4P−7　汎用日本語形態素解析ツールの開発

西森裕司（日立西部ソフトウェア）
木山忠博　絹川博之（日立製作所　システム開発研究所）

1．はじめに

日本語形態素解析をツール化し、自然言語処理の共通基盤とすることで、様々な自然言語処理アプリケーションの構築を容易にすることができる。このような日本語形態素解析ツールの具備すべき要件について述べる。併せて、報告者らの作成した汎用日本語形態素解析ツールのプロトタイプについて述べる。

2．汎用日本語形態素解析ツールの具備要件

ツールとしての日本語形態素解析においては、以下の要件が重要である。まず表現の自由度(flexibility)に関しては、(1)多様な表現・文型・文種が受理可能であること、(2)受理可能な語彙が多いこと、である。次に多種分野への適応性(transportability)に関しては、(3)アプリケーション（以下ＡＰと略す）での利用が容易であること、(4)処理速度や辞書容量が実用的であること、である。以下、これらの具備要件とプロトタイプでの実装について述べる。

2.1 受理可能な表現・文型・文種

汎用日本語形態素解析は、広範囲の漢字仮名混じり文が受理可能でなければならない。これには、完全な文章になっていないもの、断片的な言い回し、文法から逸脱した文、などが含まれる。また、書き言葉と話し言葉の両方が受理可能であることが望ましい。

どのような品詞が定義されているかということは、日本語形態素解析の精度を規定する。具体的な例としては、「健康だ」を形容動詞と考える立場と、〈名詞〉＋「だ」と考える立場があることなどである[1)]。さらに、日本語形態素解析の品詞の定義は、自然言語処理ＡＰの能力も規定することになる。

日本語形態素解析における共通の問題として、「拡張された付属語」がある。「について」のような語は、文法上の定義からは付属語ではないが、意味的には付属語として扱ってよい。さらに、「ことがない」は、

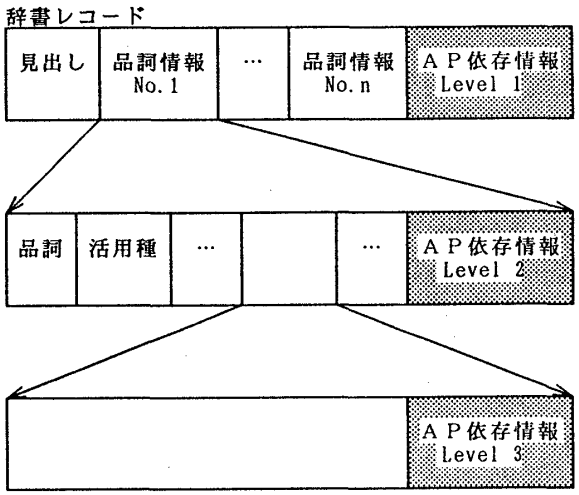

図１．辞書レコードへのＡＰ依存情報格納例

〈形式名詞〉「こと」＋〈格助詞〉「が」＋〈形容詞〉「ない」と分割されるが、全体で形容詞的な接尾語として扱うことも可能である。多くの自然言語処理では、この「拡張された付属語」を利用して処理の効率を上げている。また、自立語においては、「（１）」や「５００万」をどのように分割するかが問題となる。

本プロトタイプでは、原則として最も細かい分割を出力とし、それらの組合せの解釈はＡＰに任せるという立場をとることにした。すなわち、「ことがない」は、「こと」＋「が」＋「ない」に分割される。但し、漢数字については例外とし、前述の「５００万」の「万」を数詞接尾語とするのではなく、〈漢数字〉「５００万」にまとめた。

2.2 語彙

日本語形態素解析が汎用的に利用できるためには、大規模な辞書が必要である。多様な文章を解析可能にするために、またシステムを評価する際に信頼できるデータを得るために、大語彙辞書が必要である。汎用日本語形態素解析では、５万語の語彙を持つ基本語辞書を提供する。

また、全ての語を辞書登録することは不可能であることから、辞書に登録されていない語、すなわち未知語を適切に処理できることも重要である。未知語の範

Toolkit of General Purpose Japanese Morphological Analysis

Hiroshi NISHIMORI (Hitachi Seibu Software)
Tadahiro KIYAMA, Hiroshi KINUKAWA (Hitachi)

