日本語形態素解析における効率的な動詞活用処理

1Q－5

久光 徹　新田 義彦
日立製作所 基礎研究所

1．はじめに

データ構造の工夫や主記憶の大規模化により辞書引きが大幅に高速化された現在，最尤解抽出部の効率化は形態素解析の効率改善において重要な位置を占めるに至った．我々はこの問題を，従来あまり深刻に考察されたことのない動詞の活用処理に焦点を当て，非サ変動詞活用処理に用いる辞書見出しの側面から考察する．以下では，従来方式（音韻論的扱い，及び学校文法に準ずる扱い2種類）を簡単に示した後，動詞の音韻的語幹の末尾子音を屈折接辞先頭側に付加した見出しを用いる新手法を提案し，計算効率を含む種々の観点から従来方式に対する優位性を示す．提案法は，最も一般的な活用語尾分割方式の辞書にわずかな変更を加えるだけで実現できる．

2 活用処理に関する既存の手法

2.1 音韻論的手法

動詞活用の音韻論的分析は[1]に始まる．この立場では，学校文法と異なり表記単位でなく音素単位で動詞活用を記述する．このため日本語の非サ変規則動詞は，音韻上の語幹末尾により子音動詞（いわゆる五段動詞.例えば'消す'の音韻的語幹は'kes'である）と母音動詞（いわゆる一段動詞）に分類される．例えば子音動詞については，否定，受身，完了等を表わす形態素（これらは動詞語幹に後接する屈折接辞とみなされる）は音韻単位で 'ana', 'are, 'ita'等としてとり扱われ，これらが動詞語幹に後接することにより，いわゆる'活用'が生じる．

この立場を実際の形態素解析に応用する場合，'ana', 'are, 'ita' 等の辞書見出しを解析対象文字列中に顕在化させる必要があるため，前処理により入力文字列中の平仮名部分をローマ字表記に変換した後，形態素解析を行うという前提で，表1のような辞書を用いることになろう：

entry	comments
消s	stem
:	:
ana	Negative
are	Passive
ita	Past
:	:

表1
音韻論的分析に基づく辞書見出し

例えば，'消さなかった'は次のように解析される：

消さなかった --> 消sanakatta --> 消s / ana / katta .

音便処理まで考慮すると，例えば我々の用いている非サ変系動詞2805個を含む辞書の場合，約1600個の異形態を登録する必要がある．

この枠組によれば動詞活用を極めて合理的に記述できるが，仮名部分のローマ字化により解析対象文字列の長さが増大し，解析効率の面で明らかに不利である．生成に用いるのが適切な手法といえよう[2].

2.2 学校文法的手法

学校文法では表記単位に基づいて活用を記述するため，例えば子音動詞の場合，音韻上の語幹から末尾子音を除いた部分を語幹とし，この末尾子音と，後続する屈折接辞の先頭母音をひとまとめにして'活用語尾'と考えることになる．活用語尾の扱いにより，活用処理手法は二つに分かれる：

2.2.1 活用形展開方式

最も単純な手法は，活用形をすべて辞書見出しとして登録することである．例えば，'消す'は，{消さ，消し，消す1，消す2，消せ1，消せ2，消そ}に展開して登録される．2.1の例の解析は次のようになる：

消さなかった --------> 消さ/なかっ/た.

2.1で述べた我々の辞書の場合，動詞実数より約12000個多い活用形を登録する必要がある．日本語の動詞活用は極めて規則的なため，従来この手法はほとんど利用されていない．

2.2.2 活用語尾分離方式

実際上ほとんどの場合に利用されている手法である．学校文法での動詞語幹，活用語尾を分割して辞書見出しとする．動詞実数に対する辞書見出しの増加は，活用語尾 100個未満であり，前述の2手法と比べて格段に少ない．しかし代償として，同じ屈折形の解析に1分割余計に必要である：

消さなかった --------> 消/さ/なかっ/た.

3 提案方式

2節で3種類の手法を概観した．音韻論に基づく手法は簡明であるが，実際の日本語表記とのずれのため効果が減殺される．活用形展開方式は，分割手続が簡単になるが，動詞実数とかけ離れた数の活用形を登録しなければならない（これを看過しても，後述する他の観点から必ずしも最良の手法とは言えない）．活用語尾分離方式は，辞書見出しの数は最小にとどまるが，分割効率が劣る．本節では，これらの長所を併せもつ手法を示す．

学校文法では，屈折接辞の先頭母音（例えば否定接辞'ana'の'a'）を音韻的語幹（例えば'消s'）側にまとめて，'活用形'（例えば'消さ'）と呼んだ．一方，

An Efficient Treatment of Japanese Verb Inflection for Morphological Analysis
Toru Hisamitsu and Yoshihiko Nitta
Advanced Research Laboratory, Hitachi, Ltd.

