2重マルコフモデルを用いたべた書きかな文の仮文節境界の推定方法

荒木哲郎[†1] 池原悟[†2]
土橋潤也[†3] 笹島伸一[†4]

べた書きかな文のかな漢字変換精度を向上させるためには，変換の過程で正解を漏らさないように，辞書から，かな文字列に含まれる単語候補をすべて抽出して組み合わせて評価することが必要であるが，文の長さが長くなるにつれて単語候補の組合せの数が増大し解析が困難となる問題がある．従来，べた書きの漢字かな混じり文の場合は，字種の変化点に着目して仮文節境界を決定する方法が提案されているが，この方法は字種が，かな文字に限定されるべた書きかな文には適用できない．かな文の場合も，何らかの方法で仮文節境界を見つけることができれば，解析の困難さの問題は解決できると期待される．本論文では，かな文字列の連鎖確率の変化点に着目した仮文節境界の推定法を提案する．具体的には，マルコフ連鎖確率モデルによる仮文節境界の推定法を，(1) 文節境界の学習の有無，(2) 連鎖確率の変化点の再評価の有無，および (3) マルコフ連鎖確率の適用法の違いの3点に着目して，8通りに分けて評価した．その結果，文節境界を学習したデータを用いて連鎖確率の落ち込む点を抽出し，その点に文節境界の存在を仮定して再評価する方法が最も優れていること，また，その際，マルコフ連鎖確率は前方向，後方向を組み合わせて使用するのが良いことが分かった．この方法によって推定された仮文節境界の精度は，適合率94.0%，再現率76.8%で，従来，漢字かな混じり文の解析で使用されている仮文節境界推定法（字種の変化点に着目する方法）の精度よりも良いことから，提案したマルコフ連鎖確率モデルの方法はべた書きかな文の解析に有効と判断できる．

A Method of Finding the Provisional Boundaries of "Bunsetsu" for Non-segmented "Kana" Sentences Using 2nd-order Markov Model

TETSUO ARAKI,[†1] SATORU IKEHARA,[†2] JYUNYA TUCHIHASHI[†3] and SHINICHI SASAJIMA[†4]

In order to improve the precision to translate from the non-segmented "Kana" sentences into "Kanji-Kana" sentences, it is necessary to examine all of the word candidates extracted from the dictionary for the sentence. However, the amount of computer memories required for the translating processing explodes in many times, because the number of the combinations of candidates for "Kanji-Kana" words grows rapidly in propotion to the length of the sentence. The memory explosion can be prevented if a sentence is separated into "bunsestu". Therefore, a method to correctly find the boundaries of bunsetsu are considered to be a key technique to improve the precision of "Kana"-"Kanji" translation. However, the useful method to find them are not known yet. This paper proposes a new method of finding provisional boundaries of "bunsetsu" for non-segmented "Kana" sentences using 2nd-order Markov model. "Precision factor" and "Recall factor" for provisional boundaries of "bunsetsu" determined by this method, were experimentally evaluated using the statistical data for 70 issues of a daily Japanese newspaper.

†1 福井大学工学部電子工学科
Department of Electrical Electronics Engineering, Faculty of Engineering, Fukui University
†2 鳥取大学工学部知能情報工学科
Department of Information and Knowledge Engineering, Faculty of Engineering, Tottori University
†3 NTT法人営業本部
NTT Business Communications Headquarters
†4 株式会社SRA
Software Research Associates, Inc.

1. はじめに

漢字かな混じり日本文を計算機に入力する方法として，かな入力された日本文をかな漢字変換によって，漢字かな混じり文字列に変換する方法が行われている．また，音声認識においても，音節単位に音声処理を行うような場合は，音節表記から，かな漢字混じりの日本文に変換することが必要である．かな漢字変換の方法としては，最長一致法[1]，二文節最長一致法[2]，

格文法を応用した方法[5]，連語解析による方法[6]，最近使用語の優先学習方式[12]などが提案されているが，数十文字以上の長い文字列単位の変換では，十分な精度を得ることは難しく，さらに変換精度の向上が期待されている．高い変換精度を得るには，変換の過程で正解を漏らさないようにすることが必要で，辞書からかな文字列に含まれる単語候補をすべて抽出してそれを組み合わせ評価することが要求されるが，文が長くなるにつれて単語候補の組の数が幾何学的に増大し，解析が困難となる問題がある．従来，この問題を避けるため，ワープロなどを使用した日本文入力では，いくつかの文節のまとまり程度の長さに区切って入力されたかな文を変換の対象としている場合が多い．しかし，かな文字列でべたに入力された文の場合でも，適当な長さ（数文節程度）ごとに，文節境界に相当する区切りを自動的に発見することができれば，文単位のかな漢字変換を高精度に実現できる可能性がある．したがって，今後，文の単位でのかな漢字変換の精度を向上させるための技術の1つとして，文節境界を自動的に決定する技術が重要と考えられる．

一方，漢字かな混じり日本文の形態素解析技術について見ると，係り受け解析を行う方法[4),7]，語基と接辞を組み合わせたマルコフモデルを用いる方法[8]など，多くの技術が開発されており，最近では，解析精度の一層の向上を狙って，解析結果自動書き替えによる解析誤り回復処理の研究[11]が行われている．かな漢字変換の場合と同様，文字列の長さが大きくなると，単語解釈の組の数が増大し解析困難となる問題があるが，形態素解析では文字列上の字種（漢字，ひらがな，カタカナ）の変化点に着目して仮文節境界を定め，それをもとに解析の単位を決めることにより，この問題を解決している．字種の変化点に着目した仮文節境界は，必ずしも正しい文節境界とはいえず，かなりの頻度で誤りを含むが，形態素解析ではそれを高い精度で補正する方法が実現されている．

この仮文節境界の補正技術は，かな漢字変換にも応用できると期待されるから，文単位のかな漢字変換のために必要とされる文節境界の精度は，漢字かな混じり文の場合と同様，必ずしも高精度である必要はないと考えられる．経験的には再現率50％以上で，適合率90％程度の仮文節境界が得られれば，単語辞書引きや品詞接続属性情報，ならびに従来の漢字かな交じり文に対する形態素解析技術などを用いることにより，誤った文節境界を自動的に補正することができるものと期待される．

