自然言語処理 93－5
(1993. 1. 13)

帰納的学習によるべた書き文のかな漢字変換の有効性

高橋 祐治　荒木 健治　桃内 佳雄

北海学園大学

〒064 札幌市中央区南26条西11丁目

現在、日本語を計算機に入力する手段としては、区切りを一切いれないべた書き文のかな漢字変換が主流となっている。しかし、この手法には様々な問題が残されている[1]。我々は、このような問題を解決する一手法として、べた書き文とその正しい変換結果から帰納的に語を学習し、確実性の高い順に変換を行う、べた書き文のかな漢字変換手法を提案している[2]。本手法は、人間が未知の読みと、その表記を比較する事により語を獲得するという過程に基づいている。

本報告では、我々が以前決定した語候補を選択するための尤度評価関数の係数の値[3]をシステム上で実際に用い、大量実験を行った結果と本システムにおける辞書の階層化の有効性について述べる。

The evaluation of Kana-Kanji transformation of non-segmented Japanese Kana sentence by inductive learning

Yuji Takahashi, Kenji Araki, Yoshio Momouchi

Hokkai-Gakuen-University

S 26 W 11, Chuo-ku, Sapporo, 064, JAPAN

We have proposed the new system of Kana-Kanji transformation of non-segmented Japanese Kana sentence. This system acquires the words by comparing the non-segmented Japanese Kana sentence and the result of correct transformation of it. The dictionary of this system is composed of four classes. And this method is based on the acquisition process of words by human. We have confirmed the validity of this system by some experiments.

We determined the some good coefficients for transformation in this system. By the previous experiments. In this paper, we evaluate the experimental results of transformation and learning by applying their coefficients, and the validity of the stratified dictionary.

1．はじめに

日本語を計算機に入力する方法として、文節分かち書き入力方式、字種指定入力方式、べた書き文入力方式など種々の方式が考案されている[1]。しかし、前者2つの入力方式は、利用者側に文節に関する文法的な知識や字種指定情報の入力の負担などを要求する。このため、利用者側に負担が少ないべた書き文入力方式が、日本語文を計算機に入力する方法として現在主流となっている。べた書き文によるかな漢字変換手法は、入力時に利用者側の負担を軽減させるが、一方では、辞書作成の労力、未登録語の処理、変換精度などに依然として問題が残されている。

そこで、我々は、以前から、これらの問題点を解決する1つの手法として、べた書き文とその正しい変換結果である漢字かな混じり文から帰納的に語を学習し確実性の高い語から順に変換を行う手法を提案している[2]。本手法は、人間が未知の読みとその表記から語を獲得し、その獲得した語を用いてかな漢字変換を行う過程に基づいている。本手法の特徴は、学習により語を自動的に獲得し辞書を作成するので、辞書を作成する労力がいらず、また、専門用語や未登録語なども個人が使用する範囲で自動的に学習を行うので、辞書の容量が小さくて済み、学習とフィードバック学習を併せて行うことで、利用者や分野が変わる場合などにも動的に対応でき変換精度を向上させるので、多数の利用者、多分野に高い精度で用いることができる。これまでの実験により、本手法の有効性が確認されている[2]。

本稿では、先に決定された各係数の値[3]を用いて、大量実験を行った結果と階層化された辞書の有効性について述べる。

2．システム概要

処理の流れを図1に示す。

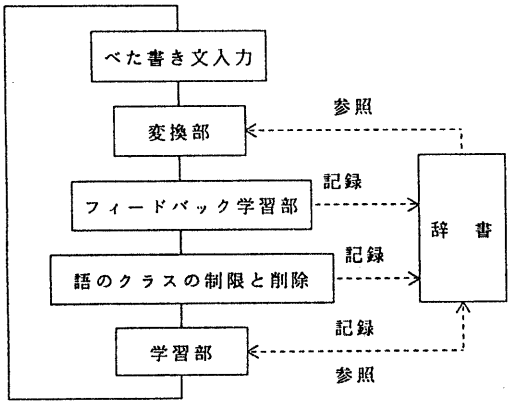

図1　処理の流れ

まず、入力されたべた書き文は、変換部に行き、学習部で語が獲得されたときの状況とフィードバック学習による語の精度によって確実性の高い語から変換を行う。次に、変換された語の正誤を判定する。この判定は、システム側で正しい変換結果との比較をとることにより自動的に行われる。

