

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

荒木 哲郎+ 池原 悟++ 土橋 潤也+

+ 福井大学工学部

++ NTT情報通信網研究所

本論文では、べた書き日本語文（音節表記、漢字かな表記）を文節単位に分割する問題に対して、m重マルコフ連鎖確率を用いて、文節切りした時の仮文節先頭位置を、推定する為の新しい方法を提案する。またそのアルゴリズムの有効性を評価する為に、べた書きの新聞記事200文を入力文として、2重マルコフ連鎖確率を用いて、仮文節先頭位置を推定する実験を行った。その結果、文節境界の学習を行わない場合には、音節文節では適合率31%、再現率89%、また漢字かな文節では適合率61%、再現率70%に対し、文節境界の学習を行った場合には、音節文節では適合率83%、再現率78%、また漢字かな文節では適合率93%、再現率95%に向上する効果が確認された。

A method for hypothetically deciding of bunsetu boundaries in solid japanese sentences using 2nd-order markov model

Tetuo ARAKI+ Satoru IKEHARA++ Junya TUTIHASI+

+ Faculty of Engineering, Fukui University

++ NTT Network Information Systems Laboratories

In this paper we propose a new method which hypothetically decides bunsetu boundaries in solid japanese sentences using 2nd-order markov model. This method is based on the property that values of 2nd-order markov probability for characters within any bunsetu are usually great, but the values of markov probability for characters between two consecutive bunsetus are usually low.

The experimental result using 200 sentences of japanese newspaper articles, shows that the method using markov model introduced learning of all correct bunsetu boundaries in all sentences is useful for hypothetically deciding bunsetu boundaries in solid sentences.

1. はじめに

日本語文の解析には形態素解析、構文解析、意味解析などの各種レベルがあるが、べた書きの日本語文（漢字かな交じり文）に対して最初に単語や文節などの単位に分ち書きを行う形態素解析が基本的な処理として重要である。従来、べた書きの漢字変換における単語や文節の分ち書きとしては、最長一致法 [1]、文節数最小法 [2]、連語解析 [3] や単語共起関係 [4] を用いる方法があり、また漢字かな交じり文の単語（主に複合語）分割としては、係り受け解析を用いる方法 [5] [6]、接辞と語基を組み合わせたマルコフモデルを用いる方法 [7] がある。特に [5] では新聞記事の複合語に対象に、99.8%の高精度で分割するなど実用的なレベルに至っており、これを用いた単語分割、文節分割などの日本語解析処理技術が実現されている [6]。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わせを考慮して解析を試みようとする、一般に探索木が爆発するという問題が生じる。

このような問題を解決するために、本論文では、[6] において導入された仮文節の考え方を用いる。すなわち探索木の爆発を防ぎ、一定の処理時間内で日本語文の解析を終える為に、かな漢字変換を含めた形態素解析の対象となる範囲を仮文節として一次的に定め、その範囲内でまず単語候補列の生成・解析処理を行い、次に辞書引きや品詞接続テーブルによる接続検定等によって、仮文節境界を補正して正しい文節境界を設定する方法である。しかし漢字かな交じり文においては、字種（漢字、ひらがな、記号等）が変化する点に着目して仮文節境界を設定できるが、べた書きのかな文においては仮文節境界の手がかりが一般に見当たらず、これを如何に求めるかが重要な問題となってくる。

本論文では、音声入力などの曖昧な日本語文に対しても解析可能な日本語処理技術を確認することを目的に、先の音節認識候補の絞り込み [8] 及び漢字かな候補の絞り込み [9] において有効な役割を果たした2重マルコフ連鎖確率を用いて、音節文並びに漢字かな交じり文の両方に適用可能な仮文節境界の推定法を提案する。ここで提案する文節分割法は、音節の2重マルコフ連鎖確率を用いるもので、これは従来のようにかな漢字変換処理を必要とせず、比較的簡単な処理で行え、比較的高い正解率が得られること、更に単語辞書引き等を組み合わせると、更に高精度化が図れる可能性があること、さらに同様に2重漢字かなマルコフ連鎖確率を用いて、漢字かな交じり文の文節分割が可能であり、かなり高い文節分割正解率が得られるという特徴がある。この方法は、マルコフ連鎖確率情報が、文字間の結合力の強さを表していることに着目し、文中の各文字の2重マルコフ連鎖確率値が設定された足切り値を下回るか否かによってべ

