の組織化と利用法

ヒューリスティックな知識の利用において，本システムには次の四つの特徴がある．

（1） Bottom-up オートマトン

ヒューリスティックな知識を各特徴語ごとの規則に書き，Bottom-up ATN (B. ATN) と呼ぶ形式で記述する．処理するとき，システムは入力文における各特徴語の B. ATN を独立に起動して，可能な断片をすべて抽出する．

（2） 断片競合の検出と解決

独立に抽出された各断片は，互いに矛盾する(競合と呼ぶ)可能性がある．ここで断片の位置関係のチェックなどにより競合の検出を行う．競合の性質によって，競合の両方の統合処理を行うか，あるいは一方の選択処理を行う．選択処理は断片集合から，各断片の優先度によって競合のない，最尤部分集合を優先的に選び出す．

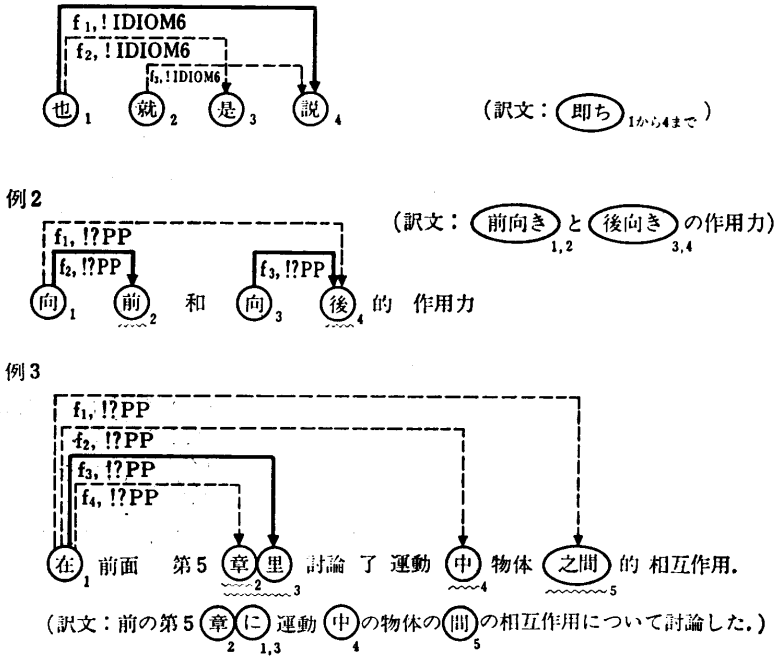


図 6 断片の範囲決定における曖昧性

Fig. 6 The ambiguity of fragment scope decision.

（3） 優先度（スコア）

断片の最尤さに関するヒューリスティックな知識は優先度設定として組み込む．優先度は断片パターンの精密度，完全さ，および断片の種類などに依存するものである．優先度はスコアによって数値化され，各断片につけられる．スコアは定数のみに限らず，関数式でも計算され，各断片の構造を反映するものである．

（4） 正しさの保障（バックトラッキング）

ヒューリスティックな知識を利用して，多くの場合正しいと思われる結果が先に得られるようにする．同時に，残りの可能性も保存しておき，先の結果が誤っていた場合，バックトラッキング方式により，前処理ルーチンはいつかは正しい解を出力することを保障する．