そこで本論文では，マルコフ連鎖モデルを用いて，べた書きされたかな文の文節境界を推定する方法を提案する．マルコフ連鎖確率モデルは，音声認識における音節認識候補の絞り込み[9]や，漢字かな混じり文節候補の絞り込み[10]に効果のあることが知られている．これらの方法では，文字列間の結合の強度に着目し，結合強度の強いものを正解率の高いものとして抽出している．これに対して，仮文節境界の場合には，逆に結合強度の弱いところが抽出の対象となる．すなわち，文字列の結合強度は，単語内では強く，単語間，文節間の順に弱くなると考えられるから，かな文字列上，ある適当な結合強度以下の結合強度を持つ位置が文節境界であると推定される．そこで本論文では，かな文の文字連鎖の強度を以下の3つの観点から評価する．

まず第1の観点は，文節境界《に関する学習の》有無である．仮文節境界の判定で，べた書きかな文の標本から得られた文字連鎖確率を使用する場合（学習なし）と文節境界付きのかな文の標本から得られた文字連鎖確率を使用する場合（学習あり）を比較する．

第2の観点は，文節境界を判定する方法の違いである．かな文の連鎖確率が単にある値以下に落ち込んだ位置を文節境界とする方法と落ち込んだ位置と落ち込まない位置との違いを学習データを用いて再評価する方法を比較する．

第3の観点は，文字連鎖確率の適用の方向の違いである．従来，日本語誤字訂正候補の抽出等においてマルコフ連鎖モデルを使用する場合，前方連鎖確率よりもむしろ後方連鎖確率を適用する方が精度良い候補が抽出できることが知られている[3]．そこで，ここでも，文字連鎖確率の適用方向の違いによる文節境界推定精度の差を評価する．

以上から，具体的には，これらの3つの観点の違いによる仮文節境界推定精度の違いを8つのモデルによって実験的に比較評価し，最適な仮文節境界推定法を決定する．また，同時に，その方法で得られる仮文節境界の適合率と再現率を求める．

2. マルコフモデルを用いた文節境界の推定法

日本語文の文字列 $\alpha = x_1x_2\cdots x_n$ を考える．x_i $(1 \leq i \leq n)$ は，日本語文字で，ひらがな，漢字，カタカナ，記号のいずれかである．ただし，ここでは記号を除く日本語文字を，字種上かなと漢字の2種類に分けて扱うこととし，ひらがなを単に《かな文字》，漢字とカタカナを《漢字》と呼ぶ．次に日本語文 α に含まれるすべての文字 x_i が，すべてかな文字である場合 α を《べた書きかな文》と呼ぶ．

日本語文は，通常文節と呼ばれる統語上の単位に分

けることができる．ここでは，以下のように文節を1つの自立語と0個以上の付属語から構成された単位と考える．

文節 = { 自立語 } { 付属語 }

ただし，自立語 = { 名詞，動詞，副詞，形容詞，形容動詞，接続詞等 }

付属語 = { 助詞，助動詞，接頭語，接尾語 }

2.1 文と文節マルコフ連鎖確率

2.1.1 マルコフ連鎖確率を用いた仮文節境界推定の基本的な考え方

文字列の結合強度は，単語内では強く，単語間，文節間の順に弱くなると考えられる点に着目して文節境界を推定する．そのため，まず，大量のべた書きかな文データ（テキストデータ）から文字列の連鎖確率を求め，それを用いて，与えられたかな文の各文字位置の連鎖確率を評価する．その結果，文字連鎖の確率がある足切り値以下の連鎖確率を示す（落ち込む）文字の位置の直前を，仮文節の境界と判定する．

2.1.2 文節境界の学習を考慮した仮文節境界推定法の考え方

文字連鎖確率を評価するとき，文節境界が分かっている標本データ（文節境界記号が付与されたかな文）から得られた連鎖確率を用いれば，より精度良く，文節境界を判定できると期待される．

文節末の文字は，助詞や助動詞などの比較的限定された文字となる場合が多いから，これらの特定の文字の後が文節境界となる確率は，その他の文字が来る確率よりも大きいと考えられる．したがって，与えられたかな文字列の連鎖確率値を，文節境界の学習を行ったデータを用いて評価すると，文節境界の位置では，2.1.1 項の文節境界の学習なしデータを用いる場合に比べて，より大きく落ち込むことが予想される．また，連鎖確率値が落ち込んだところが文節境界であることを確認するために，文節境界区切り記号を付加した後，学習を行ったデータを用いて再評価すれば，文字連鎖確率は，通常の文節内の文字連鎖確率に比べてより大きい値に立ち直ると期待される．逆に，文節境界以外の位置に文節区切り記号を挿入した場合は，連鎖確率値は文節内の文字の連鎖確率より小さな値になると推察される．

学習ありのモデルでは，これらの特徴を利用して文節境界を推定する．まず，あらかじめかな文の文節境界に文節の区切り記号を挿入した標本データからマルコフ連鎖確率を求めておくこととし，文節境界を以下の2つのステップで推定する．第1のステップでは，与えられたかな文字列に対して今求めたマルコフ連鎖確率を適用して連鎖確率の落ち込んだところを抽出する（図2(1)）．

次に，第2のステップでは，すべての文字位置を対象に，順に文節の区切り記号を入れて，連鎖確率値の値が立ち上がるか，逆に下がるかを評価する（図2(2), (3)）．第1のステップで連鎖確率が落ち込んだところの中で，第2のステップで立ち上がりのあったところを文節境界と判断する．

2.1.3 マルコフ連鎖確率の定義

文節境界の学習を行わない場合と，学習を行う場合について，マルコフ連鎖確率モデルを定義する．

日本語文を $x_1x_2\cdots x_n$ と表すとき，文の先頭，末尾位置に，文の境界を示す空白文字を付加したもの，すなわち $\sqcup x_1x_2\cdots x_n\sqcup$ およびを空白文字付き文と呼ぶ．文の場合と同様に，日本語の文節を $x_ix_{i+1}\cdots x_{i+k}$ と表すとき，文節の先頭，末尾位置に，文節の境界を示す空白文字を加したもの，すなわち $\sqcup x_ix_{i+1}\cdots x_{i+k}\sqcup$ を空白文字付き文節と呼ぶ．

