1E－4

日本語形態素解析システムにおける部分的再試行機構の導入とその効果

尾嶋 基　宮崎 正弘

新潟大学大学院自然科学研究科

1 はじめに

日本語形態素解析において発生する種々の曖昧性を解析精度を低下させずに、可能な限り抑止することは、高精度で効率的な解析を実現する上で重要である。日本語形態素解析用の辞書には使用頻度の低い語を含め多くの語が収録されている。本稿では単語の使用頻度には大きな偏りがあることに着目し、使用頻度の低い表記の語や固有名詞などを用いず形態素解析を行ない、解析に失敗した場合等に必要に応じて、このような語を用いて解析を部分的にやり直す再試行機構を導入した形態素解析を提案し、その有効性を論じる。

2 日本語形態素解析システム Maja

本稿で使用している日本語形態素解析システム**Maja**[1]（以下、**Maja**）は、選択的辞書引き機構を用いて解析を行っている。そのため、入力文に対して一般語辞書、固有名詞辞書、数詞辞書を必要に応じて検索し、入力文字列に含まれる単語候補を抽出し、それを CYK 表に格納する。次に接続辞書 [1] を参照し、拡張 CYK 法 [2] による解析を行う。文末まで解析が成功しなかった場合には、固有名詞解析、未知語解析を行い、文末まで解析を成功させる。未知語解析では部分的再試行機構により未知語の抽出と品詞推定を行う。最後に複合名詞解析 [3] を行ない結果を出力する。同形語や単語分割の曖昧性はコスト最小法 [4] により絞り込む。

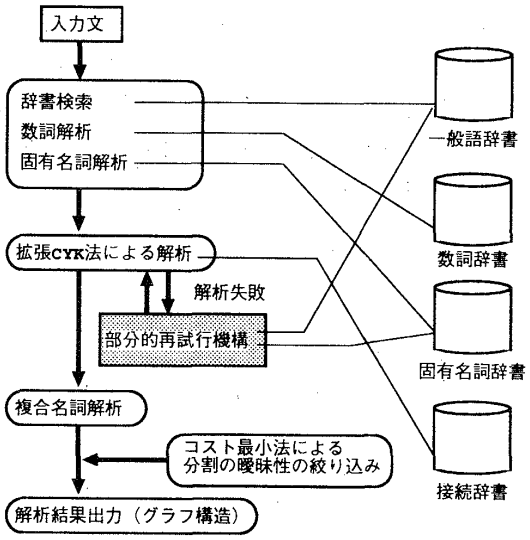

図 1: 日本語形態素解析システム Maja

3 部分的再試行機構

Majaでは解析が文末まで成功しなかった場合、すなわち形態素の接続が途中で失敗した場合に再度別の方法で辞書検索を行い、解析を行う部分的再試行機構が採り入れられている。

3.1 一般語辞書

Majaで使われている一般語辞書には、形態素解析における単語分割の曖昧性を抑止するために、平仮名表記の語に関しては、よく使われる平仮名表記の語のみ辞書に収録してある。そのため、表記のゆれに弱く、未知語が発生する場合がある。解析の精度を上げるためには未知語の発生は極力避けるべきであるが、発生してしまった後でも以下の方法を用いることで回避できる。

Partial Retrial Mechanism in Japanese Morphological Analysis
Hajime Ojima, Masahiro Miyazaki
Niigata University

[1] Morphological Analysis System of Japanese

3.2 平仮名未知語処理

普通に解析を行うと例えば、「たべる」「ほん」のように普通、漢字表記で表す単語を平仮名で書いた場合に未知語となりやすい。また、「さ末」「無理やり」「思いで」等のように漢字・平仮名の混ぜ書きによる表記も解析が失敗し、平仮名部分が未知語として抽出される。特に新聞や雑誌の記事はほぼ決まった形で混ぜ書きが出現するため、これらへの対処が必要である。本システムでは平仮名未知語の「表記」と辞書中の「読み」の形が同じことに着目し、読み引きを用いて再度辞書検索を行うことでCYK表に格納し、解析を最後まで成功させる方法を用いている。

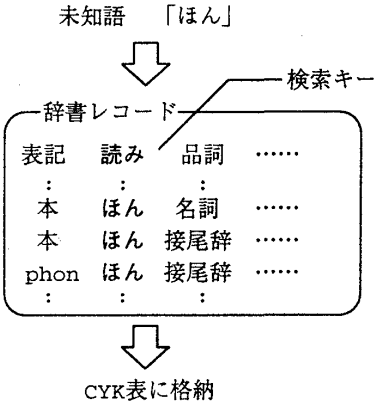

図 2: 読み引きによる辞書検索

3.3 混ぜ書きの対処

混ぜ書きの部分が未知語となってしまう場合、3.2で述べたような読み引きをそのまま使用することができない。混ぜ書きは「KH（前方漢字–後方平仮名型」「HK（前方平仮名–後方漢字）型」、そして送りがなの問題を含む「KHK（複合）型」に大きく分けられる。

- KH型
  無理やり [遣り]、思いで [出]、など
- HK型
  は [把] 握、さ [瑣] 末、きょう [筒] 体など
- KHK型
  申（し）込み、扱（か）い、呼（び）出し、など

KH型、HK型の混ぜ書きに対しては、前方あるいは後方の漢字部分を含めた単語を検索する。例をあげると「思いで」の場合、「いで」だけでは未知語解析に失敗するので、前方の漢字「思」で始まる単語を辞書から検索する。そのときにその単語の「読み」の部分で「いで」とマッチするものを探し出す。

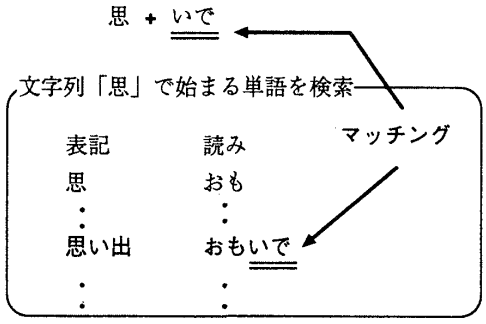

図 3: 「思＋いで」の品詞推定

4 おわりに

新聞の社説575文を対象とした実験によれば、現在の**Maja**では文単位で2.3％（13文）が未知語の原因により解析が失敗してしまう。しかし、部分的再試行機構を用いることでこれを改善できると考えられる。今後は定量的評価を行なう必要がある。

謝辞

形態素辞書の原データ、および単語意味属性体系データの使用を許可されたNTTコミュニケーション科学研究所の関係各位に深謝する。

参考文献

[1] 宮崎、白井、池原：言語過程説に基づく日本語品詞の体系化とその効用、自然言語処理、Vol.2、No.3、pp.3-25(1995)

[2] 高橋、佐野、宍倉、前川、宮崎：頑健性を目指した日本語形態素解析システムの試作、「自然言語処理における実動」シンポジウム論文集、pp.1-8(1993)

[3] 太田、前川、宮崎：規則用例融合型の日本語複合名詞構造解析法、言語処理学会第3回年次大会、pp.313-316(1997)

[4] 長尾 真：岩波講座ソフトウェア科学15、自然言語処理、岩波書店、pp126-129(1996)