5C-8

べた書きかな文の仮単語境界を用いたかな漢字変換法の精度

荒木　哲郎†　　倉野　真樹†　山田　和義†　　古川　貴康†　小越　康宏†

福井大学工学部†

はじめに

従来、べた書きかな文のかな漢字変換法としては、文節単位に分割する方法には、2 文節最長一致法[1]、文節数最小法[1]、かな文字のマルコフ連鎖モデルを用いた方法[2]などが提案されており、また、単語選択には連語解析を用いた方法[1]、格フレームを用いる方法[1]、漢字かな混じり文字のマルコフ連鎖モデルを用いる方法[3]、確率モデルを用いる方法[4]、読み情報を用いる方法[5]などがある。日本文音声入力のかな漢字変換を考えるとき、さらに変換精度の向上が望まれる。本論文では、仮単語境界推定を用いたべた書きかな文のかな漢字変換法を提案し、新聞記事データを用いた実験を行ってその有効性を定量的に評価する。

１．べた書きかな文の単語分割とかな漢字変換候補の絞り込み処理を並行して行うかな漢字変換法[3]

2.で提案されるかな漢字変換法と比較する上で、ここでは[3]で示されているかな漢字変換法の手順を示す。

(i)べた書きかな文節に対し、単語辞書の読み見出しにより検索し、先頭から順に部分列に一致する漢字かな単語候補を全て抽出する。

(ii)(i)で得られた各部分列に対する漢字かな単語候補を相互に結合して構成される漢字かな文節候補を全て生成する（漢字かな文節ラティスと呼ぶ）。

(iii)(ii)の各漢字かな文節候補に対して、漢字かな文字の 2 重マルコフ連鎖モデルを用いて評価し、最大の連鎖確率値を持つ単語候補列を最尤な単語候補列として選択する。

この手順に従って得られる漢字かな文節候補ラティスの例を図１に示す。

２．仮単語境界を用いたかな漢字変換法

べた書きかな文の文節単位の分割として、仮文節境界を推定する方法[2]が提案されていたが、ここでは、かな表記の単語マルコフ連鎖モデルを用いて仮単語境界を推定する方法を提案する。

(i)べた書きかな文節に対して、同じ読みを持つ複数の単語候補を一つかな表記で表し、文節を構成するあらゆるかな表記の単語候補列の組み合わせを網羅的に全て抽出する（かな単語候補ラティスと呼ぶ）。

(ii)(i)で抽出されたかな各単語境界候補列に対して、かな表記の 2 重単語マルコフ連鎖モデルを用いて評価し、最尤な単語境界候補を決定する。

(iii)(ii)で求めたかな単語境界に対して、単語辞書を用いて、同じ読みを持つ漢字表記の単語候補を全て抽出する。

(iv)漢字かな文字のマルコフ連鎖モデルを用いて最尤な漢字かな変換候補を決定する。

仮単語境界推定を用いたかな単語候補ラティスの例を図２に示す。

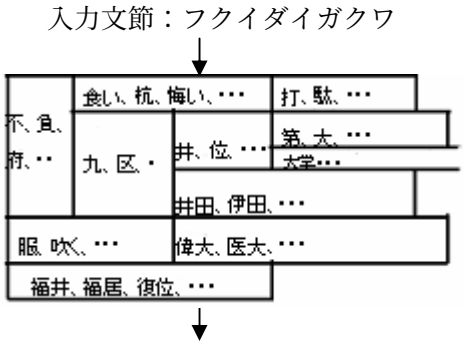

図１　方法１による漢字かな単語候補ラティスの例

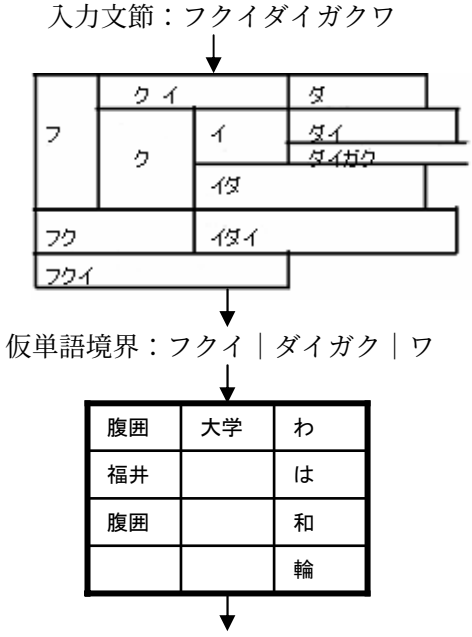

図２　方法２によるかな単語候補ラティスと漢字かな単語候補ラティスの例

An Evaluation of "Kana-to-Kanji" Conversion Method Using the Provisional Boundaries of Words for Non-segmented "Kana" Sentences

Tetsuo Araki†, Masaki Kurano†, Kazuyoshi Yamada†, Takayasu Furukawa†, Yasuhiro Ogoshi†

†Department of Human and Artificial Intelligent Systems, Faculty of Engineering, University of Fukui

