自然言語処理　33－1
（1982. 10. 22）

中国語解析システムについての考察

楊　頤明，　西田　豊明，　堂下　修司
（京都大学　工学部）

［内容梗概］
本論文では、中国語解析に適した処理機構を持つパーザとその上で働く中国語解析システムを提案する。中国語では、英語や日本語とは異なった立場からの処理が必要である。本論文は、動詞や述語の解析を中心にするパージング法と、"是字句"（"である"文）の意味解析の二つの部分について述べる。このパージング法では、解析方式は二段階からなる。まず動詞や述語などを判断して、述語動詞をキーワードとして入力文を分割し、各部分について前処理を行う。次に全体の解析を行う。意味解析では、中国語において多様な用法が行われる"是字句"を取りあげた。"是字句"の解析には、属性と用例の二種類の知識を利用する。用例知識はネットワークで表わす。入力文に関する意味情報をフレームで取り扱う。

1. 始めに

中国語を解析するときの主な問題点は、中国語に形態的な特徴が少ないことである[1]。多品詞性は、その特徴の一つである。中国語の多品詞性の多くは動詞によってひき起される。動詞と同形の語は、動名詞の全部、補助動詞の全部、"介詞"（前置詞）のおよそ半分、ある種の助動詞、および"ある種の副詞（例："比較"は動詞(比較する")と副詞（"比較的"））"である。また、中国語に日本語の形式名詞の"（である）こと"にあたる語や、動詞句と動詞句を接続する語（日本語では、前の動詞が連用型になる）などがない。したがって、動詞や動詞句の解析には、曖昧性が大量的にひき起される。それに対して特別な対策は必要であると考えられる。

一方、"是字句"は、中国語の典型的な文型であり、使用頻度も高い。"是字句"は他の文型と共に用いられると曖昧性をひき起す。"是字句"は"AはBである。"を意味するが、このとき、AとBとの間の意味的な整合性を判断することにより曖昧性を取り除くことが考えられる。ところが、従来の概念分類はこの目的には不十分であり；また、与えられた組み合わせが可とも不可とも判断しがたい場合がある。おして実際の用例やヒューリスティック情報に基づき、スコアによる相対的比較判断を行なう方が有効であると考えられる。

2. 動詞や述語の解析を中心にする中国語解析システム

2.1　システムの構成

本論文で提案する中国語解析システムの構成を図1に示す。以下では，図1に示す処理の各ステップを詳しく説明する。

(1)慣用型と固有名詞の処理：入力文に含まれている慣用型や固有名詞を認識し一つにまとめる。これにより，慣用型や固有名詞に含まれる動詞に同形の語は，以下の処理で"マスク"される。慣用型と専門用語の処理では，パーザは慣用型辞書と専門用語辞書（表1、表2）を用いる。入力文の部分リストが，ある辞書項目とマッチしたら，その部分を対応項目のノードで置き換える。名詞後綴型名詞（例えば"装置"、"系統"、"能力"）の処理では，その名詞の前に動詞か動名詞の単語があれば，それを動名詞と判断して，両者を一つの複合名詞とする。これらの処理は，単語毎の規則として辞書に記述しておく。図2のように，フレームにより，各単語の辞書項目に，その単語に関する制御ルーチンや検査用情報などを

