単語分割と単語選択に関するかな漢字変換法の評価に関する一考察

荒木哲郎　倉野真樹　山田和義　古川貴康　小越康宏　大場達哉
福井大学工学部知能システム工学科
{araki, kurano, ogoshi} @human.his.fukui-u.ac.jp

１．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するとともに、かな漢字変換法を単語分割と単語選択の観点から方式の異なる 2 つの方式について、新聞記事データを用いた実験を行って、かな漢字変換精度の評価と辞書アクセス回数の比較を定量的に行う。

２．かな漢字変換法のモデル

２．１　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]

2.2 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を方法 A とし、手順を以下に示す。

【ステップ 1】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。

【ステップ 2】ステップ 1 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（漢字かな文節ラテイスと呼ぶ）。

【ステップ 3】ステップ 2 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イスの例を図１に示す。

２．２　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。この方法を方法 B とし、以下に手順を示す。

【ステップ 1】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（かな単語候補ラテイスと呼ぶ）。

【ステップ 2】ステップ 1 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。

【ステップ 3】スペップ 2 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。

【ステップ 4】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。

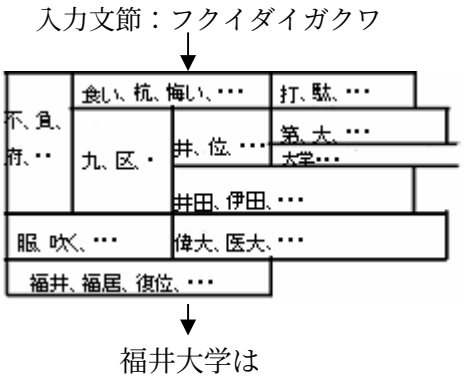

図１　方法 A による漢字かな単語候補ラティスの例

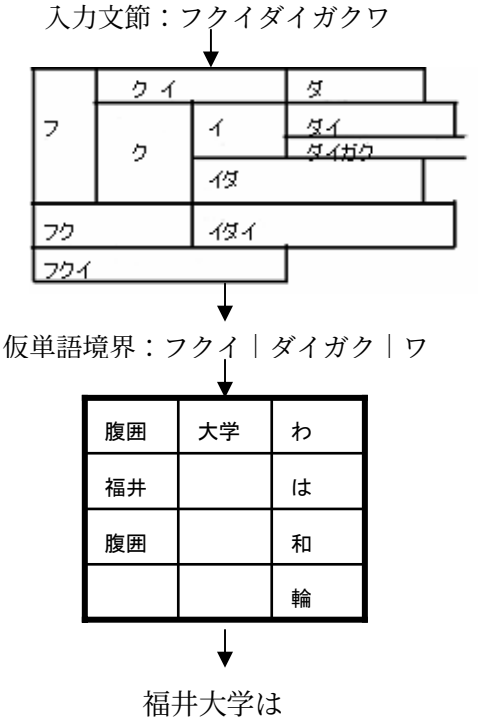

図２　方法 B によるかな単語候補ラティスと
漢字かな単語候補ラティスの例

２．３　マルコフ連鎖モデルの違いによるかな漢字変換精度の評価モデル

マルコフ連鎖モデルの違いによるかな漢字変換方法を次の観点から分類し、表１のような評価モデルを設定する。

① べた書きかな文字列を、単語辞書引きを用いて単語単位に分割する際に、どのような情報を用いて分割するか。

② ①で決定された仮名文字列の単語分割に対し、単語辞書を用いて単語候補を選択する際に、どのような情報を用いて選択するか。

このような 2 つの処理に対して、次のようなマルコフ連鎖モデルを用いてモデルを設定する。

① 情報の単位が文字（C）か、単語（W）か
② マルコフ連鎖の次数が、1 重(1)か、2 重(2)か
③ 字種が、かな文字表記(KN)か、漢字かな混じり表記(KK)か

例えば、2 重文字単位のかな文字連鎖モデルを２KNC モデルと表すことにする。

表１　かな漢字変換方式の評価モデル

モデル	単語分割	単語選択
A-1	**２KKC**	**２KKC**
A-2	**１KKW**	**１KKW**
A-3	**２KKW**	**２KKW**
B-1	**２KNW**	**２KKC**
B-2	**２KNW**	**１KKW**
B-3	**２KNW**	**２KKW**

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）　単語分割推定精度の比較

方法 A と B による単語分割の精度を、表１に示されるマルコフ連鎖モデルを用いて絞り込まれたときの第１位候補方式について比較したものを表 2 に示す。同表から、方法 B の方が高いことがわかる。これは単語分割に対して、かな表記の単語候補の数が、漢字かな混じり表記の単語候補の数より極めて少ないことによる効果が現れていると考えられる。

また、方法 B の場合についてかな文字表記の単語分割精度の 10 位までの累積正解率を図３に示すと、10 位までの累積正解率は 99.8%と非常に高い値が得られことがわかった。