フィードバック学習部では、正誤判定の結果から語の精度を更新し、語のクラスの制限と削除を行う。このため、辞書が始め空の状態であっても、学習とフィードバック学習を行うことによって、語を自動的に学習し、変換精度を向上させることができる。最後に、学習部で語を獲得し、結果を辞書に登録する。これは、入力べた書き文とあらかじめ与えられる正しい変換結果から表記文字の変化部分と不変化部分を探すことにより語を獲得する。獲得された語は、複数の語を含む場合があるので、条件に従いさらに共通部分と差異部分に分けられる。なお、辞書中に存在する語は、その抽出状況や変換率により、ＭＳ，ＣＰ，Ｓ1，ＲＳの4つのクラスに分類され、学習で語を獲得したときの状況とフィードバック学習で更新される語の精度によって分けられている。辞書のクラスの分類を図2に示す。

ＭＳ：（Most certain Segment）
　　ＣＰであり、かつ９５％以上の正変換率の語
ＣＰ：（Common Pattern）
　　Ｓ1の語について共通部分抽出により取り出される語であり、９５％未満の正変換率の語
Ｓ1：（Segment One）
　　漢字かな混じり文とその読みから獲得した語で、正変換率が、ある一定値以上の語
ＲＳ：（Remaining Segment）
　　ＣＰの抽出の際に差異部分になった語で、正変換率が、ある一定値以上の語

図2　辞書のクラスの分類

3．各モジュールの詳細

3．1　学習部

3．1．1　学習部の手順の概要

（1）入力べた書き文とその正しい変換結果から文字の変化部分と不変化部分を探すことで語の読みと表記を獲得する。
（2）獲得した語の共通部分と差異部分の抽出を行い、確実性によって語を階層化する。

（1）で獲得された語をSegment Oneと呼びＳ1と略記する。
（2）で抽出される共通部分の語をCommon PatternとよびＣＰと略記し、差異部分の語をRemaining Segment と呼びＲＳと略記する。獲得した語群に2度または、2度以上出現するということは、その語がより語としての確実性が高いと考えられるので、ＣＰはＳ1より上位の階層とし、ＲＳは差異部分であるため語としての確実性が、低いと考え、Ｓ1より下位の階層とした。ＣＰとして獲得された語の中で、フィードバック学習により語の精度が高められ、その精度が９５％以上の正変換率を越えた語をMost certain Segmentと呼びＭＳと略記し、このＭＳを最上位階層とする。以上述べたように、階層は、ＭＳ，ＣＰ，Ｓ1，ＲＳの4階層で、この順に確実性が高いことになる。

3．1．2　Ｓ1の抽出処理

Ｓ1の抽出の処理は、図3に示すように入力べた書き文とその正しい変換結果から文字の変化部分と不変

化部分を探すことで、第１段階の語の読みと表記を獲得する処理である。Ｓ１の抽出処理では、処理単位は、区切りによって挟まれる部分である。この区切りとは、句読点、空白のことで、文中に句読点、空白があればその前後で文は分けられ処理される。区切りは語ではないことが明確なため、あらかじめ、この情報を与えＳ１の抽出処理を行なっている。なお、文頭から解析を行うことを左から右への解析、文末から解析を行うことを右から左への解析と呼ぶこととする。

```
                        ↓      ↓↓
（１）べた書き文：きのうをせんていするほうほう
                              ↓    ↓↓
（２）変換結果        :  機能を選定する方法
（３）解析結果（読み  :  表記）
              （きのう    :  機能）
              （を    :  を）
              （せんてい  :  選定）
              （する  :  する）
              （ほうほう  :  方法）
```

注）↓は不変化部分を指す。

図３　左から右への解析例

図３のように語の読みと表記が一意に正しく決定される場合は問題ないが、図４，５に示すように語の読み（yomi）と表記（hyouki）の獲得に誤りを生ずる場合がある。図４は、左から右への解析の誤る例を示し、図５は、右から左への解析が誤る例を示している。この例から判るように、漢字の読みの中に図６のような助詞等と同じ文字が含まれる場合には、対応関係が多対多になってしまい、誤った対応関係が抽出されてしまう原因となる。

```
入力        :  じゅうなんなせいぎょこうぞう
変換結果    :  柔軟な制御構造
（１）左から右への解析
          ↓
      じゅうなんなせいぎょこうぞう
          ↓
      柔軟な制御構造
[yomi]    = じゅう          [hyouki] = 柔軟
[yomi]    = な              [hyouki] = な
[yomi]    = んなせいぎょこうぞう
[hyouki]  = 制御構造
（２）右から左への解析結果
              ↓
      じゅうなんなせいぎょこうぞう
          ↓
      柔軟な制御構造
[yomi]    = じゅうなん      [hyouki] = 柔軟
[yomi]    = な              [hyouki] = な
[yomi]    = せいぎょこうぞう
[hyouki]  = 制御構造
        注）↓は不変化部分を指す。
```

