自然言語処理 94－7
（1993. 3. 19）

べた書き日本語文の脱落・挿入誤りの検出法

荒木 哲郎＋　池原 悟＋＋　塚原 信幸＋

＋　福井大学工学部

＋＋　ＮＴＴ情報通信網研究所

本論文では、日本語文節（音節表記、漢字かな表記）における連続した文字の脱落誤り及び挿入誤りに対して、m重マルコフ連鎖確率を用いた誤りの位置の検出並びに、正しい日本語文節に訂正する為の新しいアルゴリズムを提案する。またそのアルゴリズムの有効生を評価する為に、１文字並びに２文字の挿入誤りと脱落誤りを埋め込んだ新聞記事800文節を用いて、誤りのタイプ及び誤り文字列長が既知の場合について、文節内の誤り位置を検出し、訂正する実験を行った。その結果音節文節では、挿入誤りに対して適合率90～95％、再現率40～50％、脱落誤りに対して適合率60～70％、再現率15～30％、また漢字かな文節では、挿入誤りに対して適合率95～100％、再現率95～100％、脱落誤りに対して適合率95～100％、再現率45～55％であり、更に誤り訂正アルゴリズムを併用すると適合率が5～30％改善されることがわかった。

Algorithm for detecting of wrongly deleted or inserted character strings in solid Japanese written sentences

Tetuo ARAKI Satoru IKEHARA Nobuyuki TUKAHARA

+ Faculty of Engineering, Fukui University
++ NTT Network information Systems Laboratories

This paper describes a new method which enables to detect the wrongly inserted or deleted characters, and to replace them by correct characters in solid japanese sentences using m-th markov model. This method is based on the property of 2nd-order markov probabilities for correct characters within any bunsetu take take continuously great values ,but the markov probabiliies for wrongly deletetd or inserted characters in bunsetus take continuously low values at the number of times characterized by the error's type and error string's length .

The experimental result using 800 bunsetus of newspaper articles shows that this methods is useful for detecting and correcting the characters wrongly deleted or inserted in string of syllables and kanji characters.

1. まえがき

日本語文を計算機に入力する方法として、漢字ＯＣＲやワードプロセッサ(ＷＰ)、さらに音声認識等が存在するが、これらを用いて入力された日本語文には、一般に誤字の他に脱落及び誤挿入が含まれることが多い。しかし従来の日本文解析では対象とする文は文法的にも、意味的にも日本文として正しい文であることが前提であり、誤った文の文章解析の技術は現状ではほとんど確立されていないため、誤りを訂正するには人手による校正に頼らざるを得なかった。最近形態素解析技術や漢字かな変換技術の発展により、正しい文を解析する精度が向上してきたため、次に誤りの自動検出と自動訂正の技術を確立することが重要な課題となり、日本文訂正支援システム(ＲＥＶＩＳＥ)[5]では誤字を対象に検出・訂正を行っているが、連続した脱落、誤挿入についてはほとんど手がつけられていない。またこれまでに漢字かな交じり文の誤字の検出、訂正を対象にした単語解析プログラム並びに１重マルコフモデルを用いる方法[1][8]、単音節認識における単語中の誤字訂正法[2][6]、文字認識における誤字訂正法[3][7]、また誤り単語間の距離を用いた脱落や誤挿入誤りを含んだ音素列からの訂正法[4]があるが、誤字の検出、訂正法の研究が中心であること、また扱える単語数に制限があり、「開かれた世界」の任意な日本語文に対する脱落、誤挿入位置の検出、及び訂正を行う有力な方法が現状ではまだ確立されていない。

