

2重マルコフモデルによる日本語文の誤り検出並びに訂正法

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

漢字OCRやWP(ワードプロセッサ)、さらには音声認識装置などの入力装置を用いて計算機入力を行った日本語文には、一般に誤字、脱落及び挿入誤りが含まれるために、これらの誤りを自動的に検出し正しい日本語文に訂正する技術が必要となる。

これまでに、日本語の誤字を対象に単語解析プログラムを用いた誤字検出法並びに1重マルコフモデルによる訂正方法があり、また、日本語文節内の連続した脱落及び挿入誤りで種別が与えられている場合については、m重マルコフ連鎖確率を用いて誤り位置の検出並びに正しい日本語文に訂正するアルゴリズムが提案されている。

本論文では、上記のアルゴリズムを、3種類の誤り種別(誤字、脱落及び挿入)がわからない場合に対して、適用可能なように拡張する。また、新聞記事の77日分の統計データを用いて2重マルコフ連鎖確率辞書を作成し誤字、脱落及び挿入誤りが混在した文節を対象に誤りの種別と誤りの位置を検出する実験を行い、その有効性を評価する。

A Method for Detecting and Correcting of Characters Wrongly Substituted, Deleted or Inserted in Japanese Strings Using 2nd-Order Markov Model

Tetuo ARAKI⁺ Satoru IKEHARA⁺⁺ Nobuyuki TUKAHARA⁺

⁺ Faculty of Engineering, Fukui University

⁺⁺ NTT Network information Systems Laboratories

In optical character recognition and continuous speech recognition of a natural language, it has been difficult to detect error characters which are wrongly deleted and inserted. In order to detect and correct these characters, up to now the method using m-th order Markov chain model when types of errors are known, is presented. In this paper, the method described above was extended to be able to detect and correct characters substituted, deleted and inserted wrongly, when the types of errors are unknown. This method is based on the assumption that Markov probability of a correct chain of syllables or "kanji-kana" characters is greater than that of erroneous chains.

This method was applied to Japanese newspaper articles. 400 erroneous chains were prepared for each of syllables of "bunsetsu" and also "kanji-kana" characters of "bunsetsu" in the experiment. "Relevance Factor" P and "Recall Factor" R for erroneous characters deleted and corrected by this method were experimentally evaluated using statistical data for 77 issues of a daily Japanese news paper.

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。これまでに、誤りの自動検出と自動訂正をめざした、日本語訂正支援システム(REVISE)^[5]が開発されているが、誤字を対象に検出・訂正を行っており、連続した脱落、誤挿入については、従来、ほとんど手がつけられていない。また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][6]}、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り、脱落、誤挿入位置及びその種別を検出することについては、現在のところ、まだ有力な方法がない。

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。

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としては、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か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されている^[12]。この方法では、誤りの種別(脱落、誤挿入)は既知であると仮定しており、誤りが未知の場合は適用できない。

本論文では[12]に示されているアルゴリズムを、誤りの種別が未知の場合に適用できるように拡張する。また、新聞記事の77日文の統計データを用いて2重マルコフ連鎖確率辞書を作成し^[9]誤字、脱落及び誤挿入誤りが混在した文節を対象に誤りの種別及び位置を検出する実験を行い、その有効性を評価する。

2. 連続した文字列の誤字、脱落及び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

における連続した誤字及び脱落・誤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 $x_e x_{e+1} \cdots x_{e+n-1}$ が誤字列である場合、 B を誤字の文節と呼び、 e を B の誤り位置、 n を誤り文字列長と呼ぶ。同様に、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 $x_{e+1} x_{e+2} \cdots x_{e+m}$ 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文節中には誤字、脱落あるいは誤挿入のうちのいずれかが1箇所しか存在しないものとする。

【定義2】 B が誤字、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、誤字、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、誤字の文節 B の誤字列 $x_e x_{e+1} \cdots x_{e+n-1}$ を正しい文字に置き換えて得られた文節脱落文節 $B' = x_1 x_2 \cdots x_{e-1} x_{e+1} x_{e+2} \cdots x_{e+n} x_{e+n+1} \cdots x_L$ を誤字の場合の正解文節と呼ぶ。また、脱落文節 B の誤り位置 e と $e+1$ の間に、脱落文字列 $x_{e+1} x_{e+2} \cdots x_{e+m}$ を付加して得られた文節 $B' = x_1 x_2 \cdots x_e x_{e+1} x_{e+2} \cdots x_{e+m} x_{e+m+1} \cdots x_L$ を、脱落の場合の正解文節と呼ぶ。

