1R－1

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

松原　雅文†　荒木　健治†　桃内　佳雄†　栃内　香次‡
北海学園大学工学部†　北海道大学大学院工学研究科‡

1．はじめに

近年の利用者の需要と技術の進歩から小型の携帯端末が登場してきた。中にはその入力を１０個の数字キーとその補助を行うわずかな特殊キーによってのみ行う携帯性に非常に優れたものがある。最近の携帯端末での電子メール利用率の増大からもこの少数のキーで軽快に日本語文章の入力ができるかな漢字変換手法の開発が望まれる。

この問題に対して我々は「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」を提案する。文字情報縮退方式[1]とは５０音のかな文字を０～9、#、*の１２個のキーそれぞれに複数対応させ日本語文を入力する手法である。かな文字１文字の入力を１ストロークで行うことができ、軽快な入力が可能である。また本手法においては、システムが帰納的学習により語を自動的に獲得するので初期辞書の作成は必要なく、使用者個人に合わせた辞書が自動的に生成される[2]。

文字情報縮退方式は、軽快な操作性実現のために母音情報が縮退し、情報が失われた形になっている。そのためかな漢字変換に際して、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。これに対して我々は、すでに開発済みの帰納的学習を用いた手法[2]による解決を試みた。この手法においては対象を限定することにより、システムが帰納的学習によって自動的にその対象に適応していく。しかし、これだけでは十分な精度は得られなかった。そこで帰納的学習による語の獲得を中心としながらも付随する他の情報を学習することによって、失われた情報の回復を試みることにした。従来の手法に比べて新しく学習する具体的な情報は、隣接文字情報と語の読み文字数である。前者は、以前に入力された数字べた書き文、校正済みの変換結果から n-gram 統計[3][4]により獲得され、かな漢字変換に利用される。これにより先行、後続文字列とのつながりを考慮した変換が可能となっている。後者は、校正済みの変換結果のかなに対応する、入力数字べた書き文中の数字列の位置の推測に利用される。これにより従来より多くの語を獲得することができる。

本稿では本手法の概要及びその有効性を確認するために行った評価実験結果について述べる。

2．概要

入力された数字べた書き文は変換処理で、すでに辞書に登録されている語と隣接文字情報により漢字かな混じり文に変換される。変換結果に誤りがある場合、校正処理を行う。学習処理では校正済みの変換結果と数字べた書き文との比較から、語を多段階に抽出し辞書に登録する。同時に数字べた書き文、校正済みの変換結果の隣接文字情報を抽出する。フィードバック処理において登録、正変換、誤変換された語はその情報を辞書に持ち、隣接文字情報と一緒に次回からの変換に役立てられる。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、使用者に合わせた辞書が生成され、次第に変換精度が向上していく。

3．処理過程

本手法における処理過程は、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。

3．1　変換処理

１２キーにより入力された数字べた書き文を漢字かな混じり文に変換する過程である。数字べた書き文に対して、階層化された辞書に登録されている語を上位層より当てはめて変換を行う。変換候補は尤度評価関数により評価され、優先度を決定して最適な語を適用する。

3．2　校正処理

変換処理において変換結果に誤りがある場合、校正処理が行われる。人手により変換結果を訂正

Non-Segmented Kana-Kanji Translation Method
Using Inductive Learning
with Degenerated Keyword Input
Masafumi Matsuhara †　Kenji Araki †
Yoshio Momouchi † and Kouji Tochinai ‡
Faculty of Engineering,Hokkai-Gakuen University †
Graduate School of Engineering,Hokkaidou University ‡

する過程である。

3．3　学習処理

語を獲得し、辞書に登録する過程である。従来の帰納的学習と同様に、校正済みの変換結果と数字べた書き文との共通、差異部分を多段階に抽出することにより語を獲得する。獲得できない語が存在した場合、位置推測処理により推測した位置を基準として、同様に語を獲得する。またここで数字べた書き文と校正済み変換結果に含まれる全文字列の n-gram 統計を抽出する。