これまでに曖昧な音節候補列における絞り込み[10]や、漢字かな交じり文の絞り込み[11]において、２重マルコフ連鎖確率の有効性が知られている。本論文では、音節及び漢字かなの２重マルコフ連鎖確率がそれぞれ個々の音節文字や漢字かな文字の組み合わせ(３字組)における結合力の強度を表すことに着目し、音節及び漢字かな交じり文節中における連続した脱落・誤挿入文字の誤り位置を検出する方法並びに、正しい日本語文に訂正する方法について述べる。具体的には、日本語文節(音節表記、漢字かな表記)における連続した文字列の脱落及び誤挿入を自動的に検出並びに訂正する方法を検討する第一ステップとして、特に文字間の結合力を表すマルコフ連鎖確率値が、脱落及び誤挿入の所で連続して減少すること、また誤りタイプと脱落や誤挿入文字列長によって、マルコフ連鎖確率の減少回数が識別可能となる性質があることに着目することにより、ｍ重マルコフ連鎖確率を用いた誤りの位置の検出並びに、正しい日本語文節に訂正するための新しいアルゴリズムを提案する。実際に新聞記事データを高精度に文節分割し、音節変換する日本語解析プログラム[9]を用いて、２重マルコフ連鎖確率辞書の作成を行い、音節文節並びに漢字かな交じり文節における脱落並びに誤挿入位置の検出及び訂正実験を行いその有効性を評価する。

2. 連続した文字列の脱落と誤挿入とマルコフ連鎖確率モデルによる検出・訂正方法

ここでは、音節文節及び漢字かな交じり文節中における連続した脱落・挿入文字の誤りの位置を検出する方法並びに、正しい日本語文に訂正する方法について述べる。

2.1 日本語文節と誤りのタイプ

本論文では、日本語文として音節表記の文節(音節文節)と漢字かな交じり文節を扱い、検出並びに訂正の対象とする誤りのタイプを、次のように定義する。

【定義１】日本語の文節を$B = x_1 x_2 \cdots x_L$と表す。ここで各x_i ($1 \leq i \leq L$) が、全て音節文字のときにBを音節文節、またx_iが全て漢字文字であるかまたはかな文字の時Bを漢字かな交じり文節と呼び、Lは文節Bの長さを表す。文節B中の位置eからe+n-1 (但し、$1 \leq e \leq L$) までの文字列$x_e x_{e+1} \cdots x_{e+n-1}$が誤挿入の文字列である場合、Bを誤挿入の文節と呼び、eをBの誤り位置、nを誤り文字列長と呼ぶ。また文節B中の位置eとe+1の間に、文字列$x_{a1} x_{a2} \cdots x_{am}$が脱落しているとき、Bを脱落の文節と呼び、eを脱落文節Bの誤り位置、nを脱落文字列長と呼ぶ。

(定義終)

なお、本論文では、１文節中には誤挿入か脱落かのどちらか一方が１箇所にしか存在しないものとする。

【定義２】Bが誤挿入の文節または脱落文節の時、B中の位置eに、長さがnの文字列$x_e x_{e+1} \cdots x_{e+n-1}$が、誤挿入または脱落していることを自動的に検出することを、文節Bにおける誤り位置の自動検出と呼ぶ。また、誤挿入の文節Bから誤挿入文字列$x_e x_{e+1} \cdots x_{e+n-1}$を削除した文節$B' = x_1 x_2 \cdots x_{e-1} \cdots x_{e+n} \cdots x_L$を正解文節と呼び、BからB'を求めることを自動訂正と呼ぶ。同様に、脱落文節Bの誤り位置eとe+1の間に、脱落文字列$x_{a1} x_{a2} \cdots x_{am}$を付加して得られた文節$B' = B = x_1 x_2 \cdots x_e x_{a1} x_{a2} \cdots x_{am} x_{e+1} \cdots x_L$を、脱落の場合の正解文節と呼ぶ。

(定義終)

【定義３】誤挿入または脱落文節中の、誤挿入または脱落誤りの位置を自動検出する場合の誤り位置の適合率及と再現率を式(1)及び(2)によって、また自動訂正された正解文字列の適合率及び再現率を式(3)と(4)により定義する。

$$\text{誤り位置の適合率} = \frac{\text{誤り位置が正しく検出された件数}}{\text{検出された誤りの総件数}} \quad \text{---(1)}$$

$$\text{誤り位置の再現率} = \frac{\text{誤り位置が正しく検出された件数}}{\text{埋め込まれた誤りの総件数}} \quad \text{---(2)}$$

$$\text{正解文字列の適合率} = \frac{\text{誤り文字列が正しく訂正された件数}}{\text{検出された誤りの総件数}} \quad \text{---(3)}$$

$$\text{正解文字列の再現率} = \frac{\text{誤りの文字列が正しく訂正された件数}}{\text{埋め込まれた誤りの総件数}} \quad \text{---(4)}$$