(定義終)

【定義3】誤字、誤挿入または脱落文節中の誤

字、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）内に誤字、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（誤字、脱落または誤挿入）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。

【誤字及び誤挿入位置の検出手順】

誤挿入の文節を $B = x_1 x_2 \dots x_e \cdot x_{e+1} \dots x_{e+n}$ 、 $x_{e+1} \dots x_{e+n}$ 、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 から $e+n-1$ に存在する長さ n の誤字、あるいは位置 e に誤挿入された長さ n の文字列 $x_e x_{e+1} \dots x_{e+n-1}$ は、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 ($i=e-1$ または $i=e+n+m$) において、

$$P(x_i | x_{i-m} \dots x_{i-1}) > P_k \quad (1)$$

(2) 任意の位置 j (但し、 $e \leq j \leq e+n-1$) に

おいて、

$$P(x_j | x_{j-m} \dots x_{j-1}) < P_k \quad (2)$$

但し、 x_e 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】

脱落の文節を $B = x_1 x_2 \dots x_e x_{e+1} \dots x_l$ 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 $x_{e+1} x_{e+2} \dots x_{e+n}$ 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、

$$P(x_i | x_{i-m} \dots x_{i-1}) > P_k \quad (1)$$

(2) 任意の位置 j (但し、 $e+1 \leq j \leq e+2$) において、

$$P(x_j | x_{j-m} \dots x_{j-1}) < P_k \quad (2)$$

(手順終)

図1に2文字誤字または誤挿入がある場合の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文字の誤字または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、2重及び m 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 n 個の文字列の連続した誤字、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、脱落文字の長さは判定することができない事及び、誤字と誤挿入の区別が出来ない事に注意する。

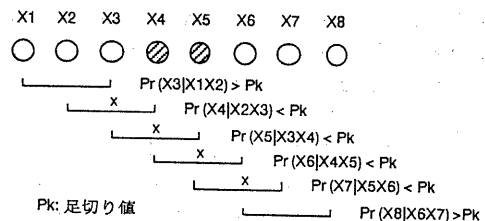


図1 2文字挿入または誤字の場合の2重マルコフ連鎖確率

表1 n個連続した誤字、脱落及び挿入誤りの検出方法

	1重マルコフ	2重マルコフ	m重マルコフ
挿入誤り	1個 連続2回落ち込み	連続3回落ち込み	連続m+1回落ち込み
	2個 連続3回落ち込み	連続4回落ち込み	連続m+2回落ち込み
	⋮	⋮	⋮
脱落誤り	n個 連続n+1回落ち込み	連続n+2回落ち込み	連続m+n回落ち込み
	1個 1回落ち込み	連続2回落ち込み	連続m回落ち込み
	2個 1回落ち込み	連続2回落ち込み	連続m回落ち込み
誤字	⋮	⋮	⋮
	n個 1回落ち込み	連続2回落ち込み	連続m回落ち込み
	1個 連続2回落ち込み	連続3回落ち込み	連続m+1回落ち込み
誤字	2個 連続3回落ち込み	連続4回落ち込み	連続m+2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	連続m+n回落ち込み

表2 n個連続した誤字、脱落及び挿入誤りの訂正方法

	1重マルコフ	2重マルコフ	m重マルコフ
挿入誤り	1個 1回改善	連続2回改善	連続m回改善
	2個 1回改善	連続2回改善	連続m回改善
	⋮	⋮	⋮
脱落誤り	n個 1回改善	連続2回改善	連続m回改善
	1個 連続2回改善	連続3回改善	連続m+1回改善
	2個 連続3回改善	連続4回改善	連続m+2回改善
誤字	⋮	⋮	⋮
	n個 連続n+1回改善	連続n+2回改善	連続m+n回改善
	1個 連続2回改善	連続3回改善	連続m+1回改善
誤字	2個 連続3回改善	連続4回改善	連続m+2回改善
	⋮	⋮	⋮
	n個 連続n+1回改善	連続n+2回改善	連続m+n回改善

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、誤字、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 P_k 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。

