論　文

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数字漢字変換手法の有効性について

松原　雅文[†]　荒木　健治[††]　桃内　佳雄[†]　栃内　香次[††]

Effectiveness for Non-Segmented Number-Kanji Translation Method Using Inductive Learning with Degenerated Keyword Input

Masafumi MATSUHARA[†], Kenji ARAKI[††], Yoshio MOMOUCHI[†], and Koji TOCHINAI[††]

あらまし　本手法は，文字情報縮退方式により入力された数字列を漢字仮名交じり文に変換するものである．文字情報縮退方式により一つの数字にあ行，か行など 50 音の仮名 1 行を対応させることにより，迅速な入力を可能としている．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．入力が母音情報の縮退した数字の列であるため，仮名に比べてあいまいさが生じている．このあいまいさを解消するため，本手法においては，隣接文字列情報，最上位階層語，位置推測処理を利用している．最上位階層語の利用と位置推測処理により，獲得される語数の増大を図っており，隣接文字列情報により，語のつながりを考慮した変換が可能となっている．

キーワード　帰納的学習，数字漢字変換，隣接文字列情報，最上位階層語，位置推測

1. まえがき

近年，携帯端末の性能が飛躍的に進歩している．携帯性を重視して洋服のポケットに入るほど小さな端末もあり，その中にはインターネット，電子メールを意識して，通信機能を有しているものがある．このような端末は，携帯性を重視しているのでその大きさに制約があり，大きなキーボード，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方式においては，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，精度の良い仮名漢字変換を行うために，大きな辞書をも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，仮名 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．このような小型の携帯端末で，電子メールなど日本語を入力する機会が増え，また迅速に入力したいというような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．従来より，我々は，人間の言語及び知識獲得能力の解明とその工学的応用を目的として，「帰納的学習を用いたべた書き文の仮名漢字変換」を提案している [1], [2]．この手法においては，帰納的学習により語を自動的に獲得するので，使用者，または対象にシステムが動的に適応し，適応した辞書を自動生成することができる．この手法のもつ適応能力は，個人使用が多く，記憶容量にも制約がある携帯端末には適していると考えられる．

そこで，この手法を応用し，小型の携帯端末での日本語入力を想定した工学的なシステムの実現に向けて，「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」を本論文で提案する [3]〜[5]．本手法においては，迅速な入力を可能とするため，入力には文字情報縮退方式による数字列を用いており [6]，この入力された数字列を漢字仮名交じり文に変換する．

[†] 北海学園大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkai-Gakuen University, Minami 26 Nishi 11, Chuo-ku, Sapporo-shi, 064–0926 Japan
[††] 北海道大学大学院工学研究科，札幌市
Graduate School of Engineering, Hokkaido University, Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060–8628 Japan

仮名を入力して漢字仮名混じり文に変換する「仮名漢字変換」に対して，入力が数字である本手法の変換を「数字漢字変換」と呼ぶ．変換を誤った語はそのゆう度を下げ，正しく変換した語はそのゆう度を上げることにより，文脈に依存した語を次回から優先的に当てはめることができる．また，本手法においても，入力された数字列と，人手により訂正された校正済み変換結果との比較から，帰納的学習により語を獲得するので，辞書が全く空の状態からでも，文脈に依存した辞書が自動生成される．よって，辞書の見出し語そのものを学習し，使用者，または対象に根本的，かつ動的に適応することができる [1], [2]．文字情報縮退方式による入力は，母音情報が縮退しており，数字 1 文字の情報量は，仮名 1 文字の情報量と比べて少ない．そのため，縮退により失われた情報をいかにして回復するかが問題となる．この問題に対して，本手法では帰納的学習による語の獲得に加えて，隣接文字列情報，最上位階層語，位置推測処理を利用することとした．最上位階層語の利用と位置推測処理により，語の獲得時，情報縮退のために発生するあいまいさを極力解消し，獲得される語数の増大を図っている．隣接文字列情報は n-gram 統計により抽出される [7]〜[10]．この情報により変換候補となる語単体のゆう度だけでなく，先行，後続の文字列とのつながりを考慮した変換が可能となっており，同文中に出現する同音異義語にも対処できる．これらはヒューリスティックスであるが，統計的な情報から得られるものなので，本手法の汎用性を保持することができると考えられる．

本論文では，本手法の概要，及び，本手法に基づくシステムが帰納的学習を用いて種々の対象に適応できることを，実験により確認した結果から述べる．

2. 概　　要

本手法において使用者は 12 個のキー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．数字と仮名の対応関係を表 1 に示す．数字の 1 にあ行，2 にか行のように，一つの数字に複数の仮名が割り当てられている．そのため，少数のキーのみを使うにもかかわらず，1 ストロークで仮名 1 文字が入力でき，迅速な入力が可能である．本手法による変換例を図 1 に示す．このように，使用者が意図する日本語文の仮名に対応した数字列を入力し変換を行う．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仮名混じり文に変換される．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，階層構造を有しており，上位階層の語から優先的に当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，変換は単純な語の当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，語を抽出する．ここで抽出できなかった語については，最上位階層語の利用と位置推測処理により語の獲得を試みる．これにより，獲得できる語数の増大を図っている．抽出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，更にそれらを共通，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もち，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．

表 1　数字と仮名の対応関係
Table 1　Correspondence of number to kana.

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:やゆよゃゅょ	9:らりるれろ
∗:（半）濁音	0:わをん	#:句読点

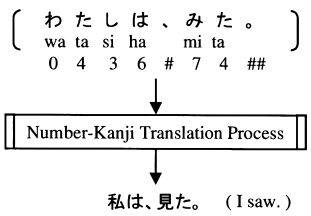

図 1　変　換　例
Fig. 1　Example of translation.

3. 処 理 過 程

本手法における処理過程を図 2 に示す．このように，変換処理（Translation Process），校正処理（Proofread Process），学習処理（Learning Process），

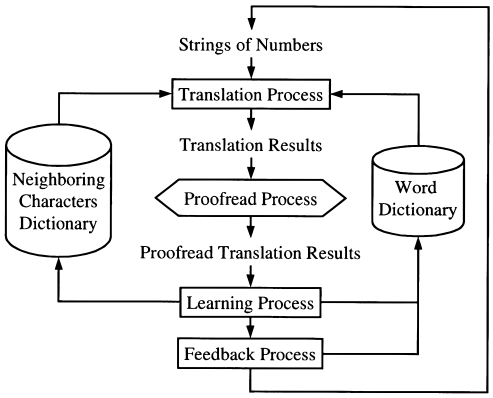

図 2　処 理 過 程
Fig. 2　Procedure.

