漢字かなのTRIGRAMをもちいたかな漢字変換方法

村上仁一
（ATR自動翻訳電話研究所）

1　はじめに

かな漢字変換は、自然言語処理研究の一つの分野として従来から多くの方法が研究されている。これらの方式は、単語を構文的、意味的に分類して、これらを接続ルールや係受け情報などのルールで記述しておくことによって、かなから漢字かなまじり文に変換されるときに生成される曖昧性を減らしている。しかし、実際の文章では単語の多品詞性や曖昧な係受けなどが存在するために、正確なルールを書くことは困難であり、そのためかな漢字変換の問題点が完全に解決されたとは言いがたい。

一方、言語処理は音声認識の分野においても認識性能の向上をめざして研究がされている。そして、この分野では言語の統計的手法、特にマルコフモデルが有効であることが実験的にしめされている [1][2]。 そこで、本論文では漢字かなのマルコフモデルをもちいたかな漢字変換方式を提案し、これの変換精度を調べた。この結果、漢字かなの 2 重マルコフモデルを使用したとき、オープンデータで 88%クローズドデータで 98%の高い変換精度が得られることが示された。

2　マルコフモデルを利用したかな漢字変換方法

2.1　一般的なかな漢字変換

一般のかな漢字変換は、次の 2 つの処理に分割される [3]。

（A）切り出し処理

入力されたひらがなベタ書き文に対し、単語辞書を利用して単語候補を出力する。 ただし、かなに一致する単語を単語辞書内において、すべて探索したとき、大量の単語候補が出力される。このため文節数最小法や最長一致法などのヒューリステックな方法を用いることにより、生成される単語候補の数を減らしている [3]。

（B）　変換処理

複数の単語候補を組み合わせて、漢字かなまじり文を出力する。この処理において、単語候補を選択するために、従来は単語の接続ルールや係受け情報を用いていた。

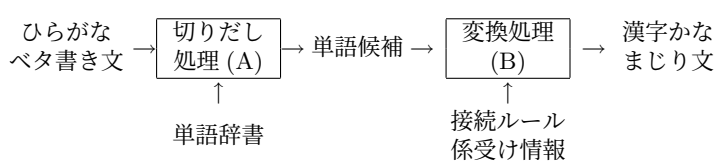

図 1: 一般的なかな漢字変換

2.2　マルコフモデルを利用したかな漢字変換方法

本論文で提案するかな漢字変換方式は、変換処理（B）において単語候補の中から漢字かなのマルコフモデルにおいて出現確率が最大になる漢字かなまじり文を選択し、かな漢字変換結果とするものである。

例として「おおくらしょうは」が入力され、切り出し処理において図 2 のような単語候補が生成されたとする。変換処理（B）は、図中の単語候補の組み合わせ (2×3×2)12 通りの漢字かなまじり文の中から、漢字かなのマルコフモデルにおいて出現確率が最大になる、1 つの漢字かなまじり文を選択する。

この場合、漢字かなまじり文 “大蔵商は” の出現確率は、漢字かなの 1 重マルコフモデルにおいては次のように計算される。

P {“大蔵商は”} =P(“大”|start)×P(“蔵”|“大”) ×P(“商”| “蔵”)×P(“は”| “商”) ×P(end| “は”)

ただし P(“蔵”| “大”) は “大” の漢字の後に “蔵” の漢字が続く連鎖確率である。このような出現確率を “大蔵省は” “大蔵しょうは” など 12 通りの全ての組み合わせの漢字かなまじり文に対して計算する。そして、このうち確率のもっとも高い漢字かなまじり文を選択し、かな漢字変換結果とする。この結果 “大蔵省は” が選択されることが期待される。

大蔵 大倉	商 省 しょう	は 羽

図 2: 単語候補（入力 「おおくらしょうは」）

2.3　かな漢字変換における Viterbi アルゴリズム

図 2 のようなマトリクスから出現確率が最大の漢字かなまじり文（単語列）を選択するには、原理的には、すべての単語候補を漢字かなまじり文に展開して、漢字かなのマルコフモデルにおける出現確率を計算する必要がある。しかし、Viterbi アルゴリズムを使用することによって、全展開をした場合に比べ計算量を大幅に減らすことができる。