3.1 前処理の構成

図7は，前処理系の構成を示す．前処理は次のように行われる．

（1） スキャン：左から右へ入力系列における特徴語の検出を行う．

（2） 断片抽出：特徴語があれば，その特徴語の処

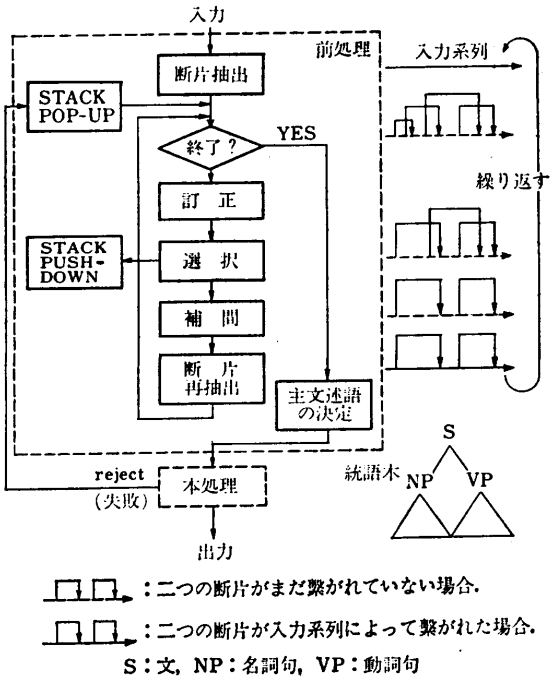


図 7　前処理系の構成

Fig. 7　Structure of preprocessing system.

理規則を記述したオートマトンを起動して，入力系列にパターンマッチングを行って断片を抽出する．

（3）　訂正：訂正規則にマッチする断片があれば(3.3 節の(2)と図 10 を参照)，訂正を行う．

（4）　選択：断片間の位置関係（重なっているかどうか）をチェックしながら，選択処理と残る可能性の保存を行う．

（5）　補間：選択処理の結果（いくつかの断片の集合）に対し，入力系列により補間する．すなわち，断片と断片の間の繋がっていない部分を入力系列の対応部分によって繋ぐ．

（6）　繰返し：補間の結果を再び入力系列として(1)へ．以上の処理を新断片が生じなくなるまで行う(このように，断片抽出は入力系列に対して繰り返し行われるので，前処理の各部への入力系列は単語の系列に限らず，前の処理ステップですでに抽出された断片を含んでいることもありうる)．

（7）　主文の述語の決定：以上の処理結果に基づいてさらに主文の述語動詞句を抽出する．

（8）　バックトラック：後続の処理ステップで断片抽出の結果が誤りと判断されたとき，バックトラッキングを行い，次の可能な結果を作り出す．

3.2　Bottom-up ATN (B. ATN)

B. ATN は各特徴語ごとの規則を記述する．この B. ATN は非決定性であり，OR ノードと DET ノードの 2 種類のノードがある．OR ノードで非決定性の処理を行う．すなわち，OR ノードからのすべてのブランチを独立にたどる．DET ノードでは決定性の処理を行う．すなわち，DET ノードではブランチ（互いに排他的なもの）を順番にたどって，最初に出られるところから出ていく．

処理するとき，システムは入力系列を左から右へスキャンして，特徴語を探しあてると制御を B. ATN に移して，その特徴語に関する局所的な解析を行い，断片の構造を抽出する．B. ATN の処理が済んでから，制御をもとのレベルに戻して，次の位置からスキャンを継続する．

B. ATN において，スキャナは左へバックしたり右へ飛んだり自由に検索できる．状態遷移の矢印に付けられている数字は，スキャナの動きを制御するための情報である．たとえば，「−1」は現ポイントから左へ 1 ステップバックする，「+2」は右へ 2 ステップ進めることを指示する．

図 8 に特徴語「在」の B. ATN を示す．図 8 には，前処理システムのスキャンが文中の「在」の位置まで来ると処理の制御はこの B. ATN に入る．状態②から③への遷移では次のことを記述している．

もし単語 W（「在」の左の単語）が動詞の部分集合 D_3 の要素ならば，「W 在」を複合動詞 V_2 の統語構造をもつ断片として抽出し，同時ににその断片のスコアを設定する（スコアについては 3.4 節で説明する）．それからスキャナが 1 ステップとばして（「在」の右の単語を見て），制御は状態③に移る．

3.3　競合の検出と解決

競合処理では，競合の検出，誤り拒否と訂正，および選択処理を行う．

（1）　競合の検出

以下の場合，断片の間に競合があるとする（その他の場合，競合がないとする）：

（i）　入力系列が禁止パターンにあたるとき．

（ii）　断片が重なったとき．

例を**図 9** に示す．例 1 では，主文の述語を抽出するとき，禁止パターンを検出しており競合が発見される．例 2，3 と例 4 では，断片が重なったことによって，競合が検出される．