フィードバック処理（Feedback Process）の順である．

3.1 変 換 処 理

入力は 12 個のキーを用いて，べた書きの数字列で行う．本処理過程では，この入力された数字列を漢字仮名混じり文に変換する．

3.1.1 語の当てはめ

入力された数字列に対して，語情報辞書に登録されている語を当てはめて変換を行う．語情報辞書は，語が獲得された状況と，その変換精度により，階層構造を有している．**3.3**，**3.4** で述べるように，確実性の高い階層から，MS（The Most Certain Segment），CS（Common Segment），S1（Segment One），RS（Remained Segment），LS（The Least Certain Segment）である．語の当てはめは上位の階層より，語の読みにあたる数字列の長さが 2 文字以上のものから行う．これは 1 文字語のもつ情報量が，2 文字以上の語のもつ情報量と比べて極端に少ないからである．この情報量の少なさを補うために，1 文字語の変換はその両側の語が確定済みのときにのみ行うものとする．変換候補が重複した場合，**3.4** で述べる語の正変換率 CR，誤変換率 ER，更に，**3.3.6** で述べる隣接文字列存在度 ND を用いてゆう度評価を行い，最適な語を決定する．

3.1.2 ゆう度評価

変換候補重複の際に，正しい語を決定するゆう度評価関数 CEF（Credibility Evaluation Function）を式 (1) に示す．

$$CEF = \alpha \times ND + \beta \times CR - \gamma \times ER \qquad (1)$$

α, β, γ：係数

表 2　変換候補が重複する例
Table 2　Example of translation candidates overlap.

変換対象	238281 チーム	
変換候補	（281:今日）	（8281:野球）
先行文字列	238	23
後続文字列	チーム	チーム

この式により先行，後続の文字列に隣接する度合と正変換率が高く，誤変換率が低い語のゆう度が高くなり，優先されて選択される．

例として，「238281417」を変換することを考える．文脈に依存した正しい日本語文は「草野球チーム」である．まず，数字列が 2 文字以上の語で変換が行われる．ここでは，（417:チーム）が変換されるものとする．変換結果は，「238281 チーム」となる．このとき，辞書中に（8281:野球），（281:今日）が登録されており，変換の際に変換候補の重複が発生する．これを，表 2 に示す．語の正変換率は，それぞれ同程度である．（281:今日）を当てはめた場合の先行文字列は「238」，後続文字列は「チーム」である．同様に，（8281:野球）を当てはめた場合の先行文字列は「23」，後続文字列は「チーム」である．隣接文字列辞書よりそれぞれの文字列の隣接の度合を求めると，「今日」に「238」が先行する確率よりも，「野球」に「23」が先行する確率が高く，また，「今日」に「チーム」が後続する確率よりも，「野球」に「チーム」が後続する確率の方が高いことがわかる．よって，式 (1) のゆう度評価関数に適用すると，「野球」が優先され，正しく変換することができる．このように，隣接する文字列を考慮した変換を行っている．

3.2 校 正 処 理

変換結果に誤りが含まれている場合，校正処理が行われる．人手により変換結果を訂正する過程である．この処理により，システムは校正済み変換結果を得る．

3.3 学 習 処 理

本処理過程では，入力数字列と校正済み変換結果から語を獲得し，語情報辞書に登録する．

3.3.1 語候補の抽出

まず，入力数字列と校正済み変換結果と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仮名と，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．このようにして

表 3　S1 の抽出例
Table 3　Example of the extraction of S1.

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 S1	
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.

抽出される部分を語候補 S1（Segment One）と呼ぶ．S1 の抽出例を表 3 に示す．表 3 において，校正済み変換結果中の仮名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わかるように，(6:は)，(0:を)，(37:しむ）である．ここで，漢字数字混じり文とは，漢字仮名混じり文である校正済み変換結果の仮名を，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. 3. 2　あいまいな共通部分からの語候補の抽出

表 3 では，共通部分が一意に決定できるので問題ないが，校正済み変換結果中の仮名に対応する数字列が，入力数字列中に複数存在することも考えられる．表 4 に共通部分があいまいな例を示す．このようにあいまい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，本手法においては，あいまいさのない語だけを抽出するために，左から右方向の解析，右から左方向の解析を行い，一致したものだけを抽出している．この例を表 5 に示す．表 4 において，校正済み変換結果中の「を」に対応する数字「0」は，入力数字列中の 3 箇所に存在している．このため，表 5 の左から右方向の解析結果と右から左方向の解析結果で「野球」,「観戦」において不一致が生じている．ここで，右から左方向の解析において，校正済み変換結果中の「を」に対応する入力数字列中の「0」の位置を右から 5 番目に決定している．これは，校正済み変換結果中の仮名漢字に対応する数字が，入力数字列中に必ず存在するものとしているためである.「を」に対応する共通部分「0」の位置を入力数字列中の右から 3 番目に決定すると，校正済み変換結果中の差異部分である「観戦」に対応する数字列が入力数字列中に存在しなくなるため，共通部分の位置を右から 5 番目に決定しているのである．このような解析により，両方向からの解析結果が一致している，すなわち対応関係が一意に決定している語のみを獲得する．よって，ここで獲得される語は，(29:彼)，(6:は)，(0:を)，(39:する）である．ここで獲得されなかった語は，他のあいまい性のない文中より獲得されると考えられる．しかし，より早い段階で獲得された方が変換精度の向上に役立ち，また，入力文字列の情報縮退によりこのような獲得されない語は増加すると考えられるので，変換結果の最上位階層語を利用した語の獲得を試みる．

表 4　共通部分があいまいな例
Table 4　Example of the common segments with ambiguity.

入力数字列	29682810203039
校正済み変換結果 （漢字数字混じり文）	彼は野球を観戦する 彼 6 野球0観戦 39

下線部分は共通部分の候補を表す.

表 5　あいまいな共通部分からの S1 の抽出例
Table 5　Example of the extraction of S1 from the ambiguous common segments.