図４　左から右への解析が誤る例

このためＳ１の抽出処理では、Ｓ１を抽出する際、左から右への解析と右から左への解析を行い（双方向解析と呼ぶ）、その結果抽出されたＳ１の抽出個数が同じならば、解析結果の出現順が同じで、読みおよび表記が一致するもののみ辞書に登録し、それ以外の語は、辞書への誤登録をしてしまう恐れがあるため、登録しないものとする。

```
入力        :  あたらしいめいれいをくわえる
変換結果    :  新しい命令を加える
（１）左から右への解析
          ↓↓      ↓    ↓↓
    あたらしいめいれいをくわえる
          ↓↓    ↓   ↓↓
        新しい命令を加える
[yomi]    = あたら          [hyouki] = 新
[yomi]    = しい            [hyouki] = しい
[yomi]    = めいれい        [hyouki] = 命令
[yomi]    = を              [hyouki] = を
[yomi]    = くわ            [hyouki] = 加
[yomi]    = える            [hyouki] = える
（２）右から左への解析
                  ↓    ↓    ↓↓
        あたらしいめいれいをくわえる
                  ↓    ↓   ↓↓
                新しい命令を加える
[yomi]    = あたらしいめ    [hyouki] = 新し
[yomi]    = い              [hyouki] = い
[yomi]    = れい            [hyouki] = 命令
[yomi]    = を              [hyouki] = を
[yomi]    = くわ            [hyouki] = 加
[yomi]    = える            [hyouki] = える
        注）↓は不変化部分を指す。
```

図５　右から左への解析が誤る例

格助詞　：「が」、「の」、「を」、「に」、「へ」
副助詞　：「さへ」、「だけ」、「など」、「は」、「も」
接続助詞：「から」、「けれど（も）」、「て（で）」、「のに」、「ば」
終助詞　：「か」、「ぞ」、「な」、「ね」、「よ」

図６　日本語文中に用いられる助詞の例

３．１．３　一字語の処理

学習の例外的な処理として一字語の連続に対する処理がある。この処理は、「書き込む」などのように文字の変化部分と不変化部分の読みと表記が１字である語が連続して現れる場合、双方向解析で抽出される語は、「か（書）」、「き（き）」、「込（こ）」、「む（む）」となるが、これらを１語にまとめる処理である。変換部では、読みが１字の語が変換される場合、変換精度を上げるためにその両側の語が確定済みでなければならないという条件があるためで、上記の例のような場合の１字語の連続による未変換を防ぐために行う処理である。

３．１．４　ＣＰ，ＲＳの抽出処理

この様にして抽出された第１段階の語であるＳ１は、複数の語を含む可能性があるので、さらに共通部分と差異部分の抽出を行う。この抽出例を図７に示す。

しかし、このように抽出処理を行うと、辞書中の語が単表記のレベルまで分割されてしまうので、ＣＰ，ＲＳを抽出する際の共通部分となるための条件を図８

のように定める。

（１）抽出されたＳ１
　　れんぞくおんせい　：　連続音声
　　おんせい　　：　音声
（２）抽出されたＣＰ，ＲＳ
　　おんせい　　：　音声　　（ＣＰ）
　　れんぞく　　：　連続　　（ＲＳ）

図７　ＣＰ，ＲＳの抽出の例

（１）文字数による制限
　　Ｄ＝読みの文字数－表記の文字数
　①Ｄ＝０の場合　：　表意文字でない字種を近似
　　３文字以上表記が一致する。
　②Ｄ≠０の場合　：　表意文字を近似
　　２文字以上表記が一致する。
（２）共通部分の状況による制限
　　一方が他方を完全に含む形で重複する。

図８　ＣＰ，ＲＳの抽出条件

①の条件では分割されない語も存在するが、漢字は表記が２字の語が多いことから語の未分割より過分割を防ぐことに重点をおいた条件であり、逆に、②の条件は、分割され過ぎる語も存在するが、語の過分割より未分割を防ぐことに重点をおいた条件である。なお、ＣＰ，ＲＳの抽出が行われた場合、抽出元となった語は、辞書から削除される。このような手順を経て、辞書の語候補が登録される。

３．２　変換部

３．２．１　変換部の手順の概要

語の変換は、左から右への解析を行い、確実性の高い語から順次変換を行う。最も確実性が高い語は、ＣＰで後述する正変換率が、ある一定値を越えて上位の階層に昇格したＭＳの語である。次いで、ＣＰ，Ｓ１，ＲＳの順に確実性が高い語である。また、読みの文字数によっても語の確実性が異なるので、始めは、ＭＳ，ＣＰ，Ｓ１，ＲＳの順に各クラスの読みが２文字以上の語による変換を行い、次いでＭＳ，ＣＰ，Ｓ１，ＲＳの順に読みが１字の語による変換を行う。但し、読みが１字の語による変換を行う条件は、語の情報量を考慮し、誤変換を少なくするために両側の語が確定済みの場合のみである。

