3N－9

辞書先読み方式のかな漢字変換におけるメモリ利用効率に関する考察

隈井　裕之　　畑谷　茂樹　　中島　晃
㈱日立製作所　映像メディア研究所

１．はじめに

ワードプロセッサに代表されるＯＡ機器の日本語入力では、かな漢字変換による文字入力が主流となっている。文字入力ではユーザに対する応答性の確保が重要課題となる。

かな漢字変換の高速化については、辞書やアルゴリズムの改良等[1]、いろいろ考えられるが、今回我々は、辞書先読み方式のかな漢字変換を考案し、ユーザの変換指示から結果表示までの高速化を図った。

辞書先読み方式のかな漢字変換では、ユーザの文字と文字の入力の合間に、先行して辞書検索を行うことで、ユーザの変換指示から結果表示までの応答時間を短縮できる。しかし、辞書先読み方式を単純に適用すると、後で不要になる可能性のある単語でも、一旦辞書検索及び登録を行うため、メモリ使用量が増加するという問題が生じる。

この問題を解消するために、我々は、一旦登録した単語が無意味となった場合に、その単語を削除する方法を開発し、メモリ使用量を削減した。

２．辞書先読み方式

我々のかな漢字変換は、文節数最小法[2]に基づき次のステップでかな漢字変換を行う。

(1)辞書検索処理
　入力されたかな読みに基づき辞書から単語を検索する。

(2)形態素解析木作成処理
　辞書検索した単語間の接続を検定し、形態素解析木を作成する。

(3)変換結果作成処理
　作成された形態素解析木から、文節数が最小になるパスを選択し、結果を出力する。

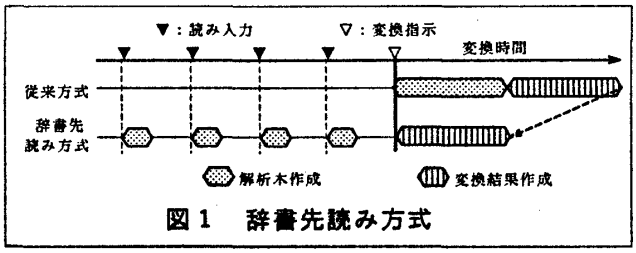

図１　辞書先読み方式

上記処理のどこまでを先行させるかは、内部構成や、CPUの処理能力等を勘案して決めることになる。特にユーザの文字入力を妨げないように、文字と文字の入力の合間に処理可能であることが重要である。

我々の、従来のシステムで各処理の実行時間を測定した結果、(1)(2)の全体(約2秒/20文字)に占める割合はほぼ50%であることがわかった。これらの処理は、１文字入力毎に処理を分散させることが可能である。また、(3)の処理は、形態素解析木が作成済みであることが前提となるため、ユーザの文字と文字の入力の合間(約0.2秒)に終えることは困難である。そこで、今回は、図１に示すように、上記(1)(2)までの処理を先行して行うこととした。

３．メモリ使用量の削減

辞書先読み方式により、ユーザの「ほうほう」という文字入力毎に形態素解析木を作成

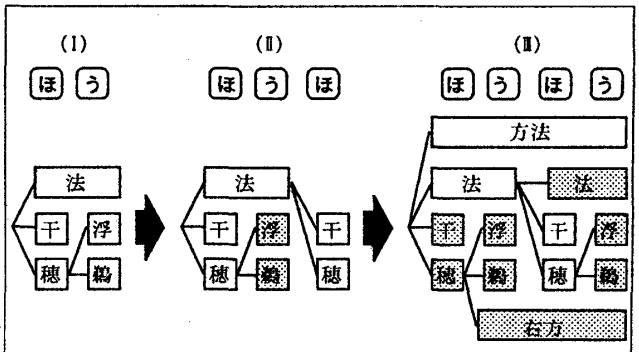

図２．形態素解析木作成過程

A Method of memory reduction in High-Speed Kana to Kanji Conversion .
Hiroyuki KUMAI, Shigeki HATAYA, Akira NAKAJIMA,
Hitachi, Ltd.