左から右方向の解析	
入力数字列	29682810203039
校正済み変換結果 （漢字数字混じり文）	彼は野球を観戦する 彼6野球0観戦39
左から右方向の解析結果	
共通部分	差異部分
（6:は） （0:を） （39:する）	（29:彼） （8281:野球） （2030:観戦）
右から左方向の解析	
入力数字列	29682810203039
校正済み変換結果 （漢字数字混じり文）	彼は野球を観戦する 彼6野球0観戦39
右から左方向の解析結果	
（39:する） （0:を） （6:は）	（30:観戦） （828102:野球） （29:彼）
獲得される S1	
(29:彼)，(6:は)，(0:を)，(39:する）	

3. 3. 3　最上位階層語の利用

変換に用いられた最上位階層語は，確実性の高い語であるから，この語の表記と数字列との対応が確実であるものとして，システムは入力数字列と校正済み変換結果から語を獲得する．最上位階層語とは，語情報辞書の最上位の階層である MS 階層に登録されている語である．表 4 において，共通部分を一意に決定することができない部分である（828102030:野球を観戦）に対して本処理を適用する．ここでは，MS 階層の語である（2030:観戦）が変換に使われていたものとする．この場合，入力数字列中の「2030」と校正済み変

換結果中の「観戦」の対応が確実であるものとして決定され，残りの部分である（82810:野球を）から「野球」,「を」を語として獲得することができる．変換結果中に MS 階層の語が存在する場合には，このように語の獲得を行うことが可能である．しかし，変換に MS 階層の語が使われていない場合も考えられるので，その場合には位置推測処理により，語の獲得を行う．

3.3.4 位置推測処理

変換に最上位階層語が使われなかった場合，若しくは，最上位階層語を用いても獲得できない語が存在した場合には，語情報辞書に登録されているすべての語を利用して，対応する共通部分の位置を推測する．推測は，語の数字列の平均長を利用して行われる．語の数字列の平均長は，語情報辞書に登録されているすべての語から表記文字数ごとに求められる．表 4 において，共通部分を一意に決定することができない部分である（828102030:野球を観戦）に対して本処理を適用する．ここでは，MS 階層の語は変換に使われていないものとする．校正済み変換結果中の「を」に対応する数字「0」が，入力数字列中で一意に決定できないので，この位置を推測する.「野球」に対応する数字列の長さが決定できると,「0」の位置を決定することができるので，ここでは表記の文字数が 2 である語の数字列の平均長を，語情報辞書中のすべての語から求める．数字列の平均長 4 が得られると,「野球」に対応する数字列の長さを 4 と推測できる．これにより，入力数字列中の共通部分「0」の位置が左から 5 番目に決定され,（8281:野球),（2030:観戦）を獲得することができる．ここで獲得される語は必ずしも正しいとは限らないが，誤って獲得された語は **3.4** で述べるフィードバック処理によりそのゆう度が下がり，次第に淘汰されていく．

3.3.5 語候補からの語の獲得

抽出された S1 は，更に共通部分，差異部分に分解され，辞書に登録される．これは抽出された語が複数の語により構成されている可能性があるからである．ここでの共通部分とは，一方の語がもう一方の語を完全に含んでいる場合の含まれている部分であり，差異部分とは，その残りの部分である．共通部分を CS（Common Segment），差異部分を RS（Remained Segment）と呼ぶ．獲得された語は，それぞれ階層 CS，RS に登録される．CS, RS に分解された S1 は辞書中より削除される．表 6 にその例を示す．表 6 において,（8281:野球）は（82813∗81:野球場）に数字列，表記文字列と

表 6　CS，RS の抽出例
Table 6　Example of the extraction of CS and RS.

対象となる S1	
S1	（野球場：82813∗81）
S1	（野球:8281）
共通部分	
CS	（野球：8281）
差異部分	
RS	（場：3∗81）

もに完全に含まれている．よって，共通部分 CS として（8281:野球），差異部分 RS として（3∗81:場）が抽出され，辞書のそれぞれの階層に登録される．

3.3.6 隣接文字列情報の獲得

同時に，ここで使用した入力数字列，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．隣接文字列辞書から獲得される文字列の隣接の度合は，**3.1.2** で述べたゆう度評価に利用される．隣接文字列存在度 ND（Degree of Neighboring Character Strings）を，式 (2) に示す．文字列 a_x に対する隣接の度合である．

$$\begin{aligned} &\text{文字列} \quad \cdots a_{x-1} \cdot a_x \cdot a_{x+1} \cdots \\ &ND(a_x) = len(a_{x-1}) \times P_{l(a_x)}(a_{x-1}) \\ &\qquad\qquad + len(a_{x+1}) \times P_{r(a_x)}(a_{x+1}) \qquad (2) \end{aligned}$$

ここで，$len(X)$ は，文字列 X の長さを表し，$P_{l(X)}(Y)$ は X の左確率分布の Y の値，$P_{r(X)}(Y)$ は X の右確率分布の Y の値をそれぞれ表している．隣接文字列 a_{x-1}，a_{x+1} は長いほど多くの隣接情報を有しており，有用度が高いと考えられるので，ここでは，隣接文字列辞書に存在する文字列のうち，最も長い文字列を適用するものとしている．また，長い文字列は短い文字列に比べて出現頻度の点において不利になるので，これを補うために，文字列の長さを重みとして掛けるものとしている．このように a_x に対して，先行文字列 a_{x-1}，後続文字列 a_{x+1} が存在する確率を求め，隣接文字列の長さに対応した重み付けをして，足し合わせた値により，先行，後続する文字列とのつながりの確からしさの度合を抽出している．

3.4 フィードバック処理

語を語情報辞書に獲得する際に，その語の情報をともに獲得する過程である．獲得する情報は，正変換，誤変換のそれぞれの度数である．これらをそれぞれ CF（Frequency of Correct Translation），EF（Frequency of Erroneous Translation）と呼ぶ．CF，EF より求められる語の正変換率 CR（Rate of Cor-

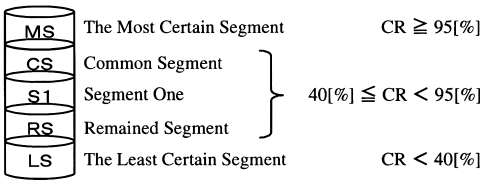

図 3　語情報辞書の階層
Fig. 3　Ranks in the word dictionary.

rect Translation），誤変換率 ER（Rate of Erroneous Translation）を，式 (3)，(4) に示す．

$$CR = \frac{CF}{CF + EF} \tag{3}$$

$$ER = \frac{EF}{CF + EF} \tag{4}$$

ここで，辞書中の語のうち，正変換率が 95.0[%] 以上になったものを MS（The Most Certain Segment），40.0[%] 未満のものを LS（The Least Certain Segment）と呼び，それぞれの階層に移動する．また，MS 階層の語のうち正変換率が 95.0[%] 未満，LS 階層の語のうち正変換率が 40.0[%] 以上となったものは，もとの階層に戻される．語情報辞書の階層構造を図 3 に示す．このように語情報辞書の階層は，上位の階層より，MS，CS，S1，RS，LS となっており，辞書の効率化，活性化を図っている [1], [2]．CR，ER は，**3. 1. 2** で述べたゆう度評価にも利用される．