（2）　競合の解決—誤り拒否，訂正と選択処理

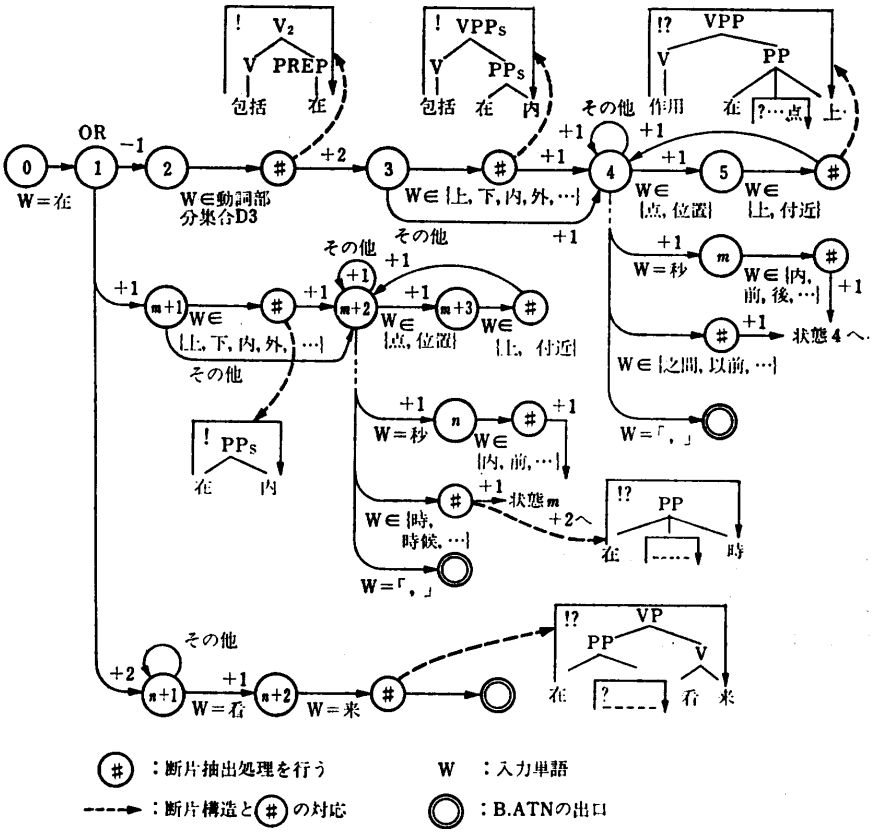


図 8 特徴語「在」の Bottom-up ATN (概要)
Fig. 8 Bottom-up ATN of characteristic word.

競合が検出されると，ある場合には誤り拒否や訂正を行う，その他の場合には選択処理を行う．

選択処理では，競合のある断片集合を，競合を含まない部分集合に分けて，別々に処理する．母集合に k 個の断片があれば，部分集合は 2^k 個ありうる．ここでは，2^k 個の部分集合をすべて作り出すことはしないで，優先度（スコア）の高い断片を優先に選択しながら，競合のない最大集合を先に作り出す．

上の例では，図9の例1の場合は入力系列を棄却するが，例4は訂正を行い，例2と例3では選択処理を行う．**図10**に，その訂正処理（断片合成）を示す．

図11には選択処理の1例を挙げた．図11の例では，断片競合の組 $\{f_1, f_3\}$，$\{f_2, f_3, f_4\}$ がある．断片の優先順で，f_2 のスコアがいちばん高いので，先に選択される．ゆえに f_3 も f_4 も選択できず（f_2 と競合するので），最終結果は，母集合の $\{f_1, f_2, f_3, f_4, f_5, f_6, f_7\}$ から f_3 と f_4 を除いた部分集合 $\{f_1, f_2, f_5, f_6, f_7\}$ が先に選ばれる．

