利用者による調節が可能な高速日本語形態素解析

5B－4

颯々野 学　難波 功
富士通研究所

1　はじめに

さまざまな自然言語処理の応用を考える上で、日本語の形態素解析の技術は最も基本的なものである。しかし、従来の形態素解析システムでは、特定のアプリケーションに依存し過ぎていることや、処理速度が遅いことが問題になり、形態素解析の応用を広げる障害になっていた。

そこで、筆者らはこれらの障害を乗り越えるために以下の特徴を持つ形態素解析システムを開発した。

- 辞書や文法の定義、評価関数のパラメータ、出力形式などを利用者が調節したりカスタマイズできる。
- トライを用いた辞書アクセスルーチンを使って、非常に高速に処理を行なう。

本稿では、この形態素解析システムの概要と解析速度の実験結果について述べる。

2　開発の基本方針

さまざまな応用の土台となるように、次の点に重点をおいてシステムを開発した。

- 利用者による調節やカスタマイズができ、保守性が良いこと。
- 高速であること。

以下では、これらを満たすために採用したアルゴリズムと実現方法に触れながらシステムの概略を述べる。

3　日本語形態素解析システム

3.1　基本アルゴリズム

形態素解析の基本的なアルゴリズムは動的計画法を用いたものである。隣接する形態素が連接可能かどうか検査するために接続表を使う。コスト関数を使って解を順序付けする。コスト関数の枠組は [2] に準じ、式(1)で表される。

$$(\text{接続コスト}) = (\text{形態素コスト}) + (\text{連接コスト}) \quad (1)$$

(形態素コスト) は品詞などに基づき決定されるコスト、(連接コスト) は隣接する形態素の連接コストである。通常の動作ではもっともコストの低い解を一つ出力する。このアルゴリズムを採用した理由は、きめ細かな尤度付けができ、計算量の点でも優れているからである。

利用者がコスト関数を調節するときには、一位の解を見るだけでは不十分である。上位 n 位の解を得ることや、指定したコスト以下の解を得ることが必要になる。この機能を A*アルゴリズムを使って実現した [3]。

3.2　調節できる機能項目

本システムが持っている調節可能な機能項目のうち主なものを説明する。

辞書と文法

辞書や文法はアプリケーションの要求によって異なる。利用者が辞書や文法を自由に定義できるように、解析エンジン部は特定の辞書や文法には依存させずに実現した。個々の形態素が接続情報 (前接コード、後接コード) と品詞コードを持ち、接続表がある辞書ならば基本的には動作可能である。

また、文法の保守性を高めるために、前接形態素と後接形態素を連接規則として記し、それをコンパイルして接続表を作るツールも作成した。

コスト関数

辞書を交換し易く、しかも単語登録を容易にするため、次のことを行なった。

- 辞書とは独立に解析エンジン内部に品詞体系を想定する。それに基づき、式(1)の形態素コストを計算する評価関数を定義する。評価パラメータは利用者が変更できる。
- 個々の形態素には優先度などを持たせない。

利用者は評価パラメータを用いて解析精度向上や未登録語の切り出し方の変更ができる。

また、利用者が品詞体系の異なる辞書を使おうとする場合には、その品詞体系とエンジン内部との品詞体系を対応付ければ良い。

Development of a User Customizable High-Speed Japanese Morphological Analyzer
Manabu Sassano, Isao Namba
Fujitsu Laboratories Ltd.
1015, Kamikodanaka Nakahara-ku, Kawasaki 211, Japan

出力形式

処理を高速にするために出力情報を選択できるようにした。不要な項目を出力させることは速度の低下を招くからである。二次記憶にアクセスする場合は特に速度が低下する。

自立語や未登録語、指定した品詞だけを出力させることができる。また、読みの有無、品詞情報の有無も選択できる。

品詞コードと表示形式の分離

辞書中で定義する品詞と出力時に表示される品詞とは独立に指定できるようにした。この機能と、内部品詞を設定したことにより、品詞の追加や修正が容易になる。例を図 1 に示す。

```
# 辞書中での品詞,    内部品詞,    表示される品詞
形容動詞,            形容動詞,    状態性名詞
                       ⋮
助詞相当語 1,        接続詞,      助詞相当語
助詞相当語 2,        助詞,        助詞相当語
                       ⋮
```

図 1: 辞書中の品詞と内部品詞、表示品詞名の対応例

3.3 辞書アクセスルーチン

辞書を使う日本語の形態素解析では、辞書のアクセスが全体の処理時間の非常に大きな部分を占める。高速な日本語形態素解析を実現するには、辞書アクセスルーチンを高速にすることが重要である。

入力文字列から最左部分列を切り出す作業が必要な日本語形態素解析では、辞書の構造としてトライが最も適している。キーの長さに比例して常に高速な検索が実現できるからである。

トライを効率良く実現するためにダブル配列 [4] を使う辞書アクセスルーチンを作成した。外部記憶へのアクセスを最小限に押えるため、インデクスと解析に必須な情報 (接続情報や品詞コードなど) はメモリ上に読み込み、読みなど解析動作に必要ない情報は二次記憶上に置いた。

表 1: 辞書の見出し語数とインデクス部のサイズ

	50 万語	27 万語	10 万語
サイズ (KB)	11,344	7,056	2,784

辞書の見出し語数とインデクス部 (インデクス+解析に必須な情報) のサイズの関係を表 1 に示す。メモリ上に置くことが十分可能であることが分かる。

4 実験結果

新聞記事 61 万字に対して解析速度を調べた。JUMAN 2.0 と速度を比較した。本システムは C 言語で記述してある。実行環境は Sun4-670(メモリ 64M バイト) である。その結果を表 2 に示す。

表 2: 新聞記事 61 万字の解析速度の比較

	A	B	C	D
CPU 時間 (秒)	4406.9	356.5	178.8	153.5
文字/秒	137.9	1705	3400	3961
msec/文字	7.25	0.587	0.294	0.252
速度比	1	12	25	29

	システム	辞書	読みの出力
A	JUMAN 2.0	JUMAN 2.0	あり
B	本システム	JUMAN 2.0(*)	あり
C	本システム	JUMAN 2.0(*)	なし
D	本システム	独自 (10 万語)	なし

(*) JUMAN の辞書と連接定義を変換して使用した。

[5] で掲げられた解析速度の目標 1,000 文字/秒を上回る結果が得られた。B の場合でも JUMAN 2.0 より 12 倍高速である。

5 おわりに

以上、利用者による調節が可能な高速日本語形態素解析システムについて述べた。

今後は、本システムを実際のアプリケーションで利用していく予定である。また、精度の向上に取り組みつつ、更に扱い易い連接定義の手法なども検討したい。

参考文献

[1] 松本裕治, 他:“日本語形態素解析システム JUMAN 使用説明書 version2.0” (1994).

[2] Toru Hisamitsu and Yoshihiko Nitta:“A Uniform Treatment of Heuristic Methods for Morphological Analysis of Written Japanese”, *Proc. of 2nd Japan-Australia Joint Workshop on NLP* (1991).

[3] 永田昌明:“前向き DP 後向き A*アルゴリズムを用いた確率的日本語形態素解析システム”, 自然言語処理研究会, 情報処理学会, NL 101-10, pp. 73–80(1994).

[4] 青江順一:“ダブル配列による高速ディジタル検索アルゴリズム”, 電子情報通信学会論文誌, J71-D, 9, pp. 1592–1600 (1988).

[5] 長尾真, 他:“自然言語処理技術のこれからの課題”, 「自然言語処理の技術動向」 調査報告書 (1994).