4. 評価実験

処理概要に基づき，実験システムを作成した．字面情報から獲得できる語の有効性の確認を目的とし，評価実験を行っている．

4. 1　実験データ及び実験手順

実験に用いたデータを表 7 に示す．大きな対象として，UNIX オンラインマニュアルと論文の分野があり，その中にそれぞれの項目が含まれている．実験は，辞書が空の状態から表 7 に示される 1〜7 の順に行っている．「論文 1」は文献 [11] であり，「論文 2」(注 1)，「論文 3」(注 2) は「論文 1」と同一筆者により，同一研究に関して書かれた論文である．よって，分野「UNIX」，「論文」は，ともに対象が限定されているといえる．

初期辞書に語を登録しておいた場合，登録されている語が対象となる実験データに適合しているときには，実験の早い段階から高い変換精度を得ることができる．しかし，実際には，どのような対象のデータが入力されるかは予測不能であり，辞書に登録されている語に適合しない異なった対象のデータが入力された場合，迅速な適応の妨げとなり，適応能力の評価に影響を与えると考えられる．よって，初期状態を一定にし，種々の対象に適応可能な本手法の適応能力の確認のため，実験は辞書が空の状態から行うものとした．1 文単位で図 2 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，誤変換率，未変換率により行う．

表 7　実験データ
Table 7　Data of the experiment.

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
		50,000
論文		
5	論文 1	23,000
6	論文 2	32,000
7	論文 3	17,000
		72,000
合計		122,000

$$\text{正変換率} = \frac{\text{正変換文字数}}{\text{入力文字数}} \tag{5}$$

$$\text{誤変換率} = \frac{\text{誤変換文字数}}{\text{入力文字数}} \tag{6}$$

$$\text{未変換率} = \frac{\text{未変換文字数}}{\text{入力文字数}} \tag{7}$$

それぞれ，入力文字数に対する入力数字列中の正変換文字数，誤変換文字数，未変換文字数の割合である．また，ゆう度評価関数である式 (1) の係数は予備実験より以下の値を用いた．

$$\alpha = 10,\ \beta = 1,\ \gamma = 5$$

4. 2　予備実験

式 (1) の係数を決定するため，予備実験を行った [12]．実験に用いたデータは，分野「UNIX」の項目「ftp」の 10,000 文字である．$\beta = 1, \gamma = 5$ として，α の値を変化させて実験を行った．辞書は空の状態から実験を行っている．実験結果を表 8 に示す．表 8 より，平均正変換率が最大となった $\alpha = 10$ を適用した．

(注 1)：題名は，「学習型機械翻訳手法 GA-ILMT における帰納的学習を用いた淘汰手法の有効性について」である．
(注 2)：題名は，「旅行用英会話文を用いた GA-ILMT の性能評価」である．

表 8　予備実験結果
Table 8　Result of the preliminary experiment.

α	1	2	5	10	20	30	50
平均正変換率 [%]	52.0	51.9	52.7	53.6	52.8	52.7	52.6

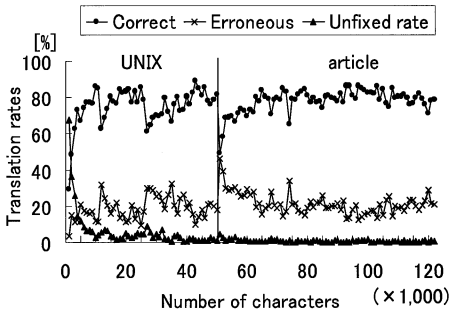

図 4　変換率の推移
Fig. 4　Changes in the translation rates.

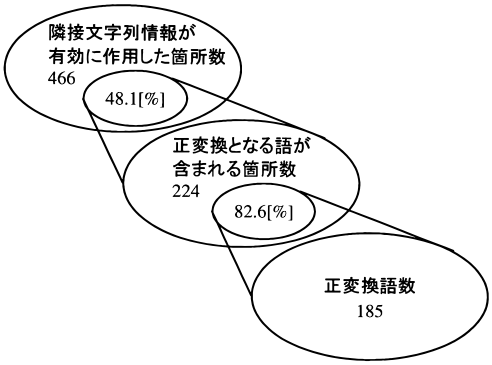

図 5　隣接文字列情報を利用した変換語数
Fig. 5　The number of the translated words by using neighboring character strings.

表 9　平均変換率
Table 9　Means of the translation rates.

UNIX オンラインマニュアル			
項目名	平均変換率 [%]		
	正変換	誤変換	未変換
1　ftp	68.9	13.8	17.3
2　mail	78.5	17.5	4.0
3　cc	69.5	26.1	4.4
4　csh	78.9	19.5	1.6
	75.1	18.7	6.2
論文			
5　論文 1	73.5	25.3	1.2
6　論文 2	80.7	18.8	0.5
7　論文 3	78.5	21.1	0.4
	77.8	21.4	0.8
全平均	76.7	20.3	3.0

表 10　獲 得 語 数
Table 10　The number of the acquired words.

分野	MS	CS	S1	RS	LS	累計
UNIX	86	243	435	255	505	1,524
論文	173	352	756	341	970	2,592

表 11　最上位階層語を利用した獲得語数
Table 11　The number of the acquired words by using MS.

	語数	割合 [%]
正獲得	240	100.0
誤獲得	0	0.0
合計	240	100.0

4.3 実 験 結 果

各変換率の推移を図 4 に示す．また，各平均変換率を表 9 に示す．図 4 からわかるように，辞書が空の状態から入力文字数の増加に伴い，徐々に正変換率が上昇している．項目や分野など，対象が変化するときに正変換率はいったん下降するが，その後，再び正変換率は上昇する．それぞれの分野において，85[%] 程度までの正変換率の上昇が確認された．また，表 9 からわかるように，未変換率は全体を通して非常に低い値となっている．全体の平均未変換率は 3.0[%] であった．変換時に変換候補が重複しゆう度評価を行う際に，隣接文字列情報が有効に作用した箇所数を図 5 に示す．これは，変換候補として語 w_1，w_2 が重複した際に，語単体の変換精度では w_1 の優先度の方が高いが，本手法における隣接文字列情報を加味することにより，w_2 の優先度が高くなるような箇所である．図 5 からわかるように，重複した変換候補の中に正しい語が含まれている場合，82.6[%] の割合で正しい語を選択することができている．

それぞれの分野の終わりまでに，語情報辞書に登録された語数の累計を表 10 に示す．一般的な仮名漢字変換システムの辞書に比べて，登録されている語数が非常に少数であるのがわかる．最上位階層語の利用により，システムが獲得した語数を表 11 に示す．また，位置推測処理により，獲得した語の延べ数を表 12 に示す．最上位階層語を利用した語の獲得において，正しい語を獲得できる精度は，100.0[%] であり，位置推測処理においては，88.9[%] であった．獲得された語

表 12　位置推測処理による獲得語数
Table 12　The number of the acquired words by position prediction.

