5Q－4

置換誤りを含んだべた書きかな文の仮文節境界の推定法

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

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、べた書き日本語文に対して最初に単語や文節などの単位に分か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、漢字かな混じり文に対する分かち書き処理としては、高精度な技術が確立されている[1]。しかし、べた書きかな文の場合は、総当たり法で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と一般に探索木の爆発が生じる。

このような問題を解決するために、[2]では２重マルコフモデルを用いて、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時的に定める方法が提案され、その有効性が示されている。本論文では、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３重マルコフモデルに拡張し、その有効性・向上効果を示すとともに、それらの手法を置換誤りタイプの音節ラティスから得られる音節文に適用し、仮文節境界推定の有効性を定量的に評価する。

A Method to Decide Provisional Boundaries of Bunsetsu in Non-segmented Japanese Kana Sentences Included Characters Substituted Wrongly.

Tetsuo Araki+　Satoru Ikehara++
Junya Tsutihasi+　Sinichi Sasajima+
+:Fauculty of Engineering,Fukui University
++:NTT Network Information Systems Laboratories

2. 仮文節境界推定法

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、マルコフ連鎖確率値が小さいほど、文字間の結合力が弱いという性質を用いて仮文節を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、表１のような各種推定法を定義し、仮文節境界推定の有効性を定量的に評価する。

表１　仮文節境界の判定方法

方法名	条件式
ＬＭ法	$P(X_j\|X_{j-2}X_{j-1})<T$
ＢＬＭ法	$P(b\|X_{j-2}X_{j-1})>T$
２ＢＬＭ法	$P_1(X_j\|X_{j-2}X_{j-1})<T_1$ かつ $P_2(b\|X_{j-2}X_{j-1})>T_2$
２ＣＢＬＭ法	$P_1(b\|X_{j-2}X_{j-1}X_j)>T_1$ かつ $P_2(b\|X_{j+2}X_{j+1}X_j)>T_2$
(2,3)-2BLM法	$P_1(X_j\|X_{j-2}X_{j-1})<T_1$ かつ $P_2(b\|X_{j-2}X_{j-1})>T_2$

3. 実験条件

種類：新聞記事

文種：①正しいべた書きかな文
②置換誤りタイプの音節ラティスから２重マルコフにより最尤に選ばれたべた書きかな文（マルコフ最尤型置換誤りと呼ぶ）
③ランダムな置換誤りを含んだべた書きかな文（ランダム置換誤りと呼ぶ）

総文章数：200文（1458文節）

総文字数：7272文字

誤り文字数：マルコフ最尤型置換誤り　＝ 874文字
ランダム置換誤り　＝ 921文字

4. 実験結果

LM法及びBLM法により、仮文節境界を推定した結果を図１（正しい文）、図２（置換誤りを含んだ

文）に対して示す。同図より、文節境界を表す特殊文字（空白文字）を用いたBLM法により、最大で再現率79％、適合率75％の結果が得られ、マルコフ最尤型置換誤りに対しても、最大で再現率73％、適合率67％の結果が得られた。

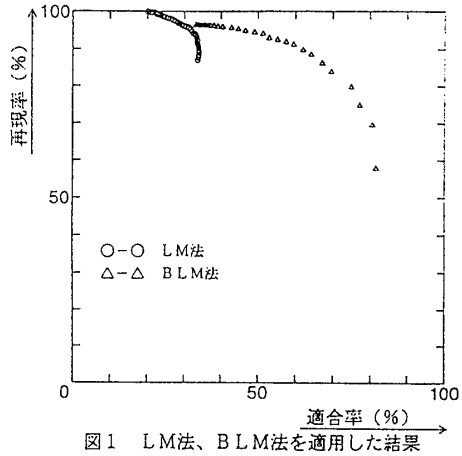

図1 LM法、BLM法を適用した結果
（正しいべた書きかな文）

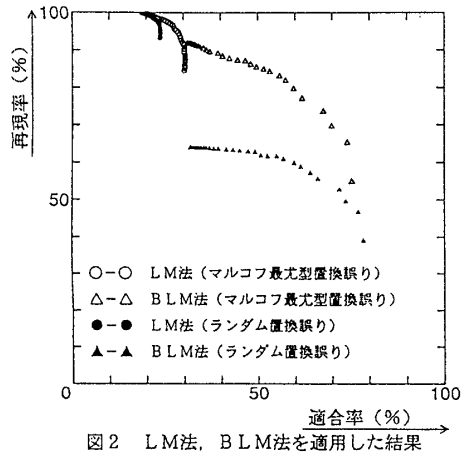

図2 LM法，BLM法を適用した結果
（置換誤りを含んだべた書きかな文）

次に、2BLM法、2CBLM法、(2,3)-2BLM法により、仮文節境界を推定した結果を図3（正しい文）、図4（置換誤りを含んだ文）に対して示す。同図より、2種類の足切り値を用いることで、最大で再現率79％、適合率87％の結果が得られ、マルコフ最尤型置換誤り文に対しても、最大で再現率71％、適合率80％の結果が得られた。

5. おわりに

3重マルコフ連鎖確率による仮文節境界推定法を提案し、それを用いて正しいかな文及び置換誤りを含んだかな文に対する仮文節境界推定の定量的評価を行った。その結果、正しいかな文に対しては最大で再現率79％、適合率87％の結果が得られ、従来の2重マルコフ連鎖確率による推定法と比較して、再現率で1％、適合率で4％の向上効果が得られた。また置換誤りを含んだ文の場合にも、正しい文と同様な方法で、最大で再現率71％、適合率80％の結果が得られた。

今後の課題としては、このように推定された仮文節境界の補正を行い、置換誤りと文節境界を区別し、更に高精度な文節境界判定を実現することが挙げられる。

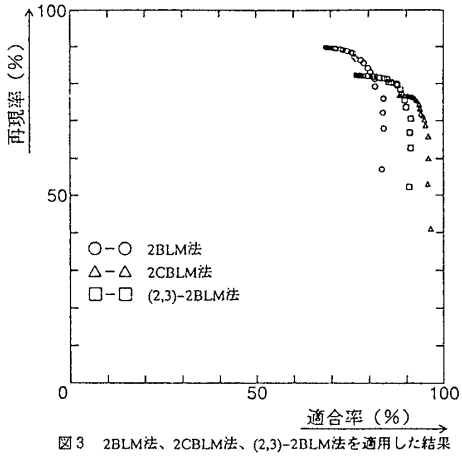

図3 2BLM法、2CBLM法、(2,3)-2BLM法を適用した結果
（正しいべた書きかな文）

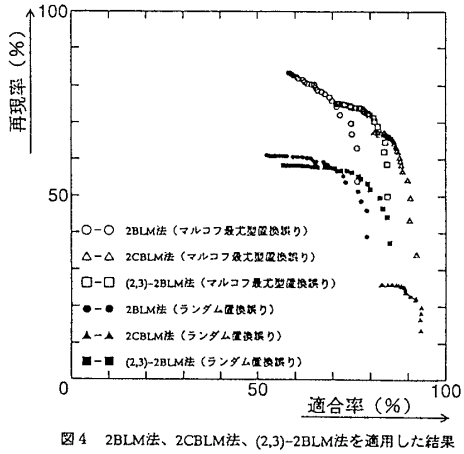

図4 2BLM法、2CBLM法、(2,3)-2BLM法を適用した結果
（置換誤りを含んだべた書きかな文）

参考文献

[1]宮崎、大山:"日本文音声入力のための言語処理", Vol.27, No.11, pp1053-1061(1986)

[2]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)