する過程を、図２に示す。

辞書先読み方式では、あらかじめ、すべての読みが判っておらず、現在までに入力されたかな読みで辞書検索を行い、形態素解析木を作成していくため、次の文字が入力されたときには、無意味となる単語(図１網かけ部)も、解析木に一旦登録してしまう。このために、全てのかな読みが入力されてから解析を行う従来方式に比べ、メモリ使用量が増大する。

そこで、我々は辞書先読み方式のかな漢字変換の(1)(2)の処理に無意味な単語を削除する処理を加えた。

無意味な単語の判定は、以下のルールに基づく。

R1：日本語文法に基づく接続ルールに従って、後続する単語がないと判断された単語。
例)語尾が見つからなかった動詞語幹(図１－(Ⅲ)「干」)

R2：文節数最小法に従って、最適解となる可能性が無くなった単語。
例）既に文節数最小のパスが存在し，それよりも文節数が多くなる単語の組合せ(図１－(Ⅲ)「右方」)。

４．試作・評価

以上の考え方に基づき、辞書先読み方式のかな漢字変換を試作し、評価を行った。

無意味単語の削除の効果を確認するために単語削除処理の有無による比較実験を行った。

実験は、文献からサンプル文例をランダムに抽出し、１文字入力毎に作成した形態素解析木のメモリ使用量を測定することにより行った。測定の結果を表１に示す。測定結果から、無意味単語の削除処理を行うことによって、削除処理を行わない場合に比べて、メモリ使用量は平均約40%で済むことが分かった。

表１　メモリ使用量計測結果

文例	文字数	メモリ使用量(最大値) a.削除有	b.削除無	a/b(%)
1	33	6.0 KB	15.0 KB	40.0%
2	38	5.1()	17.3	29.4
3	28	6.9(7.9)	16.9	40.8
4	23	6.1	13.0	46.9
5	26	6.9	15.7	43.9
平均	－	－	－	40.2

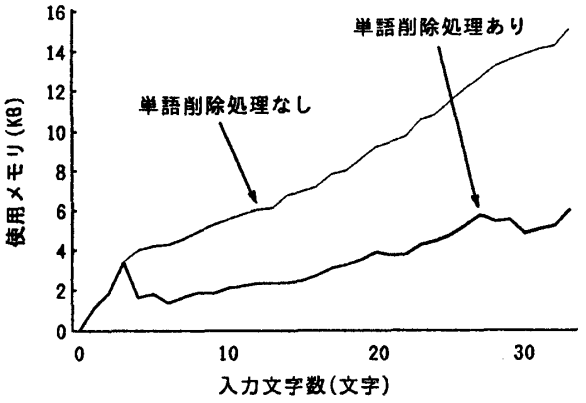

図３．メモリ使用量の推移

図３に文例１におけるメモリ使用量の推移を示す。最初の数文字は、先頭単語のみの解析木作成であり、単語削除の条件が発生しないため、削除の効果は現れない。文字数が増えるに従い、削除の効果が現れ、メモリ使用量の割合(削減率)はほぼ一定の値に落ち着く。

また、別に行っている変換精度の評価では、全体として影響は殆どなかったが、削除処理によって過度に削除されたために生じた誤変換が数件検出されている。これは、条件R2の基準が厳しいことによるもので、適用条件の再検討が必要である。

また、削除処理の平均実行時間は、16ビットCPUでも約4msであり、ユーザの文字と文字の入力の合間に十分吸収可能な時間である。

５．まとめ

かな漢字変換のユーザに対する応答性を向上させるために、辞書先読み方式のかな漢字変換を考案し、そのメモリ利用効率について検討した。無意味単語の削除処理を加えることで、辞書先読みに伴う使用メモリを40%に抑えることができた。

参考文献

[1]小松：「コスト最小法に基づく逐次確定型・形態素解析」，情報処理学会第４７回全国大会 6M-2(1993)

[2]吉村：「文節数最小法を用いたべた書き日本語文の形態素解析」,情処学論,24(1983)