自 然 言 語 処 理 100−16
（1994. 3. 18）

置換誤りを含んだべた書きかな文の仮文節境界の推定方法

荒木 哲郎+ 池原 悟++ 土橋 潤也+ 笹島 伸一+
+福井大学 ++ＮＴＴ情報通信網研究所

べた書きの漢字かな交じり文の文節切りについては，これまでに高精度な技術が確立されている．しかしべた書きかな文の場合には，あらゆる単語候補の組み合わせを考慮して解析しようとすると，一般に探索木が爆発するという問題が生じ，これまでに有効な方法が得られていない． このような問題を解決するために，正しい日本語文に対して，仮文節の考え方を導入し，音節の２重マルコフ連鎖モデルを用いて仮文節境界を推定する方法が提案され，その有効性が示されている．

本論文では，これらの仮文節境界を音声認識装置から出力された文に適用可能とするために，正しいかな文(音節文)に対する従来の2重マルコフモデルによる仮文節推定法ならびにこれらの方法を拡張して新しく提案する推定法を，音節候補ラティスからマルコフ連鎖確率を用いて評価されたマルコフ最尤型置換誤りと一定区間内に高々１つの置換誤りがランダムに生じるランダム型置換誤りに適用し，仮文節境界推定法の有効性を定量的に評価する．また新聞記事を用いた実験を行うことにより，次のような知見を得た．

1. マルコフ最尤型置換誤りを含んだ文の場合には，正しい文と同様な方法により，最大で再現率72%，適合率76%の結果が得られた．正解文の場合と比較して，再現率で5%，適合率で6%程度の低下であり，このタイプの置換誤りは正解文にかなり近いことが分かった．

2. ランダム型置換誤りを含んだ文の場合には，順方向と逆方向の２重マルコフモデルを組み合わせて誤りの位置を考慮した新しい手法により，最大で再現率77%，適合率69%の結果が得られ，本手法が有効であることが分かった．

A method to Decide Provisional Boundaries of "Bunsetu" in Non-segmented Kana Sentences included Characters Substituted Wrongly

Tetsuo ARAKI+ Satoru IKEHARA++ Junya TUTIHASHI+ Shinichi SASAJIMA+
+Fukui University ++NTT Network Information Systems Laboratories

As Japanese sentenses are usually written using thousand kinds of characters, especially "kanji" characters, it is not easy to input them into computer files. There are many reseaches on the method which translates the non-segmented "Kana" sentences into the "kanji-kana" sentences. 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 te increasing of the length of the sentence. The memory explosion can be preventred if a sentences is separeted into "bunsestu". Up to now, in order to this problem an useful mehod for finding the provisional boundaries of "bunsestu" for correct Kana sentences using 2nd-order Markov model has been proposed.

This paper proposes a method of finding the provisional boundaries of "bunsetsu" by 2nd-order Markov model, for two types of Kana sentences including characters substituted wrongly.

I.はじめに

べた書きの日本語文を解析する為には，最初に単語や文節単位に分かち書きする処理が必要であり，これまでに漢字かな交じり文に対しては，高精度に分割する技術が確立されている[1][2].

しかしべた書きかな文の場合に，総当たり法でかな漢字変換して得られるあらゆる単語候補列の組み合わせを考慮して解析を試みると，一般に探索木の爆発が起こる.

このような単語候補の探索木の爆発を防ぐ為に，かな仮名漢字変換を含めた形態素解析の対象範囲を文節境界候補として仮に設定し（仮文節境界と呼ぶ），その範囲内でまづ単語候補列の生成・解析を行い，その後で辞書引きや品詞接続テーブルによる接続検定等を行うことによって仮文節境界を補正し，正しい文節境界を決定する方法が考えられる.

これまでにべた書きかな文に対して，２重マルコフ連鎖モデルを用いてこのような仮文節境界を見つける方法が提案されており，その有効性が示されている[5]-[8]. このようなべた書きかな文は，ワープロ入力において得られるが，かな漢字変換などでは通常正しいかな文が対象になる．一方，日本語文音声入力においては，音声認識装置から出力される音節候補文には，通常誤りが含まれることが多い．このような誤りを含んだ音節文から漢字かな交じり文を生成するためには，仮文節境界を設定する技術が必要となる.