	語数	割合 [%]
正獲得	2,577	88.9
誤獲得	323	11.1
合計	2,900	100.0

が既に語情報辞書に登録されている場合，追加は行われないため，ここでの延べ数は表 10 の辞書登録語数とは一致しない．

本手法においては，入力数字列のあいまいさの解消のために，隣接文字列情報，最上位階層語，語の平均長を利用したヒューリスティックスを組み込んでいる．これらを組み込まない場合の実験も行ったが，その場合，全体の平均正変換率は 3.3 ポイント減少していた．

5. 考　　察

5.1 未変換率

実験の全体を通して，未変換率は非常に低い値であった．これは，本手法の入力が仮名の母音情報が縮退した数字列であり，仮名に比べてあいまいさが増しているためである．例として,「20281（けんきゅう）」を仮名漢字変換，数字漢字変換する場合を考える．ここで，それぞれの辞書には（かんきょう：環境),（20281: 環境）が登録されているものとする．仮名漢字変換では,「かんきょう」は「けんきゅう」に当てはまらないので未変換となる．しかし，数字漢字変換においては,「20281」が当てはまるので,「環境」に変換され誤変換となる．このように，仮名漢字変換においては未変換となるような語でも，数字漢字変換においては誤変換ではあるが，変換が行われる語が数多く存在するため，未変換率が低くなっているのである．

5.2 辞書登録語数

実験結果からわかるように，高い変換精度を示しているにもかかわらず，本システムの辞書登録語数は非常に少数であった．これは，本手法の学習能力により，文脈に依存した必要最小限の語を獲得しているためである．このように，文脈に依存した適応型の辞書を自動生成することにより，処理量の増大，変換候補重複の増大を抑えることができる．変化した対象に動的に適応できることも確認され，本手法の適応能力の高さが確認された．また，表 11，表 12 からわかるように，最上位階層語の利用，位置推測処理により，高い精度で正しい語が獲得されており，これらの処理の有効性が確認できる．

表 13　適応前の変換例
Table 13　Translated example before adaptation.

入力数字列
251422*2381589221608238615129 14*042192*93*754281
変換結果
[この][移][区切][こ][使用][により] [囲][ー][翻訳][者][は][いない] [クリア]4*0[使][新][が][らず][みに] [適用]
校正済み変換結果
帰納的学習による機械翻訳手法における 遺伝的アルゴリズムの適用

表 14　適応後の変換例
Table 14　Translated example after adaptation.

入力数字列
251422*2381589221608238616 14*042192*93*7042813924
変換結果
[帰納的][学習][により][機械][翻訳] [手法][は][遺伝][的][アルゴリズムを] [適用][し][ること]
校正済み変換結果
帰納的学習による機械翻訳手法へ 遺伝的アルゴリズムを適用すること

5.3 対象への適応

大きな対象としての分野が変化した直後の項目「論文 1」の実験データ，及び変換結果の例を表 13 に示す．入力文字数 50,000〜51,000 文字中のデータである．この段階で，システムは変化する前の対象である「UNIX」に適応しており，辞書に登録されている語は，現在の対象である「論文」に適合していない．そのため，表 13 のように誤った変換が数多く行われている．しかし，その後，変化した対象に適合した語を学習していく.「論文 1」の最後のデータを表 14 に示す．入力文字数 72,000〜73,000 文字中のデータである．このように対象が変化した直後は正しく変換できなかった語が，システムが変化した対象に動的に適応し，正しく変換できているのが確認できる．他の対象に関しても，同様である．

5.4 隣接文字列情報を利用した変換

隣接文字列情報の利用により，変換を誤る例を表 15 に示す．入力文字数 84,000〜85,000 文字中のデータである．入力数字列の変換の過程を見てみると,［ヒューリスティックスの」,「データ」,「いる」の順に変換が行われており，この時点での変換結果は，以下のとおりである．

表 15　隣接文字列情報を利用した変換例
Table 15　Example of translation using neighboring character strings.

入力数字列
68193414235744*1419
変換結果
[ヒューリスティックスの][メタ][データ][いる]
校正済み変換結果
ヒューリスティックスに基づいている

表 16　位置推測処理による学習例
Table 16　Example of learning by position prediction.

入力数字列
203*0531303109
変換結果
[関]3*05[送信][し][オン][り]
校正済み変換結果
完全に送信し終わる
獲得される語
(3:し),(31:送信),(031:終)

[ヒューリスティックスの]74[データ][いる]

次に,(74:メタ) の変換が行われる．このとき，正しい語である (74:基) も変換候補となるが，先行文字列「ヒューリスティックスの」，後続文字列「データ」に隣接する度合が高い「メタ」が優先されている．なお，ここで，後続文字列として「データ」が適用されているが，これは,「データいる」,「データい」などの文字列が隣接文字列辞書に登録されていなかったためである．変換結果において，先行文字列中の「の」，後続文字列の「データ」は誤った変換である．しかし，正しい語である「基」にこれらが隣接する度合よりも，「メタ」にこれらが隣接する度合の方が高くなるため，ゆう度評価関数により「メタ」が優先され，誤変換となる．本手法における隣接文字列情報を利用した変換処理は，人間が行う同様の変換処理を模倣したものとなっているので，このような誤りも人間に近いものと考えられる．

