16

帰納的学習による数字漢字変換手法における最上位階層語を利用した学習の有効性

松原雅文†*　荒木健治‡　桃内佳雄†　栃内香次‡
(北学園大工)†　(北大工)†

1　はじめに

近年，携帯端末の性能は飛躍的に進歩しており，小型化，高性能化が進んでいる．さらに，電子メールを意識して，通信機能を有した携帯端末もある．このような小型の端末は，その大きさに制約があるため，多数のキーを備えることができない．一般的な日本語入力方式であるローマ字入力においては，ローマ字を入力するために多数のキーが必要となる．また，精度の良いかな漢字変換を行うために，大きな辞書を持っているのが普通であり，端末の大きさから辞書容量にも制限がある小型の端末には不向きであると考えられる．少数のキーのみで入力が可能な方式として，現在の携帯電話等で用いられている文字循環指定方式がある．しかし，文字循環指定方式においては，かな１文字の入力に複数回の手数が必要となり，迅速な入力は難しい．小型の端末から日本語を入力する機会は増え，また迅速に入力したいという要求も高まっていることから，少数のキーのみで迅速な日本語入力が可能な手法が望まれる．

そこで，我々は従来より，小型の携帯端末での日本語入力を想定し，「文字情報縮退方式を用いた帰納的学習による数字漢字変換手法」を提案している[1][2][3]．我々が提案した手法は，文字情報縮退方式により入力された数字列を，漢字かな混じり文に変換する手法である．文字情報縮退方式により１つの数字にあ行，か行など５０音のかな一行を対応させることにより，１２個の数字キーのみで迅速な日本語入力を可能としている[4]．本手法においては，入力された数字列と人手により校正された校正済み変換結果から帰納的学習により語を獲得する．よって，辞書が空の状態からでも文脈に依存した語を獲得し，動的に対象に適応することができる[5][6]．本手法において入力に用いる数字列は，かなの母音情報が縮退しており，かなに比べて曖昧性が増している．このため，学習処理における語の獲得の際に，校正済み変換結果中の語の表記と，入力数字列中の数字列との対応関係を一意に決定できない場合が数多く存在し，システムが獲得できない語が増大する．対応関係が曖昧なために獲得されない語は，曖昧性のない他の文中から獲得可能であるが，より早い段階で語を獲得したほうが，変換精度は向上すると考えられるので，本手法においては，位置推測処理による語の獲得を行っている．しかし，この場合，誤って獲得する語が増大するという問題がある．

*matsu@ai.eli.hokkai-s-u.ac.jp
†札幌市中央区南 26 条西 11 丁目北海学園大学工学部
‡札幌市北区北 13 条西 8 丁目北海道大学工学部