【誤字及び誤挿入の誤り訂正手順】

文節 $B = x_1 x_2 \dots x_{e-1} x_e x_{e+1} \dots x_{e+n-1} x_{e+n} \dots x_L$ 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 $x_e x_{e+1} \dots x_{e+n-1}$ が誤字あるいは、誤挿入文字である事が検出されたとき、

(i) 誤挿入と仮定し、以下の操作を行う。

文節 B から文字列 $x_e x_{e+1} \dots x_{e+n-1}$ を取り除いて得られる文節を $B' = x_1 x_2 \dots x_{e-1} x_{e+n} \dots x_L$ とする。 B' の位置 i ($e+n \leq i \leq e+n+m-1$)でm重マルコフ連

鎖確率が全て条件

$$(1) P(x_i | x_{i-m} \dots x_{i-1}) > P_k$$

を満足するとき、文字列 $x_e x_{e+1} \dots x_{e+n-1}$ を挿入誤りの文字候補とする。

(ii) 誤字と仮定し、以下の操作を行う。

文節 B において文字列 $x_e x_{e+1} \dots x_{e+n-1}$ を文字列 $x_{e_1} x_{e_2} \dots x_{e_m}$ に置き換えて得られる文節を $B' = x_1 x_2 \dots x_{e-1} x_{e_1} x_{e_2} \dots x_{e_m} x_{e+n} \dots x_L$ とする。 B' の位置 i ($i=e_1, e_2, \dots, e_{n+m}, e+1, \dots, e+m+n-1$)でm重マルコフ連鎖確率が全て上記条件(1)を満足するとき、文字列 $x_{e_1} x_{e_2} \dots x_{e_m}$ を誤字候補とする。

(iii) (i)、(ii)の結果、挿入誤りの文字候補と誤字候補のうちマルコフ連鎖確率の高い方を取る。すなわち、もし誤字候補の方が連鎖確率が高ければ誤字とし、 B' を誤字の訂正された文節と呼ぶ。

(手順終)

【脱落の訂正手順】

文節 $B = x_1 x_2 \dots x_e x_{e+1} \dots x_L$ 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 $x_{e_1} x_{e_2} \dots x_{e_n}$ を付加して得られる文節を $B' = x_1 x_2 \dots x_e x_{e_1} x_{e_2} \dots x_{e_n} x_{e+n} \dots x_L$ とする。 B' における位置 i ($i=e_1, e_2, \dots, e_{n+m}, e+1, \dots, e+m+n-1$)でm重マルコフ連鎖確率が全て条件

$$(1) P(x_i | x_{i-m} \dots x_{i-1}) > P_k$$

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文字誤字に対する訂正アルゴリズムの例を図2に示す。

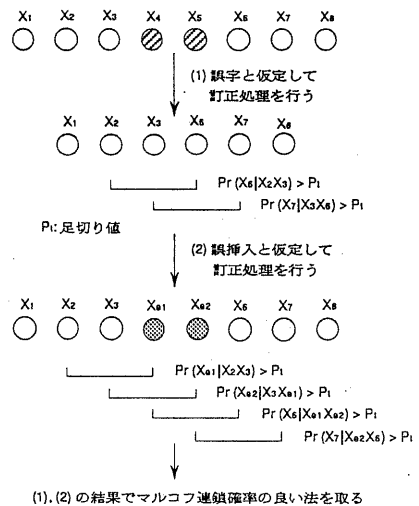


図2 2文字誤字の訂正方法

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、そのタイプは誤字、挿入誤り及び脱落誤りの3種類であり、誤字、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、文節における誤りのタイプと誤り文字列長は未知で、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=400、平均音節列長=7、平均漢字かな列長=6

3.2 実験結果

(1) 誤り位置検出実験

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の文字列の誤字、誤挿入並びに脱落に対して、誤り種別及び位置が未知であるとき、2.2の誤り検出アルゴリズムを用いてしきい値 P_x を色々変えた時の誤り種別及び位置検出の適合率と再現率の変化を、それぞれ図3及び図4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の文字列の誤り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=47~61%
再現率=44~47%
- (ii) 脱落位置検出：適合率=91~95%
再現率=16~21%
- (iii) 誤字位置検出：適合率=46~53%
再現率=40~44%

②漢字交じり文節の場合

- (i) 誤挿入位置検出：適合率=97~99%
再現率=97~99%
- (ii) 脱落位置検出：適合率=100%
再現率=57~58%
- (iii) 誤字位置検出：適合率=88~94%
再現率=88~93%