3　かな漢字変換の実験

3.1　マルコフモデルの連鎖確率

マルコフモデルを用いたかな漢字変換には、漢字かなの連鎖確率を予め推定しておく必要がある。ここでは、約 170 万文字の漢字かなまじり文を使用して、この値を計算した。具体的には、はじめに 1982 年 1 月 4 日から 3 月 31 日までの日経新聞の記事を日本文解析プログラム [4] で処理して文節ごとに区切り、次に文節内における漢字かなの出現頻度を数えて、最後に、これらの連鎖確率を計算した。ただし記号、外国語読み、数詞を含む文は対象外とした。

Kanji Trigram applied to Kana-Kanji Conversion
Jin’ichi MURAKAMI
ATR Interpreting Telephony Research Labs.

3.2 漢字かなのエントロピー

漢字かなの持つ情報量を調べるために、実験に用いた、文節単位の日本語の漢字かなの 0 重,1 重, 2 重,3 重のマルコフモデルのエントロピーを計算した。結果を表 1 に示す。なお比較のために音節のエントロピーも同時に計算した。

表 1: 漢字かなの持つエントロピー (bit)

	漢字かな	音節
0 重 (unigram)	8.15	5.67
1 重 (bigram)	4.45	4.29
2 重 (trigram)	2.87	2.94
3 重 (4-gram)	2.29	2.16

表 1 から以下のことがらが示される。

1. 漢字かなは音節と比較して数百倍の文字数からなるにもかかわらず、漢字かなの bigram,trigram,4-gram のエントロピーの値は音節と大きな差がない。したがって漢字かなのマルコフモデルは非常に高い情報量を持っていると予想される。

2. マルコフモデルの次数をあげるに従いエントロピーが減少している。これから高い次数のモデルほど高い精度のかな漢字変換が得られることが予想される。

3.3 かな漢字変換の変換精度

マルコフモデルを利用したかな漢字変換の変換精度を調べるために実験を行なった。実験は漢字かなの 1 重マルコフモデルおよび漢字かなの 2 重マルコフモデルの 2 つのモデルについて調べた。そして、３．１に示したマルコフモデルの連鎖確率に対するオープンデータとして 1982 年 1 月 1 日の日経新聞の記事を、クローズドデータとして 1982 年 1 年 5 日の記事を選んだ。入力は文節単位のひらがなとし、50 文節を実験した。使用した単語辞書の語彙数は約 6 万語である。また、計算中、連鎖確率が 0 のとき、すべての漢字かなまじり文が 0 になる可能性があるため、値を微小値 exp(-1000) に置き換えた。

漢字かなの 1 重マルコフモデルによるかな漢字変換の実験結果を表 2 に、2 重マルコフモデルの実験結果を表 3 に示す。

表 2: １重マルコフモデルによるかな漢字変換の実験結果

累積正解率 (%)	1 位	2 位	4 位	8 位
オープンデータ	60	72	86	94
クローズドデータ	88	98	100	100

表 3: 2 重マルコフモデルによるかな漢字変換の実験結果

累積正解率 (%)	1 位	2 位	4 位	8 位
オープンデータ	86	88	90	92
クローズドデータ	98	100	100	100

表 2 と 3 から示されるように、かな漢字変換に漢字かなの 2 重マルコフモデルを使用したとき 1 重マルコフモデルと比較して変換精度は飛躍的に高くなり、オープンデータで 86%、クローズデータでは 98%の正解率が得られた。

漢字かなの 1 重マルコフモデルを使用したときに、1 位に正解が出力されなかった文節の出力結果を、表４および表５に示す。表４はクローズドデータで、表５はオープンデータの実験結果である。この結果をみると、クローズドデータの誤りは表記が異なっているにしかすぎないと考えられるため、実質的にはクローズドデータの１位正解率は 100%と考えられる。また、オープンデータの誤りのなかで A,D は意味的に正しい文節であると考えられる。“デクニシ” は外国の人名であるため、これを漢字かなに変換するのは困難である。これらのことから、実際の使用における変換精度は、ここで示した値より高いと思われる。

