不完全情報を含む日本語解析システムについて

1B−5

朴 哲済　崔 卿楽　筧 捷彦
早稲田大学

1　はじめに

自然言語処理システムにおいて、言語に関する情報の完全性は、そのシステムの性能に大きな影響を与えるが、膨大な言語に関する情報を系統的に作成することは困難な問題である。未知語処理や推論の研究分野で、既知の問題を未知の領域に適用する方法や解析の過程で最もらしい解を導出する方法が提案されているが、分野の限定や速度の向上が必要とされている。

本稿では、依存文法による日本語解析システムにおいて、不完全な辞書情報に対して推定機能を用いた日本語解析システムについて述べる。解析システムの文法は基本的に依存文法を採用しており、各単語の意味素性の決定により意味処理を行っている。辞書は、辞書情報推定システム[2]によって作られた不完全情報を持つ仮想辞書と完全な形態素及び文法情報が登録されている完全情報辞書と二つに大別される。未知語を含んでない文については、解析結果、その係受けの木構造を出力するし、完全情報辞書にない単語については仮想辞書を検索し、その言語情報の形態素及び文法情報推定を行い、可能な解析結果を求める日本語解析システム開発を行っている。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

システムは、文に含まれる単語間の接続可能性を記述した依存関係規則をもち、必要な品詞間の関連関係をそこから取り出しながら、入力文の係受け構造の判定を行う。次に不完全情報から評価関数を使用して、評価値を計算する。本解析システムでは、形態素候補が存在するか否かにかかわらず、常に、不完全情報の可能性を考慮し、構文解析システムと辞書情報推定システム間の循環機能を導入する。図1にシステムの構成を示す。

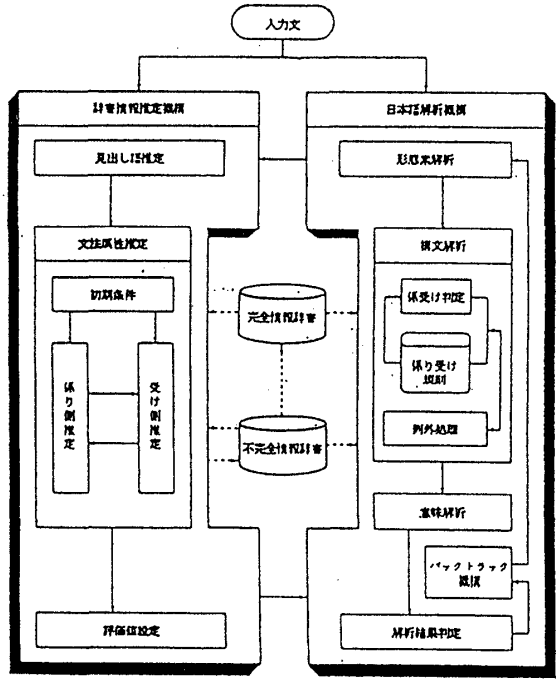

図1　システム構成

3　解析手法

日本語解析のアプローチ方法は、単語間の係受け関係を分析するところで得られた形態素の評価値を基に、優先的木構造をスタックにおいて最適木を求める。その結果は辞書情報推定システムに渡され、不完全情報の最適解を求める。入力文に対して、単語の信頼度が1になるまで繰り返される。

3.1　形態素解析

形態素解析は、入力文字列に対して、字種切り法と最長一致法により完全情報辞書引きを行ない、切

A System of Japanese Sentence Analysis Including Incomplete Information
Chul-Jae Park , Kyong-Rak Choi , Katsuhiko Kakehi
Waseda University

り出された文字列を形態素として認定する。失敗した場合には、不完全情報辞書引きを行ないながら、各形態素候補の持つ評価値を計算する。評価は、辞書情報推定システムや解析システムの各過程からなる複数段階の結果に対応した単語の評価値を総和している。

```
Algorithm　形態素解析
  do （入力文字列 = null）
    未処理文字列の切り出し
    do （解析成功 | 推定終了）
      完全情報辞書引き処理
      if （完全情報が見つかったら）　then
        切り出された文字列を形態素として認定
      else do （推定終了）
        不完全情報辞書引き処理
        if （不完全情報が見つかったら）　then
          信頼度計算処理
          評価値をセットし、形態素として認定
        else if （処理文字列 = null & 形態素あり）　then
            解析をやり直す
          else
            最低評価値をセットし、見出し語として推定
        end-do
    end-do
  end-do
end　形態素解析
```

3.2 構文解析

構文解析はボトムアップ縦型法とし、解析過程でバックトラックが生じる場合でもそれまでの計算結果を他のファイルに保存する。解析過程で得られた情報と最終的なスタックの状態は辞書情報推定システムの入力として返す。

(1) 接続可能性の判定

入力文 S の i 番目の単語に対して、修飾関係を持つ1個以上の単語との接続可能性を依存関係規則表ー語と語の接続関係を考えて、そこに優先度を与えたものーにより判定する。失敗すれば、新たに一語取り出し同様の操作を繰り返す。特に、i 番目の単語と修飾関係がある単語が一つのみ存在する場合は、それを最適解としその単語の評価値を1にする。

(2) バックトラック制御

スタックに格納されている多義・多品詞語入力に関する情報を基に、バックトラック制御を行なう。しかし、情報の不完全性から起こる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させる。

4 実験結果及び考察

以上のような考えに従って、入力文から係受け構造のリストを出力する日本語解析システムの実験的な試作によって、不完全情報を含んだ解析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 [2]。実験は同じ文に対して、以下の3つの場合を分析する。

- 助詞のみの完全情報を用いた解析による、未知語に対する推論の精度
- 解析能力と完全情報数の相関関係
- 形態素解析アルゴリズムの字種変化と完全情報数の相関関係

5 おわりに

本稿では、自然言語処理システムにおいて未知語や情報の不完全性の問題に対して、循環機能と語の信頼に関する評価値を用いることによって、日本語文を解析する手法について述べた。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情報推定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] 田中穂積、辻井　潤一：”自然言語理解”、オーム社 (1988).

[2] 朴 哲済、崔 卿楽、筧 捷彦：”構文解析に基づく辞書情報推定に関する研究”、情報処理学会第46回全国大会 (1993).