以上より、以下のことが言える。

- i) 漢字かな交じり文節の方が音節文節の場合に比べて35%~50%程度高い。

ii) 誤り種別が既知の場合と比べて、適合率が0~50%低下するが再現率は同程度である。

iii) 誤字の検出能力は誤挿入と比べると同程度か少し劣るが脱落よりも高い。

(2) 検出及び訂正手順を組み合わせた時の誤り位置検出実験

(1)において、しきい値 P_x 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)を用いた時の誤り種別及び位置検出の適合率と再現率の変化を、それぞれ図5(音節文節)及び図6(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若干低くなるが、適合率が5~30%程度改善される効果があることがわかる。これは、誤り種別が未知であるから訂正手順を加えることで誤り種別の間違いが生じる可能性があるためであり、誤り種別が既知の場合のように、単純に改善効果が得られるとは限らない。

(3) 誤り訂正実験

(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、誤り文字訂正の適合率と再現率を図7(音節文節)及び図8(漢字かな交じり文節)に示す。

挿入誤り及び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漢字かな交じり文節の例を図9に示す。

以上より、本手順は、漢字かな交じり文節内の連続した誤字、挿入誤りの検出及び訂正に対してはかなり有効であるが、脱落誤りの検出及び訂正、あるいは音節文節に対しては本手法の他に単語辞書引き等の検証法が必要になると思われる。

2.2及び2.3で述べた挿入誤り及び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、SUN Spark 4/2上で1文節当たり約0.01~6秒程度である。

