6C−6

中国語の未知語の推論

東京農工大学　工学部　電子情報工学科

鄭皎皎・謝建明・小谷善行・西村恕彦

この研究は中国語の単語の品詞学習について考えたものである。中国語の対話系や機械翻訳などに欠かせない辞書の中にまだ登録していない単語がある場合、その単語についてどう処理すればいいのかを研究し、中国語の未知語の品詞を学習するシステムを作成した。本システムでは単語の学習方法は類推による学習（Learning of Analogy）の方法を取る。つまり入力された文に未知語があるときは、ユーザに聞くことなしにシステムが自動的にその未知語の品詞を推論し、学習する。

１．はじめに

中国語の文では、一つの単語をとりあげても、その単語の品詞は一意的に決まらないことが多い。同じ単語でも、文の中の役割によって、その単語属する品詞も変わっていく。例として次の文があげられる。

（１）　這　是　我　的　花。
　　（これは私の花です。）
（２）　我　花　了　很　多　鈔　票。
　　（私はたくさんのお金を使った。）

同じ単語「花」であるが、（１）は名詞で、（２）は動詞である。

本システムでは、どのような未知の単語についても、品詞を学習することを目指す。ここで扱う対象は、入力文は単文で、読点が含まれていない文とする。

２．単語の分類

単語は文法上の性質や働きから品詞に分類される。本研究では［２］に従い品詞を１２種類に分けた。それらは、「名詞」、「動詞」、「量詞」、「数詞」、「指示詞」、「形容詞」、「介詞」、「認定助動詞」、「副詞」、「接続詞」、「助詞」及び「感嘆詞」である。

３．構文解析

この処理系は中国語の①基本的な統辞法及び②１２種類の品詞を連続基本構造［２］を用いて構文解析をする。

①統辞法とは連続基本構造を基本文に組み合わせルールである。本システムでは、基本文は叙述文、判断文と存現文の３種とする。

②連続基本構造とは、品詞の組み合わせる規則である。それらの規則は１２種類に分けられる。例えば、「数詞」と「量詞」の順序に並んだ品詞は「数詞構造」である。

両者を組み合わせて構文の規則を作成した。

４．未知語の品詞の修得方法

単語の連続基本構造と初期辞書を使って未知語の品詞を修得する。

４．１　初期辞書

このシステムは、中国語の未知語の品詞だけを学習する。辞書は初期の解析に必要な単語だけを持っている。単語は約１５０個が登録されている。それらは、すべての指示詞、人称代名詞、一部分の助詞、副詞、動詞、形容詞、などが含まれている。

４．２　本処理系の流れ

図１に示す。

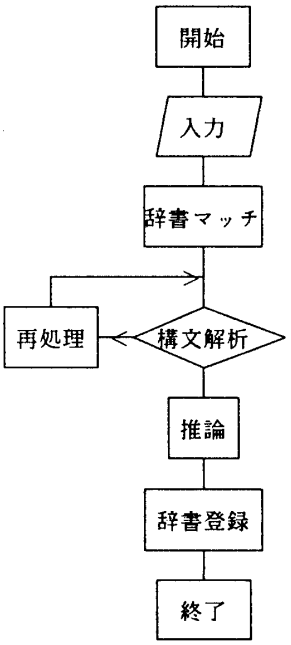

図１　システムの流れ図

①入力

入力はわかちがきによって入力する。

②辞書マッチ

入力された文中の各単語をそれぞれ辞書で参照させる。すなわち、各単語が未知語かどうかを調べる。

③構文解析

入力文をＤＣＧ(Definite Clause Grammars)の方法によって構文解析をする。

④推論

構文解析が成功すれば、構文の情報を利用して推論する。

⑤再処理

構文解析が失敗すれば、辞書で参照した単語が誤っていると判断する。その単語を未知語として、新たに構文解析をして、未知語を推論する。

Inference of Unidentified Words in Chinese
Kaukau Cheng, Kanmei Sya, Yoshiyuuki Kotani,
Hirohiko Nishimura.
Tokyo University of Agriculture and Technology