変換の手順
①クラスがＭＳで読みが２文字以上の語による変換
②クラスがＣＰで読みが２文字以上の語による変換
③クラスがＳ１で読みが２文字以上の語による変換
④クラスがＲＳで読みが２文字以上の語による変換
⑤クラスがＭＳで読みが１文字の語による変換
⑥クラスがＣＰで読みが１文字の語による変換
⑦クラスがＳ１で読みが１文字の語による変換
⑧クラスがＲＳで読みが１文字の語による変換

３．２．２　重複の処理

変換の際、同じクラスで重複する語がある場合は、尤度評価関数の値が最大の語を選択し、尤度評価関数の最大の値が同値となる語がある場合は、次の順で語を決定する。

①尤度評価関数値最大
②誤変換度数最小
③正変換度数最大
④出現頻度最大
⑤文字数最大（読み）
⑥分割位置最前
⑦辞書の登録された最も新しい語

また、尤度評価関数は、以下に示す式によって与えられ、α，βの値は、ＣＦおよびＥＦの重みとして働く係数である。

ＥＶ　＝　ＡＦ　＋　α＊ＣＦ　－　β＊ＥＦ
ＥＶ　：　Evaluation Value　：　尤度評価関数値
ＡＦ　：　Appearance Frequency　：　出現頻度
ＣＦ　：　Correctness Frequency　：　正変換度数
ＥＦ　：　Error Frequency　：　誤変換度数
α，β　：　係数

３．３　フィードバック学習部

３．３．１　正変換率による語の削除の定義

辞書の各語には、出現度数ＡＦ（Appearance Frequency）、正認識度数ＣＦ（Correctness Frequency）、誤認識度数ＥＦ（Error Frequency）が付加されており、各々、学習部やフィードバック学習部で更新される。本手法では、語の正変換率を算出し、その値が４０％より落ちると辞書の語の削除を行う。ここで、この一定条件値を削除条件と呼ぶ。その評価は、ＣＦ＋ＥＦの値が一定の値以上の時に開始することになっている。この値を評価開始値といい、先に行った実験により３である。以下に正変換率　ＣＲ（Correctness Rate）の定義を示す。

$$ＣＲ = \frac{ＣＦ}{ＣＦ＋ＥＦ} \times 100$$

フィードバック学習部では、語の正誤を判断し正しい語は尤度を上げ、誤っている語は、尤度を下げる処理を行う。語の尤度を上げるとは、語のＣＦを１増加することで、尤度を下げるとは、語のＥＦを１増加させることである。また、語の正変換率を求め所属クラスの制限と削除を行う。なお、辞書の各階層中の語の所属クラスの制限と削除は、図２を参照のこと。

３．４　辞書の語の各パラメータの更新方法

出現頻度ＡＦは、Ｓ１の抽出処理で抽出されたときと変換部で語が使用されたときに１増加させる。つまり、ＡＦはＳ１としての出現回数と変換部での語の使用回数との和となる。また、ＣＰ，ＲＳの抽出の処理では、ＡＦの引継が行われるが、ＣＰのＡＦは、ＣＰ，ＲＳの抽出の処理に用いた２語のＡＦの和となり、ＲＳのＡＦは、ＣＰ，ＲＳの抽出に用いた２語の内文字数が長い方のＡＦとなる。　正認識度数ＣＦと誤認識度数ＥＦは、３．３．１で述べたとおり、フィードバック学習部において変換結果と正しい変換結果を比較し、一致していればＣＦを１増加させ、誤っていれば、ＥＦを１増加させる。

4. 実　験

4．1　実験方法と結果

本手法に基づく実験システムをＶＡＸ８５５０上で作成し実験を行った。この実験は、先の実験によって決定された[1]、尤度評価関数の係数$\alpha=2$，$\beta=28$及び辞書の語の削除条件４０％以下、評価開始値３の各々の値を用いて変換実験および学習実験を行い、本システムの変換効率を確かめ、さらに、４つの階層に分けられている辞書の有効性の考察を行うものである。

実験方法は、べた書き文を１文入力し、そこまでの学習段階において登録された語を用いて変換し、その評価を行った後に、その１文の学習を行い、語を獲得し辞書に登録する。この処理を繰り返すことにより行った。辞書の初期状態は空にして、使用文は、情報分野より１０編（情報処理学会論文誌より抜粋），機械分野より１０編（北海道大学工学部研究報告より抜粋）を漢字かな混じり表記で１３９，４２７文字を用いた。その資料を表１に示す。また、実験結果をグラフ化したものを図９，１０，１１に示し、辞書の各階層の登録語数の推移を図１２に示す。