本論文では，正しいかな文(音節文)に対して，従来の2重マルコフモデルによる仮文節推定法を，置換誤りタイプの音節ラテイスからマルコフモデルを用いて評価された音節文（マルコフ最尤型置換誤り音節文）および一定区間内に高々１つのランダムな置換誤りを含んだ音節文（ランダム型置換誤り音節文）に適用し，仮文節境界推定法の有効性を評価する． また新聞記事データを用いて，正しい文および置換誤り文における仮文節境界推定の実験を行なう.

II. 諸定義とマルコフ連鎖モデルによる仮文節境界の設定方法

日本語の音節文（かな文）を，$\alpha = s_1 s_2 \cdots s_n$（但し，$s_i$は音節文字）で表す．また日本語の文は，自立語と付属語からなる基本的な構成単位（文節と呼ばれる）に分割される．ここでbは空白文字を表し，文または文節の語頭及び語尾に付加されるものであるとする.

置換誤りを含んだ音節文として以下の２つのタイプを定義する.

音声認識における音響処理の結果には，一般に音節認識の曖昧さと，音節区間の切り出しの曖昧さ（セグメンテーション誤り）が存在するが，ここではセグメンテーション誤りはないものとして，音節認識の曖昧さを取り上げ，これを以下のようにモデル化する．すなわち，マルコフ情報源からの一つの出力音節に同期して，音声認識装置が正解もしくはそれと類似した音節認識候補を複数個出力するもの（その中に一部の音節を除いて必ず正解候補を含む）とする．音節候補の曖昧さを表現するものを音節ラティスと呼び，次のようなラベル付きの有向グラフ$G=(V,E,\theta)$で定義する．ここで，Vは頂点集合で，Eは有向辺の集合，θはラベル付け関数である.

1. 先頭の音節候補位置をt=1として，順に離散時刻で各音節候補の出現位置を表すこととする．時刻tの音節候補の内で，上位から数えて第u番目の確からしさの値（音響処理の結果として出力される）を持つ音節候補を頂点(t,u)で表す．また便宜上，これを単にv=(t,u)と表現し，頂点の有限集合を，

$V=\{ v \mid v=(t,u), 0 \leqq t \leqq L+1, 1 \leqq u \leqq K(t)\}$ で表す．K(t)は時刻tにおける音節の候補数を表わし，Lは各音節候補を組み合わせて得られる文節単位の音節列の長さを表す．特に時刻tの頂点の集合$V_t=\{ v \mid v=(t,u), 1 \leqq u \leqq K(t)\}$と表すと，音節ラティスGの全ての頂点集合は，各時刻毎の頂点集合V_t(t=0,1,L+1）の和集合として，次のように表される．

$$V=\bigcup_{i=0}^{L+1} V_i$$

2. 各セグメント相互間の接続関係を表す為に，時刻t及びt'（但し，t'> t）の頂点及びv'（即ちv∈Vt及びv'∈Vt'）が接続されるとき，vからv'の向きに辺（有向辺と呼ぶ）を設定し順序対(v,v')で表す．またGの有向辺の集合をEで表す．
3. 音節ラベルの集合をSとする．このときラベル付け写像θ:V→Sは，V_tの各頂点に，Sの一つの音節を1対1に割り当てる写像である．但し両境界位置を表す時刻t=0及びt=L+1の頂点集合V_0とV_{L+1}には，文節境界を示す空白記号bが一意に正しく割り当てられる．

【定義1】　セグメンテーション誤りの無い音節ラティスは，次の条件を満たすラベル付き有向グラフG＝(V，E，θ)で表される．任意な時刻t,t+1及びt+k（但し，1≦t≦L, 1≦k≦L-t+1）における頂点集合をそれぞれV_t, V_{t+1}及びV_{t+k}とする時，

1. 任意な頂点v∈V_tとv'∈V_{t+1}に対して，
   (v,v')∈E
2. 任意な頂点v∈V_tとv''∈V_{t+k}に対して，
   (v,v'')∈E
3. 任意な頂点v∈V_tに対して，θ(v)＝1　■