5.5　位置推測処理による学習

位置推測処理により，誤った語が獲得される例を表 16 に示す．入力文字数 4,000〜5,000 文字中のデータである．校正済み変換結果の共通部分は,「に」,「し」,「わる」である．ここで,「し」に対応する数字「3」の位置が数字列中であいまいであるため,(313031:送信し終) から語を獲得することができない．変換に最上位階層語も使われていないため，位置推測処理が行われる．語情報辞書中で，表記の文字数が 2 である語の数字列の平均長を求めると，2 となっていた．よって,「送信」に対応する数字列の長さが 2 と推測され，3 番目に位置している「3」が「し」に対応すると決定される．このようにして，共通部分 (3:し)，差異部分 (31:送信),(031:終) が S1 として獲得される.(31:送信) は誤って獲得された語であるが，語情報辞書に登録され，今後の実験において変換に用いられる．既に，語情報辞書中の S1 階層には (3130:送信) が登録されていたが，数字列は短い方が変換に用いられる割合は高くなるので,(31:送信) が当てはめられる機会が増える．しかし，誤って獲得された語を変換に用いると，フィードバック処理において誤変換と判断されるので，そのゆう度は次第に低下する．入力文字数 5,000 文字までの実験終了時，辞書中の (31:送信) は，正変換度数 0，誤変換度数 9 となっており，LS 階層に転落していた．このように，誤って獲得された語は，本手法のフィードバック処理により，次第に変換に使われなくなっていくのが確認できる．

しかし，誤って獲得された語のゆう度の低下は次第にであり，はじめから辞書中に存在しない方が，変換精度には良い影響を与えると考えられる．よって，位置推測処理による語の獲得の確実性を高めるために，**3.3.2** で述べた処理と同様に，ここでも両方向からの解析を行い，一致したものだけを取り出す方法が考えられる．しかし，本手法における位置推測処理は，システムが獲得する語数の増大を目的としたものであり，この目的の妨げとならないかについて調査が必要である．

6.　少量データにおける有効性

小型の携帯端末などでの実用性を考慮すると，マニュアルや論文のような長い文章を入力する機会は少ないものと考えられる．そこで，短い単位で文章の内容が頻繁に変化するようなデータにおける本手法の有効性を確認するために，実験を行った．

6.1　実験データ及び実験手順

入力データとして，本論文の第 1 著者の送信メールを用いた．実際の送信日時順に入力を行った．このデータは，内容については公的な文章から私的な文章まで様々であり，また，文章の長さについても数十文字の短いものから数百文字のものまで，様々である．すなわち，短い単位で対象が頻繁に変化するデータである．入力文字数 100,000 文字に対して，**4.** と同様の実験を行った．この実験も初期辞書は空の状態から

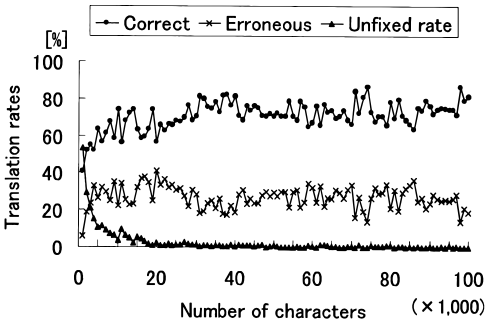

図 6　電子メールにおける変換率の推移
Fig. 6　Changes in the translation rates for e-mail.

行っており，使用者により校正された校正済み変換結果を用いて学習を行っている．

6.2 実験結果

各変換率の推移を図 6 に示す．全体の平均正変換率は 70.7[%] であった．図 6 からわかるように変換精度は次第に向上しており，最終的に 80[%] 程度までの上昇が確認された．また，100,000 文字の実験終了時の辞書登録語数は，3,640 語であった．

6.3 考　　察

図 6 からわかるように，正変換率の上昇率は **4.** の実験に比べて鈍くなっている．**4.** の実験の全体の平均正変換率は 76.7[%] であったので，約 6 ポイントの平均正変換率の減少であった．これは，対象の変化に伴い，その時点で辞書に未登録の語が数多く出現するためである．このように，対象の変化に伴い多数の未登録語が出現するが，そのような語も対象の変化にかかわらず，本手法の学習機能により獲得していき，徐々に変換精度は上昇している．**4.** の実験の 100,000 文字までの辞書登録語数は，2,320 語であったので，1,000 語以上多くの語が辞書に獲得された．

n-gram を用いた純粋な確率的手法による仮名漢字変換手法も提案されている [10]．この手法においては，解析的な知識は使わずに，大量のサンプルデータのみから高い仮名漢字変換精度を実現している．しかしながら，特定使用者の電子メールのように対象が頻繁に変化する場合，変換精度の向上のために，それぞれの対象において大量のサンプルデータが必要となるが，そのようなデータの収集は困難であると考えられる．これに対して，本手法においては，対象の頻繁な変化にも追従し，電子メールのような対象であっても約 80[%] の精度で変換でき，本手法の有効性が確認された．これは，少量のサンプルデータからでも，抽象度の異なる様々な変換規則をシステム自身が自動的に獲得することによるものであると考えられる．この獲得した知識を用いて次回からの変換を行うことにより，種々の対象への動的な適応を可能にしている．

7. 入力打鍵数について

4.，**6.** の実験においては，本手法の適応能力の確認を目的としており，校正済み変換結果は何らかの方法で使用者により校正されたものとしてきた．しかし，実際には校正処理も 12 キーで行われるので，その場合の校正処理を含んだ本手法全体に要する打鍵数を評価し，有効性を確認するために実験を行った．

7.1 校正処理について

校正処理は，変換結果に対して使用者が訂正箇所を指定し，再変換することにより行う．再変換は文字循環指定方式による仮名漢字変換により行う．この変換手法は現在の携帯電話等で一般的に用いられている手法であり，容易に実現可能であると考えられる．

仮名漢字変換辞書は，校正処理でのみ用いている．これは，本手法の数字漢字変換においては，システムが独自の単位で語を獲得し変換を行っているためである．再変換箇所を指定する際，使用者が意図する単位は，システム独自の単位とは異なる場合があるため，校正処理においては汎用的な仮名漢字変換辞書を用いるものとしている．一方，仮名漢字変換においては，個人，あるいは分野を限定すれば，3,000 語程度の辞書で十分であり，分野にシステムが適応することにより精度が向上することが確認されている [13], [14]．よって，すべての使用者が数万語の辞書をもつことは，不要な単語候補や同音異義語の増加による精度の低下及び処理量の増大という点から決してよいことではないと考えられる [1]．このような理由から，本手法の数字漢字変換においては，汎用的な辞書は使わず，帰納的学習により文脈に依存した語を獲得していくものとしている．