5．考察

5．1　変換率と占有率の定義

グラフ中の変換率および占有率の各定義について述べる。これは、このシステムを評価するもので、変換精度や辞書中の各階層が、全体に占める割合を示すものである。変換率は、変換された、漢字かな混じり文が、正しい変換であるか否かを評価するものであり、百分率で示す。占有率は、変換の際にあてはまった各クラスの語数が全体の文字数に占める割合を示すものである。また、グラフ上の情報分野と機械分野の切れ目である累積文字数の値は、８３，９３７文字である。

5．2　総合の変換率について

総合の変換率は、情報分野の前半では、未登録語が多いため、それほど高くはないが、学習が進むに連れて、未変換が少なくなり、ＭＳによる正変換の文字の占有率が高くなる情報分野の後半では、誤変換が少なくなり、変換率が９０％台に上昇している。機械分野では、情報分野全体をとおした変換率の推移が各資料毎に現れる。これは、未登録語が多いためと、逆に語の個数が多いためであると考えられる。未登録語が多く語の個数が多いと、未登録の語の部分に現在辞書に登録されている語が対応してしまう確率が高くなるため誤った変換を起こすことにつながる。また、情報分野の１０編の資料の学習と変換を行ったため、安定した単語の数が多く、尤度評価関数値が高い語が機械分野で学習した語による変換を妨げる結果を引き起こしたと考えられる。また、総合の正誤の変換率の推移は、Ｓ１の正誤の変換率に推移と酷似しており、このことは、Ｓ１の語数が他のクラスの語数と比べて非常に多いことで上述した変換の妨げを起こす主要な要因となっていると考えられる。変換部では、語の候補を選択する最初の基準は、尤度評価関数値であり、この値に大きな影響を与えるのは、係数α，βである。この実験で使用した尤度評価関数の係数α，βは、$\alpha=2$，$\beta=28$という大変厳しい重みとなる値を用いている。図９のグラフを見ると機械分野の変わり目で、変換率が急に低下するが、そのあとすぐに、その分野または著者に適応し、変換率が復帰していることがわかる。つまり、学習がかなり進んだ段階でも、一度誤変換となった語は、次回の出現率が低化し、同じ語候補は出ずらくなる。この作用によってすぐに変換率が復帰していると思われる。このことから、尤度評価関数の係

表　1　実　験　資　料

情報分野
出典：情報処理学会論文誌，Vol.23，No.1～4

著者名	論文題目と総文字数
前島 ほか	高集積マイクロコンピュータに適したマイクロプログラム制御方式（１０９４６）
山本 ほか	COBOLマシンとその設計思想-ハードウェア構成について（８１８４）
松山 ほか	フーリエ変換を用いたテクスチュアの構造解析（７５１７）
木村 ほか	日本語入力用カタナカ語検出規則とオンライン国語辞典の一分析（８４８５）
有田 ほか	インテリジェント・コンソール-ＯＳの機能拡張の一方法（９０７４）
田村 ほか	ポータブル画像処理ソフトウェア・パッケージＳＰＩＤＥＲの開発（９２６９）
高藤 ほか	グラフィク・ディスプレイ・ターミナルのための端末作画システム（６２７９）
長岡 ほか	オペレーティング・システムのファームウェア化対象選定法（７１７９）
酒井 ほか	プログラム階層構造の生成，処理，文書化能力を有するText･Editer（７６９８）
中所 ほか	パステストに本質的な分岐に着目した網ら率尺度の提案（９３０６）

機械分野
出典　：　北海道大学工学部研究報告，96号～108号

著者名	論文題目と総文字数
有江 ほか	格子乱流中における垂直平板の流力測定（４８７５）
園田 ほか	暖房用ストーブの燃焼性能に関する研究（第１報）（７２７４）
園田 ほか	暖房用ストーブの燃焼性能に関する研究（第２報）（５３２１）
飯田 ほか	任意に調節可能な座標線密度を持つ流れ場内格子点網の創成法（５９６４）
岸田 ほか	円弧切欠きと荷重端の干渉について（２０９０）
伊藤 ほか	換気回数の低い室内における都市ガスの燃焼と一酸化炭素の発生（７０９７）
金内 ほか	チェーンソーの振動におよぼす切削条件の影響（６７６４）
知名 ほか	境界層剥離の近似的な推定法（４０４６）
入江 ほか	任意形状を有する四辺形膜の自由振動（３８９６）
奏 ほか	疲れ強さにおよぼす加工硬化および残留応力の影響（８１６３）

数の厳しい重みが、辞書の活性化を促進していることがよくわかる。

5．3　辞書の各階層の考察

5．3．1　ＭＳ

情報分野では、変換率が一貫してほぼ１００％である。分野の変わり目では、９０％程度に落ち込むが、すぐに９５％程度に達し機械分野の中盤から後半にかけては、ほぼ１００％に回復している。クラスがＭＳである語を最も確実性の高い語と定義したが、実験結果の変換率が一貫してほぼ１００％であることから、その定義の正当性が示された言える。分野の変わり目で変換率が低下したのは、未登録の語があるため語の重複が多くなり、誤変換が多くなったと考えられる。しかし、学習が進むに連れて未登録語が減り、また、他のクラスに比べ、ＣＰであり、９５％以上の正変換率の語といったＭＳの所属クラスの制限が厳しいため、変換で語の重複が起きても誤った語はＭＳからＣＰに落ち、速やかに変換率が回復することからも高い変換率を示すことはよくわかる。

