高速日本語形態素解析ソフト「SuperMorpho-J」

多田智之　金岡秀信
オムロン株式会社　IT研究所
{tada, kana}@ari.ncl.omron.co.jp

1　はじめに

最近のインターネットの発展、パソコンの普及に伴って、利用可能な日本語電子化テキストが爆発的に増加している。そして、その膨大なテキストをコンピューター・システムで限られた時間内に処理するという要望が高まってきている。

日本語形態素解析ソフトウェアは日本語処理システムのベーシックな部品として重要な位置を占めているため、オムロンは高速で使いやすい組込み型の日本語形態素解析 SuperMorpho-J を開発してきた。今回、機能・性能を改善したので その概要を報告する。

2　処理速度の高速化

最近のテキスト量の増加で、数 GB のテキストを処理することも珍しく無くなってきた。しかも実用的には数時間で、または、一晩以内に処理が終了することを要求される。この状況に答えるため形態素解析の処理速度 1.2GB/時間を実現した。測定プラットフォームは Sun Ultra-2 (200MHz)である。

2.1 解析処理速度の高速化

形態素解析のアルゴリズムはコスト最小法を用い、コスト最小解のみを採用することにした[1]。辞書検索は処理時間全体に占める割合が小さいため、非常にシンプルな構造、探索アルゴリズムを用いた。N 分木のトライ辞書を各ノードで 2 分探索で探索する、というものである。このシステムでは辞書検索自体よりも検索結果を辞書から作業領域にコピーする時間のほうが大きい。

既に 1GB/時間を達成する処理速度を持つ日本語形態素解析が報告されている[2]。そのシステム「すもも」との測定条件での最大の違いは辞書の単語数であるため、SuperMorpho-J でも同等の単語数での測定も行なった。すなわち、標準で装備されている単語数は 17 万語であるが、比較のため 35 万語でも測定した。その結果を表 1 に示す。

表 1：形態素解析の処理速度

	単語数	処理速度
SuperMorpho-J	17 万語	1.20GB/時間
SuperMorpho-J	35 万語	1.08GB/時間
すもも	35 万語	1.05GB/時間(注)

プラットフォーム：Sun Ultra-2 (200MHz)
データ量：4,426,199 バイト
測定方法：UNIX time コマンドのユーザ時間とシステム時間の和。
(注) 論文[2]の「SUN Ultra-2 (200MHz)では 290K から 300K バイト/秒を得ている」という記述の 300K バイト/秒から算出した値。

SuperMorpho-J の高速性はインターネット検索エンジン Infoseek[3]でも実証されている。インターネット検索エンジンでは数十 GB に及ぶホームページをインデキシングし、多数のユーザが同時にアクセスしている。 SuperMorpho-J はその両方の処理で活用されている。

2.2 マルチスレッドセーフ構造

テキストを処理するシステムにおいては、形態素解析の処理フェーズは並列性が高いと思われる。文書をインデキシングする場合は文書ごとに独立して形態素解析が可能であるし、複数ユーザの自然言語入力を一度に受付ける場合はユーザの入力ごとに独立している。

そのため、API 関数をマルチスレッドセーフな構造にすることにより、アプリケーション・システムは複数の形態素解析処理を同時に実行できるようになり、全体の性能が向上することが見込まれる。

1 つのシステムでは、通常同じ辞書・文法規則を使用し、作業領域のみをスレッドごとに確保すれば良い。一般的には辞書への書込みはほとんど無いため、解析中はロックして占有する資源が無く複数スレッドが完全に並列で動作できる。

3　ユーザカスタマイズ機能

これまで幾つかのシステムへ形態素解析を組込んだ経験、また組込んだユーザからの要望をもと

に、エンドユーザが簡単に機能をカスタマイズできる工夫をした。その概要について述べる。

3.1 未登録語処理のユーザカスタマイズ機能

これまで最も多くあったカスタマイズの要望は未登録語の処理である。ユーザ定義の単語を登録する方法では解決しきれない問題があるためである。その主なものを示す。

カタカナ未登録語処理のカスタマイズ

未登録語のうちカタカナ語が占める割合が高い。カタカナ語は文字種によって単語の範囲が決定できるため比較的扱いが容易と思われているが、カタカナ語で構成される複合語を考慮するとあまり簡単ではない。例えば、「ラン」「キング」「システム」は登録語、「ランキング」「バンキング」は未登録語とすると従来の SuperMorpho-J では、

- 「バンキングシステム」は「バン」「キング」「システム」と分割するか、「バンキングシステム」と1単語になってしまう。「バン」は一致する単語が無い。
- 「ランキングシステム」は「ラン」「キング」「システム」と分割され、未登録語を含むことも発見できない。

形態素解析のレベルでこの問題を完全に解決することはできず、ユーザはシステムの性質を考慮して妥協点を設定せざるを得ない。その妥協点の相違がユーザカスタマイズ要求となって現れる。その要求は、

要求 1.ある程度細かく分割したい。単語が一致しない部分だけを1単語とすれば良い。「バン」「キング」「システム」でよい。

要求 2.カタカナ語は未登録語、複合語関係なく1単語とすれば良い。「ランキングシステム」でよい。更に区切り文字も含めて1単語としたい。「バンキング・システム」も1単語。