（定義終）

2．2　マルコフ連鎖確率モデルによる誤り位置の検出手順

音節文字または漢字かな交じり文節内の文字間の結合力は、一般に誤挿入または脱落の文字列がある場合には、正解文字列の場合に比べて弱くなる性質があるので、以下の仮説を設ける。

【仮説】音節または漢字かな交じり文節（一般には文）内に脱落または誤挿入の文字列が存在するときは、m重マルコフ連鎖確率が一定区間だけ連続してあるしきい値以下の値を取る。

（仮説終）

この仮説が成り立てば、誤りのタイプ（脱落かまたは誤挿入）とその誤りの文字列長が存在する位置を決定する手順が次のように求まる。

【誤挿入位置の検出手順１】

誤挿入の文節を$B = x_1 x_2 \cdots x_{e-1} x_e x_{e+1} \cdots x_{e+n-1} x_{e+n} \cdots x_L$、またm重マルコフ連鎖確率をPとするとき、Bの中の位置eに誤挿入された長さnの文字列$x_e x_{e+1} \cdots x_{e+n-1}$は、次の手順によってで検出可能である。すなわち、m重マルコフ連鎖確率のしきい値P_kによって、

（１）位置i（i＝e-1またはi＝e+n+m）において、

$$P(x_i \mid x_{i-m} \cdots x_{i-1}) > P_k \qquad (1)$$

（２）任意の位置j（但し、e≦j≦e+n+m-1）において、

$$P(x_j \mid x_{j-m} \cdots x_{j-1}) < P_k \qquad (2)$$

但し、x_uでu<0のときは、X_uは空白文字とする。

（手順終）

【脱落位置の検出手順１】

脱落の文節を$B = x_1 x_2 \cdots x_e x_{e+1} \cdots x_L$とするとき、Bの中の位置eとe+1の間に、長さnの文字列$x_{e1} x_{e2} \cdots x_{en}$が脱落していることを、次の手順によって検出可能である。すなわち、m重マルコフ連鎖確率のしきい値P_kによって、

（１）位置i（i＝eまたはi＝e+3）において、

$$P(x_i \mid x_{i-m} \cdots x_{i-1}) > P_k \qquad (1)$$

（２）任意の位置j（但し、e+1≦j≦e+2）において、

$$P(x_j \mid x_{j-m} \cdots x_{j-1}) < P_k \qquad (2)$$

（手順終）

図１に２文字挿入がある場合の２重マルコフ連鎖確率値の変化を示す。同図より、仮説に従えば、２文字の挿入誤りがある場合には、２重マルコフ連鎖確率値が４回連続して落ち込むと言える。１重及び２重マルコフ連鎖確率を用いた場合の誤挿入及び脱落の誤り位置検出法を、表１に示す。以上より、マルコフ連鎖確率の連続した落ち込む回数を調べる事により、文節Ｂ内の長さがnの文字列個の連続した誤挿入及び脱落の誤り位置が検出可能となる。但し、長さnの文字列の脱落の場合には、マルコフ連鎖確率の落ち込み回数が常に２回と一定であるために、本手順では脱落文字の長さまでは判定することができない事に注意する。