3.4 優先度設定

前処理では，各断片に異なる優先度をもたせる．優先度設定には，まず各種の断片をいくつかの優先度レベルに分けて，各レベル内部をさらに断片長さなどの情報によって順序づけする．優先度を二つの値の組（スコア1，スコア2）で表わす．スコア1は優先度レベルの順位を示す，スコア2は同レベル内の各断片の優先度を示す．**表2**にスコアの設定一覧を挙げた．ここでスコアの小さいほうが優先度が高いとする．

次に，スコア設定について詳しく述べる．

（1） 優先度レベル（スコア1）の設定

優先度レベルの設定には単純な方法がなく，ここで断片の各種の性質を合わせて経験的に決めた．その設定は次のような基準にしたがうものである．

（i） 精密度：断片パターンの型により精密度は異なる．連続する特徴語のパターンは統語カテゴリを含んでいるパターンより精密であるが，それらのいずれも“…”（前処理の段階で指定しない部分）を含むパ

ターンより精密である．また，全パターンは部分パターンより精密であると考えられる．ここで精密度の高いほうを高いレベルにおく．

（ⅱ） 断片種類：ある種の断片は他種の断片より頻繁に使われる．そのようなパターンにより高い優先度を与える．たとえば図9の例3に示したように，複合動詞断片 V_3 と V_1 が重なった場合，V_3 のほうはつねに正しいので，V_3 を V_1 より優先度の高いレベルにおく．

以上の基準と実例の考察結果に基づいて，現段階において，本システムでは断片（26種類）を七つの優先度レベルに分けた．

（2） 同レベル内部の優先順序設定（スコア2）

スコア2は断片パターンのなかに含まれる特徴語列の長さ（長いほうを優先する）および"…"部分の長さ（短いほうを優先する）によって設定される．たとえば，

$$\underbrace{<\text{特徴語列1}>}_{l_0} \underbrace{\cdots}_{l_1} \underbrace{<\text{特徴語列2}>}_{l_2}$$

のような断片パターンの優先度は $(l_0-l_1+l_2)$ に比例する．

具体的には，レベル A, CA, CCA（表2を参照）の各種断片パターンは特徴語列になっている．全パターンを部分パターンより優先するためには，それらのスコア2は断片長さ (l) によって次のように計算される．

$$\text{SCORE}_2=k_1\times(L_1-l)+\delta \quad (L_1>l)$$

この式では，l の大きいほうが SCORE_2 が小さく，優先度が高い．L_1 は SCORE_2 が負数にならないように定めた定数である．δ は SCORE_2 の修正値，k_1 は定数である（δ と k_1 について本節（ⅲ）で説明する）．

ほかの例では，レベルCと CC における各種の断片は

＜特徴語列1＞…＜特徴語列2＞

のようなパターンをもつものである．そういうタイプの断片に対して，特徴語間の距離（"…"部の長さ）が短く，しかも特徴語列の長いほうを優先にする．たとえば，前置詞句 PP のスコア2の計算は次のようである：

$$\text{SCORE}_2=k_4\times(L_4+l_1-l_2)+\delta$$

ここで L_4 は SCORE_2 が負数にならないように定めた定数，l_1 は前置詞から呼応語までの距離，l_2 は呼応語列の長さである．$l_2=1$ の場合，上式は l_1 の最小のほうを優先するようになる（δ と k_4 については（ⅲ）を参照．また，例については，図6の例3を参照）．

（ⅲ）「δ 修正」（左優先）：

優先度レベルも長さも同じ断片が抽出された場合，スコア δ を付加して断片の入力系列上の位置を区別し，左のほうを優先する．この「δ 修正」を（ⅰ）や（ⅱ）で決めた優先度に影響させないために，スコア2の計算式に係数 k_i を用いる．たとえば，文のなかでレベルも長さも同じ断片が10個を越えないとすると，δ はたかだか10を越えない．それに対しては k_i を10と設定すればよい．

