自 然 言 語 処 理 105－2
（1995．1．20）

マルコフモデルを用いて漢字かな混じり文候補を選択する方法

荒木 哲郎[+] 池原 悟[++] 真田 陽一[+] 芳永 寛司[+++]
+福井大学
++NTT情報通信網研究所
+++松下電器産業

Abstract

これまでに、文節単位のべた書きかな文字列から変換された大量の漢字かな混じり候補を、文節内の漢字かな文字のマルコフモデル(文節マルコフモデルと呼ぶ)を用いて絞り込む方法が提案され、その有効性が示されている。

本論文では、従来の文節マルコフモデルを用いて絞り込まれた漢字かな混じり文節候補を、相互に組み合わせて作られる文候補ラテイスから、最も確からしい文候補を絞り込む方法(文マルコフモデルと呼ぶ)を提案する。実際に、43万語単語辞書を参照して、新聞記事77日分の統計データを用いた実験を行った結果、次のような知見を得た。

1. 文節マルコフモデルと文マルコフモデルを用いた絞り込み方法は、文マルコフモデルを単独で用いる方法より、累積正解率で5－10%程度優っていることがわかった。
2. 正規化された文マルコフモデルを用いる方法は、正規化しない文マルコフモデルを用いる方法より劣っていることがわかった。
3. 文マルコフモデルを用いる方法は、第一位の文節候補を全て用いて文候補を生成する方法に比べて、累積正解率で5%(標本外データ)－18%(標本内データ)改善されることがわかった。

A Method of Correctly Selecting Candidates of "kanji-kana" Character Strings Using Markov Model

Tetsuo ARAKI[+] Satoru IKEHARA[++] Hiroshi Yoshinaga[+++] Youichi Sanada[+]

+Fukui University
++NTT Network Information Systems Laboratories
+++Matushita Electric Industorial Company Limited

Up to now, it is known to be useful to select the most suitable phrase (called "bunsetsu" in Japanese) candidates of "kanji-kana" character strings, which are translated from non-segmented kana character strings, using Markov model of "kanji-kana" characters within "bunsetsu" (called *"Bunsetsu" Markov Model*).

This paper proposes a new method to correctly select sentence candidates of "kanji-kana" character strings, which are made from the combination of "bunsetsu" candidates of "kanji-kana" character strings determined by *"Bunsetsu" Markov Model* using Markov model of "kanji-kana" character in sentence (called *Sentence Markov Model*).

From the results of experiments, it is concluded that Sentence Markov Model is useful for selecting sentence candidates.

1 はじめに

日本語文は、通常約3000種の文字(特に漢字文字)を用いて書かれるため、コンピュータのファイルに入力することが容易ではない。べた書きのかな文を漢字かな混じり文に変換する方法については、これまでにもいろいろ研究されており、かな文については最長一致法による方法[1]、文節数最小法[2]、前後の連接文字を利用した方法[3]、格文法を用いる方法[4]、連語解析を用いる方法[5]、単語共起の関係を用いる方法[6]等があるが、同音語による曖昧さと分かち書き処理の曖昧さを同時に解決しなければならず、現在のところではまだ十分な精度を得るには至っていない。

総当たり法でかな漢字変換等の処理により生成される、あらゆる単語候補列の組み合せを考慮して解析を行うと、一般に探索木が爆発する問題が生じるが、文が文節に分割されているときには、このような探索木の爆発を防ぐことができる。これまでに2重マルコフモデルを用いて、べた書きかな文の仮文節境界を推定し、補正する方法が提案されており、その有効性が示されている[11]-[14]。

また文節単位のべた書きかな文字列から変換された大量の漢字かな混じり候補を、文節内の漢字かな文字のマルコフモデル(文節マルコフモデルと呼ぶ)を用いて絞り込む方法が提案され、その有効性が示されている[9][10][15][16]。