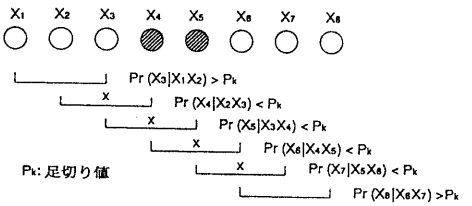

図１　２文字挿入の場合の２重マルコフ連鎖確率値

表２　n個連続した脱落誤りまたは挿入誤りの訂正方法

		１重マルコフ連鎖確率	２重マルコフ連鎖確率
挿入誤り	１個	確率値の連続１回改善	確率値の連続２回改善
	２個	確率値の連続１回改善	確率値の連続２回改善
	n個	確率値の連続１回改善	確率値の連続２回改善
脱落誤り	１個	確率値の２回改善	確率値の連続３回改善
	２個	確率値の３回改善	確率値の連続４回改善
	n個	確率値の(n+1)回改善	確率値の連続(n+2)回改善

表１　n個連続した脱落誤りまたは挿入誤りの検出方法

		１重マルコフ連鎖確率	２重マルコフ連鎖確率
挿入誤り	１個	確率値の連続２回落ち込み	確率値の連続３回落ち込み
	２個	確率値の連続３回落ち込み	確率値の連続４回落ち込み
	n個	連続(n+1)回落ち込み	連続(n+2)回落ち込み
脱落誤り	１個	確率値の１回落ち込み	確率値の連続２回落ち込み
	２個	確率値の１回落ち込み	確率値の連続２回落ち込み
	３個	確率値の１回落ち込み	確率値の連続２回落ち込み

2．3　マルコフ連鎖確率モデルによる誤り文字の訂正方法

2.2の手順により検出された、誤挿入及び脱落の誤り位置に対して、更にm重マルコフ連鎖確率のあるしきい値Ptを用いて、その誤りを訂正することが可能でありその手順を述べる。

【誤挿入の誤り訂正手順】

文節$B = x_1 x_2 \cdots x_{e-1} x_e x_{e+1} \cdots x_{e+n-1} x_{e+n} \cdots x_L$において、誤挿入の誤り位置検出手順により、位置eからe+n-1の文字列$x_e x_{e+1} \cdots x_{e+n-1}$が誤挿入である事が検出されたとき、文節Bから文字列$x_e x_{e+1} \cdots x_{e+n-1}$を取り除いて得られる文節を$B' = x_1 x_2 \cdots x_{e-1} x_{e+n} \cdots x_L$とする。B'の位置i（e+n≦i≦e+n+m-1）でm重マルコフ連鎖確率が全て条件

（1）　　$P(x_i \mid x_{i-m} \cdots x_{i-1}) > P_t$

を満足するとき、文節B'をBから誤挿入の訂正された文節と呼ぶ。

（手順終）

【脱落の訂正手順】

文節$B = x_1 x_2 \cdots x_e x_{e+1} \cdots x_L$において、脱落の誤り位置検出手順により、位置eとe+1の間に脱落がある事が検出されたとき、文節Bに長さnの文字列$x_{e1} x_{e2} \cdots x_{en}$を付加して得られる文節を$B' = x_1 x_2 \cdots x_e x_{e1} x_{e2} \cdots x_{en} x_{e+1} \cdots x_L$とする。B'における位置i（$i = e_1, e_2, \cdots, e_{n+m}, e+1, \cdots, e+m+n-1$）でm重マルコフ連鎖確率が全て条件

