1P−8

2文字間の連接を利用した仮名漢字変換用辞書

堤 豊† 下園 幸一‡ 菅沼 明‡ 牛島 和夫‡
†有明高専 ‡九州大学

1. はじめに

仮名漢字変換の，変換率を上げるためには変換用辞書の項目を増やさなければならない．従って辞書をいかに小さくするかが課題となっている．一方，アクセス速度の面では，逐次検索に適しているTRIE構造の辞書が使われることが多い．

本稿では，辞書の見出し語を圧縮し，TRIE構造で表現する手法と，それを用いた辞書の構成法について述べる．この方法を採用すると，7万語程度の辞書の見出し語を主記憶上に置くことができるので，高速な検索を行なうことが可能となる．

2. 従来の辞書

仮名漢字変換システムは製品化されていることもあり，辞書の研究が進んでいる．辞書の圧縮に関しては1文字単位のハフマン符号化方式[1]などが提案されている．また，辞書の構造としては，TRIE構造がよく使用されている[2]．TRIE構造は図1に示すように，文字をポインタで接続したものである．仮名漢字変換のように，入力文字列を前から順に辞書検索するような場合，1文字入力されるごとにポインタをたどっていけばよい．また，先頭からの部分的なマッチングも取れるので，辞書の構造としては，TRIE構造は仮名漢字変換に適していると考えられる．しかし，TRIE構造の辞書は，ポインタを多用するため，何も工夫しない普通のテキスト形式で格納した辞書よりもサイズが大きくなるという欠点がある．このため，TRIE構造の辞書のうち，分岐がないポインタについてはポインタを省略する方法などが提案されているが，見出し語そのものを圧縮するわけではないので圧縮効率は高くない．

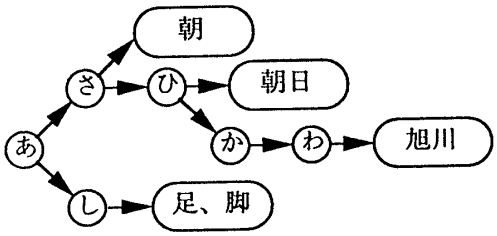


図1: TRIE構造の例

3. 辞書圧縮方法1

仮名漢字変換用の辞書は，仮名による見出し語と品詞番号と漢字文字列とから構成されている．このうち，仮名による見出し語の部分は，文字数は多いが，含まれる文字種が少ないため圧縮効率が高いと期待される．

表1: 仮名漢字変換辞書見出し語における2文字連接頻度

連接	出現数	連接	出現数
ん-	15315	し-	4373
う-	14534	と-う	4301
い-	10921	せ-い	3783
よ-う	10603	し-ゅ	3287
く-	7029	か-ん	3067
き-	6061	か-い	2861
ゅ-う	5654	か-	2784
り-	4901	う-し	2781

まず，われわれの研究室で利用可能な機械可読の自立語辞書[3]の見出し語(約74,000語)についてひらがな2文字間の連接の頻度を計数した．この結果を表1に示す．表1で連接の項目に1文字しかないものは，その文字で単語が終っているものの総数である．辞書引きでは，単語の区切りを検出する必要があるため，単語区切りも1つの文字と考えて連接個数を計算している．出現頻度数の合計は622171であった．連接頻度が多いもの上位40の合計で出現頻度が158505あった．これは，全体の1%で約25%の出現頻度を占めていることになる．方法1では，表1をもとに，ひらがな2文字単位でハフマン符号化した．符号は最短5ビット，最長24ビットであった．この符号化により，一度のマッチング処理で2文字の照合ができる．ひらがなの文字種は，「あ」「ゃ」「が」「ば」，および単語区切りも含めて84あるので，復号化には84×84の表を必要とする．

4. 辞書圧縮方法2

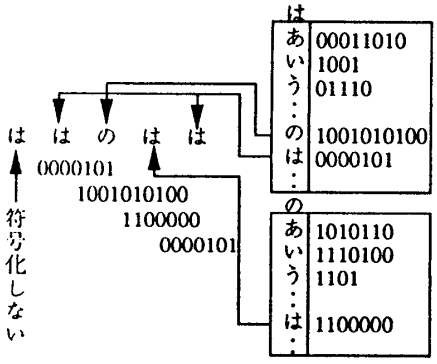
方法1では，2文字単位で符号化したため，奇数文字位置でのポインタの分岐が複雑になり，圧縮効率が良くない．そこで，1文字単位の圧縮を考える．圧縮方法は方法1と同様にハフマン符号化を採用する．方法2では直前の文字ごとにコード体系を変更する．言い替えれば，ひらがな84文字それぞれについて，その次に続く文字の出現頻度に対応して符号化を行なう．そしてその文字をその符号で置き換える．図2(a)に符号表を(b)に符号化の例を示す．この場合の符号は最短2ビット，最長10ビットであった．この方法でも，復号化のための表は方法1と同じく84×84の大きさが必要である．