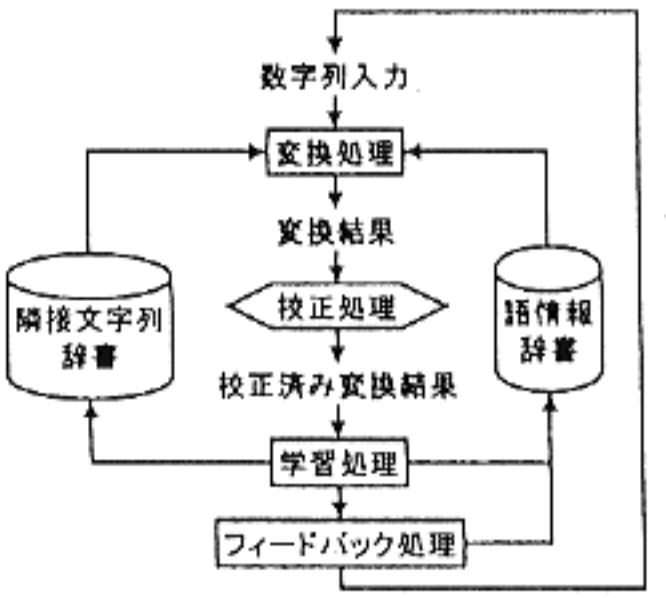

Fig. 1　処理過程

この問題を解決するため，最上位階層語を利用した語の獲得処理を本稿において提案する．最上位階層語とは，変換に用いられる階層化された辞書中の，最上位階層に登録されている語である．最上位階層は，変換精度 95[%] 以上の語が登録されている階層であり，最も確実性が高い階層である．最上位階層語が変換に使われた場合，最上位階層語の表記と数字列の対応関係が，校正済み変換結果と入力数字列中において確実なものとして決定され，語の獲得が行われる．これにより，学習時に発生する曖昧さが解消され，表記とそれに対応する数字列を正しく獲得できるので，変換精度の向上につながると考えられる．本稿では，最上位階層語を利用した語の獲得処理の概要と，本処理を行うことにより，最上位階層語を利用しなかった場合に比べて変換精度が向上することを，実験により確認した結果から述べる．

2　帰納的学習による数字漢字変換手法

本手法の処理過程を Fig. 1に示す．変換処理，校正処理，学習処理，フィードバック処理の順である．本手法において，使用者は１２個のキーのみによる文字情報縮退方式を用いて，日本語入力を行う．１２キーは数字で構成されている．数字とかなの対応関係を Table 1に示す．数字の１にあ行，２にか行のように，１つの数字に複数のかなが割り当てられている．これにより，少数のキーのみを使うにもかかわらず，１ストロークでかな１文字を入力でき，迅速な入力が可能である．本手法による変換例を Fig. 2に示す．Fig. 2に示されるように，本手法においては，まず，入力する日本語文のかなに対応した数字列を入力する．入力された数字列は，変換処理で語情報辞書と隣接文字列辞書を用いて漢字かな混じり

Table 1 数字とかなの対応関係

1:あいうえおー	2:かきくけこ	3:さしすせそ
4:たちつてと	5:なにぬねの	6:はひふへほ
7:まみむめも	8:やゆよやゆよ	9:らりるれろ
*:(半) 濁音	0:わをん	#:句読点

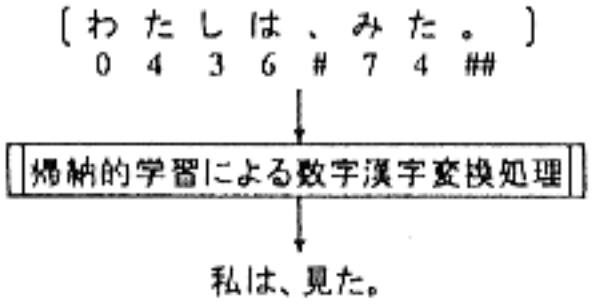

Fig. 2 変換例

文に変換される．語情報辞書は，語の獲得された状況とその変換精度によって，階層構造を持っており，上位階層の語から優先的にあてはめられ，変換が行われる．変換候補が重複した場合，語の正変換率，誤変換率，隣接文字列情報を利用して，最適な語を決定する．このように，変換は単純な語のあてはめだけではなく，隣接する文字列を考慮したものとなっている．変換が正しく行われなかった場合，校正処理を行う．人手により変換結果を訂正する過程である．学習処理では，入力数字列と校正済み変換結果との比較から，語を獲得する．この際に，本稿で提案する最上位階層語を利用した語の獲得を行う．最上位階層語を利用しても獲得できない語が存在した場合，位置推測処理により語の獲得を試みる．位置推測処理においては，語情報辞書中のすべての語から計算される，語の表記に対応する数字列の平均長を利用して語を獲得する．獲得された語は複数の語から構成されている可能性があるので，さらにそれらを共通，差異部分に分解し，語として辞書に登録する．同時に数字列，校正済み変換結果の全文字列を隣接文字列辞書に登録する．この登録された情報により，隣接する文字列を考慮した変換が可能となっている．フィードバック処理では，正変換，誤変換された語はその情報を語情報辞書に持ち，次回からの変換に役立てられる．また，正変換率により語が所属する階層を移動し，辞書の活性化を図っている．このように，変換処理，学習処理，フィードバック処理を繰り返し，変換精度が向上するのと同時に，対象，または使用者に合わせた辞書が生成されていく．