5．3．2　ＣＰ

情報分野の始めでは、高い変換率を示すが、情報分野の中盤では、８０％台の前半にまで変換率が落ち込む。そして、情報分野の終わりでは、９０％台に回復する。分野の変わり目では、８０～９０％程度で、機械分野の中盤では７０％台に落ち込む。しかし、機械分野の後半では、９０％台に回復する。変換率が各分野の中盤で、落ち込むのは、ＣＰとして登録された語が、変換精度が高いとＭＳに昇格してしまうためと、ＭＳによる変換で認識できなかった文字列に対して変換を行うためであると考えられる。しかし、ＣＰによる誤変換文字の占有率は、１％程度で総合の変換率には、ほとんど影響を及ぼさない。また各分野の後半で変換率が上がっているのは、ＭＳによる変換文字数が増加したためで、これは、Ｓ１，ＲＳについても言えることだが、上位クラスの変換が増えることで、変換できなかった文字列部分の句切りが、より確からしくなり、重複する単語の個数を減らし変換率を上げることにつながっていると思われる。

5．3．3　Ｓ１

情報分野の前半では、９０％台であるが、情報分野の中盤から機械分野にかけては、各論文の始めの章では、７５％程度に低下し後半の章では、９０％台に回復する。Ｓ１には、他のクラスと比べると非常に多い語数が登録されているため上位クラスでの変換が少ない場合、語の重複が起こりやすく誤りが多くなると考えられる。また、Ｓ１は、学習が進んだ場合、正誤の占有率が高いため、変換率に最も影響を及ぼし易い。しかし、我々の先の実験において、各階層の語の削除条件や評価開始値を決定したが、この値が４０％厳しい条件であり、評価開始値が３といった早い段階に評価されることから、変換率の大きな変動は見られなかったと思われる。

5．3．4　ＲＳ

情報分野の始めでは、９０％台であるが、情報分野の終わりでは、６０％台にまで落ち込む。機械分野の前半では、５０％台であるが、機械分野の後半にかけて各論文の始めの章では、６０％台で、後半の章では、９０％台に達している。ＲＳの語は、ＣＰ，ＲＳの抽出の際の残留部分であることや、上位３クラスの変換で変換できなかった文字列に対して変換を行うのでその変換率が低くなるのは、当然の結果である。しかし、実際には、占有率が１～２％と低いため変換率には、ほとんど影響が見られなかった。

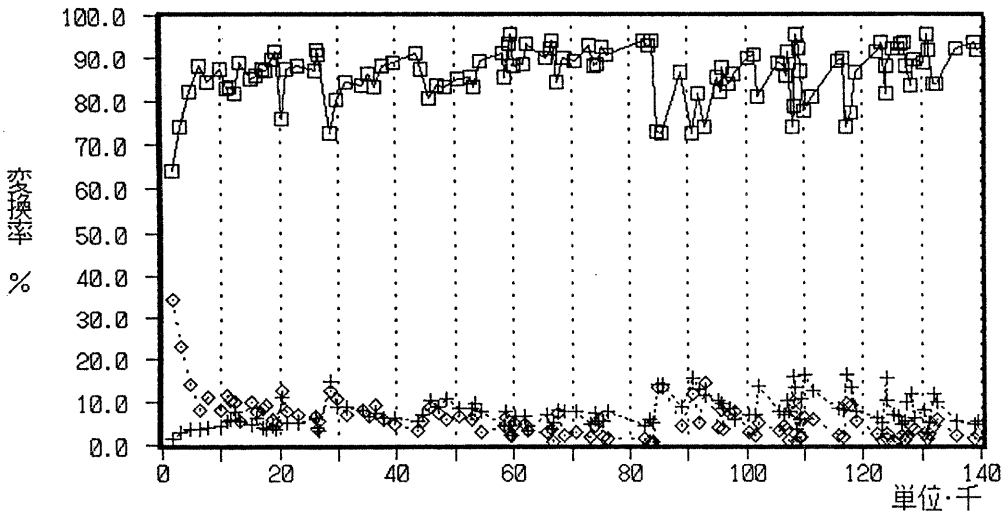

図９　総合の変換率

5．3．5　未変換

未変換の正変換、誤変換は未変換部分に対応する正しい変換結果中に、ひらがな、英数字記号といった読みと表記が同じ文字がどの程度含まれていたかを示すもので、語を認識するという立場からみれば、本来、すべて未変換と見なされるべきであるが、読みと表記が同じ語は、見かけ上、変換されたように見えるためこの場合の変換率のことを未変換という。よって、率が５０％前後でかなり上下が激しい。率が高い場合は、数式のような文が、広範囲にあった場合であり、逆に低い場合は、未登録語が多かったことを示している。つまり、規則性はないが、論文の筆者がどの程度ひらがな、英数字記号を用いるのかを知る指標には成りうる。