【定義2】　音節ラテイスおいて定義される全ての音節列を，2重マルコフ連鎖確率値の大きい順に並べて得られる最上位の音節列を，マルコフ最尤型置換誤り文と呼ぶ．■

【定義3】　一定区間内（誤り間は最低4文字以上の距離）に高々1つのランダムな置換誤りを含んだ音節文を，ランダム型置換誤り文と呼ぶ．■

本推定法は，マルコフ連鎖確率が文字間の結合力を表すことに着目し，マルコフ連鎖確率が小さいほど，文字間の結合力が弱いと言う性質を用いて推定する方法である．マルコフ連鎖モデルのタイプにより，次のような各種推定法を定義し，仮文節境界推定法の有効性を定量的に評価する．

【定義4】　文節境界の学習有りおよび無しの2重マルコフ連鎖確率の集合を，それぞれLSMP, NLSMPと表す．これら2つのモデルに対して，更に順方向の2重マルコフ連鎖確率の集合をそれぞれF-LSMP，F-NLSMP，また逆方向の2重マルコフ連鎖確率の集合をそれぞれB-LSMP，B-NLSMPと表す．■

これらのマルコフ連鎖確率の集合に従って，仮文節境界を設定する5つの方法を次のように定義する．すなわち，

1. 空白文字無しのLSMPおよび空白文字ありのLSMPを各々単独に用いる方法を，それぞれLM-法，及びBLM-法と呼ぶ．
2. 空白文字無しのLSMPと空白文字有りのLSMPの組み合わせによる方法を2BLM-法と呼ぶ．
3. 空白文字無しのLSMPと空白文字有りのF-LSMPおよび空白文字有りのB-LSMPの組み合わせによる方法を3BLM法と呼ぶ．
4. 空白文字無しのF-NLSMP,F-LSMP,B-NLSMP,B-LSMPの組み合わせにより，全セルに対する誤り位置を決定し，次に，2重

マルコフ列に対する誤りの位置によって，順方向および逆方向の2BLM法を使い分ける方法をC-2BLM法と呼ぶ．

次にこれらの方法によって仮文節境界の位置jが，如何に設定されるかについて定義する．

【定義４】　定数Tに対して，次の条件を満たす位置jが仮文節境界として判断される．

(1) 空白文字無しのLSMP(LM-法)の場合：

$P(s_j|s_{j-2}s_{j-1}) < T$，但し，P,T∈LSMP

(2) 空白文字有りのLSMPと１つの定数T(BLM-法)の場合：

$P(b_j|s_{j-2}s_{j-1}) > T$，但し P,T∈LSMP

(3) 空白文字有りのLSMPと２つの定数T_1およびT_2(2BLM-法)の場合：

$P_1(s_j|s_{j-2}s_{j-1}) < T_1$ & $P_2(b_j|s_{j-2}s_{j-1}) > T_2$，

但し P_1,P_2,T_1,T_2∈LSMP

(4) 空白文字有りのF-LSMP，B-LSMPと３つの定数T_1，T_2およびT_3(3BLM-法)の場合：

$P_1(s_j|s_{j-2}s_{j-1}) < T_1$ &

$(P_2(b|s_{j-2}s_{j-1}) > T_2$ or $P_3(b|s_{j+2}s_{j+1}) > T_3)$

但し P_1,P_2,T_1,T_2∈F-LSMP,P_3,T_3∈B-LSMP

(5) 空白文字無しのF-LSMP,F-NLSMP,B-LSMP,B-NLSMPにより，誤り位置を決定し，２重マルコフ列に対する誤りの位置によって，順方向および逆方向の2BLM法を使い分ける(C-2BLM法)の場合：