3 語の獲得

語の獲得は，学習処理で行われる．入力された数字列を，変換処理において変換した結果に誤りがある場合，校正処理において人手による訂正が行われ，システムは校正済み変換結果を得る．システムは，この入力数字列と校正済み変換結果との比較から語を獲得する．

Table 2 語の抽出例

入力数字列	296828104537
校正済み変換結果 (漢字数字混じり文)	彼 は 野球 を 楽 しむ 彼 6 野球 0 楽 37
抽出される語	
共通部分	差異部分
(6:は) (0:を) (37:しむ)	(29:彼) (8281:野球) (45:楽)

下線部分は共通部分を表す．

Table 3 共通部分が曖昧な例

入力数字列	828102030 39
校正済み変換結果 (漢字数字混じり文)	野球 を 観戦する 野球 0 観戦 39
変換結果	8281 0 [観戦] 39

下線部分は共通部分の候補を表す．
[] は最上位階層語を表す．

3.1 曖昧さのない共通部分からの語の獲得

まず，入力数字列と校正済み変換結果との字面上の比較から，共通部分，差異部分を抽出する．ここで，共通部分とは，校正済み変換結果中のかなと，それに対応する入力数字列中の部分数字列である．差異部分とは，共通部分に挟まれる部分である．共通部分，差異部分の入力数字列と校正済み変換結果，それぞれにおける対応関係は，その出現順に決定される．

共通部分が一意に決定できる場合の，語の抽出例をTable 2に示す．Table 2において，校正済み変換結果中のかなと，それに対応する入力数字列中の部分数字列は，漢字数字混じり文から分かるように，(6:は)，(0:を)，(37:しむ) である．ここで，漢字数字混じり文とは，漢字かな混じり文である校正済み変換結果のかなを，それに対応する数字に置き換えたものである．差異部分は，共通部分に挟まれている (29:彼)，(8281:野球)，(45:楽) となる．このように，数字列と表記文字列の対応関係は出現順に決定される．

3.2 最上位階層語を利用した語の獲得

共通部分が曖昧な例をTable 3に示す．Table 3において，校正済み変換結果中の「を」に対応する数字「0」の位置が，入力数字列中に3箇所存在し，共通部分が曖昧となっている．このように曖昧さを含んでいる場合，共通部分を決定できないので語を抽出することができない．

そこで，最上位階層語を利用した語の獲得を試みる．Table 3の変換結果中の [観戦] は辞書中の最上位階層の語である．よって，校正済み変換結果中の「観戦」と入力

Table 4 実験データ

UNIX オンラインマニュアル	
項目名	文字数
1 ftp	11,000
2 mail	15,000
3 cc	8,000
4 csh	16,000
合計	50,000

Table 5 平均正変換率 [%]

項目名	最上位語利用	
	あり	なし
1 ftp	68.9	60.7
2 mail	78.5	74.1
3 cc	69.5	63.5
4 csh	78.9	68.9
全平均	75.1	67.8

数字列中の「2030」の対応関係が確実であるものと決定することにより，差異部分 (82810:野球を) から「野球」，「を」を抽出することができる．このようにして，本処理により，曖昧さを解消し，システムが獲得する語数の増大を図っている．