表 4: クローズドデータにおける誤り

	正解	1 位出力
A	ときだけに	時だけに

表 5: オープンデータにおける誤り

	正解	1 位出力
A	しようと	使用と
B	デクニシ	出荷市区
C	国務省内	国務相ない
D	音楽好きの	音がく好きの
E	反核集会は	反核収かいは
F	きわみに	きわ味に

4 考察

4.1 単語のマルコフモデル

本報告で提案した方法は、単語のマルコフモデルにおいても原理的に動作可能である。単語は、漢字かなに比較して意味的にまとまった単位であるため、かな漢字の変換精度は、さらに向上すると予想される。しかし、日本語において単語の概念は曖昧であり、また単語は、漢字かなと比べるとマルコフの連鎖確率の収集において大量のデータが必要であるなどの問題点がある。

4.2 未知語の存在しないかな漢字変換

今回の実験では、単語候補の曖昧性を減らすために漢字かなの 2 重マルコフ連鎖確率のみを使用した。したがって、単語辞書のかわりに漢字かな 1 文字に対して読みが附属された漢字 1 文字辞書を使用しても、原理的には同じ方法でかな漢字変換が可能である。この場合、利点として、かな漢字変換において大きな問題点である未知語は存在しなくなるが、かな漢字変換の精度は下がると予想される。

4.3 連鎖確率の推定問題

このかな漢字変換で使用する漢字かなの連鎖確率は、サンプルデータが無限大にあったときに得られる “真の連鎖確率” を仮定している。しかし、実際には限られた数のサンプルデータから連鎖確率を推定する。この場合、もっとも問題になるのは、サンプルデータ中に出現しない連鎖の連鎖確率を、どのように推定するかということである。この実験では 2 重連鎖がないとき、連鎖確率を計算上の都合からも非常に小さい値 exp(-1000) に設定した。一方、連鎖確率を、 1 重連鎖確率 ×0.0001 と設定して同様な実験を行なったところ、オープンデータにおいて 1 位正解率は 90%まで上昇することが観測された。限られたサンプルデータから、連鎖確率を推定する方法としてとして、他に deleted interpolation[5] などが知られているが、今後この問題は検討する余地があると思われる。

5 まとめ

本論文では、漢字かなのマルコフモデルを用いたかな漢字変換の手法を提案し、実際のデータを用いてその変換精度を調べた。この結果、50 文節に対する実験では、漢字かなの 2 重マルコフモ

デルを使用したとき、1位正解率はオープンデータにおいて86%、クローズドデータにおいて98%が得られた。この結果単純な漢字かなの2重マルコフモデルを用いることによっても、高い変換精度を持つかな漢字変換が得られることが示された。

謝辞　この論文を書くにあたりATR自動翻訳電話研究所の嵯峨山氏および海木氏から、多くのコメントをいただきました。厚くお礼を申しあげます。

参考文献

[1] 荒木, 村上, 池原, “2重音節マルコフモデルによる日本語の文節音節認識候補の曖昧さの解消効果,” 情報処理学会論文誌, Vol.30, No.4, pp. 468-477 (1989)

[2] 村上, 荒木, 池原, “2重マルコフ連鎖確率モデルを使用した単音節音声入力の改善,” SP88-29, pp. 63-70 (1988)

[3] 長尾 真, “日本語情報処理,” 社団法人電子通信学会, pp. 63-64

[4] 宮崎, 大山, “日本文音声出力のための言語処理方式,” 情報処理学会論文誌, Vol.27, No.11 (1986)

[5] 花沢, 川端, 伊藤, 鹿野, “HMM音韻認識における音節連鎖統計情報の利用,” 日本音響学会講演論文集, 3-3-9, pp. 87-88 (1990)