［ステップ１］誤り位置の決定

$P_1(s_j|s_{j-2}s_{j-1}) < T_1$ & $P_2(s_j|s_{j-2}s_{j-1}) < T_2$ &

$P_3(s_j|s_{j+2}s_{j+1}) < T_3$ & $P_4(s_j|s_{j+2}s_{j+1}) < T_4$

但し，P_1,T_1∈F-NLSMP,P_2,T_2∈F-LSMP,

P_3,T_3∈B-NLSMP,P_4,T_4∈B-LSMP

［ステップ２］仮文節境界の決定

２重マルコフ列に対する誤り位置が

(1)含まれない場合

$P_5(s_j|s_{j-2}s_{j-1}) < T_5$ & $P_6(b|s_{j-2}s_{j-1}) > T_6$

(2)jの位置に存在する場合

$P_5(s_j|s_{j-2}s_{j-1}) < T_5$ & $P_6(b_j|s_{j-2}s_{j-1}) > T_6$

(3)j-1の位置に存在する場合

$P_7(s_{j-1}|s_{j+1}s_j) < T_7$ & $P_8(b|s_{j+1}s_j) > T_8$

(4)j-2の位置に存在する場合

$P_7(s_{j-1}|s_{j+1}s_j) < T_7$ & $P_8(b|s_{j+1}s_j) > T_8$

(5)複数の位置に存在する場合

仮文節の設定は行わない

但し P_5,P_6,T_5,T_6∈F-LSMP，P_7,P_8,T_7,T_8∈B-LSMP

■

ここで，上記LM法，BLM法および2BLM法は従来提案されている推定法であり，特にマルコフ最尤型の置換誤り文に適用する．また，3BLM法およびC-2BLM法は，今回新たに提案する推定法であり，マルコフ連鎖確率値が，一般に文節境界に限らず誤り位置においても小さな値をとる（文字間の結合力が弱くなる）という性質に着目した方法であり，特にランダム型置換誤り文に適用する．誤りを含んでいる場合は，図１に示されるように，正しい文の場合の２重マルコフ列に対する文節境界位置の組み合わせが，３ケースから12ケースに増加することから，仮文節境界を推定する際に，誤り文字の位置を区別することが重要となる．

また，これらの条件は表１にまとめられる．

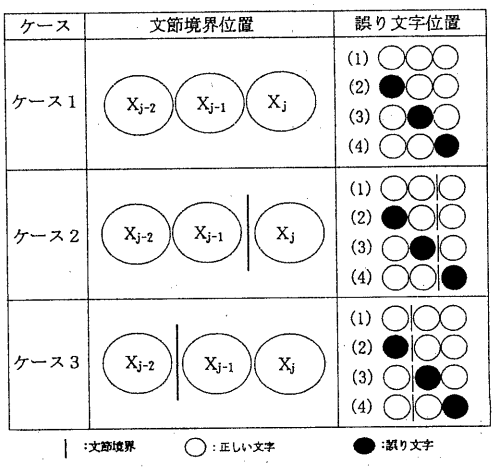

図１　２重マルコフ列に対する文節境界位置

【定義５】仮文節境界の適合率P及び再現率Rを次のように定義する.

$$P \equiv \frac{\text{仮文節境界に含まれる正しい文節境界の数}}{\text{仮文節境界の総数}}$$

$$R \equiv \frac{\text{仮文節境界に含まれる正しい文節境界の数}}{\text{全ての正しい文節境界の数}}$$

■

表１　仮文節境界の設定法

推定方法	条件式
LM法	$P(s_j\|s_{j-2}s_{j-1}) < T$
BLM法	$P(b\|s_{j-2}s_{j-1}) > T$
2BLM法	$P_1(s_j\|s_{j-2}s_{j-1}) < T_1$ & $P_2(b\|s_{j-2}s_{j-1}) > T_2$
3BLM法	$P_1(s_j\|s_{j-2}s_{j-1}) < T_1$ & $(P_2(b\|s_{j-2}s_{j-1}) > T_2$ or $P_3(b\|s_{j+2}s_{j+1}) > T_3)$
C-2BLM法	【ステップ１】誤り位置の決定 $P_1(s_j\|s_{j-2}s_{j-1}) < T_1$ & $P_2(s_j\|s_{j-2}s_{j-1}) < T_2$ & $P_3(s_j\|s_{j+2}s_{j+1}) < T_3$ & $P_4(s_j\|s_{j+2}s_{j+1}) < T_4$ 【ステップ２】仮文節境界の決定 $(P_5(s_j\|s_{j-2}s_{j-1}) < T_5$ & $P_6(b_j\|s_{j-2}s_{j-1}) > T_6)$ or $(P_7(s_{j-1}\|s_{j+1}s_j) < T_7$ & $P_8(b\|s_{j+1}s_j) > T_8)$

III. 置換誤りを含んだ音節文におけるマルコフ連鎖確率値の変化

ここではマルコフ最尤型置換誤り音節文と，ランダム型置換誤り音節文におけるマルコフ連鎖確率値の変化を示す.

