自然言語処理 68－9
（1988. 9. 16）

中国語入力における文法解析技術の応用

俞 士汶
（北京大学）

野口 喜洋
（松下電器産業）

文および句の単位で入力パラメータから中国語への変換を行う中国語入力システムの研究開発を行った。単語ごとの句切りをユーザに入力してもらい、単語辞書を検索して得た複数の候補単語について、文法解析を応用した品詞・小分類の絞り込みを行うことにより、ユーザによる単語の選択なしで８５％以上の正解率を得ている。また、文法解析処理を高速化するため、適用する文法について二段階の枝刈りを行うことにより実用的な処理速度を実現した。本稿では入力方法の概要、文法公式の適用原理、文法枝刈りのアルゴリズムについて論述する。

Application of Grammar Parsing Technique in Chinese Language Input

Yu Shiwen
Institute of Computer Science
and Technology
Peking University
Beijing, China

Yoshihiro Noguchi
Matsushita Electric
Industrial Co.,Ltd.
1006 Kadoma, Osaka,
571 Japan

A method of inputting Chinese sentences or phrases based on words has been developed. A user inputs the retrieval features of every word from keyboard. The retrieval features of each word is followed by a space. To reduce the troubles in choosing one word out of others characterized by the same feature, grammar parsing technique is applied and good results have been achieved. The hitting accuracy is more than 85 percent. Besides, a way to speed up parsing has been found. We call it a branch-cutting to the grammatical trees. This article describes the outline of the method, the principle of applying grammatical formulas and the branch-cutting algorithm.

1．はじめに

日本語ワードプロセッサでは既に語句変換技術が広く採用されており，「ぺきんはちゅうごくのしゅとである」→「北京は中国の首都である」のような複数文節単位のかな漢字変換による日本語入力が一般的である．

中国語には文節の概念はなく，入力の単位としては漢字，単語，句，文などがありうるが，中国語ワードプロセッサでは漢字単位の入力が多く採用されており，一，二年前までは単語入力は漢字入力の飾りものにすぎないと考えられていた．それは当時利用できるハードウェアの性能が不十分でかつ高価だったこと等による．現在ではその状況も好転し，システムに数万語の単語辞書を搭載することができるようになったため，単語入力が取入れられるようになってきている．単語入力の方法や入力するパラメータは多種多様である．たとえば，単語と1対1対応するコードを入力することで単語入力を行う方法がある．しかしこの方法は漢字単位のコード入力の延長にすぎず，無数にある単語の入力方法としては，根本的な困難を持つと言える．

北京大学計算機科学研究所では，機械翻訳および自然言語処理の基礎研究の一環として，文および句の単位で入力パラメータから中国語への変換を行う中国語入力システムの研究開発を行った．これは松下電器産業からの研究開発委託に基づくものである．

中国語の構文上・意味上の基本単位は単語であるから，自然言語処理等のフロントエンドに用いるためには，中国語入力は単語を単位にすべきである[1]．また，単語単位で入力した場合，漢字単位の入力に比べ選択肢が少なく，目的の出力を得る確率，すなわち正解率が高い[2]．中国国家標準GB2312-80にある6763漢字中，声調（各漢字の発音表記に付随するイントネーション．一声，二声，三声，四声，軽声の別がある）を考慮しない場合，99.8%の漢字に同音異字が存在する．たとえば，拼音綴り（中国語のローマ字発音表記）が"JI"である漢字は115字にもなる．これに対し，同音異義単語の数はかなり少ない．本システム中の33671語の2漢字単語を例に挙げると，同音異義単語はわずか45%であり，同音異義単語数は一般に2～3単語しかない．3漢字単語，4漢字単語ではさらに少なくなる．したがって　音による単語入力は，漢字入力に比べて著しく候補数が減少するので有利である．

しかし，単語を1つ1つ単独で入力する方法では，ユーザが複数の候補の中から目的の単語を選択するという手順は避けられない．そこで，より大きな区切り，である句や文を単位として入力し，文法規則および意味情報に基づいて句中または文中の各単語候補についての絞り込みを行えば，同一のパラメータに対応する複数の候補から目的の単語を自動的に選択・確定することが可能になる．

ところで，中国語には日本語の「てにをは」のような助詞がないため，日本語に比べ単語区切りの自動化が困難である．本システムでは，文法解析過程に重点を置き，開発を1年の限られた期間内に遂行するため，後述する虚詞のような例外を除き，単語区切りの入力を前提にした入力方法を採用した．

2．文および句単位の変換に基づく中国語入力方法

システム中に単語辞書と文法辞書を搭載する．同一パラメータに対応する複数の候補から目的の単語を選択するという処理を，文法規則に照らして自動的に解

決することを目的にしている.

