構文解析に基づく辞書情報推定に関する研究

4A−2

朴 哲済　崔 卿楽　筧 捷彦
早稲田大学

1　はじめに

本稿では、依存文法による構文解析システムを利用して、自然言語の形態素及び文法属性を推定し、辞書に登録・管理する辞書管理システムについて述べる。現在の自然言語処理システムは言語に関する情報が辞書に存在することを前提として構築されているが、実際、膨大な言語に関する情報を持つ辞書を構築することは不可能に近いことである。そこで、我々は構文解析システムの実行結果を基に日本語の形態素及び文法属性を仮想辞書に登録し、解析システムが動くことによりその情報の精度を上げて自動的に辞書を拡張管理するシステム開発を行っている。本研究の目的は、辞書構築手法を模索する一つの試みとして解析段階で未知語が見つかった場合、推論機構を通して獲得された情報に基づいて、辞書情報としての利用可能性を探索することである。従って、ここで、文法属性というのは解析システムの環境[2]に違反しない範囲で再構成され、文の中で語の役割を重視する文法知識であることを前提とする。ここでは、その基本アルゴリズムと実験結果について報告する。

2　システム概要

本辞書情報推定システムでは、入力文の見出し語とその文法属性の推論を第一の課題と考え、推定の効率を高めるため、特定な品詞情報を完全情報として用いることにした。また、不完全情報を含む日本語解析システム[2]の実行過程で得られた情報の役割を推定する手法は、語と語との依存関係に基づくようにした。これを考慮し、特定な品詞情報は助詞を与えることにする。依存関係に関しては自然言語の構文解析技術が利用できる。言語に関する知識は、構文解析から直接的に収得できることにする。図1にシステムの構成を示す。

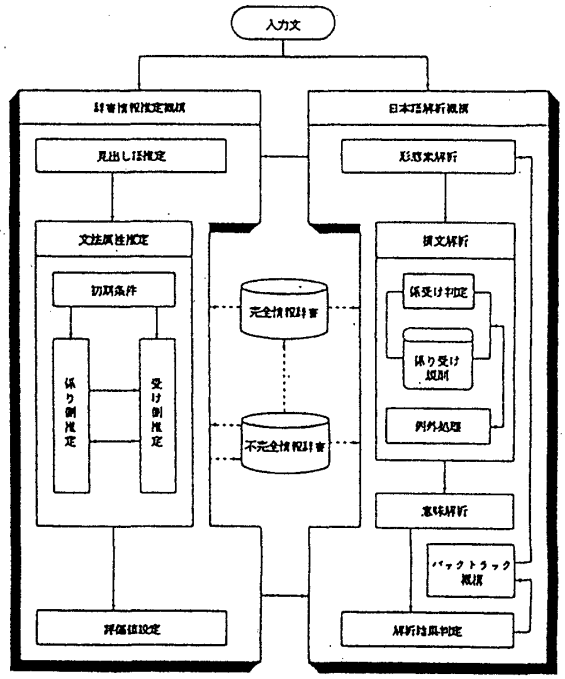

図1　システム構成

3　辞書情報推定手法

辞書情報推定のアプローチ方法は、文に含まれる単語間の係り受け関係を、優先順位別に分析することにある。その結果は仮想辞書に登録され、解析システムで用いられる。ここでは、その辞書の見出し語及び文法属性推定方法について述べる。

3.1　見出し語推定

見出し語推定は、字種切り法をヒューリスティックスとして右方向最長一致法で決定する。字種には、漢字、片仮名、平仮名、アルファベット、数字、区切り符号を用いる。まず、入力文字列に対し

A Study on Dictionary Information Inference Based on Parsing
Chul-Jae Park , Kyong-Rak Choi , Katsuhiko Kakehi
Waseda University