TRIE構造では，各文字間がポインタで結ばれるため，単にTRIE構造の各文字を符号化しただけでは，圧縮効率はさほど高くならない．そこで，TRIE構造を図2(c)のように，リストの形式で持つことにすると，ポインタは不要となり，その分だけ圧縮が可能である．しかし，リスト形式では

Kana-to-Kanji dictionary using statistic data between two hiragana characters
Yutaka TSUTSUMI†, Koichi SHIMOZONO‡, Akira SUGANUMA‡, Kazuo USHIJIMA‡
†Ariake National College of Technology
‡Kyushu University

直前の文字	符号化する文字	符号
あ	ふ	1100001011
あ	と	1100001010
あ	お	1100000
ぃ	っ	00
か	な	1100000
か	え	01111

(a) 符号表の一部



(b) 符号化の例

(あ さ (朝)
(ひ (朝日)
(か (わ (旭川))))
し (足、脚))

(c) 図1のリスト表現

図 2: 圧縮法の概要

検索が遅くなるため，検索時には復号化とともにポインタ形式に復元する必要がある．

5. 辞書の構成

仮名漢字変換では，辞書サイズの問題から辞書を外部記憶装置に置くのが一般的である．そのため辞書検索の回数が多いほど，変換時間がかかる．ここでは，辞書の構成を見出し語と内容に分割する方式を採用し，外部記憶装置へのアクセスをできるだけ減らすようにしている．

見出し語は方法2で圧縮されたTRIE構造で構成され，候補となる品詞名と，漢字候補文字列へのポインタを持っていて，主記憶上に配置される．一方，候補となる漢字列は外部記憶装置上に配置され，見出し語からは，ポインタにより指されている．また，符号化された見出し語中の文字間のポインタについては，先頭から2文字までを従来の通りポインタ形式で表現し，3文字目以降をリスト形式で持たせている．これは，すべてをリスト形式で持たせると，復号化のために速度が犠牲になるためである．逆にすべてをポインタ形式で持たせると，符号化のメリットがなくなり，主記憶上で検索することが困難となる．

仮名漢字変換プログラムでは，文字列について候補となる品詞名だけで接続関係をチェックし，接続が可能な場合に限り外部記憶装置上の漢字文字列を取り出し表示する．これにより，外部記憶装置へのアクセスを大幅に減らすことが可能である．図3に辞書構成を示す．

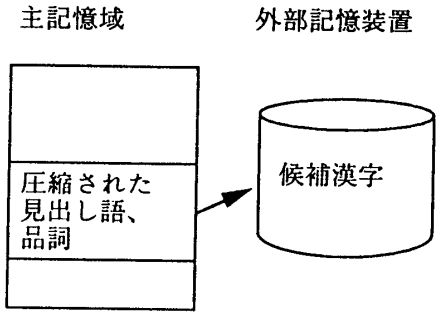


図 3: 辞書構成

表 2: 辞書圧縮の結果

辞書形式	文字のみ	見出し語全体
テキスト形式	1012256	1012256
TRIE 構造	403424	802808
方法 1	117741	322587
方法 2	98393	303239

単位:byte

6. 評価

表2に辞書圧縮の結果を示す．圧縮結果は，方法1より方法2が優れている．方法1では，文字数が奇数の場合に，圧縮を有効に使えないことも考えると，方法2の方が実用的である．見出し語全体について言えば，方法2では圧縮しないTRIE構造に比べて半分以下の大きさである．

7. まとめ

仮名漢字変換に限定すれば，辞書の見出し語はひらがなだけであるため，本稿で述べた2文字間の連接頻度を利用した符号化による辞書圧縮は，かなり有効であることが確かめられた．しかし，符号の長さが文字ごとに異なっているため，辞書を効率良く格納するためには，ビット操作が必要となり，速度が犠牲となる．6節で述べたように，圧縮しないTRIE構造に比べて，記憶容量が半分以下であるが，ポインタの格納方法を工夫することによりさらに圧縮効率を上げることが可能であろう．

また辞書構成についても見出し語と候補漢字を分離することで，外部記憶装置にアクセスする回数を減らすことができる．辞書見出し数が増えてもTRIE構造自体はそれほど大きくはならない．しかし，候補漢字は見出し語数に比例して多くなるため，この部分だけを外部記憶装置に格納することは仮名漢字変換システムでは有用である．

謝辞

この研究を行なうにあたり，機械可読辞書の利用に便宜を図って頂いた九州大学の日高達教授に感謝いたします．

参考文献

[1] 福島「大語彙かな漢字変換」，情処43回全国大会3H-9, 1991.

[2] 青江，森本「ダブル配列法によるトライ検索の実現法」，情処自然言語研究会91-NL-85-3, 1991.

[3] 吉田，日高他「公用データベース日本語単語辞書の使用について」，九州大学大型計算機センター広報16-4, pp.335-361, 1983.