（1）　　$P(x_i \mid x_{i-m} \cdots x_{i-1}) > P_t$

を満足するとき、文節B'をBから誤挿入の訂正された文節と呼ぶ。

（手順終）

2重マルコフ連鎖確率を用いた2文字挿入及び脱落誤りに対する訂正アルゴリズムの例を図2、図3に示す。

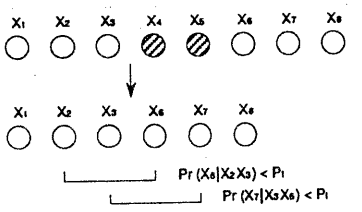

図2　2文字挿入の訂正例

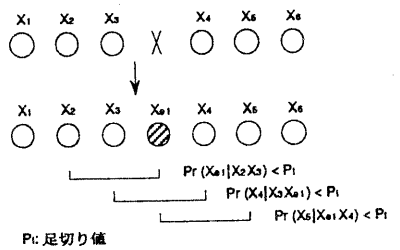

図3　脱落の訂正例

3．誤挿入及び脱落誤りの検出並びに訂正実験

3．1　実験条件

（1）日本語文の種類：新聞記事の音節表記文節と漢字かな表記文節。

（2）誤りのタイプと個数：1文節中に1箇所誤りが存在し、そのタイプは挿入誤りと脱落誤りの2種類であり、誤挿入または脱落の文字列長は、1または2である。

尚、実験では、予め文節における誤りのタイプと誤り文字列長は既知であるとして、その文節文字列（音節、漢字かな列）における誤り位置を決定するものとする。

（3）日本語文節数と平均文節長：文節数＝800、平均音節列長＝7、平均漢字かな列長＝6

3．2　実験結果

（1）誤り位置検出の適合率と再現率

音節文節及び漢字かな交じり文節中に、埋め込まれた長さ1～2の文字列の誤挿入並びに脱落に対して、2.2の誤り検出アルゴリズムを用いてしきい値P_tを色々変えた時の誤り位置検出の適合率と再現率の変化を、それぞれ図4及び図5に示す。同図より、音節文節と漢字かな交じり文節における、長さ1～2の文字列の誤挿入位置検出の適合率と再現率はそれぞれ次の通りである。

①音節文節の場合

(i)誤挿入位置検出：適合率＝90～95%
　　再現率＝40～50%

(ii)脱落位置検出　：適合率＝60～70%
　　再現率＝15～30%

②漢字交じり文節の場合

(i)誤挿入位置検出：適合率＝95～100 %
　　再現率＝95～99 %

(ii)脱落位置検出　：適合率＝95～100%
　　再現率＝45～55%

以上より、漢字かな交じり文節の方が音節文節の場合に比べて30%～45%程度高いことがわかる。

（2）検出及び訂正手順を組み合わせた時の誤り位置検出の適合率と再現率

（1）において、しきい値P_tを色々変えて誤り検出アルゴリズムを用いるとともに、2.3の誤り訂正アルゴリズム（但し、しきい値$P_t = 10$とする）を用いた時の誤り位置の適合率と再現率の変化を，それぞれ図6（音節文節）及び図7（漢字かな文節）に示す。同図より、誤り検出アルゴリズムと誤り訂正アルゴリズムを併用することにより、誤り位置の再現率はほとんど改善されないが、適合率が5～30%程度改善される効果があることがわかる。

（3）訂正手順による正解文字列の適合率と再現率

（1）で得られた誤り位置の適合率と再現率の積が最大となるときのしきい値を設定し、さらに2.3の誤り訂正アルゴリズムによるしきい値P_tを色々変えたときにマルコフ連鎖確率値が最小となるときの文字列候補を、それぞれ挿入または脱落させる方法で求めた正解文字列の適合率と再現率を図8（音節文節）及び図9（漢字かな交じり文節）に示す。

（4）しきい値の変化と適合率・再現率の関係

2.2及び2.3で述べた挿入誤り並びに脱落誤りの検出・訂正アルゴリズムを適用して、誤りのタイプ及び誤り文字列長の判定する際には、しきい値を下回るマルコフ連鎖確率値の個数が大きく関係している。そのためしきい値P_kまたはP_tを減少させていくと、必ずしも適合率が減少し、再現率が向上するという一般的な傾向が現れず、しきい値を下回るマルコフ連鎖確率値の個数が増大するために、逆に再現率が減少する場合も生じてくる。

以上より、漢字かな交じり文節内の連続した挿入誤り、及び脱落誤りの検出訂正に有効であることがわかる。挿入誤り及並びに脱落誤りの検出・訂正アルゴリズムを用いて得られた誤り位置並びに正解文字列の音節文節、漢字かな交じり文節の例を図10に示す。

2.2及び2.3で述べた挿入誤り及並びに脱落誤りの検出・訂正アルゴリズムを適用するのに要する処理時間は小さく、SUNSpark 4/2上で1文節当り約0.01～6秒程度である。