て未処理文字列の左端から字種が一最初推定の場合は1回、解析システムの実行結果を基に推定する場合は4回一変化する部分まで切り出す。そして、切り出された文字列について、辞書引きを行いながら最長一致法で推定する。見出し語推定アルゴリズムは文献[2]に詳しい。

3.2 文法属性推定

文章の構成要素を W_i とし、P_i を品詞としよう。要素 W_i と対応する品詞 P_i との関係を $(W_i\ \ P_i)$ とすると要素 W_i を中心とする品詞の推定は次のように考えられる。

$T_f = Cond(W_i\ \ P_{fixed})\ ([W_{i+1}] * C_{i+1}[P_j])$

$T_b = Cond(W_i\ \ P_{fixed})\ ([W_{i-1}] * C_{i-1}[P_j])$

T_f とは $(W_i\ \ P_{fixed})$ の条件下で次に受ける要素、W_{i+1} の品詞を推定することであり、$C_{i+1}[P_j]$ とは要素 W_{i+1} の対応可能な品詞の種類である。同様に T_b とは $(W_i\ \ P_{fixed})$ の条件下で前に係る可能性がある品詞の種類である。（ただし、P_{fixed} は与えられた既知の品詞であり、ここではそれを助詞にする。）

このような品詞の推定方法は盲目的探索よりは効率的であると予想されるが、その理論的基礎と可能性については文の文法的構造と品詞の分類、及び、その係り受けの関係から導かれると期待される。しかし、システム構築という立場から3つの原則、すなわち、データの抽象化、停止性の保証、計算量の現実性を満足しなければならない。

(1) 文法属性推定の最悪計算量

完全情報に対する最短一致法と最長一致法の最悪計算量は、$O\ (2^n)$ として知られてる。本手法の計算量 $f(n)$ を計算すると以下のようになる。

$$f\ (n) = (N\ (P_j))^{n}$$
$$= O\ [\ (N\ (P_j))^{n}]$$

(n：文構成要素の数、P_j：ある品詞の次に依存する品詞の数、$N\ (P_j)$：P_j の最大数)

上記のような爆発的計算量を防ぐため、評価値計算によって依存関係が一番強い品詞（最優先度を持つもの）を優先的に次回の係り受け関係の品詞と対応するようにすれば、計算量 $g(n)$ は、

$$g\ (n) = N\ (P_j)$$
$$= O\ (n)$$

4 実験結果及び考察

以上のような考えに従って、実際の文から、係り受け構造のリストを出力する日本語解析システムと辞書情報推定システムの実験的な試作によって、見出し語及び文法属性を推定する辞書管理システムを評価する。システムはＫＣＬ（Kyoto Common Lisp）を用いて書かれており、Sun4(Sparc)上でインプリメントされている。評価例文は、情報分野の本から抜き出した100文（約2500語）であり、完全情報として辞書は、最初88個の助詞のみを有する[1]。

実験は同じ文に対して、以下の場合を分析する。

- 助詞のみの完全情報を用いた推定による、見出し語及び文法属性推論の精度
- 推定能力と完全情報数の相関関係
- 解析回数と辞書信頼度の相関関係

5 おわりに

本稿では、自然言語処理システムにおいて辞書構築や言語知識の獲得問題に対して、循環機能と推論機構を用いることによって日本語の見出し語及び文法属性を推定する手法について述べた。文法属性推定機構では、依存関係規則の優先度を使うことにして、推定結果を入力文の係り受け構造を求める解析システムにフィードバックすることにより、辞書データの精度を向上させると共にその文の解析が可能になる。現在は、その評価を行なっている段階で、今後、解析システムと辞書管理システムのインターフェース項目、及び処理回数と実行速度を高める方法の検討を続けて行く予定である。

参考文献

[1] 日本電子化辞書研究所：”TR-018 日本語単語辞書”、日本電子化辞書研究所（1990）.

[2] 朴 哲済、崔 卿楽、筧 捷彦：”不完全情報を含む日本語解析システムについて”、情報処理学会第46回全国大会（1993）.