校正処理を含む本手法全体の処理手順を以下に示す．

① 数字列入力
② 数字漢字変換
③ 再変換箇所を指定
④ 文字循環指定方式による仮名入力
⑤ 仮名漢字変換
⑥ 学習

意図する日本語文を得るまで，③〜⑥を繰り返す．例として，「野球を観戦する」を考える．

① 表 1 に従い，数字列を入力する．
82810203039
② 変換結果は次のようになる．
野球脇 030 する
③ 次のように下線部分を指定する．
野球0̲2030 する
④「を」を入力する．
⑤「を」が意図する表記なので，そのまま確定
⑥（0：を）の対応関係が確定する．
③ 更に，再変換箇所を指定する．
野球を2030する
④「かんせん」を入力する．
⑤ 仮名漢字変換が行われる．
候補として感染，観戦，幹線,... が存在
意図する「観戦」に決定
⑥（2030：観戦）の対応関係が確定する．

校正箇所を使用者が指定していることにより，⑥の学習効率が上昇している．すなわち，文単位の比較ではあいまいである（0：を）の位置が，使用者により指定されるので，正しく対応関係を獲得することが可能である．このようにして，意図する日本語文の入力を行う．

7.2 評価方法

校正処理に文字循環指定方式による仮名漢字変換を組み込んだ本手法と，文字循環指定方式のみによる仮名漢字変換手法において，手法全体の打鍵数の評価を行う．

7.2.1 文字循環指定方式のみによる仮名漢字変換

携帯電話等で用いられている日本語入力方式である．文字循環指定方式により，仮名を入力し，意図した日本語文に変換を行う．入力が仮名であるので，数字に比べて有する情報量が多い．そのため，変換精度は一般的に高いと考えられる．しかし，入力においては，仮名を明示するために多くの打鍵数を必要とする．「野球を観戦する」を入力する場合を考えると，入力文字列は「やきゅうをかんせんする」となる．この 11 文字の入力に要する一般的な打鍵数は，$1+2+5+3+2+1+3+4+3+3+3=30$ である．

7.2.2 校正打鍵数

数字漢字変換結果の校正は，変換を誤った箇所を指定することにより行われるので，校正に要する打鍵数は，校正対象となる文字数で近似できると考えられる．NC 文字に対して数字漢字変換を行った結果，校正の対象となる文字数が NC_k であれば，これに要する校正打鍵数 $PR_n(NC_k)=NC_k$ となる．例えば，1,000 文字に対して数字漢字変換を行った結果，校正対象文字数が 200 であれば，$PR_n(200)=200$ である．

仮名漢字変換結果の校正も，使用者が訂正箇所を指定することにより行われるので，これに要する打鍵数も校正対象となる文字数によって近似する．NC 文字に対して，変換精度 r_k で仮名漢字変換を行うと，校正対象文字数は $NC\times(1-r_k)$ であり，これに対し更に再変換を行うので，これに要する校正打鍵数は $NC\times(1-r_k)^2$ となる．よって，校正に要する打鍵数 PR_k は以下の式で表される．

$$PR_k(NC)=NC\times(1-r_k)+NC\times(1-r_k)^2+\cdots$$
$$=\left\lceil\frac{NC\times(1-r_k)}{r_k}\right\rceil=\lceil NC\times\lambda\rceil \quad (8)$$
$$\lambda=\frac{1-r_k}{r_k} \quad (9)$$

ここで，打鍵数は整数であるので，このように切り上げた値を用いている．例えば，$r_k=0.90$, $NC=1,000$ であれば，λ，PR_k は以下のようになる．

$$\lambda=\frac{1-0.90}{0.90}=0.1111$$
$$PR_k(1000)=\lceil 1,000\times 0.1111\rceil=112$$

このように，仮名漢字変換結果に対して校正を要する打鍵数 PR_k は，それに要した文字数と変換精度から近似的に求めるものとする．

7.2.3 仮名漢字変換精度

仮名漢字変換精度を決定するため，予備実験を行った．実験に用いたデータは，**4.** の実験における分野「UNIX」の項目「ftp」の 10,000 文字のべた書き仮名文である．仮名漢字変換システムとして Microsoft IME98(注3)を用いて，実際に仮名漢字変換を行った．入力文字数に対する正変換文字数で評価を行っている．この結果から，仮名漢字変換精度 $r_k=0.97$ とした．よって，λ の値は以下のようになる．

$$\lambda=\frac{1-0.97}{0.97}=0.0309$$

7.2.4 手法全体の打鍵数

それぞれの手法において，変換が行われる文字数を NC とする．文字循環指定方式のみによる仮名漢字変換手法全体の打鍵数 NP_{cycle} は以下の式により評価される．NC 文字の入力に要する文字循環指定方式による打鍵数を $IN_k(NC)$ とする．

(注3)：Microsoft 社の日本語入力システムである．

$$NP_{cycle}(NC) = IN_k(NC) + PR_k(NC) \quad (10)$$

本手法全体の打鍵数 NP_{our} を考える．NC 文字の入力に要する文字情報縮退方式による打鍵数を $IN_n(NC)$ とする．打鍵数 1 で 1 文字の入力が可能なので，$IN_n(NC) = NC$ である．NC 文字に対して数字漢字変換を行った結果，NC_k 文字に対して校正を行う．すなわち，NC_k 文字に対して文字循環指定方式による仮名漢字変換を行うことになる．

$$NP_{our}(NC) = IN_n(NC) + PR_n(NC_k) + NP_{cycle}(NC_k) \quad (11)$$

7.3 評価実験

実験に用いたデータは **4.** と同様である．IN_n，IN_k，PR_n の値は，**4.** の実験より抽出し，それぞれの手法全体の打鍵数を評価した．例えば，**4.** の 10,000〜11,000 文字の実験データにおいて，入力文字数 1,000 文字に対して，文字循環指定方式において要する打鍵数は 2,932 であり，また，数字漢字変換結果に対する校正対象文字数は 151 文字で，これに対して文字循環指定方式において要する打鍵数は 606 であった．よって，式 (11)，式 (10) により，それぞれの手法全体の打鍵数は，以下のように評価される．

$$\begin{aligned} NP_{our}(1000) &= IN_n(1000) + PR_n(151) \\ &\quad + NP_{cycle}(151) \\ &= 1,000 + 151 + 611 = 1,762 \\ NP_{cycle}(1000) &= IN_k(1000) + PR_k(1000) \\ &= 2,932 + 31 = 2,963 \end{aligned}$$

7.4 実験結果

実験結果を図 7 に示す．図 7 からわかるように，文字循環指定方式のみによる仮名漢字変換手法全体の入力打鍵数に比べて，本手法全体の入力打鍵数は減少している．平均入力打鍵数で，30[%] 程度の減少が確認された．

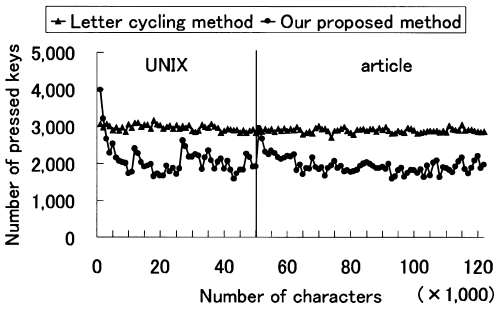

図 7　打鍵数の推移
Fig. 7　Changes in the number of pressed keys.