3． 実験結果

3．1 実験条件

（1） 入力データ

① 文の種類：日本経済新聞記事

② 字種：べた書きかな文節

③ 総文節数：標本外データ 668 文節（100 文）文節平均 6.9 文字，文平均 47.4 文字

（2） 使用辞書

① 40 万語の単語辞書

② 単語境界推定に用いる文節単位のかな表記単語 2 重マルコフ連鎖確率辞書

③ かな漢字変換候補の絞り込みに用いる漢字かな混じり文字の 2 重マルコフ連鎖確率辞書

3．2 実験結果と考察

（1） かな漢字変換精度

① かな文節のかな漢字変換結果

方法 1 と方法 2 による、べた書きかな文節のかな漢字変換結果を図 3 に示す。同図より、方法 2 の精度の方が、方法 1 と比べて、10 位までの累積正解率で約 2%高い結果となった。

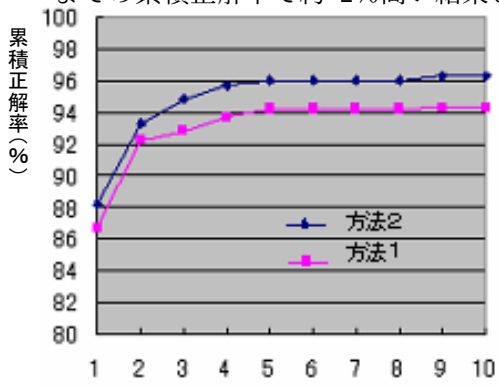

図 3 文節のかな漢字変換精度

② かな文のかな漢字変換結果

方法 2 を用いて文節ごとにかな漢字変換した結果から文を生成する方法を、方法 3-1 とし、方法 2 をかな文に拡張したものを方法 3-2 とする。これらのかな漢字変換結果を図 4 に示す。同図から、方法 3-1 方が、方法 3-2 よりも、10 位までの累積正解率で 7.5%高い結果となった。

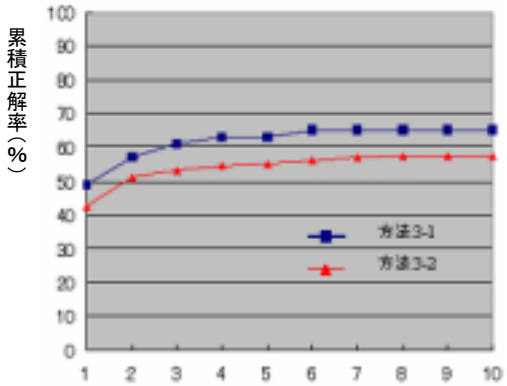

図 4 文のかな漢字変換精度

さらに、Microsoft 社の IME2000 を用いた実験では、第 1 位正解率が 44%となり、本手法の方が 5%高いという結果となった。尚、実験条件については、変換辞書は初期化し、初期変換モードは一般、学習は無しである。

（2） かな漢字変換に要する処理時間の評価

かな漢字変換に要する処理時間を、辞書へのアクセス回数を用いて評価する。方法 1 と方法 2 を用いた、文節のかな漢字変換時の辞書へのアクセス回数を、図 5 に示す。平均文字数である 7 文字のとき、方法 2 は約 $1/10^2$ に短縮することができた。

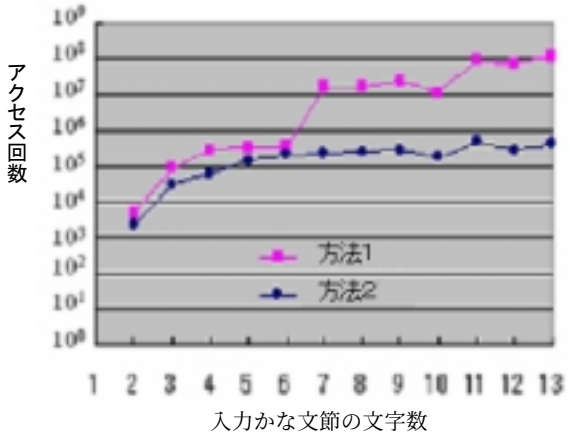

図 5 辞書へのアクセス回数

4． まとめ

文節の実験では、仮単語境界を用いる方法 2 の方が、方法 1 に比べて 10 位までの累積正解率で約 2%高い結果を得た。また、文のかな漢字変換精度は 10 位までの累積正解率で 65%であった。また、辞書へのアクセス回数は、7 文字以上のとき約 $1/10^2$ に縮小されることがわかった。

文献

[1]斉藤裕美，川田勉:”仮名漢字変換アルゴリズム”，進学誌，vol.70，No.8，pp679-687,(1986)

[2]荒木哲郎,池原悟,橋本昌東，三品尚登:”2 重，3 重のマルコフ連鎖モデルを 2 段階に使用したべた書き仮名文の文節境界推定法”，信学論，Vol.J83-D-II，No.12pp2745-2754，(2000.12)

[3]村上仁一，荒木哲郎，池原悟:”日本文音節入力に対して 2 重マルコフ連鎖モデルを用いた漢字仮名交じり文節候補の抽出精度”，信学論，Vol.J75-D-II，No.1，pp11-20,(1992.1)

[4]森信介，土屋雅稔，山地治，長尾真:”確率的ンモデルによる仮名漢字変換”，情処論，Vol.40，No.7，pp2946-2953，(1999.7)

[5]荒木哲郎，池原悟，真田陽一，横川秀人:”読み情報を用いた仮名漢字変換の精度向上効果の推定”，信学論，Vol.J84-D-II,No.2，pp351-361，(2001.2)