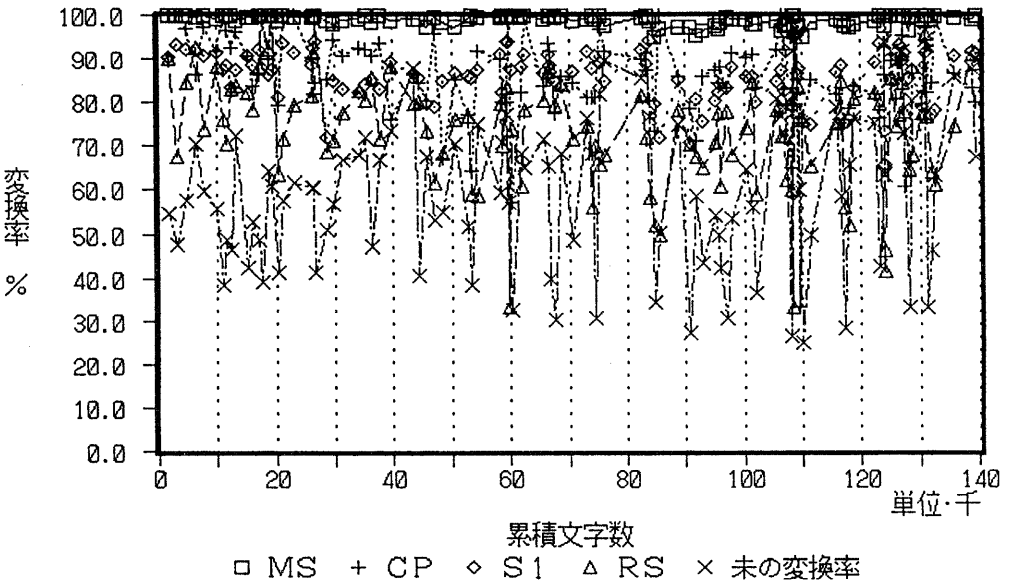

図１０　各階層毎の変換率

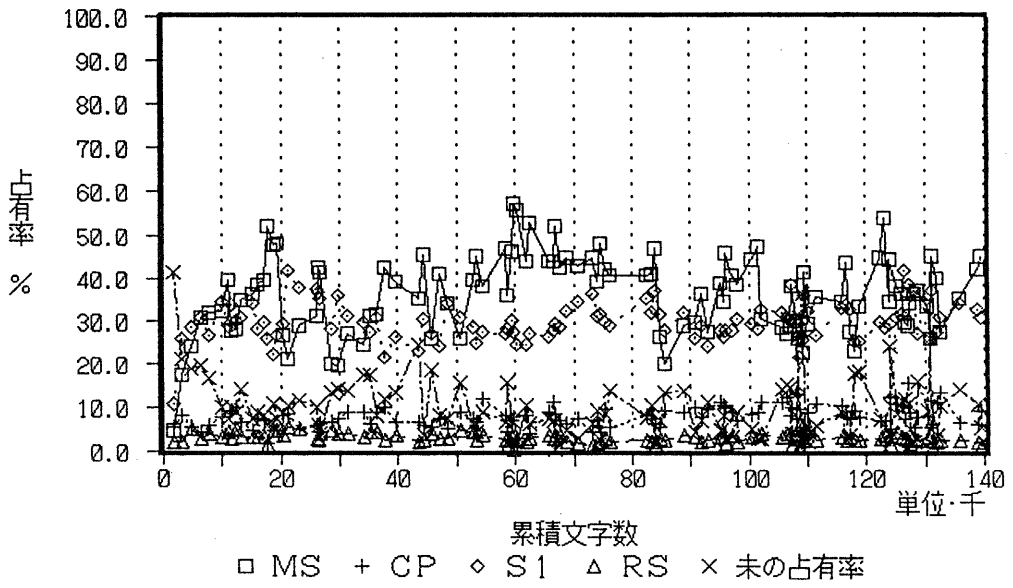

図１１　各階層毎の占有率

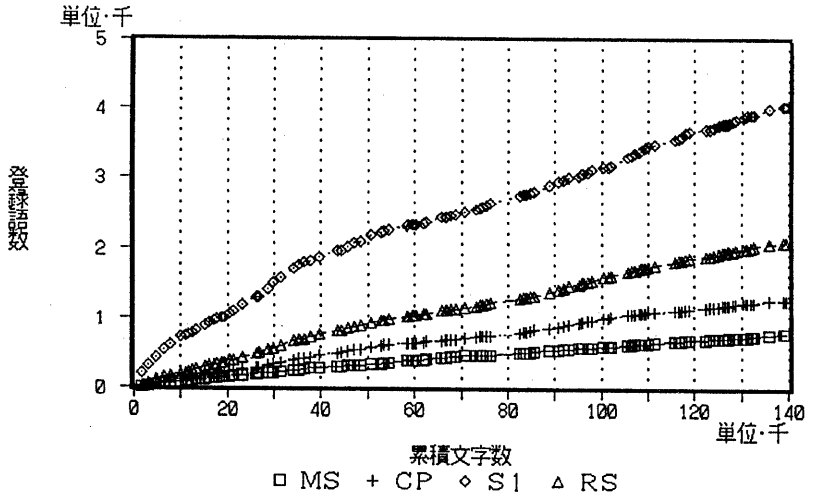

図１２　各階層の登録語数の推移

５．８　全体的な考察

本実験は、先に決定された各係数の値を実際に用いて大量実験を行ったものである。総合の変換率から、削除条件が４０％以下の正変換率とした事による、未登録語の増加が見られたものの、評価開始値が３という早い段階でクラスの制限が行われることから、分野及び著者への適応が敏速に行われている事がわかった。このことは、先に決定した係数の値が、汎用的なものであることを示していると考える事ができる。

次に、辞書の階層について考えてみる事にする。

１つ目として、本手法では、ＭＳ，ＣＰ，Ｓ１，ＲＳと辞書が４つに階層化され、確実性の高い順に変換処理を行う事により、変換率を高めているが、実験結果からも明らかなように、確実性の高い順に、変換を行う事により入力されたべた書き文を正確に変換単位に区切れる事ができ、変換率の向上に貢献している事がわかる。このことは、本システムの４つの辞書の階層の有効性を示していると考えられる。ＭＳでは、９５％以上の正変換率である。階層であるので、登録語は、頻繁に用いられる語や専門用語が多い、ＣＰも９５％以下の語ではあるが、ＭＳと同様の傾向がみられる。Ｓ１は、学習部で獲得された語が登録されている事から、図１２からもわかるように、全階層中で最も登録語数が多い。また、登録語の中に共通部分がない場合は、そのままＳ１の階層に残留してしまうことから、比較的、語の文字数（長）が大きい。ＲＳは、ＣＰの抽出処理の残留部分である事から、文字数（長）の小さい語が多く。単漢字レベルまで語が、分割されているものが見受けられる。