本論文では、従来の文節マルコフモデルを用いて絞り込まれた漢字かな混じり文節候補を、相互に組み合わせて作られる文候補ラティスから、最も確からしい文候補を絞り込む方法(文マルコフモデルと呼ぶ)を提案する。

実際に、43万語単語辞書を参照して、新聞記事77日分の統計データを用いた実験行ってその有効性を評価する。

2 基本的な定義と2重マルコフモデルを用いた文候補絞り込み方法

2.1 基本的な定義

文字列 $\Gamma = s_1 s_2 \cdots s_l$ によって表現される日本語文は、そのすべての要素 (s_i) がかな文字であるとき、かな文と呼ばれる。また文字列 のすべての要素 (s_i) が漢字またはかな文字であるとき、その日本語文を漢字かな混じり文と呼ぶ。かな文 Γ に対する日本語の漢字かな混じり文は、$\Lambda(\Gamma) = t_1 t_2 \cdots t_m$ によって表現され、その全ての要素 (t_j) がかな文字または漢字文字である。

日本語の文は、文節と呼ばれる構文の単位に分割できるとする。文節の定義は、[11] に従う。

かな文節から、漢字かな混じり文節候補の生成方法は、以下のようになされる。単語辞書はかな文字列をキー見出しとして、漢字かな表記の単語を読み出せるものとする。(1) かな文節を重複のない部分列に分割する。(2) 各々の分割されたかな文字部分列をキーとして辞書引きを行ない、すべてのかな部分列に対する単語候補が存在するとき、単語境界の整合 (もとのキーとなるかな文字列相互間に重複がない) がとれた単語を順次接続して得られた文節候補を、漢字かな混じり文節候補とする。これを存在するすべての単語候補の組合せに対して行なう。(3)(2) で漢字かな混じり文節候補が存在する場合のかな文節の分割の中で、最小な分割数 (これを最小分割数と呼ぶ) より一つ大きな分割数のあらゆる組合せ ((1) の分割) に対して、(2) を行なう。

日本語のかな文 Γ が文節 $\gamma_1, \gamma_2, \cdots,$ および γ_n (すなわち、$\Gamma = \gamma_1 \gamma_2 \cdots \gamma_n$), に分割されるとき、対応する漢字かな混じり文 $\Lambda(\Gamma)$ はそれぞれ文節 $\lambda(\gamma_1), \lambda(\gamma_2), \cdots,$ および $\lambda(\gamma_n)$ (即ち、$\Lambda(\Gamma) = \lambda(\gamma_1)\lambda(\gamma_2)\cdots\lambda(\gamma_n)$). に分割される。ここで日本語文の各文節は、一般に長さが2以上の漢字かな混じりの部分列であると仮定できる。

最初に、漢字かな混じり文節候補または文候補の最適性を評価するのに、用いられる2つのタイプのマルコフモデルを定義する。

[定義 1] 各文節が、$\lambda(\gamma_k) = t_{v(1)}t_{v(2)}\cdots t_{v(r)}$ によって与えられるとき、全ての文節に対する2重マルコフ連鎖確率 $p(t_j|t_{j-2}t_{j-1})$ ($v(1) \leq j \leq v(r)$) の集合を BMPと呼ぶ。. また各漢字かな混じり文が、$\Lambda(\Gamma_k) = t_1t_2\cdots t_m$と与えられたとき、全ての文に対する2重マルコフ連鎖確率 $p(t_j|t_{j-2}t_{j-1})$ ($1 \leq j \leq m$) の集合を SMPと呼ぶ。

ここで、2つの空白記号 (⊔ と表す) が、各文および各文節の先頭、末尾にに付加される。 ∎

BMP および SMP の例を、図1に示す。.

