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」

1C－6

鷲坂光一　山崎憲一　廣津登志夫　尾内理紀夫
NTT 基礎研究所

1　はじめに

これまで日本語形態素解析は，機械翻訳などを行うための最初のステップとして用いられることが多かった．ここでの解析間違いは当然のことながら後のステップの解析精度に影響を及ぼす．このため，これまでは形態素解析の精度の向上を主眼として研究が進められてきており，計算量についての研究なども幾つか見られるものの[1)]，実システムの速度はそれほど重視されていなかった．しかし，形態素解析を情報検索のような大量の文書処理に適用する場合，処理の高速性が極めて重要となる．本論文では，応用の立場から形態素解析の速度向上の重要性を指摘するとともに，我々が実装したシステム「すもも」における高速化技法を述べ，最後に評価結果を示す．

2　大量文書処理のための形態素解析

インターネット技術の普及にともない，電子的にアクセス可能なデータは増え続けており，例えばWorld Wide Web(WWW) によりアクセス可能なページは6000万ページ以上あると言われている．仮に1ページ当たりのテキストサイズが1Kバイトだとすると，これは60Gバイトに相当する．もちろん日本語で書かれたページは一部にしか過ぎないが，今後ますます増大していくことは明らかである．WWW のページで特徴的なことはデータが日々更新されるという点である．このためWWWの空間から目的のページを探す情報検索サービスは，日々，更新されたページを処理して検索インデックスに追加する，あるいは検索インデックスを再構築する必要がある．

日本語で記述されたページから検索インデックスを作る方法は幾つかあるが，形態素解析をすれば，より高度で精度の高い検索ができる可能性がある．ここで問題となるのが，ギガバイト単位のデータ処理に要する形態素解析の時間である．上で述べたような情報検索サービスでは，形態素解析と検索インデックスの作成をデータ流入量を上回る速度で行う必要がある．後者については十分高速なシステムが存在するものの，形態素解析については少なくとも我々の知るところでは存在しない．

我々は，ギガバイト単位のデータを実用的な時間内に処理できることを目標に形態素解析システム「すもも」を設計，実装した．以下では「すもも」において用いられた高速化技法とその評価について述べる．

Sumomo – a fast morphological analyzer for information retrieval
Mitsukazu WASHISAKA (wasisaka@rouge.brl.ntt.co.jp).
Ken-ichi YAMAZAKI (yamazaki@nuesun.brl.ntt.co.jp),
Toshio HIROTSU (hirotsu@square.brl.ntt.co jp),
Rikio ONAI (onai@square.brl ntt.co.jp),
NTT Basic Research Laboratories.

3　形態素解析の高速化

形態素解析のアルゴリズムには，コスト最小法[2)]を用いた．コスト最小法は形態素解析の方法としてよく知られた手法であり，多くのシステムがこれを用いているが，実装方法によって速度は大きく異なる．

3.1　解の候補の決定方法

一般にコスト最小法を用いる場合，接続しうる形態素間にコスト付きリンクを張ったグラフを作り，その中でコスト最小のパスを捜し出す．この方法は次に小さいコストを持つパスを知る必要がある場合には意味がある．例えば，仮名漢字変換のための形態素解析では，ユーザがそのパスは正しくないと判断したら次のパスを出力しなければならない．このように形態素解析の結果を判断するには何らかの意味的な解析が必要となる．一方，情報検索においては，何ギガバイトもの解析結果を意味的に判断することは不可能である．このため，情報検索用の形態素解析システムではただ一つの解析結果を出力すれば十分だと考えられる．

「すもも」では，ある文字列が辞書にあった時，その形態素の前に接続しうる形態素のうちで，最もコストが小さいものだけにリンクを張り，それ以外を捨てている．文末まで解析が到達すると同時にコスト最小のパスが得られるため，グラフを調べ直す必要はない．

3.2　未定義語の扱い

未定義語の扱いについても，情報検索においては，正しく未定義語が取り出せたことを判断する主体がない．また，質問の中の未定義語も同じように誤解析をすれば検索には特に支障がない，と考えられる．

このため未定義語を完全に扱うことは，本応用ではそれほど重要ではない．「すもも」では，連続するカタカナ，英字，数字などを一つの形態素として扱う．これにより外来語や製品番号などの未定義語はほぼ回避可能である．また，接続しうる形態素がなくなって文末まで解析が進められないという条件が成立した場合のみ未定義語探索を行うこととした．この条件のチェックは少ない手間でできるため，速度に影響を与えない．現在の未定義語探索アルゴリズムは，接続不能となった形態素を未定義語とするという非常に単純なものである．

3.3 辞書の構造と読込み

辞書の構造としては，辞書の中を1回探索するだけですべての形態素候補を見つけることができるようにトライ構造[3]を用いた．しかし，トライ構造をどの様に構成するかによって，辞書引きの速度は大きく異なる．「すもも」の実装にあたっては，次の方法を採用した．