このとき2重マルコフ連鎖確率を次のように定義する．

[定義1] 空白文字付き文 $\sqcup x_1x_2\cdots x_n\sqcup$，および空白文字付き文節 $\sqcup x_ix_{i+1}\cdots x_{i+k}\sqcup$ において，次のように表される条件付き確率 $P_{j-2,j-1}(j)$，

$$P_{j-2,j-1}(j) \equiv P(x_j|x_{j-2}x_{j-1}) \quad (1)$$

を文字 x_j の2重マルコフ連鎖確率と定義する．空白文字付き文および空白文字付き文節のマルコフ連鎖確率をそれぞれ文マルコフ連鎖確率および文節マルコフ連鎖確率と呼び，それぞれの集合を $\mathbf{P(S)}$ および $\mathbf{P(B)}$ と表す．

ここで j は，$1 \le j \le n+1$（文の場合），また $i \le j \le i+k+1$（文節の場合）であり，文節の先頭文字および末尾の空白文字におけるマルコフ連鎖確率はそれぞれ次のように表される．すなわち，文節先頭文字が x_i の場合，前者は

$$P_{\sqcup,\sqcup}(i) \equiv P(x_i|\sqcup,\sqcup) \quad (2)$$

および文節末尾文字が x_{i+k} の場合，後者は

$$P_{i+k-1,i+k}(\sqcup) \equiv P(\sqcup|x_{i+k-1}x_{i+k}) \quad (3)$$

である☆． □

上述の文マルコフ連鎖確率は，文節境界の位置がどこにあるかをまったく考慮せずにマルコフ連鎖確率を求めているのに対して，文節マルコフ連鎖確率の場合は，文節境界の位置を表す空白文字と文節内の各文字を明確に区別してマルコフ連鎖確率を求めている．こ

☆ 文節先頭および末尾位置で2重マルコフ連鎖確率が定義可能なように便宜的に空白文字を2つ付加している．

れは前者のマルコフモデルが文節境界の位置を明確には学習させていないのに対して，後者のモデルは文節境界の位置を明確に学習させていることに相当している．

次に文字列の連結強度を詳細に調べて，さらに有効な仮文節境界推定法の可能性を比較検討していくために，通常のマルコフ連鎖確率では前から後に向かって文字の連鎖を求める（本論文ではこれを順方向と呼ぶ）のに対して，これと反対に後から前に向かって文字の連鎖を求めるマルコフ連鎖モデルを定義する．

[定義 2] 空白付き文節 $\sqcup x_i x_{i+1} \cdots x_{i+k} \sqcup$ の j 番目の文字 x_j に対して，次のように表される条件付き確率 $P_{j+2,j+1}(j)$

$$P_{j+2,j+1}(j) \equiv P(x_j | x_{j+2} x_{j+1}) \tag{4}$$

を逆方向の2重マルコフ連鎖確率と呼ぶ． □

以上より，2重の文節マルコフ連鎖確率の集合 **P(B)** を，順方向および逆方向のマルコフ連鎖確率の集合に分け，それぞれ **P(FB)** および **P(BB)** で表す．**P(S)**，**P(FB)** および **P(BB)** の2重文節マルコフ連鎖確率の例を**図1**に示す．

2.2 仮文節境界の推定法

前節で定義された文マルコフ連鎖確率の落込みを調べて判断する方法を，学習なしマルコフモデルによる推定法（NL法）と呼ぶ．同様に，文節マルコフ連鎖確率の落込みを調べて判断する方法を，学習ありマルコフモデルによる推定法（L法）と呼び，また文節の区切り記号を入れた連鎖確率値によって立ち直ることを確認する方法をBL法と呼ぶ．L法およびBL法は，さらに文字連鎖の方向が順方向である場合，それぞれFL法およびFBL法と呼び，また逆方向の場合，それぞれBL法およびBBL法と呼ぶことにする．

次に学習なしマルコフモデルによる推定法（NL法）および学習ありマルコフモデルによる推定法（FL法，FBL法，BL法，BBL法）において，仮文節境界の位置 j を，どのように推定するかを示す．

[定義 3] 定数 T に対して，次の条件を満足する位置 j を仮文節境界と定める．

(1) *NL* 法：かな文字の文マルコフ連鎖確率（**P(S)** の要素）の中から選ばれた足切り値 T_1 に対して，かな文字 x_j の連鎖確率が次の条件を満たす．

$$P_{j-2,j-1}(j) < T_1 \tag{5}$$

ただし $P_{j-2,j-1}(j), T_1 \in \mathbf{P(S)}$

(2) *FL* 法：かな文字の順方向の文節マルコフ連鎖確率（**P(FB)** の要素）の中から選ばれた足切り値 T_1 に対して，かな文字 x_j の連鎖確率が次の条件を満たす．

$$P_{j-2,j-1}(j) < T_1 \tag{6}$$

ただし $P_{j-2,j-1}(j), T_1 \in \mathbf{P(FB)}$

(3) *FBL* 法：空白文字の順方向の文節マルコフ連鎖確率（**P(B)** の要素）の中から選ばれた足切り値 T_2 に対して，かな文字 x_j の連鎖確率が次の条件を満たす．

$$P_{j-2,j-1}(\sqcup) > T_2 \tag{7}$$

ただし，$P_{j-2,j-1}(\sqcup), T_2 \in \mathbf{P(FB)}$

(4) *BL* 法：かな文字の逆方向の文節マルコフ連鎖確率（**P(BB)** の要素）の中から選ばれた足切り値 T_1 に対して，かな文字 x_j の連鎖確率が次の条件を満たす．