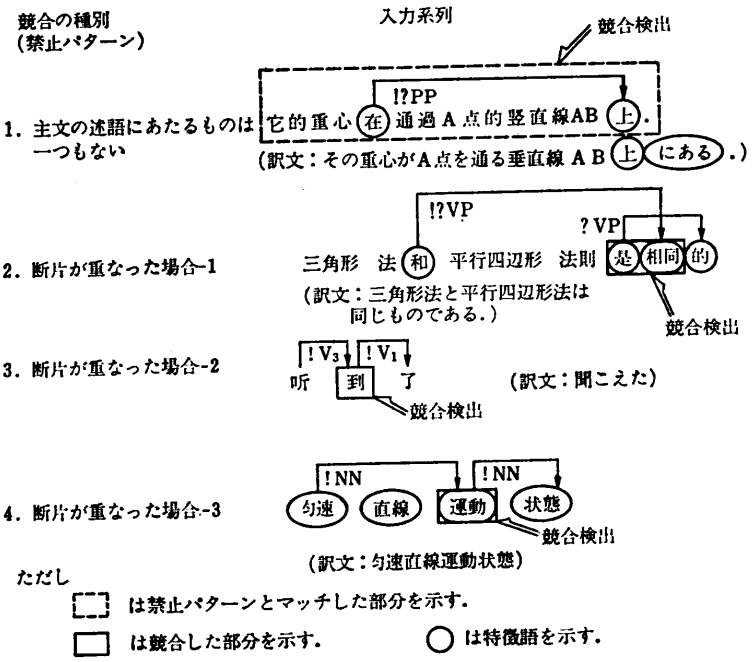


図 9　断片競合の検出の例

Fig. 9　Example of fragment conflict detection.

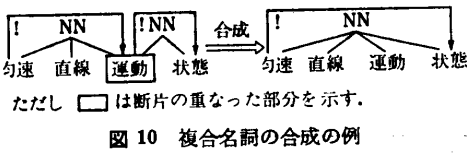
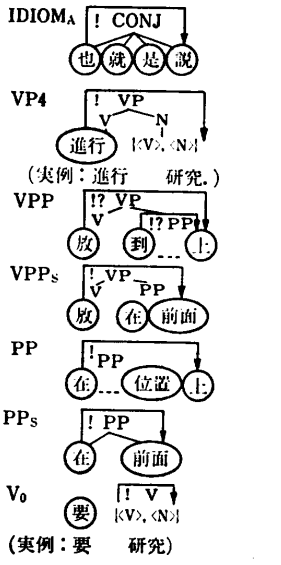


図 10　複合名詞の合成の例

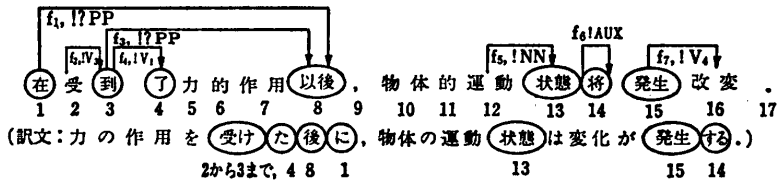
Fig. 10　Example of compound noun composition.

表 2 優先度レベルとスコア一覧

Table 2 Preference levels and their score.

優先度レベル	$SCORE_1$	$SCORE_2$	断片種類	説明	断片の例
A	1	$k_1\times(L_1-l)+\delta$	NN $IDIOM_A$ V_3 V_7	複合名詞 慣用型 複合動詞 複合動詞	
B	2	δ	VP_4 V_8 AUX AFFIX	動詞句 複合動詞 助動詞 複合名詞	
C	3	$k_2\times(L_2+l_1-l_2)+\delta$	VPP $IDIOM_C$	動詞句 慣用型	
CA	4	$k_3\times(L_3-l)+\delta$	VPP_S V_2 V_6	動詞句 複合動詞 複合動詞	
CC	5	$k_4\times(L_4+l_1-l_2)+\delta$	PP	前置詞句	
CCA	6	$k_5\times(L_5-l)+\delta$	PP_S V_1	前置詞句 複合動詞	
D	7	δ	V_0 ⋮	動詞 ⋮	