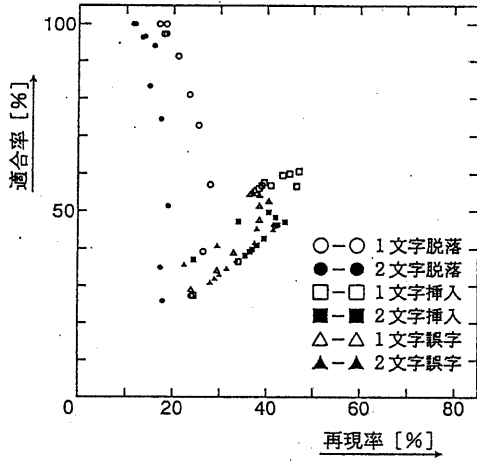


図3 誤り種別及び位置の検出結果（音節文節）

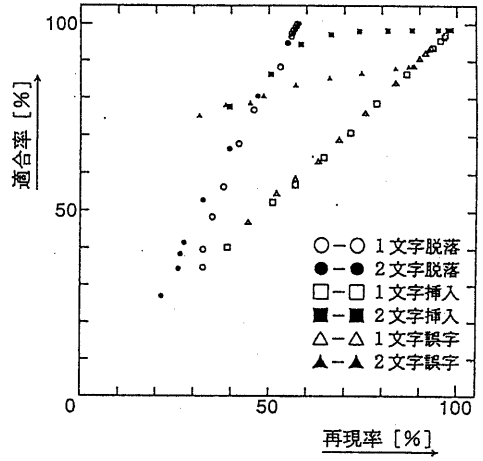


図4 誤り種別及び位置の検出結果（漢字かな文節）

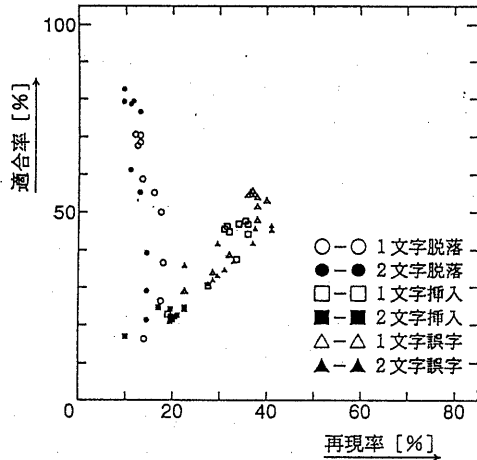


図5 検出と訂正を組み合わせたときの誤り種別及び誤り位置の検出結果（音節文節）

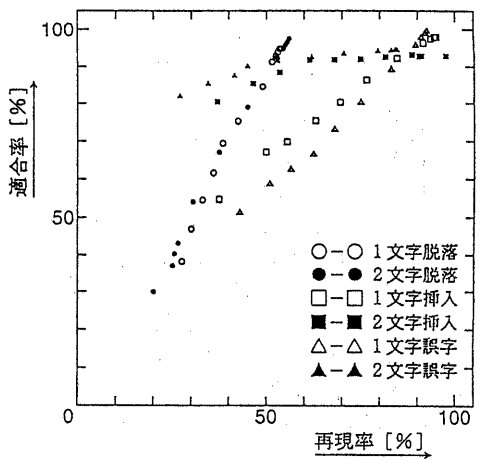


図6 検出と訂正を組み合わせたときの誤り種別及び誤り位置の検出結果（漢字かな文節）

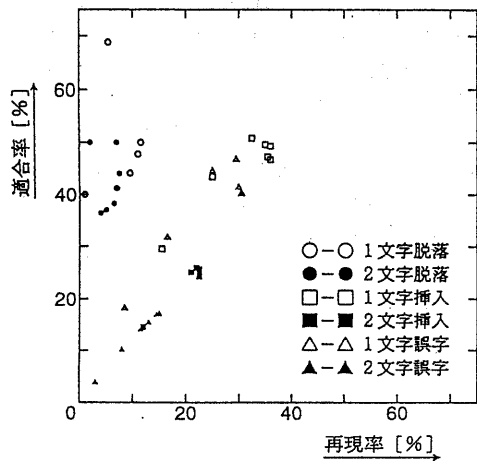


図7 誤り文字の訂正結果（音節文節）

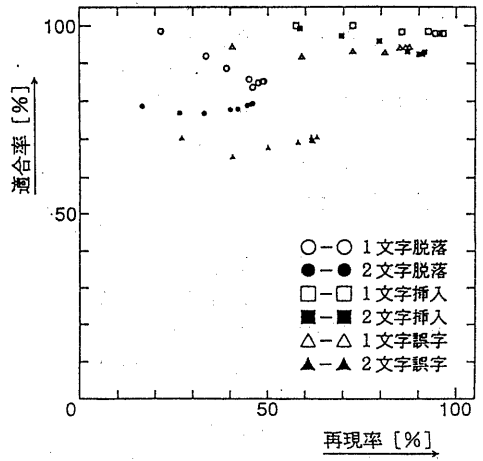


図8 誤り文字の訂正結果（漢字かな文節）

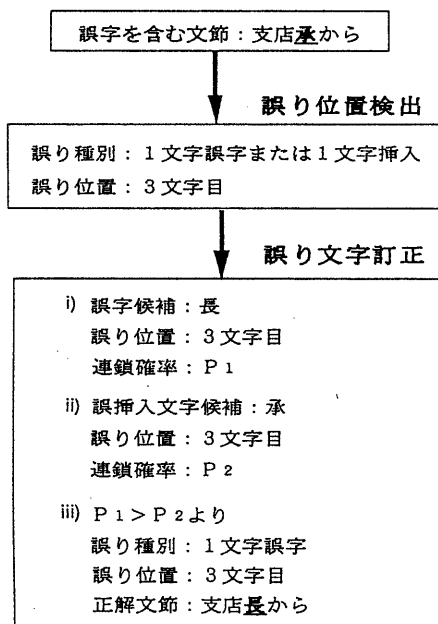


図9 1文字誤字（漢字かな交じり文節）の検出、訂正例

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誤字、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、誤字、脱落及び誤挿入文字列の誤り種別及び位置を 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、正しい日本語に訂正する方法を提案し、実際に誤字、脱落あるいは誤挿入が混在する誤り文節に対して2重マルコフ連鎖確率を用いた誤り種別及び位置の検出並びに訂正実験を行った。その結果、誤り種別が既知の場合と比べて、適合率が0~50%程度劣るが、誤字、誤挿入の検出は漢字かな交じり文節で再現率、適合率ともに88~99%と高い検出能力があることがわかり本手法の有効性を確認した。

今後の課題は、単語辞書引き方と組み合わせた効果を評価していくとともに、文節境界、単語境界などの誤り位置による検出、訂正能力等について詳細に調べることがあげられる。

<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: “文字接続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: “日本語訂正支援システム(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, 11, pp1053-1061 (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)
- (12)荒木, 池原, 塚原: “べた書き日本語文の脱落、挿入誤りの検出法”, 情処自然言語処理研究会 94-7, PP49-54 (1993)