２つ目としては、実際に、低い正変換率から削除条件である４０％以下になり削除された語の例を表２に示す。表２より、削除される語は同音異義語のものがほとんどである事がわかる。また、条件が４０％ときついためか、正しいはずの語の削除が行われている。つまり、分野や著者の適応の速度が速いことと、一長一短の関係であることを示している。よって、このような問題を解決する手段として、現在ある４つの辞書階層に、削除された語を格納する階層を追加し、その階層の語と一番最後に当てはめを行い変換し、正しく使用された語を上位の階層へ復帰させる方法が考えられる。

表２　削除語の例

階層	読み	表記	AF	CF	EF
Ｓ１	う	得	4	0	3
Ｓ１	しゅ	種	4	0	3
Ｓ１	しゅ	主	4	0	3
Ｓ１	おお	多	5	0	3
Ｓ１	おお	大	7	1	2
Ｓ１	いち	位置	4	0	3
ＲＳ	いち	一	7	1	2
ＲＳ	しん	新	4	1	2

６．おわりに

実験結果から、各係数の値は、学習がかなり進んだ段階でも、汎用的な値であることが確認された。また、本システム中の辞書階層を４つに分割することの有効性も確認できた。しかし、正しい語をまちがって削除してしまう恐れがあることから、新たな削除階層を設ける必要があることもわかった。今後は、この実験結果を基に、本手法の改良を行い、大量データによる性能評価実験を行う予定である。

謝　辞

実験の遂行にあたり、本学國在学中種々の討論及び協力をしていただいた現㈱ＮＥＣ　佐藤隆之氏、現㈱デービーソフト　富居広樹氏に感謝いたします。

＊参考文献

１）長尾：日本語情報処理，電子情報通信学会(1984).
２）荒木，坂内：帰納的学習によるべた書き文のかな漢字変換，電子情報通信学会技術研究報告，Vol 91, No. 397, pp. 9-16(1991).
３）高橋，荒木，桃内：帰納的学習によるべた書き文のかな漢字変換における尤度の評価について，電気関係学会北海道支部連合大会論文集，pp. 409-410(1992).