L_i: $SCORE_2$ は負数にならないために設定した定数，l_j: 抽出された断片の各部分の長さ，δ: L_i も l_j も同じ値をもつ断片のうち，入力系列の左のほうを優先にするための修正値，k_i: δ の修正によっても L_i と l_j で決めた優先順が影響を受けないために設定した定数，…: 前処理段階で扱わない系列，CONJ: 接続詞，{<V>, <N>}: 動詞でも名詞でもある単語



入力断片集合 $\{f_3, f_8, f_7, f_6, f_2, f_1, f_4\}$ (優先順で並んだ)

選択処理の各ステップにおける中間結果

Step	未選択の断片	現時点で見る断片	選択された断片
0			{ }
1	$\{f_3, f_8, f_7, f_6, f_5, f_1, f_4\}$	f_3	$\{f_3\}$
2	$\{f_8, f_7, f_6, f_5, f_1, f_4\}$	f_8	$\{f_8, f_3\}$
3	$\{f_7, f_6, f_5, f_1, f_4\}$	f_7	$\{f_7, f_8, f_3\}$
4	$\{f_6, f_5, f_1, f_4\}$	f_6	$\{f_6, f_7, f_8, f_3\}$
5	$\{f_5, f_1, f_4\}$	f_3	$\{f_6, f_7, f_8, f_3{}^*\}$
6	$\{f_1, f_4\}$	f_1	$\{f_1, f_6, f_7, f_8, f_3\}$
7	$\{f_4\}$	f_4	$\{f_1, f_6, f_7, f_8, f_3{}^*\}$
8	{ }		$\{f_1, f_6, f_7, f_8, f_3\}$

$f_j{}^*$: すでに選ばれた断片 f_j は現時点で見ている断片と競合した.

図 11 選択処理の一例

Fig. 11 Example of fragment selection.

4. 検討

次に本前処理システムの一般性と拡張性および有効性について検討する.

(1) 一般性と拡張性

本システムの前処理規則は，200程度の特徴語に関して101個の B. ATN で記述されるものである．断片の統語カテゴリは48種類あって，本中国語解析システム（本処理の統語・意味解析サブシステム[5]）で用いる文脈自由句構造文法の統語カテゴリ全体（72種類）の2/3をカバーしている.

中国語における典型特徴語（「虚詞」など）は言語学者によって800個程度整理されている[6]．本システムで扱う200個程度の特徴語のうち，約100個はその800個の典型特

徴語から選んだものである．他のものは本方式のために必要として加えたものであり，また，そのなかに本実験で対象とした分野（物理学のテキスト）に依存するものを約30個含む．

中国語の800個の典型特徴語のうち，本システムで用いなかったものは次のような種類のものである．

① 会話の常用語，たとえば，形容詞「了不得」(訳語：「たいした」)，「这可了不得．」(訳文：これはたいしたことだ．)

② 詞綴，すなわち，単語分ち書きレベルの特徴語，たとえば，名詞綴り「具」，「家具」(訳語：家具)，「用具」(訳語：「道具」)．

③ 文と文の間の接続詞，たとえば，「所以」(訳語：「したがって」)．

④ 意味解析の段階で重要な作用を果たす語であるが，前処理の段階で特徴を抽出しにくい，すなわち前処理で有効に利用できないもの，たとえば，前置詞「据」，「据報道」(訳文：「報道により」)．

本前処理システムの方針は，①書き言葉を解析対象にし，②単語を入力の基本単位とし，③統語解析に重点をおくことであり，目的は特徴語を利用してヒューリスティック解析を実現することである．したがって本システムにとって以上のような特徴語を現段階で考えなくてもよいし，本システムで取り扱った約100個の典型特徴語はかなりの言語現象をカバーしており，本システムは一般性をもつシステムであると考えられる．

また，現システムはまだ完全なものではないが，中国語の解析に必要な特徴語はたかだか数百程度であるから，B. ATNも数百程度ですむので，今後の拡張によっても優先度設定や前処理の有効性およびシステム上の実現に根本的な影響を与えないと考えられる．

