4C−3

ワードプロセッサにおける 地名入力用 高精度かな漢字変換方式について

伊藤 純† 隈井 裕之† 中島 晃† 谷口 茂樹‡ 柏 博文‡

† ㈱日立製作所 マイクロエレクトロニクス機器開発研究所
‡ ㈱日立製作所 多賀工場

1．はじめに

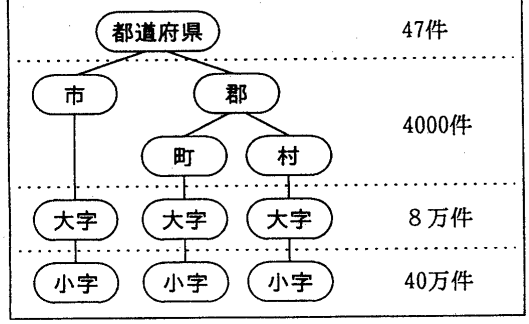
ワープロの普及に伴うユーザ層の拡大とともに誰でも簡単に入力できる文字入力方式の開発が重要なテーマとなっている。そこで我々は，「ＡＩかな漢字変換」や「モードレス入力」等を提案・開発してきた。

パーソナルワープロでは，「年賀状・はがき作成」のため，住所録機能の使用頻度が高い。ところが，従来のシステムでは，地名の変換率が低く，使いやすいとは言い難かった。原因は，一般文書のような文法構造を持たない地名の入力に，一般文章用の変換アルゴリズムや辞書を使用しているためである。そこで，今回，地名入力の変換率を大幅に高める専用辞書と変換アルゴリズムの開発を行なった。

2．地名の性質

日本の地名は，小字(こあざ)まで含めて約40万件存在し，同音異表記の地名も多い[1]。従って，単純に辞書語数を増やすのでは，実際に存在しない地名の組み合わせが候補に出力されるなど，変換率向上には限界がある。

図1　地名の階層構造



ところで，地名は図1に示すように，行政区分に従い階層構造を形成している。上位地名との接続まで考慮すると，「仙台市」「川内(せんだい)市」等のように同音異表記の地名も，ほぼ完全に正しい表記を得られることが分かった。

我々は，この点に着目し，地名入力における変換率100％の実現を考えた。

3．入力方式

本システムでは，以下の２つの地名入力方式を採用にした。

① **地名変換**：かな漢字変換を用いて地名を入力する入力方式。ワープロユーザにとって，最も自然な入力方式である。

② **郵便番号変換**：郵便番号変換は，郵便番号を入力し，郵便番号から検索される地名を表示する入力方式である。入力する文字数が少なく，簡単な入力方式である。

4．地名の高精度変換方式

上記入力方式を実現するために，(1)階層構造を保持した地名辞書，(2)地名用変換アルゴリズムを新たに開発した。

(1) 階層構造を保持した地名辞書

地名辞書は，以下の機能を満たすようにした。

・読み→地名の検索　（地名変換）
・郵便番号→地名の検索　（郵便番号変換）
・地名→郵便番号の検索
（入力した地名から郵便番号を自動表示する機能で使用）

このために，図2に示すような下位地名から上位地名を参照する構造で辞書を構築した。

(a)下位地名テーブル

郵便番号と大字(おおあざ)地名を組にして格納したテ

High-Level Kana-kanji Conversion for Address
Jun ITOH† , Hiroyuki KUMAI† , Akira NAKAJIMA† ,
Shigeki TANIGUCHI‡ , Hirofumi KASHIWA‡
† Microelectronics Products Development Laboratory , Hitachi, Ltd.
‡ Taga Works , HITACHI, Ltd.

ーブル。郵便番号順に配列したので郵便番号変換を高速に行なえる。また，かなから大字地名を検索し，対応する郵便番号を検索する事もできる。

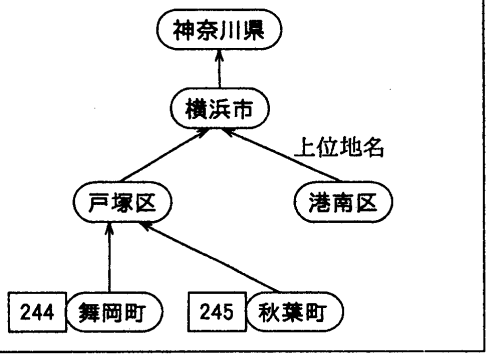
(b)読みテーブル

読みと地名を関連付けるテーブル。

(c)上位地名テーブル

県･市･区のような上位地名を，階層順に配列したテーブル。

図２　地名辞書構造 概念図



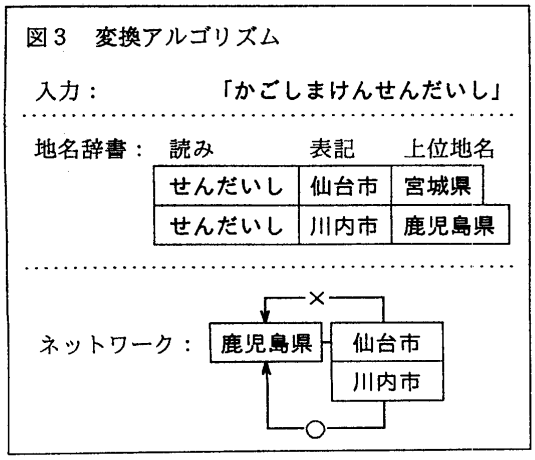
(2)地名用変換アルゴリズム

地名のかな漢字変換を高精度で行うために，以下のように変換候補を決定する。（図３参照）

日本の地名は，上位地名から入力するのが一般的である。これを用いて，先に辞書引きされている上位地名と，次の候補との接続検定を行なう。これにより，同音異表記の地名を正しく変換する。

例えば，「かごしまけんせんだいし」を入力する場合，「せんだいし」に対し「仙台市」「川内市」の候補があるとする。この時，先に入力されている候補「鹿児島県」を上位地名とする「川内市」のみが，接続チェックを満足し，候補として登録される。

図３　変換アルゴリズム



(3)課題：辞書容量圧縮

図１に示したように，小字まで含めると40万件あり，辞書容量は膨大となる。そこで地名辞書を圧縮するために，文字コードの圧縮について検討した。地名は，使用される漢字に偏りが大きい。例えば，「田」「西」「山」などは，高い頻度で地名に使用される。そこで，地名の漢字のうち，高頻度で出現する上位128字を選び，漢字コードの未使用領域を用いて１バイトに圧縮した。調査の結果，上位128字は地名で使用される全漢字の18.3％を占める。これらが１バイトコードで格納されるため，結果として９％の表記情報の圧縮ができた。

６．評価結果

大字レベル(８万件)の辞書を作成し，評価を行なった。表１は，辞書の容量別に，地名変換を適用した場合としない場合の変換率を示した表である。変換率は，全国から無作為に抽出した50件の地名を入力した結果である。辞書容量が少ない場合でも高変換率であるのは，小字の使用される頻度が低いためである。

表１　地名変換評価結果

		搭載地名数	変換率	備考
一般文書変換		4000件	75.3％	従来
高精度変換	FD用1MB辞書	８万件	95.2％	本ｼｽﾃﾑ
	HD用4MB辞書	(40万件)	(100％)	推定値

７．まとめ

地名入力に目的を絞った変換方式，及び８万件の地名辞書(１MB)を試作し，変換率95.2％を達成した。また，地名表記の偏りを利用して９％のデータ圧縮ができる事を示した。

今後は，大量データによる変換率評価と，使い勝手の評価をすすめるつもりである。

参考文献

[1]林大 他:図説日本語,角川書店,1982