4 評価実験

処理概要に基づき，実験システムを作成し，評価実験を行った．最上位階層語を利用した語の獲得の有効性の確認のため，学習処理における語の獲得の際に，最上位階層語を利用した場合と，利用しない場合での変換精度の比較実験を行っている．実験に用いたデータを Table 4に示す．実験データには UNIX のオンラインマニュアルの項目を用いている．このデータ中に出現する語は，「UNIX」に関して限られており，これらのデータは限られた対象であるといえる．実験は，辞書が空の状態から 1〜4 の順に行った．1 文単位で Fig. 1に示す処理を行い，1,000 文字単位で変換精度の評価を行っている．変換精度の評価は，以下に示される正変換率により行う．

$$\text{正変換率} = \frac{\text{正変換文字数}}{\text{入力文字数}} \quad (1)$$

入力文字数に対する正変換文字数の割合である．

5 実験結果

最上位階層語を利用した場合と，利用しない場合の正変換率の推移を Fig. 3に示す．また，それぞれの平均正変換率を Table 5に示す．Fig. 3から分かるように，ほぼ全体を通して，最上位階層語を利用した方が変換精度が高い．平均正変換率では，最上位階層語を利用した方

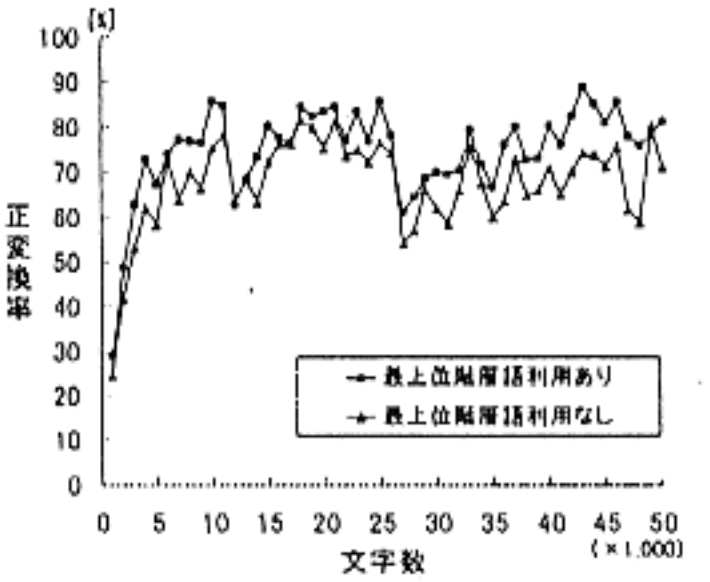

Fig. 3 正変換率の推移

Table 6 辞書登録語数

項目名	最上位語利用	
	あり	なし
1 ftp	593	585
2 mail	932	924
3 cc	1,166	1,154
4 csh	1,515	1,515

が，全体で 7.3 ポイント高くなっている．また，それぞれの項目の実験終了時までに，語情報辞書に登録された語数の累計を Table 6に示す．実験を終えた最終的な辞書中の語数は同数となっている．

6 考察

最終的な辞書登録語数は，最上位階層語を利用した場合と，利用しない場合において同数であるが，正変換率は最上位階層語を利用した方が高くなっている．これは，位置推測処理により語を獲得しているためであると考えられる．学習処理で曖昧な共通部分が存在するとき，最上位階層語を利用しない場合でも，位置推測処理により語を獲得している．しかし，位置推測処理による語の獲得．においては，語の表記とそれに対応する数字列の獲得を誤る場合がある．位置推測処理により語の獲得を誤った実験結果を Table 7に示す．入力文字数 5,000〜6,000 文字中のデータである．Table 7において，最上位階層語を利用しない学習処理の場合，校正済み変換結果中の「は」に対応する数字「6」の位置が，入力数字列中で一意に決定できないので，位置推測処理によりこの位置が決定される．このとき「6」の位置は，正しくは文頭から 7 番目の位置であるが，推測を誤り 8 番目となっていた．このため，語の獲得を誤り，(6*116:場合)，(813*80581982:標準入力) を獲得していた．このような誤って獲得した語も辞書に登録されるので，最上位階層語を利用しない場合の辞書中には，誤って獲得した語が数多く含まれている．しかし，最上位階層語を利用した場合，(6*11:場

Table 7 位置推測を誤る例

入力数字列
256*1166813*80581982
変換結果
25[場合]66 ユーザ 4 入力
校正済み変換結果
この場合 は 標準入力

下線部分は共通部分の候補を表す.
[] は最上位階層語を表す.