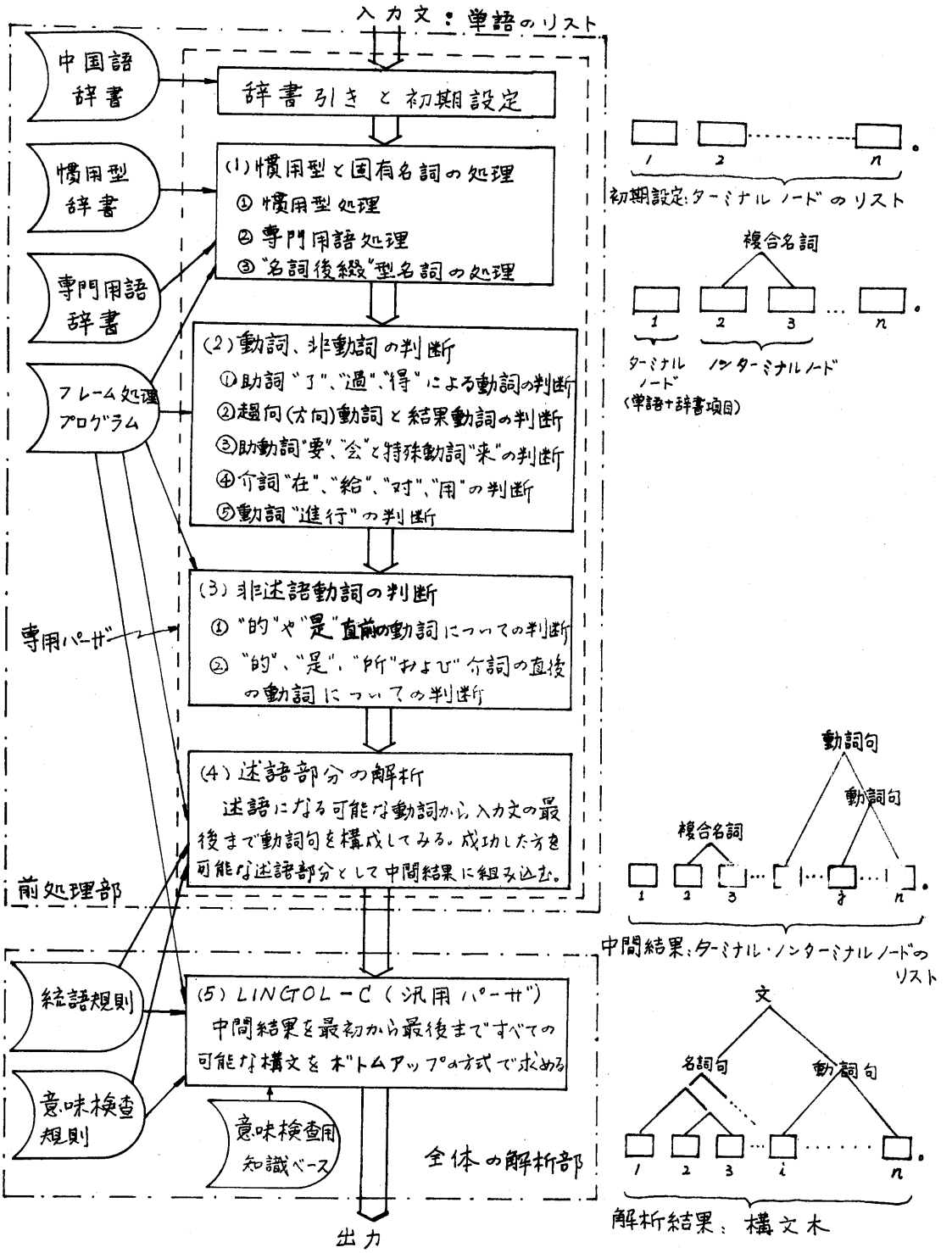


図1．中国語解析システムの構成とパージングアルゴリズム

表1. 慣用型辞書の内容

単語のリスト							リストから合成されるノード					
0	1	2	3	4	5	…	カテゴリ	T/N	構造木	フレーム	スコア	意味
⋮							⋮					
ゆ	是						副詞	N	(是)	…	…	…
ゆ	就	是	説				〃	N	(就 是説)	…	…	…
ゆ	也	就	是				〃	N	…	…	…	…
ゆ	也	就	是	説			〃	N	…	…	…	…
⋮							⋮	⋮				

ただし、ゆは入力文の始めの記号；
Tはターミナルノード；Nはノンターミナルノード。

表2. 専門用語(複合名詞)辞書の内容

単語のリスト			リストから合成されるノード				
1	2	…	カテゴリ	T/N	構造木	フレーム	スコア	意味
⋮			⋮					
抗干扰	編碼		複合語	N	…	…	…	…
干扰	功率		〃	N	…	…	…	…
出現	概率		〃	N	(出現 概率)	…	…	…
⋮								

単語"能力"の辞書項目の内容

名詞	<フレーム>	<スコア>	<意味>

OB(個別性質など)	…
FACT (制御情報)	(1)のステップで"名詞後綴"型名詞の処理ルーチンを呼び出す；前の動詞の禁止リスト=(提高 培养…)