4．2　未知語の推論

未知語の推論は構文解析で行う。

入力文を構文解析し、未知語があれば、その未知語を含んでいるいずれかの連続基本構造と一致させる。成功すれば、その未知語の品詞を推論することができる。

例文「天気很好。」について考える。図2は例文の構文木である。この場合、文の主語の位置にきた単語が未知語であるので、その主語は名詞として修得することができる。副詞「很」は、未知語「好」と合わせて、限定構造と一致したので、「好」は「形容詞」と推論する。他の品詞推論も同じようにすでに辞書にある単語と合わせて連続基本構造と一致すれば、推論して修得する。

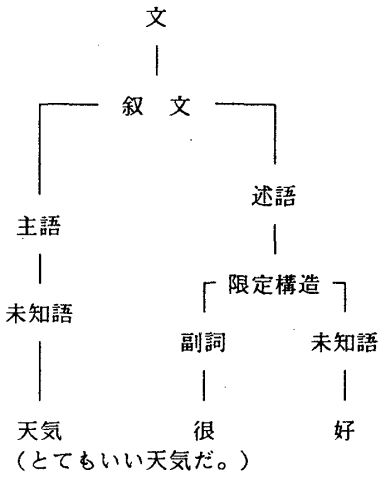

図2　構文解析の一例

5．多属性語の処理

中国語では一つの単語が複数個の属性（品詞）に属することが多い。例えば、「好」という単語の品詞は「形容詞」、「名詞」、「動詞」がある。

一旦、辞書に登録すれば、未知語でなくなる。辞書の単語の項目にその単語が推論によって学習されたものかどうかを判別できるように信頼性パラメータをつける。もし、辞書に登録された以外の品詞を発見し、その単語の信頼性パラメータが1より大きいなら、未知語として品詞を推論し、その品詞も辞書に登録する。

6．実験

現時点では、簡単な叙述文、判断文と存現文は処理することができる。従って、それらの文に含まれている未知語は推論できる。

例として次の文があげられる。

(1)叙述文：他　有　很　多　東西。
　（彼はとても多くの物を持っている。）
(2)叙述文：他　念　很　多　書。
　（彼は多くの本を読む。）
(3)叙述文：大家　都　知道。
　（みんなも知っている。）
(4)叙述文：他　準備　工作。
　（彼は仕事を準備する。）
(5)叙述文：他　準備　很　多　東西。
　（彼は多くのものを準備する。）
(6)判断文：他　是　先生。
　（彼は先生だ。）
(7)存現文：森林　里　有　很　高　的　樹。
　（森林の中にとても高いきがある。）

アンダーラインが引いている単語は未知語とする。

以上4つの文の解析結果は表1に示す。

表1をみると、簡単な文ほど推論正確率が高い。文(5)(7)はそれほど複雑な文ではないが、未知語の推論の正解回数は理想的でない。

表1 例文の推論結果

文	構文木の数	正解のもの
(1)	4	4
(2)	1	1
(3)	1	1,1
(4)	1	1
(5)	9	3
(6)	1	1
(7)	6	6,2

謝辞

本原稿執筆にあたり貴重なご意見を頂いた本学滝口伸雄助手、本学野瀬隆技官、本学のみなさんに感謝します。

参考文献

[1] Yoshiyuki KOTANI:A system of Practical Liguistic Knowledge Acquisition in Japanese Language, PRICAI'90, PP.643-648.
[2] 藤堂明保：中国語概論　大修館書店（1979）.
[3] 小谷善行：知識指向言語Prolog 技術評論社（1986）.
[4] 溝口文雄：prologとその応用2 総研出版（1985）.
[5] Ka-Wai Chui: An Information-Based Parsing Model for Resolving,PRICAI'90, p.274-279 (1990).
[6] 楊顒明、堂下修司、西田豊明：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、情報処理学会論文誌、Vol.25、No.6、pp.1044-1054（1984）.
[7] 松田晃一、高田正之、小谷善行：未知語の属性を修得する自然言語処理系、情報処理学会第31回（昭和60年後期）全国大会講演論文集（II）、pp.1201,1202.