た書き日本語文の仮文節先頭位置を推定する方法である。特に2重マルコフ連鎖モデルを用いてべた書き日本語文（音節文、漢字かな交じり文）の仮文節先頭を判定する方法として、(1) 文節境界の学習有り・無し の2つの2重マルコフ連鎖確率モデル、(2) マルコフ連鎖確率の定義に用いられる文節境界の空白文字を用いて判定する方法、(3) 順方向及び逆方向で学習された2種類のマルコフ連鎖確率値を混合して文節切りするなど各種タイプを設定して実験を行い、それらの優劣、文節境界の学習の効果などを定量的に評価し、その有効性を示す。

なお本分割方法の有効性を定量的に評価する為には、信頼のあるマルコフ連鎖確率データが必要となり、従来これらのデータの入手が難しい状況にあったが、本研究では、[6] の高精度な言語解析技術（単語分割処理、単語読み（音節）の付与処理技術など）を用いて、大量の新聞記事の音節変換、単語、文節切りデータが高精度に得られることにより、大量の信頼できる音節マルコフ連鎖確率、漢字かなマルコフ連鎖確率データが入手可能となることで解決を図った。

2. べた書き日本語文の文節とマルコフ連鎖モデルによる仮文節切りの方法

ここではべた書き日本語文の種類、文節の構造、マルコフ連鎖モデルの種類と仮文節切りの方法及び適合率と再現率について述べる。

【定義1】日本語文 $S = x_1 x_2 \dots x_n$ において、全ての x_i ($1 \leq i \leq n$) が音節文（カタカナ文字に相当）であるとき S を音節文または、全ての x_i が {漢字、カタカナ、ひらがな} のいずれかの文字である時、 S を漢字かな交じり文と呼び、 n を文 S の長さと呼ぶ。また S 中の連結した文字の部分列 $x_1 x_{i+1} \dots x_{i+q}$ が、自立語（詞）= {名詞、動詞、形容詞、形容動詞、副詞、連体詞、接続詞、感動詞、形式名詞} と付属語 = {助詞、助動詞、接辞} から構成される一つの単位を文節 B_k と呼び、 $B_k = \langle x_1 x_{i+1} \dots x_{i+q} \rangle$ と表す。一般には、自立語は必ず含まれるが、付属語は必ずしも含まれるとは限らない。また x_i を文節 B_k の先頭、 x_{i+q} を B_k の最後尾と言う。
(定義終)

【定義2】文 $S = x_1 x_2 \dots x_n$ における2つの文節を、 $B_k = \langle x_1 x_{i+1} \dots x_{i+q} \rangle$ 及び、 $B_l = \langle x_j x_{j+1} \dots x_{j+r} \rangle$ とするとき、文節 B_k の最後尾の位置 $i+q$ と B_l の先頭位置 $j=i+q+1$ の関係にあるとき、文節 B_k と B_l は連続した文節と呼び、 B_k を B_l の直前の文節、また B_l を B_k の直後の文節と呼ぶ。文 S 中の任意な連続した3文字を x_{j-2} 、 x_{j-1} 、 x_j と表すとき、

(1) 3文字 x_{j-2} 、 x_{j-1} 、 x_j が文のいかなる位置にあっても、 $p_j = p(x_{j-1} | x_{j-2} x_{j-1})$ と表される時、 p_j を文節境