3．4　フィードバック処理

語を辞書に獲得する際に、その語の情報を一緒に獲得する過程である。獲得する情報は、正変換率(CR)、誤変換率(ER)である。また n-gram 統計から対象となる語の先行、後続の隣接文字列存在確率(NR)を抽出する。これらを尤度評価関数に適用し、変換時に利用する。尤度評価関数(CEF)は次の通りである。

$CEF = \alpha \times NR + \beta \times CR - \gamma \times ER$

α、β、γ：尤度評価係数

4．評価実験

本手法と帰納的学習のみを用いた手法について評価実験を行った。入力データとして、UNIX のオンラインマニュアルより ftp の項、11,228 文字を使用した。尤度評価係数α、β、γは予備実験よりそれぞれ2、1、5とした。結果を図1に示す。本手法において初期正変換率 24%から最終的に 65%までの向上が見られる。本手法の平均正変換率は 50%、帰納的学習のみを用いた手法では 46%であった。また、両者の差の最大値は 10%であった。

5．考察

実験の前半 5000 文字程度までは、本手法の正変換率は従来の帰納的学習のみによる場合とほとんど差が見られず、最大で 4%である。これは統計的な情報である隣接文字情報の学習がまだ不十分で有効に変換に利用されないからだと考えられる。それに比べて解析的な情報を用いている位置推測処理はあまり学習量に依存しないので、この4%の差は位置推測処理によるものだと思われる。実験が進むにつれて両者の差は顕著になり、最終的に 10%の差が見られた。これは実験が進むにつれて、より多くの隣接文字情報が学習されているからであり、これが変換精度に好影響を与えている。今後も入力データの量に比例して、変換精度の向上が期待できる。このように５０音のかな文字よりも曖昧性の高い１２文字の数字においてもシステムの学習能力により、次第に対象分野に適応していくのが確認できた。

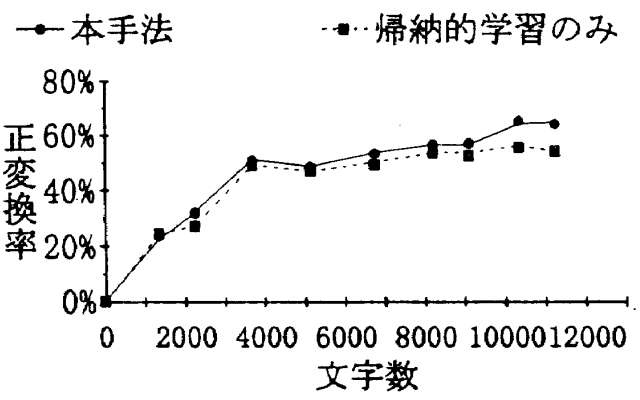

図1　正変換率の推移

6．おわりに

本稿では、文字情報縮退方式の入力におけるかな漢字変換で帰納的学習によりどの程度対象となる分野に適応できるかについて述べた。母音情報が縮退しているにも関わらず、従来の帰納的学習による語の獲得に加えそれに付随する他の情報を学習することにより、本手法の有効性が確認された。今後は入力データを増やし、隣接文字情報をさらに多く学習することによる変換精度の向上を確認する予定である。同時にデータ量に対する位置推測処理の精度も調査する。さらに、隣接文字情報だけではなく、通常の文章における格情報に相当するような情報を学習させ、さらなる適応度の向上を図りたいと思う。

謝辞

なお、この研究の一部は文部省科学研究費補助金（課題番号 09878070）により行われた。

参考文献

[1]佐藤、東田、林、奥、村上：ＰＢ電話機を利用した日本語入力方式、電子情報通信学会総合大会、D-6-6(1997)
[2]荒木、高橋、桃内、栃内：帰納的学習を用いたべた書き文のかな漢字変換、電子情報通信学会誌、Vol.J79-D-Ⅱ、No.3、pp.391-402
[3]長尾：自然言語処理、岩波書店(1996)
[4]森、長尾：n グラム統計によるコーパスからの未知語抽出、情報処理学会論文誌、Vol.39、No.7、pp.2093-2100