2．1　入力方法の概要

A．辞書

本システムの中国語単語辞書の語彙として4 万語程度を精選した．各単語には3 種類の入力パラメータ，品詞，品詞の小分類，および使用頻度を付与する．入力パラメータの選定は簡単で容易に暗記できることを主眼とした．現在採用しているのは，単語を構成する各文字の第一音素（有声母字は声母，無声母字はa，o，eをとる）と起筆，声母（子音）と韻母（母音），起筆と末筆の3 種類である．品詞は次の14種類からなる．

名詞n，動詞v，形容詞a，数詞m，助数詞q，副詞d，代名詞r，
前置詞p，接続詞c，助詞u，感嘆詞i，語気助詞o，接頭詞h，接尾詞t

各品詞をそれぞれ数種類の小分類に細分する．例えば名詞は次の小分類に細分される．

時間1，場所2，方位3，抽象4，専有5，集合6，普通7，
人間8，動物9，植物a，人名b，地名c，書名d，固定単語i

以上の小分類はある種の意味情報を担っていると考えることができる．

表1に，単語辞書のエントリの例を挙げる．

表1　単語辞書のエントリ

単語	パラメータ			品詞	小分類	頻度
	第1音＋起筆	声母＋韻母	起筆＋末筆			
得	d丿	dE	丿𠂇	u	1	1
狠	h丿	hEN	丿𠂆	d	1	0
会議	h人yi	hUI yI	人、讠𠂆	n	7	2
進行	j一x丿	jIN xING	一辶丿冂	v	7	2
老師	l一shi	lAO shI	一一丨丨	n	8	3
歷史	l一shi	lI shI	一丿丨乀	n	7	2
是	sh日	shI	日乀	v	2	0
順利	sh丿l丿	shUN lI	丿人丿刂	a	3	2
他们	t人m人	tA mEN	人口人冂	r	1	1

単語辞書中の声母，韻母，起筆，末筆はすべてASCII コード中の1 キャラクタで表現される．表1で名詞以外の品詞についての小分類の意味を挙げると，u1は結合助詞，d1は程度副詞，v7は状態転換動詞，v2は連係動詞，a3は行為状態形容詞，r1は人称代名詞である．

B．文法

本システムには使用頻度の高い中国語の句および文の構造を記述する文法を格

納している．これは文脈自由文法に属するもので，以下ＢＮＦで記述する．辞書中に含まれる意味情報も小分類として文法公式の構成要素として含まれるので，この文法も意味文法(Semantic grammar)としての特徴を備える[5].

例として，中国語で頻度の高い補語文型を表２に示す[3].

表２　中国語の補語文型

主語	動詞＋「得」	補語
会議	進行　得	很　順利
房間	収拾　得	整斉　干浄
她	打扮　得	相当　漂亮

この文型は以下の文法公式で記述できる.

<S5_BU> → <SUB1><PRED4>
<SUB1> → <r>|<n>
<PRED4> → <v><得><BU2>
<BU2> → <d1><a>|<a><a>

このうち大文字の<S5_BU>は根節点であり句あるいは文を表す．大文字<SUB1>,<PRED4>,<BU2>は非終端記号であり，文法成分を表す．小文字の<r>,<n>,<v>,<a>,<d1>等は終端記号であり，品詞および小分類である．「得」のような，強力な文法機能をもつ単語も終端記号になりうる.

C．入力および処理過程

声母と韻母からなる入力パラメータ

hUIyI jINxING dE hEN shUNlI

に対して漢字文字列

会議進行得很順利（会議は順調に進行した）

が出力されるまでの処理過程を以下で説明する.

ユーザは各単語のパラメータを単語区切りの空白とともに入力する．システムは単語辞書を検索し，各パラメータに対応する単語を検索する．表３に上記入力パラメータについての検索結果を示す.

表３　単語辞書の検索結果

パラメータ	単語	品詞	小分類	説　明
hUI yI	会意	v	3	心理活動動詞
	会議	n	7	普通名詞
	回憶	v	n	動名兼用
jIN xING	進行	v	7	状態転換動詞
	金星	n	7	普通名詞
	尽興	d	6	状態副詞
dE	的	u	1	結合助詞
	地	u	1	結合助詞
	得	u	1	結合助詞
hEN	很	d	1	程度副詞
	恨	v	3	心理活動動詞
shUN lI	順利	a	3	動作状態形容詞

