自然言語処理における
効率的な形態素解析手法に関する研究

能登研究室　　山崎貴弘 (46093)

1　はじめに

現在主要なワープロソフトでは，入力した文章についてのスペルチェック等の機能がついており，ごく自然にその文章校正機能は利用されている．しかしその単語辞書は不充分であり，正しい単語もよく間違いと指摘されることがある．そこで正しい形態素解析には，解析手法の拡張や統語辞書の充実などが重要になってくるが，それはまた検索時の負荷を増やしてしまうことになる．

本稿では，文章校正の過程での，効率性を重視した形態素解析の方法を提案し，それに伴う形態素辞書の拡張を考案し，その有用性について述べる．

2　形態素解析手法

文章校正の場合，主たるプロセスは辞書を用いた形態素解析であり，その辞書検索手法の高速化について考える．一般的リスト形態の辞書検索の手法としては，二分探索やハッシュ法が基本的である．ここでN個の単語をもつ辞書への検索を仮定し，理論上での効率性について考察する．

- 二分探索とは，検索対象の要素を使い比較を行い，探索領域をシーケンシャルにアクセスし，一回の比較ごとに半分にしていく方法であり，最大 $\log_2 N$回の計算を必要とする．
- ハッシュ法とは，ハッシュ値を用いることによって，リスト上のデータにランダムアクセスできるので，衝突が起きなければ一回の探索で対象を見つけ出す方法である．

2.1　ハッシュ法の利点と問題点

二つのアルゴリズムの原理からも，計算が有限回で終わるハッシュ法のほうが高速であり，計算の負荷も少なくなる．本研究ではその高速性からハッシュ法を用いる．実際に，英語のような分かち書きされた言語での形態素解析では，単語区切りが明白であり比較的単純，かつ，語の活用形なども少ないため，ハッシュ法が用いられることがある．しかしハッシュ法ではデータをある特定の範囲内に収まる整数のキーとして，ハッシュ値という任意の値でキーを除算して，出た値で表を検索するのであまりが同一のものも出てくる．これを衝突というが，特に日本語での形態素解析においては，辞書との照合時単語の字数が多くなるほど衝突が起こりやすく，日本語辞書は活用形が多く言語辞書の偏りもあり，そのままでは長文の形態素解析等には向かず，二分探索の方が一般的である．

2.2　一般的な衝突の解決法

そこで，一般的にはその衝突を解決するために，完全ハッシュ関数が使われる．完全ハッシュ関数とは，データのキーの値がなるべく均等に分散するように，検索対象が重複しないようにハッシュ値を設定するというものである．しかしこの方法では，自然言語辞書のような非常にデータ量が大きく，ばらついた固まりがあるリストでは，ハッシュ値が大きくなるため，使用されるキーの値にばらつきが出て，使われない無駄なキーが多くなり，あまり実用的ではない．さらにはデータの新規登録も難しくなる．

3　ハッシュ法の拡張

そこでハッシュ法の検索するリストに制約を課す．この際二分探索の探索範囲縮小を用いる．方式としてはまずリストを二分探索して行き，ある程度探索範囲が狭まり衝突が解消された後，ハッシュ法により探索し照合を行う．この方法を用いた場合の計算処理は，省略した二分探索回数 n 回，衝突の解消のための計算回数 a 回とすると，理論的には $\{(\log_2 N) - n + a\}$ 回となり，最適な n が見つかればその分計算処理が少なくなり，高速な検索が可能になる．

3.1　ハッシュ法による形態素解析

以下に本研究で提案する形態素解析の手順を示す．

- Step0：まずは解析する文から最左部分語一文字を取り出し，部分語を一文字ずつ増やしながら辞書の最長単語長まで取り出し，辞書順にソートし探索リストを作る．
- Step1：衝突がある程度起こる場合はStep2a へ．衝突があまり起こらない場合はStep2b へ．
- Step2a:二分探索を行い探索範囲を狭め，Step2b へ．(この際一度狭めた範囲は以降の探索にも反映させる．)
- Step2b:ハッシュ法を用い探索し，整合するデータを導き出す．
- Step3：結果を再リスト化し，そのリスト内で最後に全文字の照合をし，データの整合性を確かめる．
- Step4：最後に最左部分が重複した単語は，連接関係や連接コストを用いて，最適な単語を選択し，単語として辞書に該当するものがない部分はリストアップする．

4　形態素辞書の拡張

本研究で提案したプロセスを行うためには，形態素辞書からあらかじめハッシュ値を計算し，個々の単語のハッシュ値を記述したハッシュ表を作成すること，そして，その際衝突がある程度以上起こる場合は，事前に n の最適値を求めておき，リストにその情報を付加しておく．これにより，リストが小さい時にも大きい時にも，ハッシュ法の高速化の恩恵をうける事ができ，自然言語の辞書検索のような，リストにばらついた固まりがでるようなものには，それだけ有用性が増す．

5　おわりに

本研究では文書校正における形態素解析に着目し，検索効率の向上化のために，形態素辞書検索のためのハッシュ表を作りハッシュ法と二分探索の融合による文書校正の高速化を提案した．