表２　第一位正解候補の単語分割精度

A－１	62.5%
A－２	82.9%
A－３	83.6%
B	98.9%

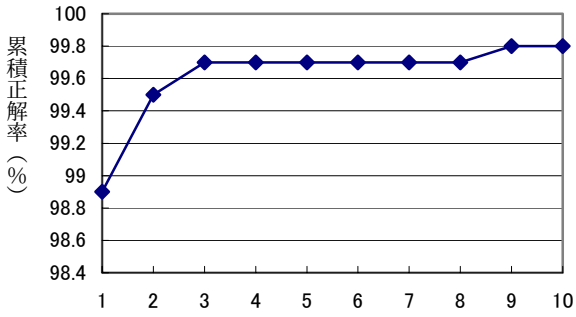

図３　方法 B のかなの単語境界推定精度

（２）　かな漢字変換精度

方法 A と方法 B による、べた書きかな文節のかな漢字変換結果を図４に示す。同図より、第一位正解率でみると、方法 B-2 が一番高く、累積 10 位までの累積正解率では方法 B-3 の精度が一番高い結果となった。いずれも、方法 A と比べて、方法 B の方が高い精度がれられることが分かった。これは、方法 B で高い単語分割精度が得られるため、単語選択で方法 B の方が、方法 A より選択候補数が少ないことが原因と考えられる。

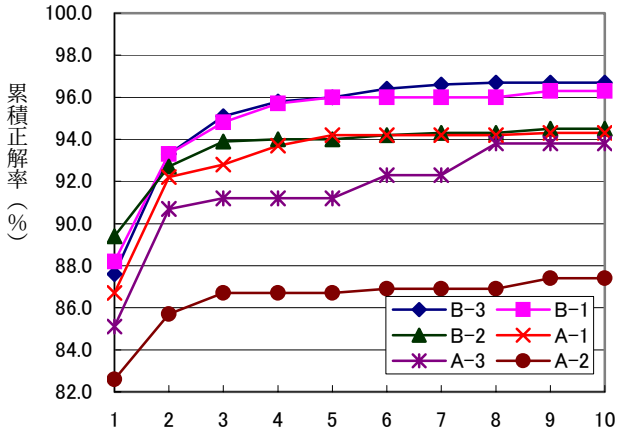

図４　かな漢字変換モデルの変換精度の比較

（３）　かな漢字変換候補の第一位文字の適合率と再現率

ここでは、かな漢字変換モデルにおける最尤候補と正解候補との最長共通部分列（longest common subsequences）[6]の文字数に基づく再現率と適合率を表 3 に示す。同表より、最長共通部分列の評価では、方法 B の方が適合率及び再現率共に若干高い値を示したが、ほとんど大差は見られなかった。

表３　最長共通部分列による各種変換モデルの最尤候補と正解候補の適合率と再現率

モデル	適合率(%)	再現率(%)
A-1	94.2	93.4
A-2	94.2	93.4
A-3	94.4	93.4
B-1	94.9	95.2
B-2	95.5	95.4
B-3	94.5	94.6

（４） 漢字変換処理に要する辞書アクセス回数の比較

処理時間を評価する上で、方式 A と B でかな漢字変換に必要な単語候補の組み合わせ数の比較結果を、1 重単語マルコフ連鎖モデル（１KKW）を用いた場合について図５に示す。

また、2 重文字マルコフ連鎖モデル（２KKC）を用いた場合の辞書アクセス回数の比較を図６に示す。同図より、方法 B がかな文字列の長さが 7 の場合で、約 100 倍アクセス回数が少ない結果となり、処理時間が大幅に削減されることがわかった。

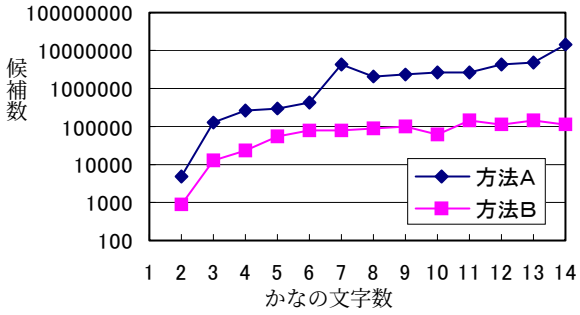

図５　１重単語マルコフ連鎖モデルの単語候補の組み合わせ数の比較

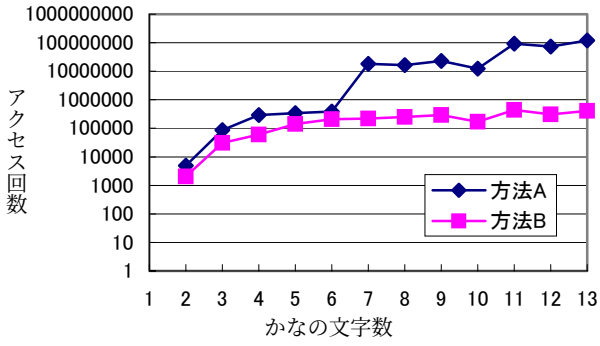

図６　２重マルコフ連鎖モデルを用いた場合の辞書アクセス回数の比較

５．まとめ

かな漢字変換法において、単語分割法と単語選択法の異なる２つの方法に対し、(1)情報単位が文字単語か,また(2)マルコフ連鎖モデルの次数が 1 重か 2 重か、及び(3)字種がかな文字表記か漢字かな混じりかについて組み合わせ、6 通りのモデルを設定した。実験により、文節単位及び最長共通部分列による比較では両方ともに余り大差は見られなかったが、仮単語境界を用いた後で単語選択を行う方法が、単語分割と単語選択を同時に行う方法に比べて若干高いことがわかった。

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。

このことから、単語分割を行った後で単語選択を行う方法が変換精度及び処理時間の上から優れていることがわかった。

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)

[6]Aho.A.V:“文字列中のパターン照合のためのルゴリズム”, コンピュータ基礎理論ハンドブッ,Vol.I, 形式的モデルと意味論, pp.263-304, Elseveir Science Publishers (1990)