[定義 2] 各漢字かな混じり文節候補 (又は漢字な混じり文候補) $\lambda(\gamma_k) = t_{v1}t_{v2}\cdots t_{vr}$ (または $\Lambda(\Gamma) = t_1t_2\cdots t_m$) であるとき、文節候補 (または文候補) に対して、BMP (または SMP) を用いて、計算される次式の値 Cは、文節コスト (または文コスト) と呼ばれる。

$$C = -\sum_{i=1}^{n+2} \log_2 p(x_i|x_{i-2}x_{i-1})$$

ここで x_j は、$j < 0$ または $j > n$ のとき、空白記号 ⊔ を表す。 ∎

各文節γ_kに対する可能な全ての漢字かな混じり文節候補の集合を $\Omega(\lambda(\gamma_k))$ と表すとき、文節コスト C 値を用いて、$\Omega(\lambda(\gamma_k))$ の中の最適な漢字かな混じり文節候補を絞り込む方法はすでに提案されている。その結果によれば、正しい漢字かな混じり文節が第一位候補に含まれる割合は、83.7(標本外データ) ー 98.2 %(標本内データ) である。

この結果を文候補の生成に適用するために、第一位から第十位までの文節候補の集合を用いて構成される文候補ラテイスを、次のように定義する。ここで文節コスト C の値を用いて、漢字かな文節 $\Omega(\lambda(\gamma_k))$ から絞り込まれた第一位から第十位までの候補の集合を、$\Omega(\lambda(\gamma_k))^{(10)}$ と表す。

[定義 3][1] 日本語のかな文 Γ が文節 $\gamma_1, \gamma_2, \cdots,$ および γ_nに分割され (すなわち $\Gamma = \gamma_1\gamma_2\cdots\gamma_n$)、かな文に対する漢字かな文 $\Lambda(\Gamma)$ が、文節$\lambda(\gamma_1), \lambda(\gamma_2), \cdots,$ および $\lambda(\gamma_n)$、に分割されているとき、第一位から第十位までの漢字かな文節候補 $\Omega(\lambda(\gamma_k))^{(10)}$ $(1 \leq k \leq n)$ の列を、漢字かな混じり文候補ラテイスとよび、次のように表す。$L(\Lambda(\Gamma))$ $= \Omega(\lambda(\gamma_1))^{(10)}\Omega(\lambda(\gamma_2))^{(10)}\cdots\Omega(\lambda(\gamma_n))^{(10)}$.

但し、正しい漢字かな混じり文節は、常に文節候補の各集合$\Omega(\lambda(\gamma_k))^{(10)}$ に含まれているものと仮定する。 ∎

文候補ラテイス $L(\Lambda(\Gamma))$ の例を図2に示す。

2.2 2重マルコフモデルを用いた文候補絞り込み方法

本章では、文候補ラテイスから得られる漢字かな混じり候補の内、最適な文を選択する3つの方法を定義する。

[文候補絞り込み法 1]

文候補ラテイスから得られる最適な漢字かな混じり文候補を、文コスト Cが最小の値を持つ文をであると定義する方法を、S-方法と呼ぶ。 ∎

[文候補絞り込み法 2] 文候補ラテイスから得られる最適な漢字かな混じり文候補を、文の中の文字数 L によって正規化された最小の文コスト Cの値 (すなわち、C/L) を持つ文をであると定義する方法を、NS-方法と呼ぶ。 ∎

[文候補絞り込み法 3] 文候補ラテイスから得られる最適な漢字かな混じり文候補を、文コスト C_1と文節コスト C_2の和が最小な値を持つ文をであると定義する方法を、$(S + B)$-方法と呼ぶ。 ∎

[1] この実験においては、日本語音声出力システムから得られる読みの情報 [16] が、文節コストの他に文節候補の絞り込みに用いられている。これらの情報を用いた文節候補の絞り込みを行なうことによって、正解率が改善され、多くの正しい漢字かな混じり文節候補が、10 位より小さい順位の候補に含まれれることに注意する。