ただし、(1)のステップとは、図1の(1)を指す。

単語"在"のフレームの内容

OB	…
FACT	(1)のステップで、慣用型処理の処理ルーチンを呼び出す。

単語"干扰"(妨害)のフレームの内容

OB	…
FACT	(1)のステップで、専門用語処理ルーチンを呼び出す。

図2. 辞書項目とそのフレームの例

①解決 了 問題 (問題を解決した)
{動名詞/動詞}

②建立 起来 (創立して来た)
{動名詞/動詞} {動詞/補助動詞}

③这里 要 研究 的… (ここで研究したいのは…)
{動詞/助動詞} {動詞/動名詞}

④对 信息 进行 量度 (情報に対して測定を行う)
{介詞/動詞} {動詞/動名詞} {動詞/動名詞}

図3. 形態素、語順関係による品詞の判断の例

記憶して、その単語を解析するときに利用する。

(2)動詞、非動詞の判断

処理方式は、(1)と同様な方式でフレーム処理によって行う。

判断のための規則を以下に述べる。

①助詞"了"、"過"、"得"の直前の単語は動詞であると判断する(図3の①)。

②中国語において補助動詞として使える動詞は趨向動詞(動作の方向を修飾する補助動詞、常用のは25ぐらいある)と、結果動詞(常用のは9個)二種類がある。ここで、そのような動詞が{動詞、動名詞}の語、あるいは{動詞、介詞}の語の直後にあると(図3の②)、それを動詞ではなく補助動詞と判断する(前後両者を1つの複合動詞とする)。

③単語"要"、"会"は助動詞でもあり動詞でもある。それが{動詞、動名詞}の語の直前にあると(図3の③)、それを動詞ではなく助動詞と判断する。

④単語"在"、"給"が{動詞、動名詞}の直後にある場合、あるいは"在"、"給"、"对"、"用"が{動詞、動名詞}の前方(直前ではなく)にある場合に、それらを動詞ではなく介詞と判断する(図3の④)。

⑤単語"進行"の直後に{動詞、動名詞}の語があると、"進行"を動詞、その後の語を動名詞と判断する(前後両者を1つの複合動詞とする、図3の④)。

(3)非述語の判断

中国語の文の構造では、一旦述語がわかれば、その述語から文の終わるまで動詞句になることもわかる。しかもその動詞句の前に付加部分としての副詞、前置詞句などが数少ないものしかないので、述語動詞を一旦判明したら、動詞句の始めも大体わかる。そこで述語動詞の判明は、動詞句の解析に重要なヒントになる。述語の判断には、直接に利用できるものが少ないので、こ

こで逆に不可能なものを可能集合から除く方法にする。

判断規則：

①助詞〝的〟、動詞〝是〟の直前の動詞（或は複合動詞）は述語でない。

②〝的〟、〝是〟、名詞前綴〝所〟および介詞（〝在〟を除いて）の直後の動詞は述語でない。

(4)述語部分の解析

(3)の処理で残った動詞を可能な述語とする。それらの動詞から最後まで統語規則によって動詞句の解析を行う。それらの解析が成功すれば（動詞句を構成したら），可能な述語部分として部分木の形で中間結果（図1）に組み込む。

(5)汎用パーザ LINGOL-C

中間結果に対して，文脈自由句構造文法で記述した統語規則によって，ボトムアップ方式で構文解析を行う（フレーム処理によって，意味解析も共に行う）。いくつかのキーワード（可能な述語）のある場合の処理を除けば，拡張LINGOLと同様に動作する。

2.2 解析方式の特徴

以上のように解析システムは専用パーザと汎用パーザからなる。

専用パーザの解析には，(1)できるだけ構文上の不能性を求める（正確性を求める方針にする）；語順関係と形態素の検査や処理を中心にする。(2)パージングは，左から右へ、あるいは右から左へどすらでも制御でき，結果間の比較と選択が簡単に行われる；各ノードに共通な規則や知識を利用するのではなく，分類された知識とノードのフレームに置かれる個別的な制御情報を利用して，必要な処理しか行わない。パーザの操作もより簡単なので（大部分は相隣関係の検査），処理スピードが速い。