7.5 考　　察

数字漢字変換の精度が低い実験の初期段階においては，本手法の打鍵数は，文字循環指定方式のみによる手法の打鍵数を上回っている．しかし，学習が進行し数字漢字変換精度が向上することにより，打鍵数が減少し，本手法の有効性が確認できる．

また，数字漢字変換精度が 0 より大きくなると，$NC_k < NC$ となる．よって，仮名漢字変換精度 r_k が低下，すなわち λ が上昇した場合，式 (8) からわかるように，$PR_k(NC)$ の増加率は $PR_k(NC_k)$ の増加率よりも大きくなる．すなわち，仮名漢字変換精度の低下による NP_{our} への影響は，NP_{cycle} に比べて少ない．解析的な仮名漢字変換においては，辞書登録語数などの変換規則の量が変換精度に大きく影響するものと考えられる．小型の端末を想定した場合，その記憶容量の制約のため，変換規則の量を増やすのが困難なことから変換精度が低下する可能性がある．しかし，このような場合でも NP_{our} への影響は少ないので，携帯電話等の変換精度が比較的低い仮名漢字変換システムを校正に用いる場合には，更に本手法は有効であると考えられる．

8. むすび

本論文では，小型の携帯端末を想定し，少数のキーのみで迅速な日本語入力が可能な「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数字漢字変換手法」を提案した．入力に，仮名の母音情報が縮退した数字列を用いており，あいまいさが増しているので，この失われた情報をいかにして回復するかが重要となる．そこで，本手法においては，帰納的学習による高い適応能力に加えて，隣接文字列情報と最上位階層語，位置推測処理を利用している．本手法の高い適応能力により，変換精度を高くすることができ，更に，種々の対象に動的に適応でき，汎用性を保持している．最上位階層語の利用，位置推測処理により，有効な語を数多く獲得することができる．実験の結果，UNIX オンラインマニュアル，論文の各分野において，85[%] 以上の変換精度が得られた．また，対象が頻繁に変化する特定使用者の電子メールのような分野においても，80[%] 程度の適応が確認された．この電子メールのよ

うな分野においては，純粋に確率的な手法では適応が困難であると考えられる．しかし，本手法においては，このような分野における対象の頻繁な変化にも追従し，有効性が確認された．

また，校正処理を含む手法全体の打鍵数の評価から，携帯電話等で一般に用いられる文字循環指定方式のみによる仮名漢字変換手法に比べて，本手法の優位性が示され，実用的にも有効性が確認された．

謝辞 なお，本研究の一部は科学研究費（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．

文　献

[1] 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，” 信学論（D-II），vol.J79-D-II, no.3, pp.391–402, March 1996.

[2] K. Araki, Y. Momouchi, and K. 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, Brisbane, Australia, April 1995.

[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, Oct. 1998.

[4] 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 10 北海道連大，pp.365–366, Oct. 1998.

[5] 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，” 情処学自然言語処理研報，vol.98 (98-NL-128)，pp.1–7, Nov. 1998.

[6] 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 信学総大，D-6-6，p.102, March 1997.

[7] 長尾　真，自然言語処理，岩波書店，東京，1996.

[8] M. Nagao and S. Mori, “A new method of N-gram statistics for large number of n and automatic extraction of words and phrases form large text data of Japanese,” Proceedings of Coling 94, vol.1, pp.611–615, Kyoto, Japan, Aug. 1994.

[9] 森　信介，長尾　真，“n グラム統計によるコーパスからの未知語抽出，” 情処学論，vol.39, no.7, pp.2093–2100, July 1998.

[10] 森　信介，土屋雅稔，山地　治，長尾　真，“確率的モデルによる仮名漢字変換，” 情処学論，vol.40, no.7, pp.2946–2953, July 1999.

[11] 越前谷博，荒木健治，桃内佳雄，栃内香次，“実例に基づく帰納的学習による機械翻訳手法における遺伝的アルゴリズムの適用とその有効性，” 情処学論，vol.37, no.8, pp.1565–1579, Aug. 1996.

[12] 松原雅文，荒木健治，桃内佳雄，栃内香次，“帰納的学習による数字漢字変換手法の性能評価，” 北海学園大学工学部研究報告，no.26，pp.427–437, Feb. 1999.

[13] 栃内香次，斉藤　康，“適応型変換辞書を用いるかな漢字変換，” 情処学論，vol.24, no.2, pp.209–213, March 1983.

[14] 栃内香次，岡沢好高，“適応変換辞書方式かな漢字変換システムの性能測定，” 情処学論，vol.26, no.4, pp.733–739, July 1985.

（平成 11 年 5 月 7 日受付，8 月 26 日再受付）

松原　雅文　（学生員）

平 8 北海学園大・工・電子情報卒．現在，同大大学院工学研究科電子情報工学専攻修士課程在学中．自然言語処理の研究に興味をもつ．情報処理学会会員．

荒木　健治　（正員）

昭 57 北大・工・電子卒．昭 63 同大大学院博士課程了．工博．同年，北海学園大学工学部電子情報工学科助手．平 1 同講師．平 3 同助教授．平 10 同教授．現在，北大・工・電子情報工学専攻助教授．自然言語の機械学習と機械翻訳に関する研究に従事．情報処理学会，日本認知科学会，人工知能学会，言語処理学会，IEEE，ACL，AAAI 各会員．

桃内　佳雄　（正員）

昭 40 北大・工・精密卒．昭 42 同大大学院修士課程了．同年(株)日立製作所入社．昭 47 北大大学院博士課程単位取得退学．昭 48 北大大学院情報工学専攻助手，昭 59 講師，昭 61 助教授．昭 63 北海学園大学工学部電子情報工学科教授．自然言語の理解と生成の研究に従事．工博（北大）．情報処理学会，日本認知科学会，計量国語学会，言語処理学会各会員．

栃内　香次　（正員）

昭 37 北大・工・電気卒．昭 39 同大大学院工学研究科電気工学専攻修士課程了．現在，北大・工・電子情報工学専攻教授．主として音声情報処理，自然言語処理の研究に従事．工博．情報処理学会，日本音響学会各会員．