界学習なしの2重マルコフ連鎖確率モデルといい、そのような確率値の集合を P_NLM と表す。

(2) x_i の2重マルコフ連鎖確率 p_i が、3文字 x_{i-2} , x_{i-1} , x_i が全て一つの文節 B_i に属するとき、 $p_i = p(x_i | x_{i-2}x_{i-1})$ 、また x_{i+2} と x_{i+1} が文節 B_i に属し、 x_i が直後の文節 B_{i+1} に属するときは、 x_i を空白 b に置き換え $p_i = p(b | x_{i-2}x_{i-1})$ と表わされるとき、 p_i を文節境界学習有りの2重マルコフ連鎖確率モデルといい、そのような確率値の集合を P_LM と表す。但し、文 $S = x_1x_2 \dots x_n$ の任意な位置において2重マルコフ連鎖確率が定義可能とするために、文または文節の語頭・語尾の境界は空白文字 b として扱っており、文節境界学習有りのモデルとは、文節境界に空白文字を挿入して学習されたモデルである。(定義終)

【定義3】文 $S = x_1x_2 \dots x_n$ の任意な文字 x_i における2重マルコフ連鎖確率 P_i を、(1) $P_i = p(x_i | x_{i-2}x_{i-1})$ 、(2) $P_i = p(x_i | x_{i-2}x_{i+1})$ によって定められるとき、(1)または(2)をそれぞれ順方向または逆方向の2重マルコフ連鎖確率モデルと呼び、そのような確率値の集合を P_FM または P_BM と表す。(定義終)

特に断りが無い場合には、2重マルコフ連鎖確率モデルは順方向であるとし、以下順方向であることを省略する。また定義2と定義3の文節境界学習有りとなしの場合の2重マルコフ連鎖確率モデル、及び順方向及び逆方向の2重マルコフ連鎖確率モデルを図1及び図2に示す。ここで文節境界学習有りのマルコフ連鎖確率の定義法は、文 S 内のどこに文節境界があるかを逐次教えられる場合の教師付き学習モデルであることに注意する。

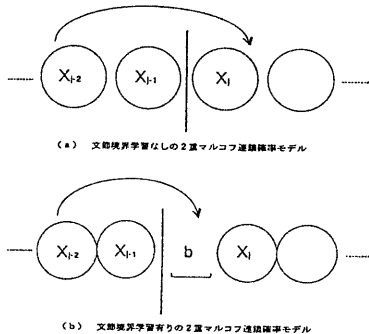


図1 文節境界学習有り、なしの2重マルコフ連鎖確率モデル

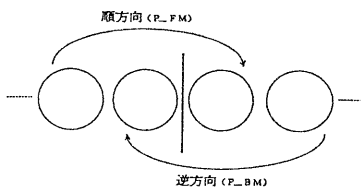


図2 順方向及び逆方向の2重マルコフ連鎖確率モデル

次に文字間の結合力は、一般に文節内では強いが文節間にまたがるときは弱くなる性質があるが、これをマルコフ連鎖確率モデルを用いて調べることににより、仮文節境界〔注〕を判定する方法を以下に述べる。

【定義4】文 $S = x_1x_2 \dots x_i \dots x_n$ において、 x_i がある文節 B_i の先頭であることを判定するのに用いる2重マルコフ連鎖確率モデル p_i の種類、及び各々の2重マルコフ連鎖確率値を比較する際に必要な足切り値の種類を T とする。

(1) 文 S の任意な x_i の2重マルコフ連鎖確率 p_i が、文節境界学習なしの2重マルコフ連鎖確率モデルによって $p_i = p(x_i | x_{i-2}x_{i-1})$ と表され(p_i は P_NLM の要素)、足切り値 T が P_NLM より選択され、文節境界の判定条件 $p_i < T$ が成り立つときに x_i をある仮文節 B_i の先頭であると推定する方法を文節境界学習無しのマルコフ連鎖モデルによる推定方法(略して NLM : Non_Learning Markov法)と呼ぶ。