[1]各置換誤り音節文の全音節位置におけるマルコフ連鎖確率値の分布

第1位のマルコフ最尤型置換誤り音節文およびランダム型置換誤り音節文の，全音節位置におけるマルコフ連鎖確率値の分布を，図２に示す．同図より，正解文ほどマルコフ連鎖確率値が1（対数値では0）の出現率が高くなっており，これは値が仮文節境界の推定に有効であることを示している．第1位のマルコフ最尤型置換誤りの音節文の分布は，正解文の分布と比較して，平均値が5.85と6.01，またマルコフ連鎖確率値が0となる割合は，17.6%と18.1%と，よく似た傾向を示すことがわかる．また，ランダム型置換誤り音節文におけるマルコフ連鎖確率値の分布は，平均値が8.47，マルコフ連鎖確率値が0となる割合が39.1%と，正解文と比較し，かなり異なる分布をとることがわかる.

[2]各置換誤り音節文の誤り音節位置におけるマルコフ連鎖確率値の分布

図２の内，特に誤り音節位置におけるマルコフ連鎖確率値の分布を，マルコフ最尤型誤りおよびランダム型誤りに関して，それぞれ図３および図４に示す．同図より，第1位のマルコフ最尤型の置換誤りの場合は，平均値が7.11，マルコフ連鎖確率値が0となる割合が22.5%と，正解文と比較し，同様な傾向を示すことが分かる．また，ランダム型置換誤りの場合は，平均値が13.14，マルコフ連鎖確率値が0となる割合が73.9%と，かなり異なる分布となることが分かる.

IV.実験条件

1. 入力データ
   ①文の種類：新聞記事
   ②字種　　：べた書き音節文
   ③総文章数：200文（1458文節）
   ④総文字数：7272文字
   ⑤誤り文字数：マルコフ最尤型1位＝874文字
   　　マルコフ最尤型10位＝975文字
   　　ランダム型＝921文字
2. 使用辞書
   ①仮文節境界設定用：音節２重マルコフ連鎖

確率（新聞記事77日分の統計データにより作成）

②仮文節境界の補正用：単語辞書（15万語），品詞接続テーブル，漢字かなの順方向及び逆方向の２重マルコフ連鎖確率（新聞記事77日分の統計データにより作成）

V. 仮文節境界推定の実験結果

マルコフ最尤型置換誤り音節文に対しては，LM法，BLM法および2BLM法，また，ランダム型置換誤り文に対しては，3BLM法およびC-2BLM法を適用して仮文節境界を求める実験を行なった．また，前者については[7]で示された補正方法によって仮文節境界の補正を行った実験結果を示す．

[1]マルコフ最尤型置換誤り音節文に対する文節推定法の効果

(1)各推定法の効果

実験結果を図5と図6に示す．同図より，仮文節推定法の中では，正解文の場合と同様に2BLM法が最も良く，BLM法，LM法の順になっている．また，正解文の場合に比べて適合率で2～6%，再現率で3～5%低下している．

(2)誤り音節文のタイプと各推定法の効果

LM法および2BLM法による各種誤り文の仮文節推定効果を，それぞれ図7および図8に示す．両法ともに，マルコフ最尤型1位誤り文，マルコフ最尤型10位誤り文，ランダム型置換誤り文の順に低下している．マルコフ最尤型置換誤り文は，ランダム型置換誤り文に比べ，正解文に近いものとなっていることが分かる．

(3)仮文節境界の補正効果

マルコフ最尤型置換誤り文において，LM法と2BLM法によって推定された仮文節境界の補正効果を図9と図10に示す．同図より，適合率で10～30%の向上が図れることが分かる．

[2]ランダム型置換誤り音節文における仮文節境界の推定効果

ランダム型置換誤り音節文に対して，2BLM法および新しく提案された3BLM法，C-2BLM法を適用して得られた仮文節境界の推定効果を図11に示す．同図より，誤り位置を検出した後で仮文節境界を推定するC-2BLM法が，3BLM法や従来の2BLM法と比較し，すぐれていることが分かる．誤りを含む場合には，順方向と逆方向のマルコフ連鎖確率値のよい方をOR条件で選ぶ方法が有効であることが分かる．

VI. おわりに

本論文では，正しいかな文（音節文）に対する従来の２重マルコフモデルによる仮文節境界の推定法を拡張し，置換誤りを含んだべた書きかな文に適用し，実験によりその有効性を確認した．その結果，以下のような知見が得られた．

1. マルコフ最尤型置換誤り文の場合には，正しい文と同様な方法により，仮文節境界の推定では，最大で再現率72%，適合率76%，また補正処理を行うことで，最大で再現率67%，適合率86%と，高い結果が得られた．
2. ランダム型置換誤り文の場合には，誤り位置を考慮した手法を用いることにより，最大で再現率77%，適合率69%と比較的高い結果が得られた．