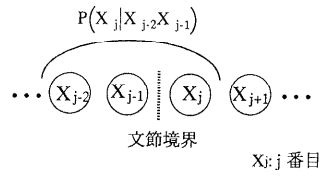
$$P_{j+1,j}(j) < T_1 \tag{8}$$

ただし $P_{j+1,j}(j), T_1 \in \mathbf{P(BB)}$

(5) *BBL* 法：空白文字の逆方向の文節マルコフ連鎖確率（**P(BB)** の要素）の中から選ばれた足切り値 T_2 に対して，文節境界文字 $\sqcup$ の連鎖確率が次の条件を満たす．

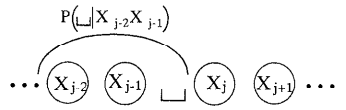
$$P_{j+1,j}(\sqcup) > T_2 \tag{9}$$

ただし，$P_{j+1,j}(\sqcup), T_2 \in \mathbf{P(BB)}$ □



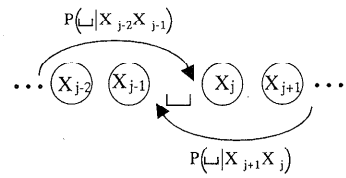
Xj: j 番目に現れる文字

(a) 2重の文マルコフ連鎖確率モデル
(a) 2nd-order Markov model of sentence



Xj: j 番目に現れる文字
␣:文節境界を表す空白文字

(b) 2重の文節マルコフ連鎖確率モデル
(b) 2nd-order Markov model of "bunsestu"



(c) 順方向および逆方向の2重文節マルコフ連鎖確率モデル
(c) Forward and backward types of 2nd-order Markov model

図1 2重の文ないし文節マルコフ連鎖確率モデル
Fig. 1 Two models of 2nd-order Markov propability of "bunsetsu".

表1 仮文節境界の推定法の概要
Table 1 Outline methods to find the provisional boundaries of bunsetsu.

方法	方法の要点	条件式	連鎖確率の種類と T_i の数
NL 法	文マルコフ連鎖確率値の落込みによる判定	$P(x_j\|x_{j-2}x_{j-1}) < T_1$ $(T_1 \in \mathbf{P(S)})$	文マルコフ連鎖確率 $\mathbf{P(S)}$ 足切り値の個数 = 1
FL 法	順方向の文節マルコフ連鎖確率値の落込みにより判定	$P(x_j\|x_{j-2}x_{j-1}) < T_1$ $(T_1 \in \mathbf{P(FB)})$	順方向の文節マルコフ連鎖確率 $\mathbf{P(FB)}$ 足切り値の個数 = 1
FBL 法	空白文字の順方向文節マルコフ連鎖確率値の向上により判定	$P(\sqcup\|x_{j-2}x_{j-1}) > T_2$ $(T_2 \in \mathbf{P(FB)})$	順方向の文節マルコフ連鎖確率 足切り値の個数 = 1
BL 法	逆方向の文節マルコフ連鎖確率値の落込みにより判定	$P(x_{j-1}\|x_{j+1}x_j) < T_1$ $(T_1 \in \mathbf{P(BB)})$	逆方向の文節マルコフ連鎖確率 $\mathbf{P(BB)}$ 足切り値の個数 = 1
BBL 法	空白文字の逆方向文節マルコフ連鎖確率値の向上により判定	$P(\sqcup\|x_{j+1}x_j) > T_2$ $(T_2 \in \mathbf{P(BB)})$	逆方向の文節マルコフ連鎖確率 足切り値の個数 = 1
FL + FBL 法	FL 法と FBL 法の 2 者の組合せによる判定	$P_1(x_j\|x_{j-2}x_{j-1}) < T_1$ $P_2(\sqcup\|x_{j-2}x_{j-1}) > T_2$ $(T_1, T_2 \in \mathbf{P(FB)})$	順方向の文節マルコフ連鎖確率 足切り値の個数 = 2
FL + FBL + BL 法	FL 法, FBL 法と BL 法の 3 者の組合せによる判定	$P_1(x_j\|x_{j-2}x_{j-1}) < T_{11}$ $P_2(\sqcup\|x_{j-2}x_{j-1}) > T_{12}$ $P_3(x_{j-1}\|x_{j+1}x_j) < T_{21}$ $(T_{11}, T_{12} \in \mathbf{P(FB)}$ および $T_{21} \in \mathbf{P(BB)})$	順方向および逆方向の文節マルコフ連鎖確率 足切り値の個数 = 3
FL + FBL + BL + BBL 法	FL 法, FBL 法, BL 法, および BBL 法の 4 者の組合せによる判定	$P_1(x_j\|x_{j-2}x_{j-1}) < T_{11}$ $P_2(\sqcup\|x_{j-2}x_{j-1}) > T_{12}$ $P_3(x_{j-1}\|x_{j+1}x_j) < T_{21}$ $P_4(\sqcup\|x_{j+1}x_j) > T_{22}$ $(T_{11}, T_{12} \in \mathbf{P(FB)}$ および $T_{21}, T_{22} \in \mathbf{P(BB)})$	順方向および逆方向の文節マルコフ連鎖確率 足切り値の個数 = 4

注) 表の (FL + FBL) 法, (FL + FBL + BL) 法および (FL + FBL + BL + BBL) 法の条件式欄において示されている各条件式は同時に成り立つこと (AND 条件) を表している.

定義 3 に述べた 5 つの方法以外に, 学習ありの方法 (FL, FBL, BL, BBL 法) を組み合わせた 3 つの推定法をまとめて**表 1** に示す.

2.3 仮文節境界の評価方法

2.3.1 評価パラメータ

仮文節境界の推定法の能力は, 以下に示す適合率と再現率の両者で評価する.