要求 3.やはりなるべく正しい1単語を検出したい。「ランキング」「システム」としたい。

などである。

要求 3 を完全に満たすことはできないため、「短いカタカナ単語がカタカナ文字列の一部に一致した場合は、偶然一致した可能性が高いので連結する」というヒューリスティックな方法で対応した。処理の概略は以下のようになる。

- カタカナ文字列のうち、前方最長一致で単語が一致しない部分文字列は1単語とする。
- その後、ユーザが設定した長さより短い単語が連接する箇所を探して、連結して1単語とする。

登録語・未登録語に関係なくカタカナ複合語の中で最も短い単語は何文字を許すか、という設定をユーザがすることになる。要求 1 に対しては数値 1、要求 2 に対しては数値 1000（など大きな数字）、要求 3 にはユーザが判断した適当な数値（例えば 3）を設定する。

要求 3 に対して数値 3、または 4 を設定した場合の正解率を調査してみた。辞書に載っていない4文字以上のカタカナ文字列 100 個を 30MB のホームページから頻度が高い順に選び、目視で判定した結果を表 2 に示す。

表 2：長さに着目したカタカナ未登録語分割

設定値 3	個数	例
誤って1単語に結合	0	
誤って複合語に分割	9	サウンド/ブ/ラスター アプレ/ット
正解率	91%	
設定値 4	個数	例
誤って1単語に結合	19	トップページ ビデオカード
誤って複合語に分割	0	
正解率	81%	

異文字種結合のカスタマイズ

一般的に未登録語は文字種の連続する範囲で単語の境界を決定する。しかし例外的に文字種の異なる文字も単語の範囲に入れて1単語にしたいという要求も多い。以下に例を示す。

カタカナに入れたい文字「・」「=」
　例) エコ・システム、バスコ=ダ=ガマ
アルファベットに入れたい文字「&」「/」
　例) AT&T、PC/AT
数字に入れたい文字「.」「,」「・」
　例) 1.5、1,200、六十・五

ユーザはどの文字種にどの文字を含めたいかを設定ファイルに設定する。ただし、指定した例外文字が含めたい文字種に挟まれる場合のみ1単語としている。

3.2 その他のカスタマイズ機能

未登録語のカスタマイズ以外にも、他の形態素解析には見られないユニークなカスタマイズ機能があるので、その幾つかを紹介する。

活用形標準化のカスタマイズ

SuperMorpho-J は活用形を標準化する機能を持つ。例えば、「走らない」「走ります」などの活用形を 1 つの形「走る」にすることである。

動詞を標準化する際には、これまで 2 通りの要求があった。

1.終止形
2.連体形

連体形として標準化したいという要望は、「連体形と動詞化した名詞を区別できないため、一緒にしてしまいたい」という背景からである。以下のような品詞の曖昧性は、現状の形態素解析レベルでは解消できていない。

例) 車の走り（動詞化した名詞）
車は走り、（連体形）

ユーザが、品詞ごとに動詞の語幹に後続する文字列を設定できるようにした。ラ行五段活用「走(る)」を連用形「走り」に標準化したい場合は、以下のような設定をする。

ラ行五段	り

空白文字除去カスタマイズ

入力テキストファイルがプレーンテキストの場合、日本語テキストでは「行分れ」と呼ばれる現象が起こる。行末で単語が分割される現象である。本来 1 単語である文字と文字の間に、改行、スペース、タブなど空白文字が挿入される。

She took out her raincoat.（彼女はレインコートを取り出した。）

では、「レインコート」の「イ」と「ン」間に改行が挿入されている。この空白文字を前処理で除去するが、英単語の間のスペースは重要な単語区切り情報なので除去しない。

しかし、ユーザによっては行分れを起こす空白文字も重要な情報として除去しないことを望む場合がある。そのため「空白文字を除去する／しない」を設定ファイルで設定できるようにした。

品詞ごとの分類カスタマイズ

解析結果の単語を品詞ごとに分類したい場合がある。例えば、検索インデックス作成時に重要性の低い「接続詞」「副詞」「連体詞」などを除外したい、または、複合名詞を検出したいので、「名詞」「人名」「接頭語」など名詞相当の単語を選び出したい、などである。

分類したい複数の品詞の ID が連続していない場合があるため、これを簡単なフラグのチェックでできるように、設定ファイルに品詞ごとにフラグを設定し、単語の属性情報としてそのフラグも出力する。以下に設定ファイルの例を示す。

名詞	1
人名	1
接続詞	2
副詞	2
連体詞	2
接頭語	1

4 おわりに

高速で使いやすい組込み型の日本語形態素解析 SuperMorpho-J は、解析処理速度 1.2GB/時間の高速化を達成した。また、未登録語処理、表記ゆれ処理などに関してエンドユーザが好みの設定ができるような構成を実現した。

形態素解析は言語処理の基本であり、機能・性能を更に改善していく必要がある。

なお、SuperMorpho-J に関する詳細な情報はオムロンソフトウェア(株)、
http://www.omronsoft.co.jp
で入手できる。

参考文献

[1] 長尾編：自然言語処理、岩波講座ソフトウェア科学 15、形態素解析、pp.117-137、1996.
[2] 鷲坂ら：情報検索のための高速日本語形態素解析システム「すもも」情報処理学会第 54 回、pp2-59,60、1997.
[3] Infoseek：インターネット検索エンジン、http://www.infoseek.co.jp