(2) 同様に p_i が文節境界学習有りの2重マルコフ連鎖確率モデル P_LM の要素であり、足切り値 T が P_LM より選ばれ、文節境界の判定条件 $p_i < T$ となるとときに、 x_i をある仮文節 B_i の先頭であると定める方法を、学習有りのマルコフ連鎖モデルによる推定方法(略して LM 法)と呼ぶ。

(3) p_i が文節境界学習有りの2重マルコフ連鎖確率モデル P_LM の要素であり、足切り値 T が P_LM より選ばれ、文節境界の判定条件 $p_i = p(b | x_{i-2}x_{i-1}) > T$ となるとときに、 x_i をある仮文節 B_i の先頭であると定める方法を、学習有りの文節境界文字付加型マルコフ連鎖モデルによる推定判定方法(略して BLM 法)と呼ぶ。

(4) p_i が文節境界学習有りの2重マルコフ連鎖確率モデル P_LM の要素であり、二つの足切り値 T_1 及び T_2 が P_LM より選ばれ、二つの文節境界の判定条件 $p_i = p(x_i | x_{i-2}x_{i-1}) < T_1$ かつ、 $p_i = p(b | x_{i-2}x_{i-1}) > T_2$ となるとときに、 x_i をある仮文節 B_i の先頭であると定める方法を、学習有り・2足切り値を併用・文節境界文字付加型マルコフ連鎖モデルによる推定判定方法(略して $2BLM$ 法)と呼ぶ。

5) 順方向の LM 法と逆方向の LM 法(即ち p_i 及び足切り値 T が P_BL の要素であり、判定条件が $p_i < T$ が成り立つとき x_i をある仮文節 B_i の先頭であると定める)を組み合わせることで、 x_i をある仮文節 B_i の先頭であると推定する方法を学習有りの組み合わせマルコフ連鎖モデル(略して CLM 法)と呼ぶ。(定義終)

(脚注) 仮文節とは、〔6〕で示されているように、単語候補数の探索木の爆発を防ぐ為に、かな漢字変換を含めた形態素解析の対象となる範囲を文節境界候補として仮に設定し(仮文節と呼ぶ)、その範囲内でまず単語候補列の生成・解析処理を行い、その後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である。

ここで定義4のNLM法とLM法は、文節境界の学習の有無による文節境界推定力の差異を表し、またBLM法及び2BLM法は、文節境界を学習した情報を有効的に用いることにより、またCLM法は順方向及び逆方向両者のマルコフ連鎖確率モデルの特徴を生かすことにより、更に仮文節先頭位置の推定力の向上を図ることを狙った方法である。

これらをまとめたものを、表1に示す。

また、定義4の各種仮文節境界判定法を評価するために必要な適合率と再現率の定義を述べる。

【定義5】2重マルコフ連鎖確率モデルによる仮文節先頭の判定方法による適合率と再現率を次のように定める。

$$\text{適合率} = \frac{\text{設定された中で正しい文節境界の件数}}{\text{設定された文節境界の総件数}}$$

$$\text{再現率} = \frac{\text{設定された中で正しい文節境界の件数}}{\text{文Sの元々の正しい文節境界の件数}}$$

(定義終)

表1 文節境界の推定判別法のまとめ

推定方法	判定条件
NLM法	$P_1 = p(x_i x_{i-2} x_{i-1}) < T$
LM法	$P_1 = p(x_i x_{i-2} x_{i-1}) < T$
BLM法	$P_1 = p(b x_{i-2} x_{i-1}) > T$
2BLM法	$P_1 = p(x_i x_{i-2} x_{i-1}) < T_1 \text{ \& } P_1 = p(b x_{i-2} x_{i-1}) > T_2$
CLM法	$P_1 = p(x_i x_{i-2} x_{i-1}) < T_1 \text{ \& } P_{1-1} = p(x_{i-1} x_{i-1} x_i) < T_2$

3. 2重マルコフ連鎖確率モデルによる仮文節切りの実験

3.1 実験の条件

(1) 仮文節境界推定方法の種類

① NLM法 (文節境界学習無しマルコフ連鎖モデルによる推定方法)