[定義 4] 仮文節境界に対する適合率 P および再現率 R を次のように定義する.

$$P \equiv \frac{\text{仮文節境界の中で正しく設定された文節境界の数}}{\text{設定された仮文節境界の総数}}$$

$$R \equiv \frac{\text{仮文節境界の中で正しく設定された文節境界の数}}{\text{正しい文節境界の総数}}$$

□

2.3.2 足切り値のタイプ, 初期値の決め方および再現率, 適合率の関係

[1] 足切り値のタイプ

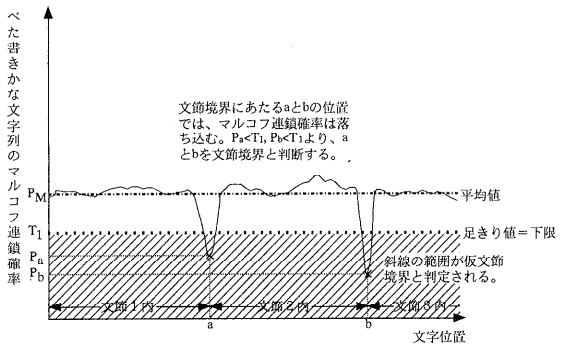
2.2 節で述べた仮文節境界の推定法では, **図 2** のように 2 種類の足切り値が用いられる.

第一のタイプは, 連鎖確率の落込みの判定に用いるもので, その足切り値より小さい場合に, 文節境界と判断する. すなわち図 2 (1) において, 文節境界をまたがる文字列 (たとえば, 文字位置 a や b) の連鎖確率値は, 文節内の文字列に対する連鎖確率値の平均値 P_M に比べて値が小さくなる (落ち込む) と考えられることから, 足切り値 (T_1) より小さな値を持つ位置 a や b を文節境界と判断する.

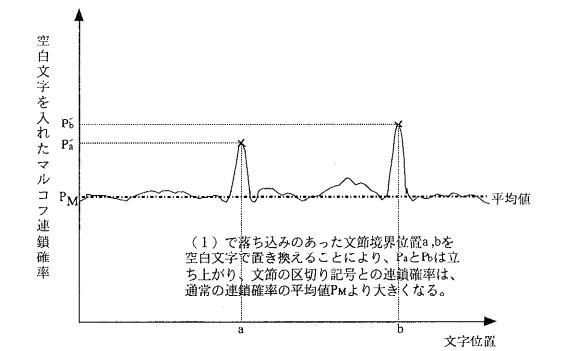
一方, 第二のタイプは, 文節の区切り記号を入れた連鎖確率の立ち上がりの判定に用いるもので, その足切り値より大きい場合に, 文節境界と判断する. すなわち図 2 (2) に示すように, 図 2 (1) で落込みがあった文節境界のところ (文字位置 a や b) に空白文字を入れて連鎖確率値を再評価し, その連鎖確率値が T_2 より大きくなる (立ち上がる) 場合, 文節境界とする.

[2] 再現率と適合率の積が最大となる足切り値の決定法

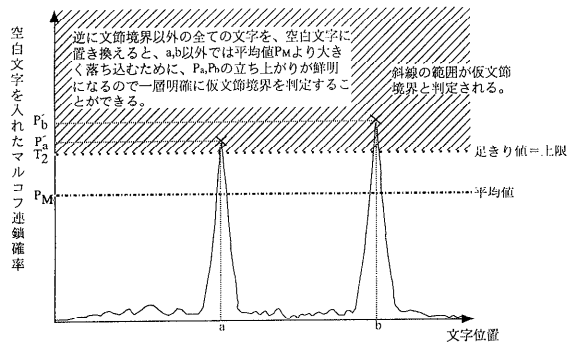
適合率と再現率の関係はトレードオフの関係にあるので, ここではその積を用いて各モデルの能力を評価する. 各モデルにおいて, 推定された文節境界の適合



(1) 仮文節境界候補の推定法（タイプ1の足切り値）
Method of finding the candidates of provisional boundaries of bunsetsu.



(2) 文節境界の認定法
Method to estimate the boundary of bunsetsu using space symbol.



(3) 仮文節境界の再評価の方法（タイプ2の足切り値）
Method of revaluating the provisional boundaries of bunsetsu.

図2 学習データを用いたマルコフ連鎖確率による仮文節境界の推定法
Fig. 2 Method of finding the provisional boundaries of bunsetsu using Markov probability obtained with taking account of learning boundary.

率と再現率の積の最大値を求めるには，文字列連鎖確率の足切り値を変化させて得られた適合率と再現率を評価すればよい．このとき，文字列連鎖確率の足切り値の初期値をいかに適切に選ぶかが問題となるが，本論文ではすでに述べた2種類の足切り値を以下の方法で決定する．

第一タイプの足切り値（上限値）は，べた書きかな文（文節境界記号のない文）の文節境界でのマルコフ連鎖確率値が基準となると考えられるので，標本データからそれを評価し，その平均値を初期値として用いることにする．同様に，第二タイプの足切り値（下限値）では，文節の区切り記号を挿入した標本データの文節境界の連鎖確率の平均値を用いることにする．

以下，各評価モデルにおいて，これらの初期値を使用して，最適な足切り値を決める方法を示す．

(1) 1種類の足切り値 T は，平均値をもとに初期値を定め，適合率 P，再現率 R，$P \times R$ の値を最大にするように T の値を決定する．

(2) 2種類の足切り値 T_1 および T_2 は，次の2つのステップにより決定される．最初に，手順1によって決定された T_2 の値を初期値として固定し，P，R，および $P \times R$ の値を最大にするように T_1 の値を決定する．次に，今決定した T_1 に対して，P，R，および $P \times R$ の値を最大にするように T_2 の値を決定する．

(3) 同様に，3種類の足切り値 T_1，T_2 および T_3 は，最初に，手順1で求めた T_3 の値を初期値として固定し，手順2と同じ方法で P，R，および $P \times R$ の値を最大にするように T_1 と T_2 の値を決定する．次に，今決定した T_1 と T_2 の値に対して，P，R，および $P \times R$ の値を最大にするように T_3 の値を決定する．

(4) 同様に，4種類の足切り値 T_1，T_2，T_3 および T_4 は，最初に，手順1で求めた T_4 の値を固定して，手順3と同じ方法で P，R，および $P \times R$ の値を最大にするように T_1，T_2 および T_3 の値を決定する．次に，今決定した3つの T_1，T_2 および T_3 の値に対して，P，R，および $P \times R$ の値を最大にするように T_4 の値を決定する．

2.3.3 かな漢字変換から見て最適な仮文節境界の足切り値の決定法

推定された文節境界の精度を考えると，かな漢字変換では未検出のものが存在することよりも，文節境界として検出されたものの精度が高いことが望ましく，再現率よりも適合率の方が重要な評価尺度であると考えられる．そこで，ここでは仮文節境界は2文節間隔程度で決定できればよいとし，再現率は50％以上の条件下で適合率が最大となるよう，以下の方法で足切り値を設定する．すなわち，2.4.2項で求めた適合率

表2 8つの方法による仮文節境界の評価結果（標本外データ）
Table 2 Evaluation of provisional boundary of bunsetsu determined by eight methods.