計算機処理の立場からは，動詞語幹の末尾子音（例えば'消s'の's'）を屈折接辞の先頭に付加して，'さな'のように屈折接辞側を展開することも平等に可能である．このとき例えば，'消s'の表記上の語幹'消'は，'s-付加屈折接辞'のみの後続を許すと辞書記載すればよい．すべての非サ変規則動詞の活用形に同様の手続を行うと，表2の様な辞書見出しを得る．

entry	comments
消	stem
：	：
さな	Negative (s + *ana*)
され	Passive (s + *are*)
した	Past (s + *ita*)
：	：

表 2
提案方式の辞書見出し

これを用いると，解析は次のように行われる：
'消さなかった' ---> '消 / さな / かった'
この分割によっても，従来と全く同一の言語情報が得られることは明らかである．子音付加により増加する辞書見出しは動詞の数によらず125個にすぎない．分割数は活用語尾分離方式より少なく，活用形展開方式に近いことは明らかである．次節では，より詳細に従来方式との比較を示す．

4 従来方式との比較

本節では，提案方式を，実際に広く利用されている学校文法に基づく手法と比較しよう．

4.1 見出し数

3節で述べたように，提案手法の見出し語数は活用語尾分割方式とほぼ同程度であり，最小に近い．

4.2 解析効率

我々は，提案法を種々の最尤解抽出アルゴリズムに適用し，従来法より最尤解抽出の効率を改善できることを確認している．ここでは，[3]で述べた作表型アルゴリズムを用い，解析表作成における主要部の効率を比較する．プログラミングの詳細に依存しない比較を行うため，指標として，(A)"文字列各位置から辞書引きにより抽出される候補形態素の数"，(B)"これらの間の接続チェックの回数"，(C)"最尤解候補の'断片'を解析表へ登録する回数"を用いる（アルゴリズムの詳細は本論では省略する）．図1は，約60000語の規模の辞書を用いて，2.2の従来方式と，提案方式を比較した結果である．

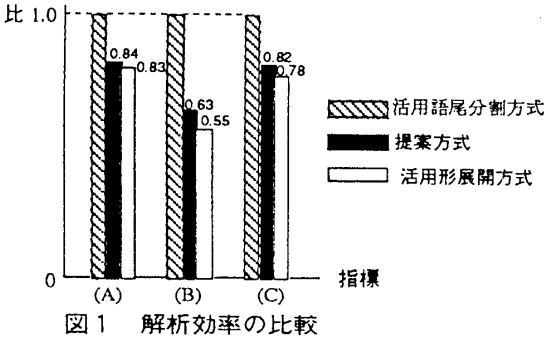

図 1　解析効率の比較

評価用の文は新聞記事100文約4500文字からなる．提案方式を用いた場合，通常の2.2.2 の手法に比べ，各指標が20～40%改善される．なお，従来問題外とされてきた2.2.1の方法は，これらの指標に関しては最良であり，総合的な優位性では劣るとしても解析効率の点からは無視できないことが，定量的に示せたといえる．

4.3 接続記述の簡明さ

提案する方式では，各活用形と，屈折接辞との間の接続が辞書見出しの中に吸収されているため，活用処理に関する接続情報は大幅に簡素化される．

4.4 確率モデル構築の容易さ

最近広く利用されている確率モデルを構成する場合，2.2.1の見出しを用いると，各品詞と各'活用形'との間の接続を考えねばならない．動詞はほとんどすべての品詞の単語に後続できるため，提案手法に比べて獲得すべきパラメータ数は数百個多くなる．

4.5 OCR後処理への応用

形態素解析の応用として重要視されているものに，OCRの誤認識修正がある．詳細は省略するが，この分野で良く利用される辞書引き法（例えば[4]）を効率化するための条件の一つに，'先頭一文字が一致する辞書見出しは少ないほうが良い'と言う条件がある．2.2.1の方法はこれに触れるため提案方式に比べ上記辞書引き法の効率が大幅に劣る．2.2.2の方式は分割数の問題からやはり提案方式に比べ効率が劣る．

4.6 その他

最も一般的な2.2.2の活用語尾分割方式から移行する場合，提案手法は少数の辞書見出しの追加と，接続表の書き換え（むしろ簡素化）により，容易に実現できる．

5 おわりに

本報告では，動詞活用形の処理のための辞書見出しとして，動詞の音韻的語幹の末尾子音を屈折接辞先頭側に付加した見出しを提案した．提案手法は，辞書見出し数，解析効率をはじめとする様々な観点から従来方式に比べ優位であると同時に，広く用いられている活用語尾分割方式からわずかな変更のみで実現できる．

参考文献

[1] B. Bloch: Studies in Colloquial Japanese, Part I, Inflection, J. of the American Oriental Society 66, (1946).
[2] 神岡 太郎 他: 述語複合体の生成とその表現, 情処論, Vol. 30, No. 4, pp457-466, (1989).
[3] 久光 徹 他: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, pp17-24 (1990)
[4] 高尾 哲康 他:日本語文書リーダ後処理の実現と評価, 情処論, Vol. 30, No. 11, pp1394-1401, (1989).