汎用パーザの解析には，すべての構文上の可能性を与える（完全性を求める方針にする）；より複雑（本格的）な意味検査を行う。

前処理と全体解析（専用と汎用）の結合方式によって，両方の機能や利点を互に補充することができる。即ち，形態素や語順関係の解析規則や知識は，完全なものでなくても、或は規則にまとめにくい個別的な知識であっても，それらを最大限に利用できる。手間のかかる複雑な意味検査などを，なるべく最小限範囲で行うようになる。部分結果（述語部分、複合名詞など）を予め処理することによって、入力文の実質的な長さを短くする効果があるので，ボトムアップ方式による全体解析にとって，探索時間と記憶容量を減らすためへ有効である。

2.3 実験

以上のパージングアルゴリズムに沿って，中国の科学技術論文から抽出した（やや修正したこともある）100個の例文について机上シミュレーションを行った。その結果を次に列挙する：

100文において，生起した動詞あるいは動詞と同形な語が324個あった。それらについての前処理の判断結果には，非動詞が77，非述語が75，可能な述語が172があった。即ち，述語である可能性の持つ動詞の数が約半分（172/324 = 52%）に減らされた。これは，中国語の構文解析の効率と曖昧性減少にかなり重要な意義を持つと考えられる。例文

〝最常用的能排除干扰的方法是抗干扰编码的採用。〟

（〝一番常用する、妨害の除ける方法は誤り訂正符号の採用である。〟）

前 一 個 字母 是 T 的 可能 性 比較 大 。

(前のシンボルがTである可能性は比較的大きい。)

(前のシンボルがTの可能性であることは、比較的大きい。)

(前のシンボルは、Tの可能性が比較的大きいと言うことである。)

図4. "是字句"の解析における曖昧性の例

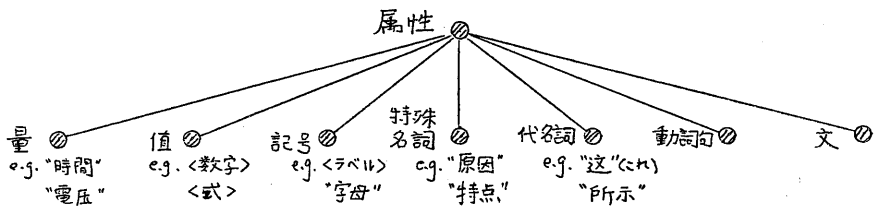


図5. "是字句"意味解析に使われる属性分類

B \ A	1 量	2 値	3 記号	4 特殊語	5 代名詞	6 動詞句	7 文
1 量	S_{AA}	?	S_3	?	S_5	×	×
2 値	S_1	S_{AA}	?	?	S_5	×	×
3 記号	S_3	?	S_{AA}	S_3	S_5	×	×
4 特殊語	?	?	S_3	S_{AA}	S_5	S_4	S_4
5 代名詞	S_5	S_5	S_5	S_5	S_{AA}	?	?
6 動詞句	×	×	×	S_4	?	S_{AA}	×
7 文	×	×	×	S_4	?	×	S_{AA}

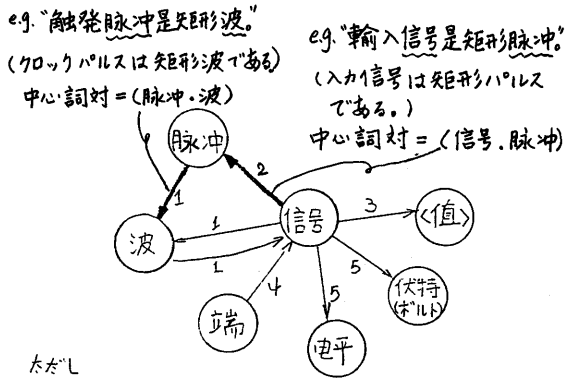
[説明] ?: 属性から"A是B"の意味上の整合性を判断できない.

S_{AA} : "A是A"のスコア

S_i : AもBも番目の属性である場合の"A是B"のスコア, $i=1,2,\cdots,7$

× : 属性からみると,"A是B"は意味的に矛盾であることを示す.