3 実験結果

3.1 実験条件

1. マルコフ連鎖確率の統計に用いられる総文節数: 70 日分の新聞記事データで、283,963 文節数
2. 一文節当たりの平均文字数:6
3. 三つの文候補絞り込み方法を、評価するのに用いられる文、漢字かな文字、文節の数:
   (a) 文の数: 229 文 (標本外データ), 235 文 (標本内データ)
   (b) 漢字かな文字数: 5583 文字 (標本外データ), 5076 文字 (標本内データ)
   (c) 平均文節数: 6.5(標本外データ), 5.7(標本内データ)
4. 辞書の単語数:430,000 語

3.2 実験結果

2 章で述べた三つの文候補絞り込み法を用いて選択された第一位から第十位までの候補の中に 、正しい漢字かな混じり文が含まれる割合 (累積正解率) を求めた実験結果を、図3(標本外データ) および図4(標本内データ) に示す。また三つの文候補絞り込み法を用いて得られた第一位から第十位までの候補の例を図6 に示す。

[1] 三つの文候補絞り込み法による累積正解率の比較:

図3 および図4 から、以下の結果が求まる。

1. $(S+B)$-方法は、S-方法よりも 5-10 %優れている。このことは文候補ラテイスから、漢字かな混じり文候補を正しく選択する際に、文コストと文節コストの両方の値の組合せて評価する方が、文コスト単独で選択するよりも有効であることを意味している。
2. NS-方法が S-方法よりも悪い。これはいろいろな長さ (文の中の漢字かな文字数) の文候補が、文コストの値を用いて評価されるとき、文コストは文の長さによって正規化されない方が有効であることを意味している。

[2] 全ての第一位文節候補より構成される文候補に対する文候補絞り込み法の改善効果

文コスト C_1 と文節コスト C_2 の両方の値に基づいた文候補絞り込み法により得られる文候補の累積正解率と、すべての第一位の文節候補 (すなわち、$\Omega(\lambda(\gamma_k))$, $(1 \leq k \leq n)$ の中の第一位の候補を用いる) から構成される文候補の累積正解率の関係を図5 に示す。本図から、文候補絞り込み法による正解率は、5 %(標本外データ) - 18 %(標本内データ) 改善されることがわかる。

これらの結果から、2 重マルコフモデルを用いた文候補絞り込み法は、曖昧な漢字かな混じり文節候補から構成される文候補ラテイスより漢字かな混じり文候補を正しく選択するのに有効であることがわかる。

4 結論

本論文は、文節マルコフモデルによって決定された漢字かな混じり文節候補の組合せから構成される漢字かな混じり文候補ラテイスより、文マルコフモデルを用いて文候補を正しく選択する方法を提案した 。

実験結果から、以下の結論を得た。

1. $(S+B)$-方法は、S-方法よりも 5-10 %優れている。
2. NS-方法は S-方法よりも 悪い。
3. 文コストと文節コストの両方の値に基づいた文候補絞り込み法により得られる文候補の累積正解率は、すべての第一位の文節候補から構成される文候補の累積正解率よりも、5 %(標本外データ) - 18 %(標本内データ) 改善されることがわかった。

これらの結果から、2重マルコフモデルを用いた文候補絞り込み法は、曖昧な漢字かな混じり文節候補から構成される文候補ラテイスより漢字かな混じり文候補を正しく選択するのに有効であることがわかった。

今後はさらに長さの長い文候補ラテイスから、漢字かな混じり文を正しく絞り込む方法について、さらに研究することなどがあげられる。

参考文献

1. 牧野, 木沢:" べた書き文の分かち書きとかな漢字変換ー二文節最長一致法による分かち書き" , 情処論,20,4,pp.337-245 (1979)

2. 吉村、日高、吉田:"文節数最小法を用いたべた書き日本語文の形態素解析"、情処論,241,pp.40-46 (1983)

3. 栃内、伊藤、鈴木:"前後連接文字を利用した同音語選択機能を有するかな漢字変換システム", 情処論,27,3,pp.313-320 (1986)

4. 大島、阿部、湯浦、武市:"格文法による仮名漢字変換の多義解消", 情処論,27,7,pp.679-687 (1986)