推定方法	べた書きかな文				備考
	足切り値(注)	適合率P（%）	再現率R（%）	積P×R（%）	（べた書き漢字かな混じり文）
NL法	$T_1 = 4.0$	33.3	88.1	29.3	
FL法	$T_1 = 7.0$	50.4	85.3	43.0	「ひらがな」から，「漢字」の字種に変化するところを仮文節境界と判断する方法によって求めた結果； 適合率P = 94.1 % 再現率R = 68.9 % 積P × R = 64.8 %
FBL法	$T_1 = 1.5$	61.5	84.1	51.4	
BL法	$T_1 = 7.5$	47.6	77.5	36.9	
BBL法	$T_1 = 2.0$	41.6	87.8	36.5	
FL + FBL法	$T_1 = 7.0$ $T_2 = 1.8$	82.3	78.0	64.2	
FL + FBL + BL法	$T_1 = 6.0$ $T_2 = 2.0$ $T_3 = 5.0$	90.6	77.7	70.4	
FL + FBL + BL + BBL法	$T_1 = 6.0$ $T_2 = 2.0$ $T_3 = 5.0$ $T_4 = 4.0$	94.0	76.8	72.2	

（注）足切り値 T_i はマルコフ連鎖確率 P_i の逆数の対数値（$T_i = -\log_2 P_i$）で表している．

と再現率の積が最大となる場合の足切り値を始点として，再現率が50%以上で適合率が最大となる場合の足切り値を求め，それをかな漢字変換に適用する際の仮文節境界推定法の足切り値とする．

3. 実験結果

3.1 実験条件

(1) 試験文用の入力データ

仮文節境界の評価に用いる試験文は，2章のマルコフ連鎖確率の辞書作成に用いた新聞記事のデータベース以外の標本外データ☆から選んだ．

(a) 日本語文の種類：日経新聞記事（株式欄，広告欄等を除く記事データの部分）

(b) 日本語文の表記：かな文字表記

(c) 試験文用のデータ量：

(i) 文数，文節の数：200 文，1,597 文節

(ii) 総かな文字数：8,005 文字

(2) マルコフ連鎖確率辞書

(a) 日本語文の種類および字種：1 の (a), (b) に同じ．

(b) マルコフ連鎖確率のタイプ：文節境界の学習ありと学習なし，また学習ありの場合はさらに順方向と逆方向のマルコフ連鎖確率

(c) マルコフ連鎖確率に用いられた標本統計データ量：日本語新聞77日分から得られた統計データ

(i) 文および文節総数：28,547 文，283,975 文節

(ii) 総かな文字数：1,409,359 文字

(d) 足切り値 T：各方法に対して，P, R, $P \times R$ の値を最大にするように実験的に決定する．

☆ 仮文節境界の評価に用いる試験文を，2重マルコフ連鎖確率の辞書作成に用いた新聞記事のデータベースから選ぶとき，これを標本内データと呼び，またこれ以外の新聞記事データから選ぶとき，これを標本外データと呼ぶ．

3.2 実験結果

2章で述べた8つの文節境界推定法のそれぞれに対して，仮文節境界の適合率 P と再現率 R の積が最大となるときの T の値と，そのときの P と R の値を**表2**に示す．それぞれの推定法で推定された文節境界の例を**図3**に示す．さらに，これらの方法のうち，FL + FBL + BL+ BBL 法の適合率 P と再現率 R の関係を**図4**に，積が最大のときの仮文節長の分布を**図5**に示す．これらの図表から以下のことが分かる．

[1] 文節境界の学習効果

文節境界を学習しない標本データを用いる方法（NL法）と，文節境界を学習した標本データを用いる方法（FL法，FBL法）を比較すると，再現率はほぼ同程度であるが，後者の方が前者よりも適合率が約20–30%高い．このことから文節境界の学習の効果が大きいことが分かる．

[2] 順方向と逆方向の2重マルコフ連鎖確率を併用し，2種類以上の足切り値を用いる効果

まず第1に，単一の足切り値の場合，連鎖確率の適用方向の違いによる文節境界推定精度の違いを比較すると，順方向に適用する方法（FL法，FBL法）が，逆方向に適用する方法（BL法，BBL法）よりも優れ

正解文	せいおうの/おんきょうききしじょうでは/にほんぜいが/はげしく/きょうそうしている
NL法	せいおうの/おんきょう|きき|ししじょうでは/に|ほん|ぜいが/はげしく/きょうそうしている
FL法	せいおうの/おんきょうききしじょうでは/に|ほんぜいが/はげしく/きょうそうしている
FBL法	せいおうの/おんきょ|うき|きしじょうでは/に|ほんぜいが/はげしくきょうそうしている
BL法	せい|おうの/おん|きょう|きき|し|じょうでは/にほんぜい|が/はげしく/きょう|そう|している
BBL法	せいおうの/おんきょうききしじょうでは/に|ほんぜいが/はげしく/きょうそうしている
FL+FBL法	せいおうの/おんきょうききしじょうでは/にほんぜいが/はげしく/きょうそうしている
FL+FBL+BL法	せいおうの/おんきょうききしじょうでは/にほんぜいが/はげしく/きょうそうしている
FL+FBL+BL+BBL法	せいおうの/おんきょうききしじょうでは/にほんぜいが/はげしく/きょうそうしている

（注）／は正しく設定された文節境界を表し、また｜は誤って設定された文節境界を示す

図3 NL法，FL法，FBL法，BL法，BBL法，FL＋FBL法，FL＋FBL＋BL法，FL＋FBL＋BL＋BBL法を用いて設定されたかな文の仮文節の例（標本外データ）

Fig. 3 Examples of provisional boundaries of "kana bunsetsu" determined by NL-, FL-, FBL-, BL-, BBL-, FL + FBL-, FL + FBL + BL-, FL + FBL + BL + BBL-method.