② LM法 (文節境界学習有リマルコフ連鎖モデルによる推定方法)

③ BLM法 (学習有リ文節境界文字付加型マルコフ連鎖モデルによる推定方法)

④ 2BLM法 (学習有リ・2足きり値を併用・文節境界文字付加型マルコフ連鎖モデルによる推定方法)

⑤ CLM法 (順方向・逆方向のマルコフ連鎖モデルの組み合わせによる推定方法)

(2) 入力データ

① 文の種類: 新聞記事

② 字種: べた書き音節文及び漢字かな交じり日本文

③ 総文章数: 200文 (1458文節)

④ 総文字数: 音節文 = 7272文字
漢字文 = 5614文字

(3) マルコフ連鎖確率辞書

① 種類: 音節および漢字2重マルコフ連鎖確率

② 学習の方向: 順方向および逆方向

③ 学習の有無: 文節境界学習有りと無し

3.2 実験結果

(1) 仮文節切りにおける文節境界の学習効果

NLM法とLM法を、音節文並びに漢字かな交じり文に適用したときの適合率、再現率の実験結果を、それぞれ図3及び図4に示す。同図より、文節境界の学習無しモデルでは、音節文の場合に適合率が最大31%のとき再現率は約89%、また漢字かな交じり文の場合に適合率が最大約61%のとき再現率は約70%である。一方文節境界学習有リモデルでは、音節文の場合に適合率が最大約51%のとき再現率は約85%、また漢字かな交じり文の場合に適合率が最大約51%のとき再現率は約95%となり、文節境界の学習によって音節文では適合率で約20%向上し、また漢字かな交じり文では適合率で約10%減少するが再現率で約25%向上することがわかった。また実際にNLM法とLM法を音節文及び漢字かな交じり文に適用した例をそれぞれ図5、図6に示す。