5. 本間、山階、小橋:"連語解析を用いたべた書きかな漢字変換", 情処論,27,11,pp.1062-1067 (1986)

6. 内山、板橋:"共起関係を利用した日本語複合名詞の分割", 情処N L研究会,91-7,pp.57-64 (1992)

7. 宮崎: " 係り受け解析を用いた複合語の自動分割", 情報処理, Vol.25, 6, pp970-979, 623-656 (1984)

8. 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, Vol.27, 11, pp1053-1061 (1986)

9. 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, Vol.30, 4, pp467-477 (1989)

10. 村上、荒木、池原: " 日本語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かな交じり候補の抽出精度", 信学論, Vol.J75-DII, pp11-20 (1992)

11. 荒木、池原、土橋: " 2重マルコフ連鎖モデルを用いたべた書き日本語文の文節先頭位置推定法の評価", 情処N L研究会, Vol.94-8, pp55-61 (1993)

12. 荒木、池原、土橋: " べた書きかな文の仮文節境界の補正方法", 情処自然言語処理研究会, 98-1, pp.1-7 (1993)

13. T. Araki, S. Ikehara and J. Tutihashi: " A New Method of Finding Provisional Boundaries of "bunsestu" using 2nd-Order Markov Model", 2nd IEEE Int. Workshop on Robot and Human Communication ,pp.114-119 (1993)

14. T. Araki, S. Ikehara and J. Tutihashi: " A Method of Correcting Provisional Boundaries of "Bunsestu", 3nd IEEE Int. Workshop on Robot and Human Communication ,pp.289-293 (1994)

15. 荒木、池原、芳永、真田: " マルコフ連鎖モデルによる文節かな漢字変換候補の絞り込み方法", 情処N L研究会, Vol.99-6, pp41-48 (1994)

16. 荒木、池原、横川、真田: " 日本語文音声出力からの読み情報を用いた漢字かな混じり文節候補の絞り込み", 情処N L研究会, Vol.102-15, pp113-120 (1994)