同一パラメータに複数の同音異義単語が対応しているが，以下で述べる文法解析の過程で一単語が選択される．まず，文法辞書中の文法公式に基づきトップ・ダウンの探索で構文木を生成する[4]．葉に達したときに，その位置にくるべき単語の品詞および小分類が確定するので，表３の検索結果中からマッチングする単語を選択する．すべての葉が検索結果中の単語とマッチングすれば，文法に符合した木が完成する．図１に上記入力パラメータに対する構文木を示す．

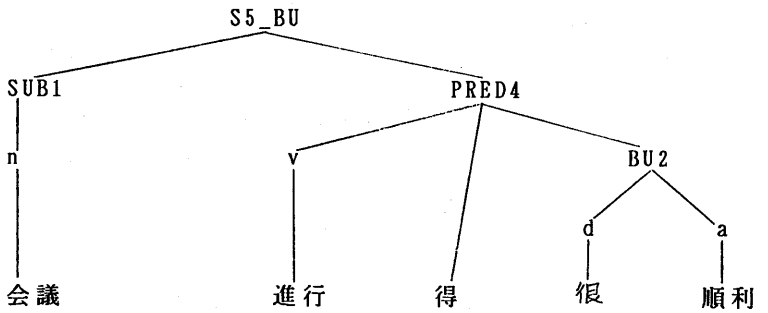

図１　検索結果の構文木

この解析の結果，目的の出力である「会議進行得很順利」という文が確定する．もし検索結果中の単語とマッチングする木がない場合は，表３の各パラメータ

に対応する単語から使用頻度の高いものが選択され，変換結果となる．

2．2　主な特徴

(1) 本入力方法は，単語のパラメータの選択，頻度順配列および文法解析を総合的に運用する事により，かなり満足できる効果を上げている．1単語の入力キー・ストローク数は2i+1ストロークで（iは単語を構成する漢字数で，1～4の値をとる），正解率は85%以上である．
(2) ユーザではなく，システムが文法解析を行うことにより同音異義単語の中から目的の単語を選択する．従来より高度な自動化の段階へ進んだと言える．
(3) 単語の前後関係のみではなく，入力パラメータ全体の文法構造を考慮する．従って，単語単位処理から句および文単位処理の方向へ一歩前進した．
(4) 入力パラメータの種類は文法解析の過程と独立に選択できる．文ごと，単語ごとに入力パラメータの切り替えができる．

3．文法解析高速化――枝刈り

3．1　文法公式設計の要点

(1) 中国語の文型は千変万化であり，厳格に語順を限定する文脈自由文法で完全に記述することは不可能である．そこで，システムの使用領域を想定し，そこで常用される句構造と文型の多くをカバーする文法公式を設計するしかない．筆者は使用領域として主にビジネスレターを想定した．また制限中国語の文法規則も目下の研究課題である[1]．
(2) 簡潔で適用範囲の広い文法公式を書くためには，非終端記号の使用が必要である．しかし文法解過程の効率を考えると文法公式のレベルは多すぎないことが望ましい．そこで公式の簡潔性と適用範囲への影響が少ない場合には，終端記号を使用した．
(3) 文法公式中に適宜意味情報を組み込み，正解率の向上を図った．例えば以下に連係動詞v2（「是」など）が構成する文の一般公式を示す．

<S3_SHI>→ <SUB1><v2><BIAO>
<SUB1> → <r>|<n>
<BIAO> → <n>|<vn>|<a>

声母と起筆を入力パラメータとして「他们是老師（彼らは先生です）」という文を入力しようとすると，変換結果は「他们是歴史（彼らは歴史です）」となる．これは「歴史」は「老師」の同音異義名詞で使用頻度が高いからである．そこで人称代名詞r1は連係動詞v2，人間名詞n8と組み合わされることに着目し，以下の文法公式を追加すれば，「他们是老師（彼らは先生である）」の変換結果を得ることができる．

<S3_SHI>→ <r1><v2><n8>

(4) 中国語の句と文は本来構造上の分岐を有するが，現在分析の対象とするのは漢字の順序ではなく，パラメータの構成する順序である．従って分岐点はさらに多くなる．文法規則に合うすべての木を探すのは時間の浪費になるだけではなく，取捨選択の決定根拠がない．現在の方法は合理的な文法木1つに絞る方法をとっているので，文法公式の配置順序は極めて重要である．専用性の高い，使用頻度の高い公式を前に配置すべきである．例えば上述したv2についての2つの公式は

次のように書き直すことができる.

<S3_SHI>→ <r1><v2><n8>|<SUB1><v2><BIAO>

本システムで使用している200 近い文法公式の配置順序は最も斟酌したところである.