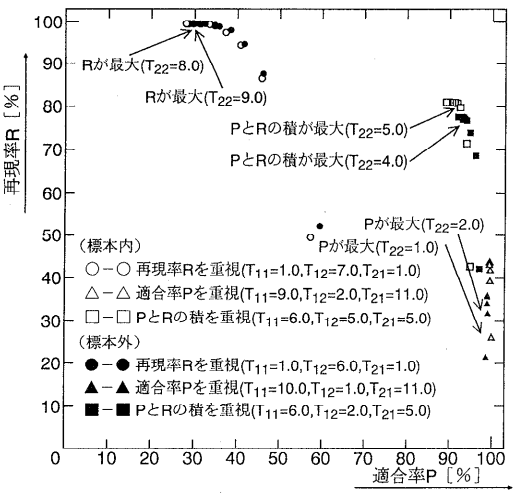


図4 FL＋FBL＋BL＋BBL法を適用したときの実験結果

Fig. 4 Experimental results by FL + FBL + BL + BBL method.

ていることが分かる．また順方向の中では，マルコフ連鎖確率の落込みによる判定（FL法）より立ち上がりによる判定（FBL法）が優れていることが分かる．

第2に，2つの足切り値を用いる場合を見ると，順方向のマルコフ連鎖による落込みと立ち上がりを組み合わせた方法（FL＋FBL法）が，順方向と逆方向の他のいかなる組合せよりも優れていることが分かる．これは，足切り値1つの場合の結果からも予測されるとおりである．

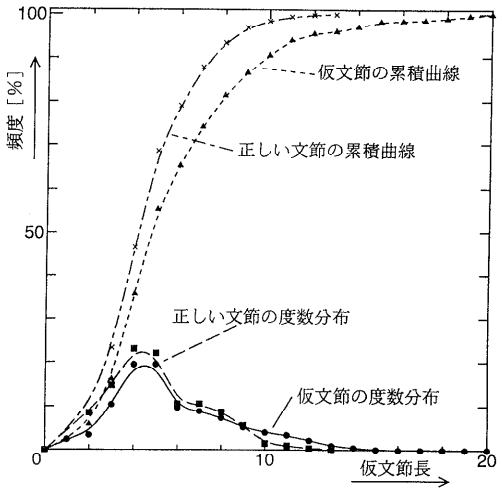


図5 仮文節長の度数分布と累積

Fig. 5 Distribution and accumulative curves of the lengths for provisional boundaries of bunsetsu.

第3に，3つ以上の足切り値の場合では，上述のFL＋FBL法にさらに，逆方向のマルコフ連鎖確率の落込みを組み合わせた方法（FL＋FBL＋BL法）が，優れていること，また4つの足切り値の場合は，これら3つの足切り値以外にさらに逆方向の連鎖確率の立ち上がりを組み合わせた方法（FL＋FBL＋BL＋BBL法）が優れていることが分かる．

[3] 最適な文節境界推定法とその精度

上記の[1]，[2]より，学習ありの場合で，順方向と逆方向のマルコフ連鎖の落込みと立ち上がりをすべて用いて判断する方法（FL＋FBL＋BL＋BBL法）が最も優れており，再現率 P と適合率 R の積が最大となるとき，R は76.8%，P は94.0%が得られることが分かる．

3.3 仮文節境界の適用性

[1] マルコフ連鎖モデルによる仮文節境界の推定精度に対する評価

今回得られたべた書きかな文に対する仮文節境界の推定精度が，かな漢字変換に適用できる精度であるかどうかを判断するため，ここではべた書き漢字かな混じり文の場合の仮文節境界の推定精度，すなわち，《ひらがな》から《漢字》の字種に変化した点を仮文節境界とする場合の推定精度と比較する．そこで，本論文で使用したかな文の標本データの元データである漢字かな混じり文を対象に，字種の変化点に着目して文節境界を決定すると，その適合率と再現率はそれぞれ，約94.1%と再現率は68.9%となる．

この値と比べると，本論文で提案した方法は，同程

度またはそれ以上の精度で文節境界が決定できるから，べた書きかな文の解析に十分適用できると期待される．

[2] 標本内データと標本外データの違いによるマルコフ連鎖確率辞書の学習度

マルコフ連鎖確率を用いる方法の実験においては，標本テキストデータの充足性（連鎖確率辞書の学習量）の問題が重要となる．図4から，標本外データの場合と標本内データの場合の精度を比較すると，両者の差が3%以内になっていることから，かな文の2重マルコフ連鎖確率辞書を学習するには，新聞記事数カ月分程度のテキストデータ量を準備すればほぼ飽和することが分かった．

3.4 仮文節境界の有効性の評価

べた書きかな文のかな漢字変換処理における仮文節境界決定の効果を評価するため，43万語の単語辞書を用いて，仮文節境界の単位に辞書引きする場合と，仮文節境界のない文を対象に辞書引きする場合について実験的に辞書アクセス回数を求めると**図6**の結果を得る．ただし，正解候補をもらさないようにするため，かな漢字変換における単語生成は，最小分割数＋1とし，辞書引きはいずれの場合も総当たり法とした．また，仮文節境界がある場合の辞書アクセス回数には，単語辞書のアクセスに加えて仮文節境界設定に要する辞書アクセス回数も含めた．

この結果から，仮文節境界のある場合は，一文全体の辞書アクセス回数が，仮文節境界のない場合の100分の1（文の長さが20文字の場合）以下となり，仮文節境界推定による辞書引き回数削減の効果の大きいことが分かった．なお，このことは，単に抽出される単語候補の数が大幅に減少するだけでなく，文解釈の候補となる単語の組合せの候補数はそれ以上の割合で減少することを意味するため，かな文の解析での効果が期待できる．

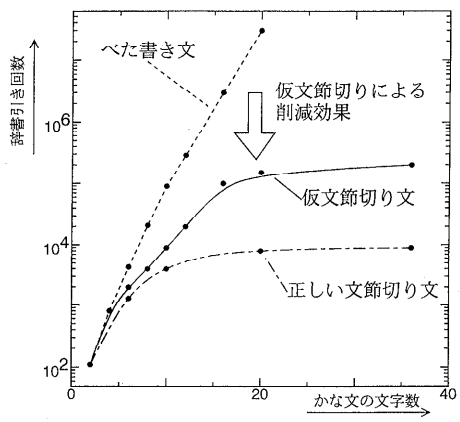


図6 仮文節境界の有無によるかな漢字変換の辞書引き回数の比較
Fig. 6 The comparison of the numbers of looking up the word candidates in a dictionary with or without taking account of povisional boundaries.

