ロバストな日本語形態素解析
－辞書依存性の低いハイブリッドアルゴリズムの提案－

1C－4

Patrick Halstead
Microsoft Corporation

奥村 薫
マイクロソフト(株)

1. 始めに

膠着言語である日本語は単語境界が自明でないため、さまざまなアプリケーションで文章を取り扱う際に、「単語」単位の操作が英語ほどには簡単でなかった。

単語の選択・置換・検索といった基本的な操作を可能にし、また更に深い自然言語解析の第一歩とするために、われわれは極めてコンパクトな形態素解析コンポーネント、愛称“T-Hammer”を研究開発した。

本稿では、共通コンポーネントとしての T-Hammer の概要、精度、および今後の課題について考察する。

現在 T - Hammer は Word 97, Encarta, Bookshelf, IIS などの製品に実装されている。

2. T-Hammer の概略

この形態素解析の特徴は、統計情報および文法情報を組み合わせて用いることにより、コンパクトでかつ十分に精度の高いコンポーネントを実現したことである。

統計情報：

文字や文字の連鎖が単語境界となる確率を用いる。例えば、「、。・」や空白、カタカナ語など、ほぼ確実に単語境界となるものもある。また、仮名から漢字への移行は単語境界となりやすい。

各文字および文字のバイグラムごとの統計情報を用いて、暫定的な単語境界を得る。

文法情報：

形態素間の推移規則とそのコストを用いて、単語の切り出しおよび品詞の同定を行なう。特に付属語部分に対しては、細かな解析が必要となる。

辞書情報：

特筆すべきこととして、T-Hammer は単語を網羅する大規模な辞書は使っていない。

通常の辞書では大部分を占める自立語については、前記の統計情報を主に利用して単語の切り出しを行なう。その精度をさらに向上させるために、限られた単語を登録した辞書を併用する。

アルゴリズムの特徴：

多くの形態素解析プログラムは網羅的な単語辞書を持ち、辞書引きからスタートする。辞書の充実度に解析精度が依存し、特に未知語があった場合には、そこで望ましくない単語切り出しが行われることも多い。

（1）左向き解析

文を切り出した後、T-Hammer は文末から文頭方向に向かって解析をする。すなわち付属語部分の解析が自立語部分の解析に先立つことになる。通常付属語のほうが可能性が限られているために、サーチ・スペースの効率化が期待できる。

（2）後段での辞書引き

我々のアルゴリズムでは辞書引きは、かなり後段に位置する。統計情報を併用するために、単語が辞書に網羅されていることを前提としないアルゴリズムとなっている。そのため未知語や入力ミスがあった場合にも、比較的安定した単語切り出しが得られる。また辞書サイズはかなり小さくてもよい。

3. 速度およびサイズ

T-Hammer の速度は次の通りである。

測定環境：100MHz Pentium、32MB メモリ
Windows95 および Windows NT4.0

A Hybrid and Robust Algorithm for Segmenting Japanese
Patrick Halstead, Microsoft Corporation: PatHal@microsoft.com
Kaoru Okumura, Microsoft Co., Ltd.: KaoruO@microsoft.com

速度： 約2,500文字／秒

コンポーネントのサイズ：
約500KB

これは必要な辞書および文法情報すべてを含んだサイズであり、非常に小さいと言えよう。現在すべてが、ひとつのDLLに含まれており、外部辞書などは必要としていない。

4. 精度

人手による単語境界が付されたテストコーパスをもとに行なった精度測定は次の通り。

	Coverage	*Precision*
総形態素	99.21%	98.88%
自立語	96.71%	96.42%

Under Break：コーパス上にある単語境界で、解析結果に単語境界がないケース。
Over Break：コーパス上では単語境界がない所を、解析結果が単語境界としたケース。

$$Coverage = 1 - \frac{Under\ Break\ 数}{総境界数}$$

$$Precision = 1 - \frac{Over\ Break\ 数}{総境界数}$$

検索のためには、Coverageの方が重要な指標となる。たとえPrecisionエラーが起こって1つの単語を2つに分割したとしても、その単語を検索することは出来る。しかしUnder Breakで単語境界がない場合には、存在する単語を検索できなくなることも多いからである。

5. T-Hammerの応用例

5.1. 文章内の単語選択(Smart Selection)

文書作成時に、単語を書き換える・削除するなどの操作を行ないたいことは多いだろう。Word 97(ワードプロセッサー製品)では、ダブルクリックした際にT-Hammerを用いて単語を選択する機能を持つ。漢字列・仮名列など同一字種の文字列を選択していた従来に比べて、自然な範囲が選択される。
これはさらにBookshelf (英和・和英・国語などの辞書製品)と組み合わせると、ダブルクリックの後右クリックで各辞書の見出し語検索が簡単にできるようになる。
また、ドラッグによる範囲選択の際に、選択している範囲が自動的に単語単位で広がるメカニズムを実装している。

5.2. キーワードおよび要約作成

同様にWord 97の要約作成およびキーワード抽出の基本情報として、T-Hammerの単語切り出し結果を利用している。

5.3. 全文検索のための単語抽出

全文検索には、文字列としての検索と単語単位での検索があり得るが、単語単位の検索を実現するためにT-Hammerが用いられている。
現在、

- Encarta (百科事典)
- Bookshelf
- Internet Index Server (Windows NT上のWebサーバー IISの機能)

において、索引作成、および検索時のクエリー文字列からの自立語切り出しを行なっている。

6. 今後の課題

精度および速度などの基本性能の向上に加えて、今後以下の課題に取り組む必要がある。

（1）辞書の見出し語形式

現在のところ、全文検索のインデックス作成や検索を行う際に、単純に解析結果の語幹部分(品詞情報なし)を用いている。しかしこのやり方だと動詞語幹と名詞(例：「弾 - く」と「弾(〜だん)」)、または活用の異なる動詞などが同じエントリーとなってしまうことがあり、検索結果の妥当性に悪影響を与えることがある。
また、Wordの自動単語選択から活用語をBookshelfで引きたい場合には、語尾を辞書の見出し語にあわせる機能を持つ必要がある。

（2）優先パラメタ付き形態素解析

形態素解析に対して、極端に速度優先を求めるアプリケーションもあり、また精度優先を求めるアプリケーションもある。それらに対して、パラメタの操作で対応する機能も、今後考慮に値する。