合) の対応関係が，入力数字列と校正済み変換結果中において確実なものと決定されるので，(6813*80581982:標準入力) を正しく獲得することができている．このように，最上位階層語を利用することにより，正しい語を数多く獲得することができる．よって，辞書登録語数は同数であるにも関わらず，最上位階層語を利用した方が変換精度が向上している．

7 まとめ

本稿では，我々が従来より提案している「文字情報縮退方式を用いた帰納的学習による数字漢字変換手法」の学習処理において，語の獲得を効率よく行うため，最上位階層語を利用した語の獲得処理を提案した．入力数字列と校正済み変換結果の字面情報のみでは誤って獲得される語が，変換結果の字面情報の最上位階層語を利用することにより，正しく，より早い段階で獲得できるようになった．より早い段階で獲得された語は，より早い段階で変換に使用され，それが正変換であれば優先度が上がる．優先度が上がり，最上位階層に移動した語は，さらに学習処理で利用され，正しい語の獲得につながる．このように，最上位階層語を利用することにより，変換精度が向上する．平均正変換率で，7.3 ポイントの向上が見られ，最上位階層語を利用した学習の有効性が確認された．

本稿で提案した処理においても，利用している情報は字面情報のみである．しかし，人手で行われる校正処理からは，より多くの情報を得ることができると考えられる．校正処理における訂正の手順などの情報を用いることにより，訂正箇所の表記と数字列を一意に決定でき，効率よく学習を行うことができると考えられる．よって今後は，校正処理における人手による訂正の情報を活用し，さらに実用性を高めたシステムを構築する予定である．

謝辞 なお，本研究の一部は文部省科学研究費 (No.09878070, No.10680367) 及び北海学園大学ハイテク・リサーチ・センター研究費による補助のもとに行われた．

参考文献

[1] 松原 雅文, 荒木 健治, 桃内 佳雄, 栃内 香次: 文字情報縮退方式を用いた帰納的学習によるべた書き文のかな漢字変換手法の性能評価, 情報処理学会自然言語処理 研究報告, Vol.98 (98-NL-128), pp.1-7(1998).

[2] 松原 雅文, 荒木 健治, 桃内 佳雄, 栃内 香次: 文字情報縮退方式を用いた帰納的学習によるべた書き文のかな漢字変換手法の変換精度, 平成 10 年度電気関係学会北海道支部連合大会講演論文集, pp.365-366(1998).

[3] 松原 雅文, 荒木 健治, 桃内 佳雄, 栃内 香次: 文字情報縮退方式を用いた帰納的学習によるべた書き文のかな漢字変換手法, 情報処理学会 第 57 回 全国大会 講演論文集 (2), pp.197-198(1998).

[4] 佐藤 亨, 東田 正信, 林 智定, 奥 雅博, 村上 仁一: PB 電話機を利用した日本語入力方式, 1997 年電子情報通信学会総合大会,D-6-6, pp.102(1997).

[5] Kenji Araki, Yoshio Momouchi and Koji Tochinai.: Evaluation for adaptability of Kana-Kanji translation of non-segmented Japanese Kana sentences using inductive learning, Conference Working Papers of PACLING-II, pp.1-7, Australia (1995).

[6] 荒木 健治, 高橋 祐治, 桃内 佳雄, 栃内 香次: 帰納的学習を用いたべた書き文のかな漢字変換, 電子情報通信学会論文誌 (D-II), J79-D-II, No.3, pp.391-402(1996).