囲を正確に決定し、未知語の存在によって他の語を誤解析しないようにする必要がある。

2.3 APからの利用

テキスト読み上げやDB検索用自然語インタフェースといった様々な自然言語処理APで使用可能であるために、辞書に格納する情報を区別する必要がある。すなわち、AP非依存の情報とAPに固有の情報とを区別する必要がある。AP非依存の情報とは、日本語形態素解析に必要である品詞に代表される文法情報などである。APに依存する情報の例として、単語の読みやアクセント、AP固有の対象世界を表す意味情報がある。APではこれら情報のすべてをアクセス可能である。

図1に辞書レコードへのAP依存情報の格納例を示す。この辞書レコードは、1つの見出しに対して複数の品詞情報が存在するといった階層構造を持ち、それぞれの階層ごとにAP依存情報を格納できる。階層ごとにAP依存情報のフィールドを設けることで、効率的にAP依存情報を格納することが可能である。AP依存情報は、データ長を属性としてもつバイナリデータとして扱う。具体的なデータの格納形式はAPの設計者が自由に決定することができる。また、AP依存情報の全てのフィールドを使用する必要はない。すなわち、特にAPに依存する情報が必要でない場合は、まったくAP依存情報を付加していない辞書を使用すればよい。

AP固有の情報については、(1)情報へのアクセスの容易性、(2)情報の付加の容易性、の2点が重要である。(1)については、アクセス関数の雛形を用意することで、(2)については、辞書エディタを用意することで対処している。

2.4 速度と容量

汎用日本語形態素解析はAPの一部として使用されるため、速度的な性能も要求される。作成したプロトタイプを、DB検索用自然語インタフェースに用いる150の例文を用いて評価した結果、1文あたりの平均処理時間は0.9秒であった。この例文は、1文平均18文字であり、平均11単語に分割された。処理は全てソフトウェアで実行し、接続条件のチェック以外にルールによる品詞決定処理も行っている。0.9秒という処理時間が実用的であるかは、一概には結論付けられない。APの応答時間に関してしばしば引用される「2秒の限界」(Miller, 1968)[2]が多くの作業に対して適当であることから、APがユーザとの応答の間に1回だけ日本語形態素解析を行うのであれば、ほぼ満足できる速度である。なお、文が長い場合に加えて、未知語を含む場合に解析時間は長くなる。

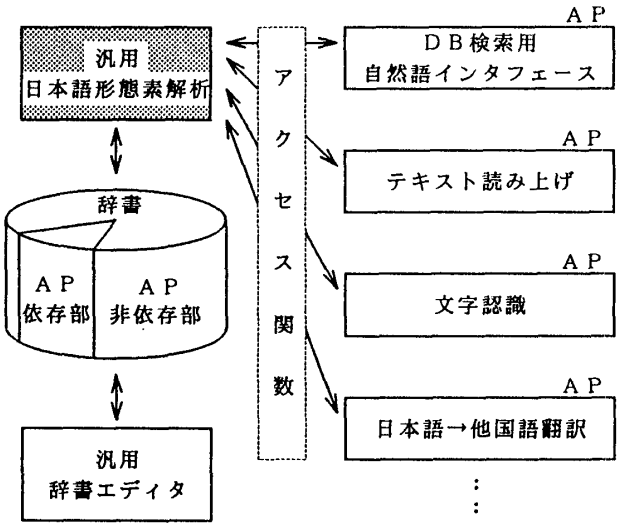

図2．汎用日本語形態素解析の位置付け

容量に関しては、辞書ファイルが約9Mバイト、文法ファイルが約3Mバイトである。なお、この辞書には、APに依存するデータは付加していない。

現在、汎用日本語形態素解析はライブラリ形式でAPと結合している。処理や辞書管理の一元化のために、汎用日本語形態素解析サーバとすることを検討している。

3．まとめ

汎用日本語形態素解析ツールの具備すべき要件について検討した。この要件を満たすべく、日立の日英機械翻訳システム[3]用の形態素解析をベースにプロトタイプを作成し、それを評価した。この結果、汎用日本語形態素解析が、既存あるいは新規の漢字仮名混じり日本語の解析を必要とするシステムの構築を簡易にし、大語彙辞書を容易に使用可能にできるという見通しを得た。現在、話し言葉の解析精度を上げるべく検討している。

汎用日本語形態素解析の位置付けを図2に示す。汎用日本語形態素解析の具体的な適用先としては、DB検索用自然語インタフェースやテキスト読み上げなどがあり、文字認識や日本語から他国語への翻訳などを適用先として検討している。

参考文献

1)宮地裕ほか:日本語6文法I,岩波書店(1976)

2)Shneiderman, Ben:Designig the User Interface, Addison-Wesley(1987)

3)梶博行ほか:日立における機械翻訳システム,情報処理 Vol.26, No.10, pp.1214-1216(1985)