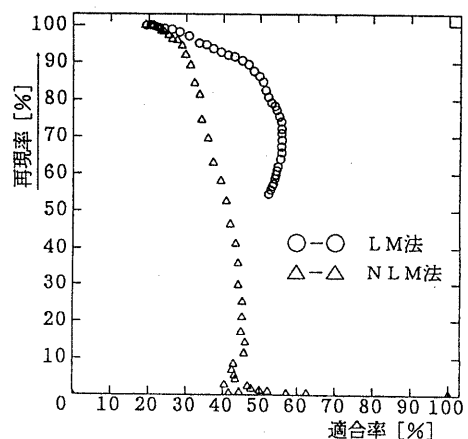


図3 NLM法及びLM法を適用したときの分割結果
べた書き音節文の場合

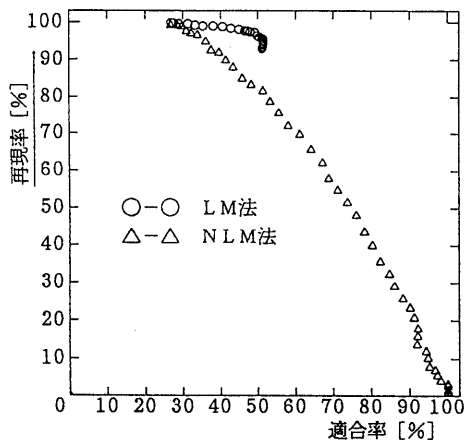


図4 NLM法及びLM法を適用したときの分割結果
べた書き漢字かな文の場合

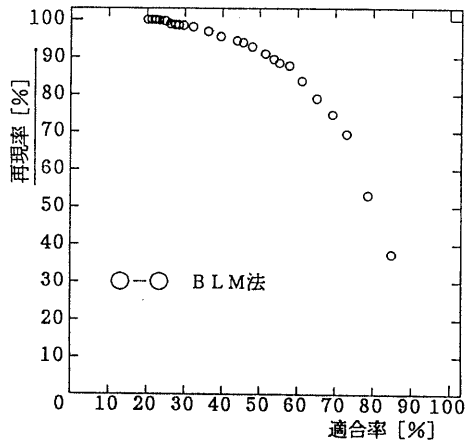


図7 BLM法を適用したときの分割結果
べた書き音節文の場合

図5 NLM法、LM法、2BLM法を音節文に適用した例

正解文	ワガクニ/ケイザイノ/コクサイカニ/トモナイ/ニホンニ/ シンシュツスル/ガイコクギンコウワ/キュウゾウシテイル
NLM法	ワガクニ/ケイザイノ/コクサイカニ/トモナイニ/ホンニ/ シンシュツスル/ガイコクギンコウワ/キュウゾウシテイル
LM法	ワガクニ/ケイザイノ/コクサイカニ/トモナイ/ホンニ/ シンシュツスル/ガイコクギンコウワ/キュウゾウシテイル
2BLM法	ワガクニ/ケイザイノ/コクサイカニ/トモナイ/ホンニ/ シンシュツスル/ガイコクギンコウワ/キュウゾウシテイル

図6 NLM法、LM法、2BLM法を漢字かな文に適用した例

正解文	わが国/経済の/国際化に/伴い/日本に/進出する/ 外国銀行は/急増している
NLM法	わが国/経済の/国際化に/伴い/日本に/進出する/ 外国銀行は/急増している
LM法	わが国/経済の/国際化に/伴い/日本に/進出する/ 外国銀行は/急増している
2BLM法	わが国/経済の/国際化に/伴い/日本に/進出する/ 外国銀行は/急増している

(2) 文節境界文字付加型の2重マルコフ連鎖確率値による仮文節切りの効果

BLM方法を音節文並びに漢字かな交じり文に適用したときの実験結果を、それぞれ図7及び図8に示す。同図より、図3及び図4の実験結果と比較して、音節文の場合には、更に適合率で約21%、また漢字文の場合には、更に適合率で約32%向上することが確認され、学習に用いた文節境界文字を仮文節切りの推定に陽に用いる効果が確認された。

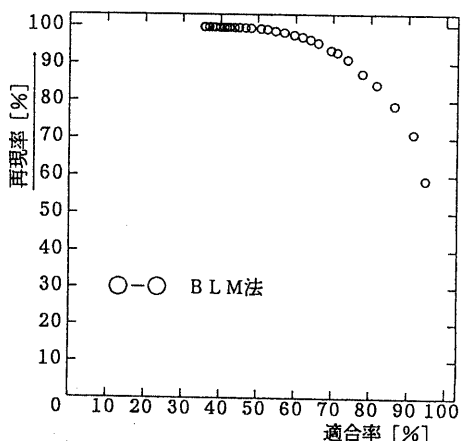


図8 BLM法を適用したときの分割結果
べた書き漢字かな文の場合

(3) 2足きり値を併用・文節境界文字付加型の2重マルコフ連鎖確率値による仮文節切りの効果

2BLM方法を音節文並びに漢字かな交じり文に適用したときの実験結果を、それぞれ図9及び図10に示す。但し、2つの足きり値の内 $P_i = p(x_i | x_{1:i-1}) < T_1$ において、再現率・適合率・再現率と適合率の積がそれぞれ最大のときで、 T_1 を固定し実験を行っている。同図より、図3及び図4の実験結果と比較して、音節文の場合には、更に適合率で約31%、また漢字文の場合には、更に適合率で約42%向上することが確認され、学習に用いた文節境界文字を文節切りの推定に陽に用いる効果が確認された。また実際に2BLM法を音節文及び漢字かな交じり文に適用した例をそれぞれ図5、図6に示す。