3．2　第1次枝刈り――単語数による

文脈自由文法は容易にトップ・ダウンの探索方法により解析を行えるが，解析処理の効率は高くない．文法公式が200 前後に達すると解析の平均所要時間が1分を越えるため，解析の高速化の方法を工夫しなければならなかった．一つの素朴な方法として，探索の枝刈りを行い，マッチングが起こり得ない文法公式の探索をできるだけ避けるという方法をとった.

本システムでは，文法公式は入力パラメータに含まれる単語数によって分類されており，解析プログラムは他の単語数の文法公式を探索しない．これにより第1次枝刈りが実現する.

3．3　第2次枝刈り――単語の性質による

第1次枝刈りの後でも十分な解析速度が得られないので，更に枝刈り条件の強化が必要であった．設計者が根節点の文法公式を大まかに解析し，最初と最後の単語の持つべき性質（品詞，小分類），および句または文中に少なくともどのような性質の単語があるべきかを根節点の文法公式の適用条件として抽出する．そして入力過程では，単語辞書の検索結果（表3 参照）を適用条件と照合し，適用条件を満たす場合にのみその根節点からの解析を行うようにする．この第2次枝刈りによって本システムの文法解析が実用的な時間で終了するようになった．CPUが8088のマイクロコンピュータで数千の例文をテストした結果，ほとんどのものは1 秒以内に解析を終了する.

解析プログラムと文法公式とを独立にするため，上記の適用条件は陽に各根節点の文法公式中に記述する．前述の2 つの根節点の文法公式は以下の形式で記述できる.

<S3_SHI>→ <n, r; v2; n, a, vn>|<r1><v2><n8>|<SUB1><v2><BIAO>

<S5_BU> → <n, r; v; a>|<SUB1><PRED4>

公式の右辺先頭が適用条件である．適用条件はセミコロンによって3 つの部分に分かれるが，それぞれ最初，2～最後-1番目，最後の単語についての条件を表し，その位置にあるべき単語の品詞を示す．セミコロンはand の関係，コンマはorの関係に当たる．上記の文法について言えば，文法<S3_SHI>の適用条件は，最初にくる単語が名詞nまたは代名詞r，中間部に連係動詞v2が含まれており，最後にくる単語が名詞n，形容詞a，あるいは動名兼用vnであることである．また，文法<S5_BU> の適用条件は，最初にくる単語が名詞nまたは代名詞r，中間部に動詞vが含まれ，最後にくる単語が形容詞a であることである.

4．おわりに

(1) 本入力方法では単語の区切りはユーザの入力に任せるが，入力上の便宜を図るため，単語処理機能をわずかに強化した．何種類かの実詞（独立語）とその後

に続く「了」「着」「地」「過」等の虚詞（非独立語）の間には単語区切りの入力は必要ない．すなわち辞書には存在しない「美麗的（美しい）」「安排了（安排した）」「迅速地（迅速に）」等を単語と同様に入力できる．
(2) 品詞の小分類を文法公式中で利用することで，ある程度の意味的な処理を行っているが，十分とは言えない．文法的ではあるが，意味の通らない変換結果を得ることがある．
(3) 現在のシステムでは，文法解析が失敗した場合，それぞれの位置で最も頻度の高い単語を選択することで変換結果とする．将来は部分的な単語間の組合せ関係を考えて，もう少し候補を絞り込めるかもしれない．たとえば，あるパラメータがただ1つの助数詞と対応しているならば，助数詞の前の位置では数詞を，後の位置では名詞か形容詞を選択するのが合理的だろう．
(4) 本入力方法は単に中国語入力の手段としてだけでなく，自然言語理解および機械翻訳のフロントエンドとしての応用も考えられる．それは入力した単語の性質や文の構造についての情報を後の処理に提供しうるからである．

本入力方法は钟耀坤，朱学锋，筆者が共同で開発したものである．入力処理プログラムおよび文法解析プログラムは钟耀坤が設計，開発したものである．開発にあたり、松下電器産業のかたがたの良い意見を吸収した．ソフトウェアは1987年10月付で松下電器産業に譲渡され，高い評価を受けている．

参考文献

[1] 陳力為:当前中文信息処理中的幾個問題　計算機世界,1987年第21期,p.34
[2] 陳西園:漢字輸入要突破旧観念　中文信息,1983年3,4月,p.22～24
[3] 呂淑湘主編:現代漢語八百詞　商務院書館,1984年12月
[4] 馮志偉:数理語言学　知識出版社,1985年8月
[5] Phillip J.Hayes and Jaime G.Carbonell:A Tutorial on Techniques and Applications for Natural Language Processing,Carnegie-Mellon University, Oct. 1983