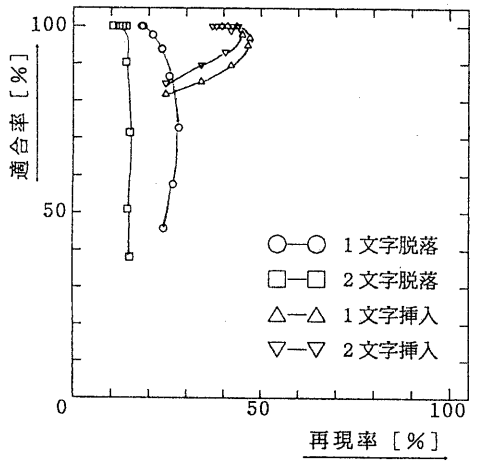

図4　誤り位置検出の適合率と再現率（音節文節）

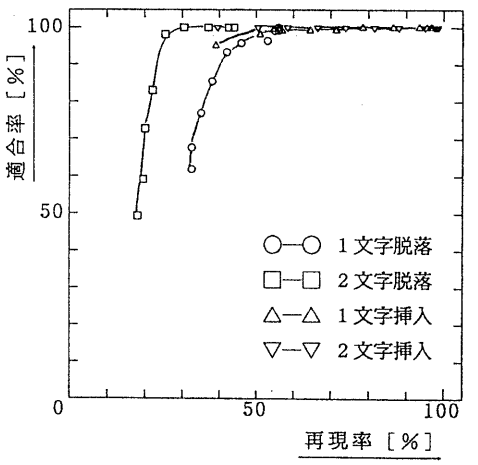

図5　誤り位置検出の適合率と再現率（漢字かな交じり文節）

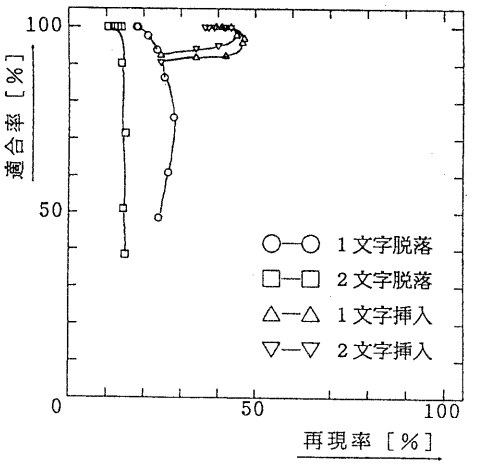

図6　誤り検出と訂正を併用した時の誤り位置の適合率と再現率（音節文節）

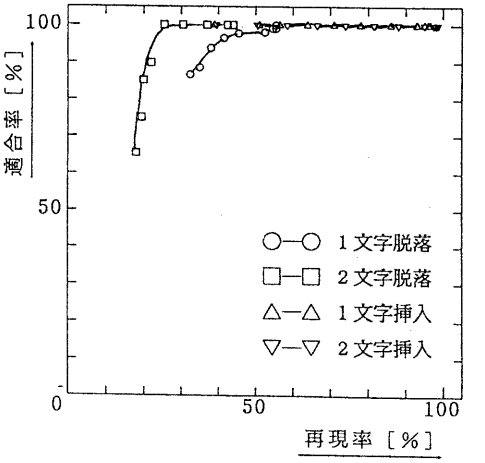

図7　誤り検出と訂正を併用した時の誤り位置の適合率と再現率（漢字かな交じり文節）

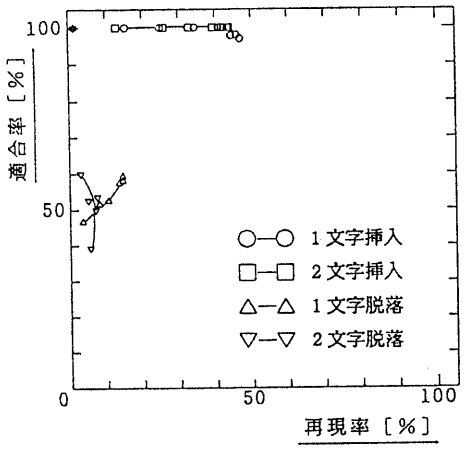

図8 正解文字列の適合率と再現率
（音節文節）

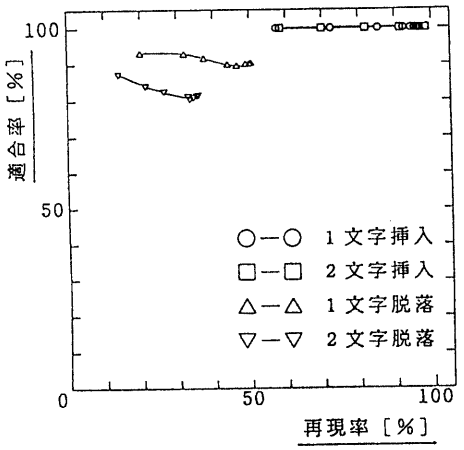

図9 正解文字列の適合率と再現率
（漢字かな交じり文節）

誤りを含む文節：経基盤を

	誤り位置	誤り文字
正解 :	1	営
検出結果：	1	不明
訂正結果：	1	営

(a) 脱落誤りの場合（漢字かな交じり文節）

誤りを含む文節：ミトメヘテイル

	誤り位置	誤り文字
正解 :	4	へ
検出結果：	4	へ
訂正結果：	4	へ

(b) 挿入誤りの場合（音節文節）

図10 誤りの訂正例

4．むすび

本論文では、「音節文節及び漢字かな交じり文節中に脱落または誤挿入の文字列が存在するときは、m重マルコフ連鎖確率が一定区間だけ連続してあるしきい値以下の値を取る」と言う仮説に従って、脱落及び誤挿入文字列の誤り位置をm重マルコフ連鎖確率モデルを用いて自動検出する方法並びに、正しい日本文に訂正する方法を新たに提案し、実際に2重マルコフ連鎖確率を用いた誤り位置の検出並びに訂正能力を新聞記事データについて評価し、その有効性を確認した。

今後の課題は、本論文で提案した検出・訂正アルゴリズムを音節文や漢字かな交じり文に拡張していくこと、また誤り種別が未知の場合の誤り文節及び文に拡張すること、更に単語辞書引き方と組み合わせた効果を評価していくことがあげられる。