␣␣契約の␣␣

p（約｜␣契）　　␣：空白記号

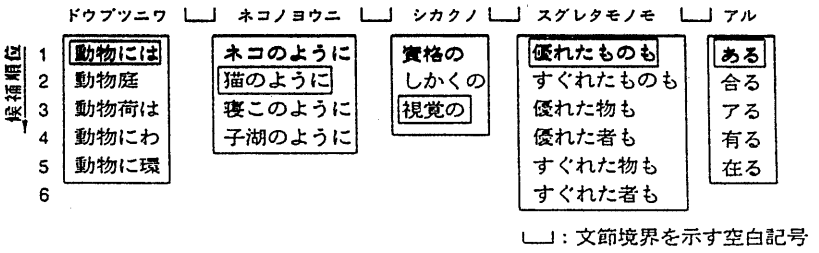
(a) 文節マルコフモデル

␣␣法と契約の社会でしょう␣␣

p（約｜と契）

(b) 文マルコフモデル

図1. マルコフモデルのタイプ



␣：文節境界を示す空白記号

図2. 漢字かな混じり文候補ラティスの例

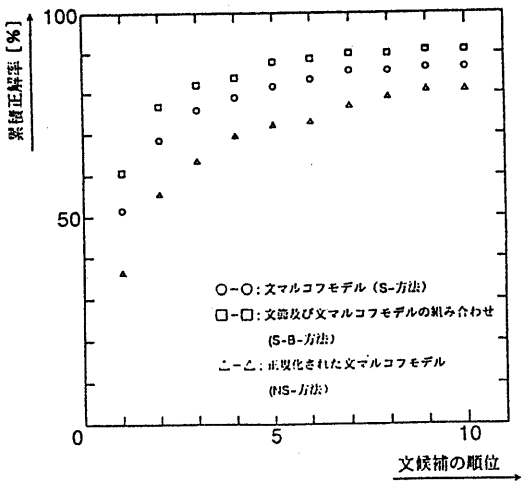


図3. 文マルコフモデルを用いた文候補絞り込みの実験結果（標本外）

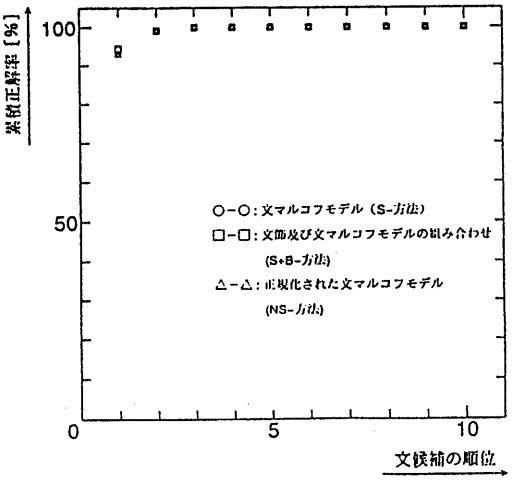


図4. 文マルコフモデルを用いた文候補絞り込みの実験結果（標本内）

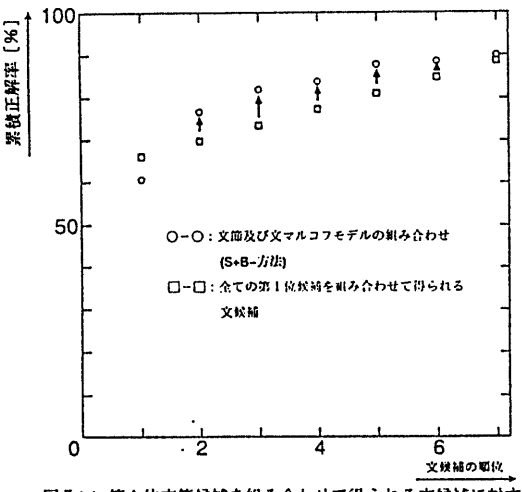


図5 (a). 第1位文節候補を組み合わせて得られる文候補に対する文マルコフモデルの改善効果（標本外）

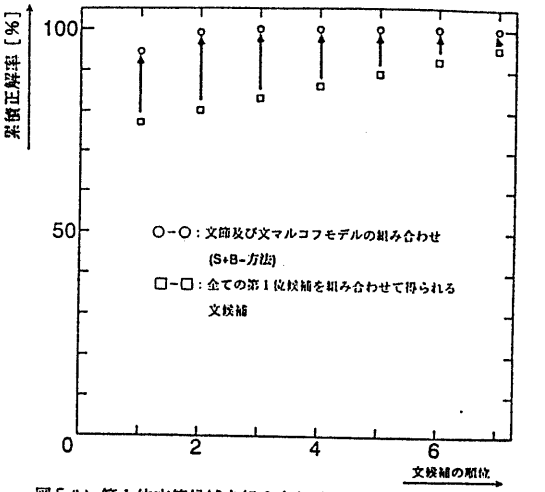


図5 (b). 第1位文節候補を組み合わせて得られる文候補に対する文マルコフモデルの改善効果（標本内）

"キュウジツニワ　カゾク　ソロッテ　サバクヘ　ピクニック"

候補順位	
1	休日には家族そろって砂漠えピクニック
2	休日には家族そろって砂漠へピクニック
3	休日には家族そろって砂漠エピクニック
4	急実には家族そろって砂漠えピクニック
5	急実には家族そろって砂漠へピクニック
6	急実には家族そろって砂漠エピクニック
7	休日には家族そろって砂漠えピクニック
8	休日には家族そろって砂漠へピクニック
9	休日には家族そろって砂漠エピクニック
10	休日にはか俗そろって砂漠えピクニック

図6.　文マルコフモデルを用いて得られる１０位までの漢字かな混じり候補の例