（2） 有 効 性

ここで述べた方式を高校物理教科書から順番にとった（やや修正したものもある）実例120文について机上で検討した（本論文で挙げた例文もほとんどそのテキストからとったものである）．120文のなかには断片が369個含まれており，そのうち競合した断片は122個あった（誤って抽出されたものを68個含んでいた）．本選択規則を使うとき，94%の場合第1位において，98%の場合第2位，100%の場合第3位選択までで正解が得られた．

選択処理の有効性をさらに確かめるために，競合解決の重点である前置詞句断片を対象として，さらに多数の実例について調べた．上記のテキストからの前置詞句を含んだ800程度の例文のうち，競合の生じた（前置詞句の範囲決定に曖昧性があった）ものは80文があった．その80文において，競合した断片は（誤ったもの101個を含めて）209個あったが，それらのスコア計算による選択処理の結果，83%の場合第1位において，90%の場合第2位，98%の場合第3位までで前置詞詞断片が正しく選ばれた．

以上に述べたような前処理を用いた中国語解析は現在システム化が進められ，一部働いているが，たとえば図1の例の曖昧さが，前処理を用いることにより，996通りのパーザ木から8通りに減少されることなど，前処理の有効性が確かめられている．このことについては後日改めて報告したい．

5. ま と め

本論文では，前処理方式において特徴語に関するヒューリスティック知識を中国語解析の曖昧性解消に利用することについて述べた．前処理では，特徴語を手掛りにして局所的な解析を行う．すなわち，入力文の特徴を抽出しやすいところから解析を展開する．前処理では，予測的な解析を行う．すなわち，部分構造の細かいところがわからなくてもそれを予測しながら解析を進める．前処理で用いられる規則は絶対正しいものとは限らず，むしろときには誤りを含んだヒューリスティックな性質をもつ．すなわち多くの場合正しいが，ときには互いに矛盾することもありえるような，各単語に依存する規則をともに利用することができる．前処理システムは規則の優先度を利用することにより競合解決の能力をももつ．このような解析方式では，常識といわれる，人間のよく使うヒューリスティックな知識がシステムに容易に組み込め，それらの最大限の有効利用が実現できる．この方式は，曖昧性の多い自然言語処理に，とくに中国語のような形態変化の規則性の弱い，統語的な曖昧性の大きい言語の解析に有効である．

本前処理システムを，文脈自由句構造文法に基づく統語・意味解析システムと結び付け，中国語解析の実用性のあるシステムを作成することが今後の研究課程である．

謝辞 末筆ながら討論していただいた本学堂下研究室の諸氏に感謝いたします．

参考文献

1) 範継淹，徐志敏（中国科学院語言研究所）：应用“扩充轉移網絡理論分解漢語”(Augmented Transition Network 理論を用いた中国語解析)，語言研究，創刊号 (1981).
2) 馮志偉(中国科技情報所)：漢語句子的多標記多叉樹形図分析法（中国語文の解析に用いる multiple-labelled and multiple-branched tree graph analysis method), Proceedings of 1983 International Conference on Chinese Information Processing (2), pp. 144-158 (1983).
3) 欧陽文道（中国科学院自動化研究所）：中文信息五維模型和分詞，析句，辨義的算法研究（中国語情報を記述するための五次元モデルとわかち書き・構文解析・意味解析の方法についての研究)，Proceeding of 1983 International Conference on Chinese Information Processing (1), pp. 153-157 (1983).
4) 李家治，陳永明（中国科学院心理所）：機器理解漢語的一些探索和設想（計算機による中国語の理解についての検討と発想)，Proceeding of 1983 International Conference on Chinese Information Processing (2), pp. 138-144 (1983).
5) 呂叔湘：現代漢語八百詞，商務印書館，北京 (1980).
6) 楊頤明：中国語解析システムに関する研究，京都大学修士論文 (1982).

(昭和59年1月27日受付)
(昭和59年6月19日採録)