4. 結　論

本論文では，べた書きかな文の文節境界を決定する問題に対して，マルコフ連鎖確率モデルを用いた仮文節境界決定の方法を提案した．具体的には，仮文節境界の推定に際して，文節境界を学習した連鎖確率データを使用するか否かの違い，推定された位置が文節境界となりうるかどうかの検証の有無，マルコフ連鎖確率の適用方向の違いの3つの観点から，8種類のマルコフ連鎖モデルを設定し，その精度を70日分の新聞記事データを用いて実験的に比較評価した．

その結果，学習データを用いる方法は，学習なしのデータを使用する場合より，大幅に優れていること，文字連鎖確率の落込みによって抽出された仮文節境界を学習データを用いて再評価すれば，さらに推定精度が向上すること，また，マルコフ連鎖確率は，順方向と逆方向を組み合わせて使用する方が良いことなどが確認された．これらの結果，マルコフ連鎖モデルによってべた書きかな文の文節境界は，適合率94.0%，再現率76.8%の精度で決定できることが分かった．この精度は，従来の漢字かな混じり文の解析で用いられている仮文節境界の決定法（字種の変化点に着目した方法で適合率94.1%，再現率68.9%）の精度より高いことから，べた書きかな文の解析に有効と判断される．

今後は，この方法で決定された仮文節境界を手がかりに，漢字かな混じり文解析の場合と同様，辞書引きによって文節境界の推定誤りを補正しつつ，かな漢字変換を行う方法を検討していく予定である．

参考文献

1) 牧野　寛，木沢　誠：べた書き文の分かち書きとかな漢字変換 — 二文節最長一致法による分かち書き，情報処理学会論文誌，Vol.20, No.4, pp.337–245 (1979).
2) 吉村賢治，日高　達，吉田　将：文節数最小法を用いたべた書き日本語文の形態素解析，情報処理学会論文誌，Vol.24, No.1, pp.40–46 (1983).
3) 池原　悟，白井　諭：単語解析プログラムによる日本文誤字の自動検出と二次マルコフモデルによる訂正候補の抽出，情報処理学会論文誌，Vol.25, No.2, pp.298–305 (1984).
4) 宮崎正弘：係り受け解析を用いた複合語の自動分割法，情報処理学会論文誌，Vol.25, No.6, pp.970–979 (1984).
5) 大島義光，阿部正博，湯浦克彦，武市宣之：格

文法による仮名漢字変換の多義解消，情報処理学会論文誌，Vol.27, No.7, pp.679–687 (1986).
6) 本間 茂，山階正樹，小橋史彦：連語解析を用いたべた書きかな漢字変換，情報処理学会論文誌，Vol.27, No.11, pp.1062–1067 (1986).
7) 宮崎正弘，大山芳史：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, pp.1053–1061 (1986).
8) 武出浩一，藤崎哲之助：統計的手法による漢字複合語の自動分割，情報処理学会論文誌，Vol.28, No.9, pp.952–961 (1987).
9) 荒木哲郎，村上仁一，池原 悟：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.467–477 (1989).
10) 村上仁一，荒木哲郎，池原 悟：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, pp.11–20 (1992).
11) 池原 悟，小原 永，高木伸一郎：文書校生校正支援システムにおける自然言語処理，情報処理，Vol.34, No.10, pp.1249–1258 (1993).
12) 酒井貴子，下村秀樹，並木美太郎，中川正樹，高橋延匡：仮名漢字変換の変換手法と学習に関する一評価，情報処理学会論文誌，Vol.34, No.12, pp.2489–2497 (1993).

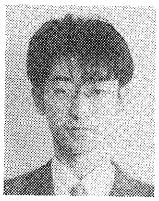
(平成8年12月2日受付)
(平成9年 4 月3日採録)

荒木 哲郎（正会員）

昭和23年生．昭和46年福井大学工学部電気工学科卒業．昭和51年東北大学大学院博士課程修了．同年電信電話公社入社．以来，横須賀電気通信研究所において通信プロトコルの試験，自然言語処理の研究に従事．平成2年より福井大学工学部電子工学科助教授．工学博士．電子情報通信学会会員．

池原 悟（正会員）

昭和19年生．昭和42年大阪大学基礎工学部電気工学科卒業．昭和44年同大学大学院修士課程修了．同年，電信電話公社に入社．数式処理，トラヒック理論，自然言語処理の研究に従事．昭和41年より，スタンフォード大学客員教授．現在，鳥取大学工学部教授．工学博士．昭和57年情報処理学会論文賞，平成5年同研究賞．平成7年日本科学技術センタ賞（学術賞），同年人工知能学会論文賞会員．電子情報通信学会，人工知能学会，言語学会各会員．



土橋 潤也

昭和46年生．平成4年福井大学工学部電子工学科卒業．平成6年同大学大学院修士課程修了．同年，日本電信電話株式会社に入社．在学中，マルコフモデルを用いた日本語処理の研究に従事．現在，NTT 法人営業本部企画部．

笹島 伸一

昭和45年生．平成5年福井大学工学部電子工学科卒業．平成8年同大学大学院修士課程修了．同年，株式会社 SRA に入社．在学中，マルコフ連鎖モデルを用いた仮文節境界の推定に関する研究に従事．電子情報通信学会会員．