<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いたします。

【文献】

(1)池原、白井："単語解析プログラムによる日本語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",情処論, 25,2,pp298-305 (1984)

(2)栗田,相沢:"日本語に適した単語の誤入力訂正方法とその大語い単語音声入力",情報処理学会論文誌, 25,5,pp831-841 (1984)

(3)杉村,斉藤:"文字連接情報を用いた読み取り不能文字の判定処理-文字認識への応用-"、信学論, J68-D,1,PP64-71 (1985)

(4)中川,義永:"誤りを含んだ音素系列からの候補単語の検索,計量言語学" ,14,8,pp327-334 (1985)

(5)池原,安田,島崎,高木:"日本文訂正支援システム（ＲＥＶＩＳＥ),研究実用化報告, 36,9,pp1159-1167 (1987)

(6)佐藤、荒木他："表層的な単語共起関係を利用した誤りを含む文の復元", 信学技報, NLC92-33,pp17-22 (1992)

(7)伊東,丸山:"ＯＣＲ入力された日本語文の誤り検出と自動訂正",33,5,pp664-670 (1992)

(8)下村、並木,中川,高橋:"最小コストパス検索モデルの形態素解析に基づく日本語誤り検出の一方式", 情処論,33,4,pp457-464 (1992)

(9)宮崎, 大山:"日本文音声出力のための言語処理,情処論",27,11pp1053-1061 (1986)

(10)荒木、村上、池原:"２重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", 情処論,30,4,pp467-477 (1989)

(11)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, D-II,J75-D-II,1,pp11-20 (1992)