図6. "A是B"の意味的(属性からみる)な対応関係



ただし
○: 詞 ; 数字: スコア
Ⓐ→Ⓑ: A是B ; ～: 中心詞

図7. "A是B"の用例ネットの一部

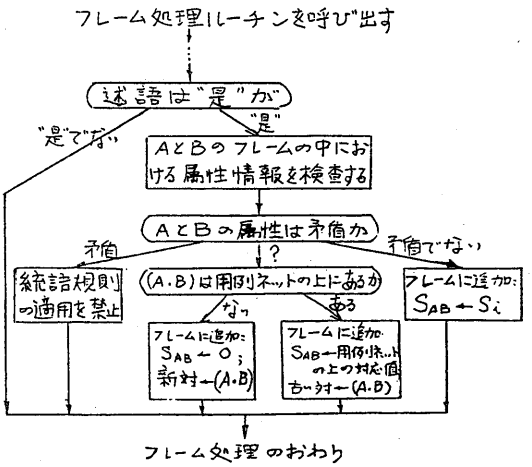


図8. "A是B"の意味検査とスコア設定のアルゴリズム

の解析には，前処理をしない方式において例文全体の解析のステップ数は，前処理をした方の1.5倍であり、その中に述語部分の解析（主語部分の解析は同様であったが）において，前者は後者の5～6倍になった。

3．"是字句"の意味解析

(1)"是字句"と意味検査：

"是"は特殊動詞であり，英語の"be"、日本語の"だ"のようなものである。"是字句"は，"是"を述語とする文である。

図4に"是字句"の統語解析に導かれる曖昧性の例を示した。それらを除くために，意味検査は必要である。

本論文では，文"A是B。"（AやBは名詞句、動詞句、文のいずれかとする）において，AとBの意味上の整合性の検査について検討する。具体的に，ディジタル回路の教科書（中国の）の一部について考察した。即ち，制限された範囲で，有限な知識を用いて意味検査を実現することを試みた。

(2)知識の利用：比較的簡単な知識を大量的に利用する方針である。属性（図5）の間の対応関係（図6）と、具体的な用例の二種類の知識を利用する。用例について，処理範囲内の典型用例と、解析されたテキストの前の部分における用例と、両方を集めて，それらの中心詞対を抽出して，"是字句"用例ネットを作る（図7）。ネットを参照して，テキストの後の部分に出現する文の正しさを判断する。

(3)意味検査の実現：AやBの意味情報をフレームで記述する。文"A是B"の意味検査は，統語規則"文→名詞句＋動詞句"（など）を起動するときに，フレーム処理によって行う（図8）。結果として，その"是字句"のフレームにスコアと中心詞対（A.B）の情報を入れて，(4)の処理のために用意する。

(4)結果の選択と用例ネットの自動修正：入力文解析の最後のステップに行う。フレームにおける情報の検査によって，スコアの一番高い"是字句"を結果として選ぶ。結果は唯一であれば，それを正しい結果とみなして，用例ネットを修正する。修正では，①中心詞対（A.B）はすでに用例ネットにあれば，その対のアク（ネットの上に）に1を増す。②（A.B）は新しい対ならば，ネットに追加して，そのスコアを1とする（ここで用例ネットは相互リンクリストによって実現する）。用例ネットの修正によってシステムの簡単な学習機能を実現した。即ち，スコアの値が，用例の重複回数にしたがって増大する。

(5)実例についての考察：以上の知識の実用性を検討するために，中国のあるディジタル教科書において順番に出現した200個の"是字句"について考察した。その200文の200個の中心詞対は，図6で示した属性関係と、40対を持つ用例ネット（最初10対を設定したが，200文の解析の終わったとき40対まで増えた）で含まれる。それらの知識を利用して図4の例文を解析すると図9の結果が得られる。

"是字句"	中心詞対	スコア
1．…字母是T	（字母・〈記号〉）	S〈記号〉＞0
2．…字母是…可能性	（字母・性）	0
3．…字母是…比較大	（字母・〈文〉）	×

↓↓スコアの比較

1の方を正解とする

図9．意味検査によって"是字句"を解析する例

4．結論

以上，動詞の解析を特徴とする解析システムと"是字句"の意味解析二つの方面から，中国語解析にとって有効な解析方式を提案した。解析システムの実現方法と有効性を検討した。解析システムは現在研究室のLISPの上に作成中である。

[参考文献] 1. 楊：＂中国語解析システムに関する研究＂，京都大学修士論文，1982年．