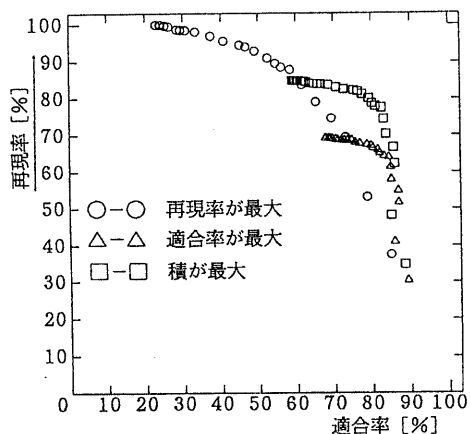


図9 2 BLM法を適用したときの分割結果
べた書き音節文の場合

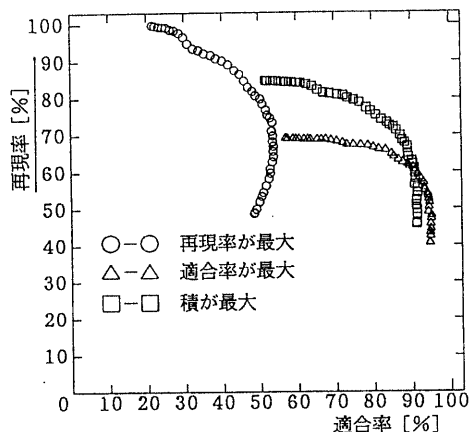


図11 CLM法を適用したときの分割結果
べた書き音節文の場合

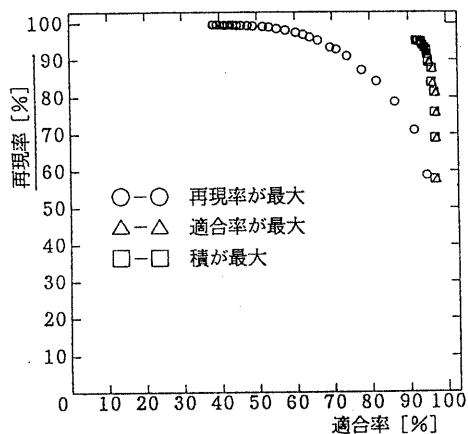


図10 2 BLM法を適用したときの分割結果
べた書き漢字かな文の場合

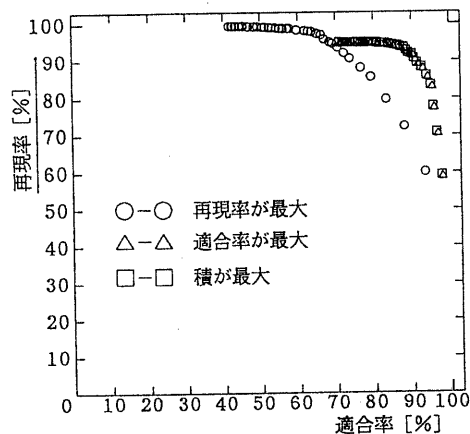


図12 CLM法を用いたときの分割結果
べた書き漢字かな文の場合

(4) 順方向と逆方向の組み合わせによる効果

順方向と逆方向の個々の2重マルコフ連鎖確率モデルによる違いを比較してみると、文頭から仮文節境界を推定する順方向の方が、より正しい推定を行っていることが定量的に確認される。次にCBM法を音節文並びに漢字かな交じり文に適用したときの実験結果を、それぞれ図11及び図12に示す。但し、2つの足きり値の内、 $P_i = p(x_i | x_{12} x_{i-1}) < T_i$ において、再現率・適合率・再現率と適合率の積がそれぞれ最大のときで、 T_i を固定し実験を行っている。同図より、図3及び図4の実験結果と比較して、音節文の場合、適合率が最大で約35%、漢字文の場合、適合率が最大で約35%向上することが確認され、順方向と逆方向の組み合わせによる効果が確認された。

(5) 推定位置の近傍に正解文節境界が存在する頻度分布

LM法を音節文及び漢字かな交じり文に適用したときの仮文節先頭位置の近傍(設定した仮文節境界を0としてその左隣り(-1,-2,...)または、右隣り(+1,+2,...))に、正解文節境界が存在する頻度分布を、それぞれ図13及び図14に示す。但し、それぞれ最大の適合率を有する設定仮文節境界に対して、調査を行っている。同図より、 ± 1 の範囲において正解文節境界が存在する割合は、音節文の場合で約92%、漢字文の場合で約86%となることが分かり、LM法等によって推定した文節境界位置を、仮文節境界として、更に ± 1 及び ± 2 の範囲に単語辞書引きを行うことにより仮文節境界を補正することが可能になる。またLM法による文節境界位置の推定能力は、