今後の課題としては，仮単語境界を用いた，誤りを含んだ文の仮文節境界の補正方法の検討および，脱落や挿入等のより複雑な誤りを含んだ文に対する仮文節境界の推定法の検討が挙げられる．

参考文献

[1]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,No.6, pp970-979 (1984)

[2] 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,No.11,pp1053-1061 (1986)
[3] 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, No.4, pp467-477 (1989)
[4] 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)
[5] 土橋, 荒木, 池原:"2重マルコフ連鎖確率を用いたべた書き日本語文の文節境界推定", 信学会春期大会, Vol.6, No.D-102, pp104 (1993)
[6] 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NL研究会, Vol.94-8, pp55-61 (1993)
[7] 荒木, 池原, 土橋:"べた書きかな文の仮文節境界の補正方法", 情処第47回後期全国大会, 2L-9, pp2-111 (1993)
[8] 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 (1993)

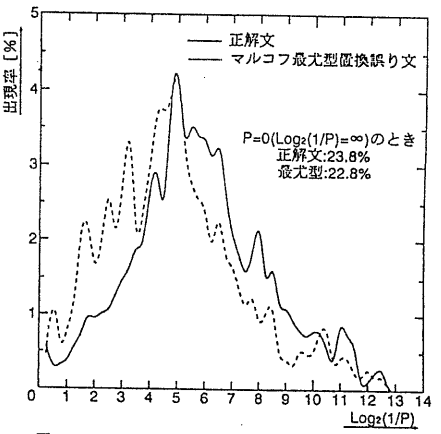