- 長さ1の形態素は，パスのすべての分岐で探索されるため，テーブルで直接アクセスする．
- 長さ2以上の形態素は，最初の2文字をもとにハッシュ表に入れ，3文字目以降をトライ構造とする．

辞書の読込みに関しても，従来のシステムでは辞書全部を読込まずに，必要な部分だけをキャッシュするようにしていた．これは，メモリの少ない計算機では使用メモリ量を抑えるという点で意味があったが，大容量メモリが計算機に搭載されるようになった現在では，解析速度を低下させる原因になっている．辞書のサイズが計算機のメモリサイズに比べて比較的小さいのであれば，辞書全部をプロセスのメモリ空間にマッピングし，読込みはページフォールトに任せた方が，キャッシュを管理する手間もなくアクセス速度も速い．このため「すもも」では，mmap関数を用いて辞書全部をマッピングする方法を採用した．

4 性能評価

4.1 解析速度

すもも，茶筅1.0b5，Juman 2.0の解析速度を比較した．使用した計算機はSun SPARCstation20(SunOS 4.1.4，主記憶224MB，クロック75MHz)，プログラムはgcc 2.7.2で-Oオプションを付けてコンパイルした．測定にはtimeコマンドを用い，10回測定した後，最高値と最悪値を除外した残りの平均値をCPU時間(ユーザ時間＋システム時間)とした．

	解析方法	辞書語彙数	辞書サイズ
すもも	コスト最小法	約35万語	約7.5 MB
茶筅1.0b5	コスト最小法	約11万語	約7.5 MB
Juman 2.0	コスト最小法	約11万語	約51 MB

表1: 比較対象プログラム

	CPU時間(秒)	解析速度(バイト/秒)
すもも	13.3	152746.2
茶筅1.0b5	159.9	12744.8
Juman 2.0	1309.1	1556.2

表2: 実行速度(テキストサイズ2037253バイト)

使用している日本語辞書の違いにより，形態素解析の出力結果は若干の異なるものの，すももでは茶筅1.05bの12倍，Juman 2.0の98倍の解析速度をあげることができた．なお，参考までにSUN Ultra-2(200MHz)では290Kから300Kバイト/秒を得ている．

4.2 解析精度

現時点ではコストの調整を行っている段階であり，以下に述べる解析精度は一時的なものである．

日本語辞書にはEDR日本語単語辞書[4]を使用しており，この辞書では各形態素に左属性と右属性が与えられている．ここでは次の2つの方法で精度を測定した．

1. 属性精度……左属性と右属性が完全に一致した割合(ただし，サ変名詞と普通名詞は区別していない)
2. 句切り精度……形態素の区切りが一致した割合

2を測定したのは，EDRの属性分類が非常に細かく，また一部に意味を意識した属性があるため，属性が完全に一致しなくても，情報検索のための形態素解析として十分な場合があるためである．

精度の測定のためEDRの日本語コーパスから任意に1000文を抽出し解析した結果，属性精度で93.5%(形態素単位)，区切り精度で84.7%(文単位)を得た．これらの精度はコスト調整により現在も向上中であり，また属性分類の変更まで含めた改善策も検討中である．

5 おわりに

情報検索サービスに用いるための形態素解析とその高速化方法について述べた．現時点での解析速度は，高速ワークステーション上で300Kバイト/秒前後である．これは約1ギガバイト/時間に相当し，当初の目標であるギガバイト単位のデータを実用時間内で処理するという目標は達成した．

さらなる速度の向上のためには，属性分類や活用の扱いといった部分が最も重要である．例えば，EDR日本語単語辞書では，他の辞書では一つの形態素として扱っているものが，意味を考慮して複数に分類されていたり，活用語尾を一つの形態素として扱っていたりしている．このため形態素候補の数が増加し，コスト比較の手間が速度低下をもたらしている．これら，辞書に関する部分は時間と労力を要する仕事であり改良の難しい点もある．しかし，形態素解析を大量文書に適用する重要性は今後ますます高まると思われ，解析速度も考慮して形態素を分類した辞書の構築も求められる．

日頃から貴重なアドバイスを頂いているNTT基礎研究所対話研究グループの皆様，石井健一郎情報科学部長に深く感謝します．

参考文献

[1] 久保，新田: 接続コスト最小法による形態素解析の提案と計算量の評価について，信学技法,NLC90-8,pp.17-24,1990.

[2] 田中: 自然言語解析の基礎，第3章形態素解析,pp.133-153,産業図書,1989.

[3] 青江: トライとその応用，情報処理,Vol.34,No.2,pp.224-251,1993.

[4] EDR電子化辞書1.5版仕様説明書，日本電子化辞書研究所,1996.