文献 [5] の方法と比べると劣るが、補正効果を加味することで更に向上が期待される。

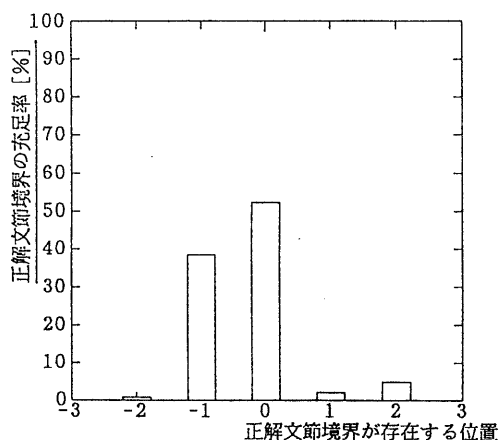


図13 推定位置の近傍に正解文節境界が存在する割合
べた書き音節文200文使用、LM法を適用

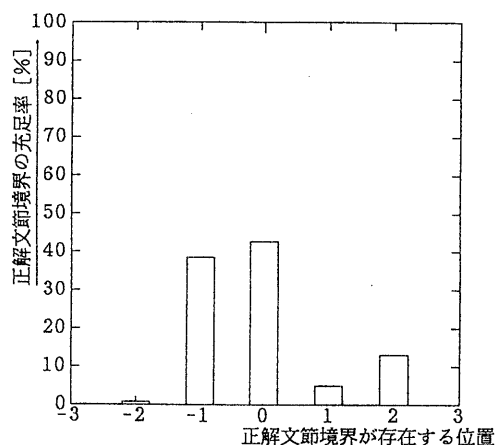


図14 推定位置の近傍に正解文節境界が存在する割合
べた書き漢字文200文使用、LM法を適用

5. むすび

2重マルコフ連鎖確率によるべた書き日本語文の仮文節切り方法をいくつか提案すると共に、文節境界を学習したモデルであるBLM法などの有効性を定量的に評価するために実験を行った。その結果、音節文の場合には最大で適合率83%のとき再現率78%、漢字文の場合には最大で適合率93%のとき再現率95%と高い正解率が得られ、文節境界の学習の有効性が定量的に確認された。

残された問題としては、今回マルコフ連鎖確率モデルによって推定された仮文節候補に、さらに単語辞書引きなどを行って仮文節候補の補正を行うことにより、更に高精度な文節境界推定方法の実現を目指すことが挙げられる。

謝辞

本研究を進めるに当たりお世話になりましたNTT情報通信網研究所の自然言語処理グループの方々に感謝致します。

参考文献

- (1)牧野,木澤:”べた書き文の分かち書きとかな仮名漢字変換-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,Vol20,No4,pp337-345 (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”,情報処理学会論文誌,Vol24,No1,pp40-46(1983)
- (3)本間,山階,小橋:”連語解析を用いたべた書きかな漢字変換”,Vol27,No11,pp1062--1068(1986)
- (4)内山,板橋:”共起関係を利用した日本語複合名詞の分割”,情処自然言語処理研究,Vol91-7,pp57-54(1992)
- (5)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, pp.970-979 (1984)”
- (6)宮崎,大山:”日本文音声出力のための言語処理方式”,Vol27,No11,pp1053-1061(1986)
- (7)武田,藤崎:”統計的手法による漢字複合語の自動分割”,情報処理,Vol.28, No.9, pp.952-961 (1987)
- (8)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”,Vol.30, No.4, pp.467-477 (1989)
- (9)村上,荒木,池原:”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かな交じり候補の抽出精度”,信学論,J75-DII,pp11-20 (1992)