図3 マルコフ最尤型誤り位置におけるマルコフ値の分布

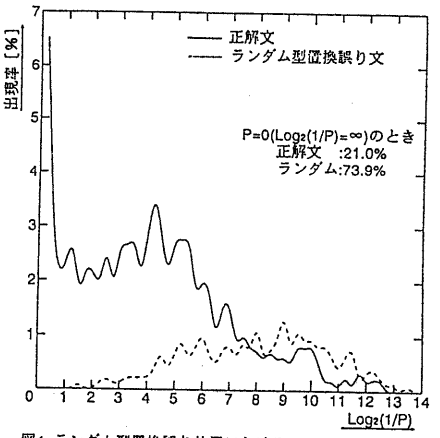

図4 ランダム型置換誤り位置におけるマルコフ値の分布

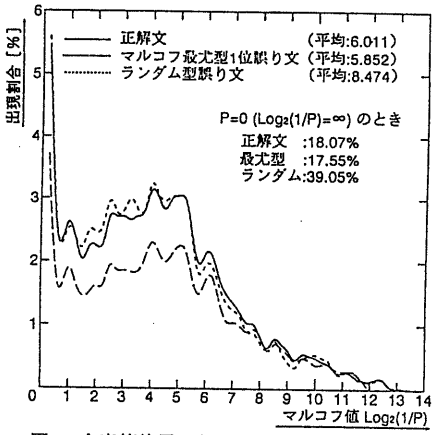

図2 全音節位置におけるマルコフ値の分布

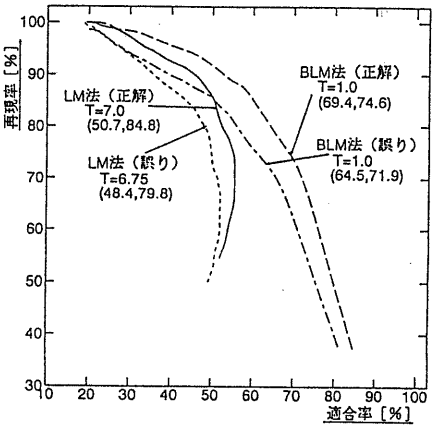

図5 マルコフ最尤型1位誤り文における実験結果

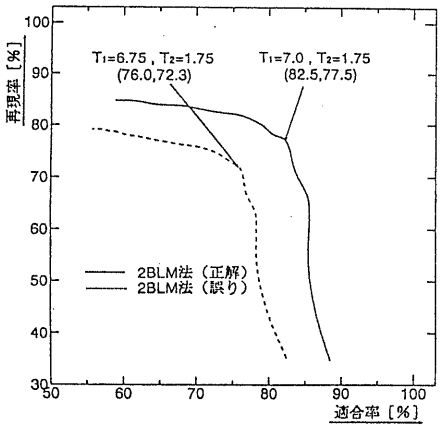

図6　マルコフ最尤型1位誤り文における実験結果

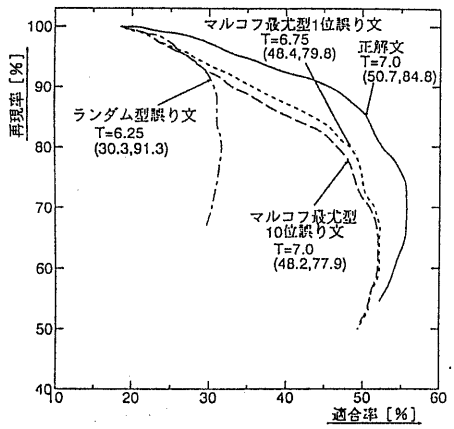

図7　誤り音節文のタイプによる仮文節境界の推定効果
（仮文節はLM法により設定した場合）

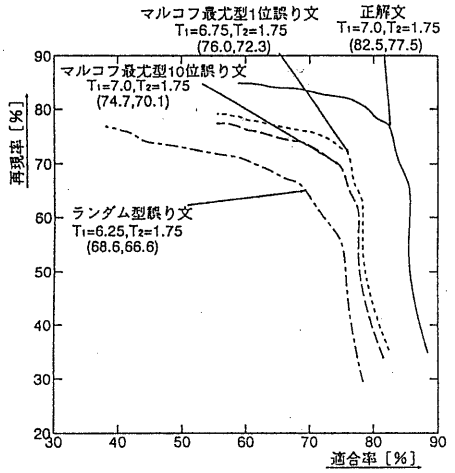

図8　誤り音節文のタイプによる仮文節境界の推定効果
（仮文節は2BLM法により設定した場合）

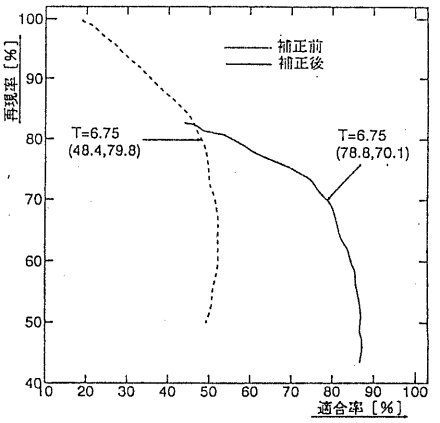

図9 マルコフ最尤型1位誤り文における仮文節境界の補正効果
（仮文節はLM法により設定した場合）

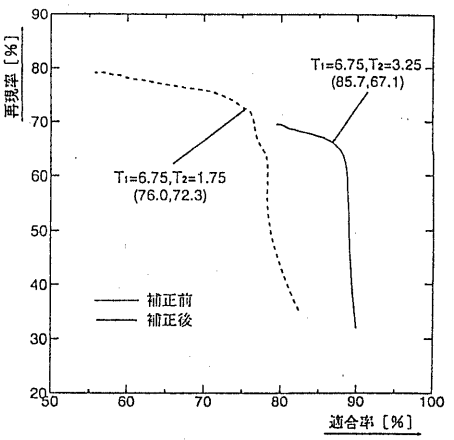

図10　マルコフ最尤型1位誤り文における仮文節境界の補正効果
（仮文節は2BLM法により設定した場合）

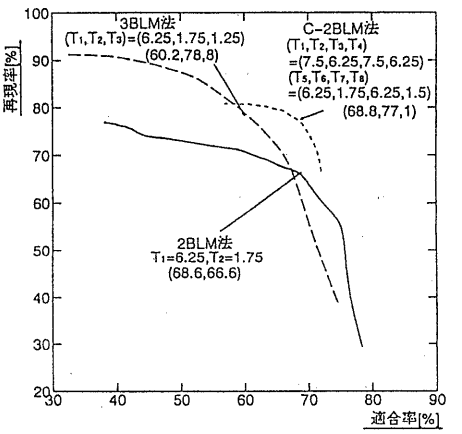

図11　ランダム型置換誤り文における仮文節境界の推定効果