日本語文の混合誤り文字列の検出訂正法の改善

荒木　哲郎＋ 池原　悟＋＋　榮代　正男＋
＋福井大学
＋＋鳥取大学

1 はじめに

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、さらに音声認識装置などが存在するが、これらを用いて入力された日本語文には、一般に置換型、挿入型および脱落型の誤りが含まれることが多く、FAX と OCR を経由して入力された文書には，置換型，挿入型，脱落型の誤りが混合して現れる場合（「混合誤り」と呼ぶ）が１割程度存在することが知られている [1]-[2]。本論文では、離れた文字間の結合力を評価するために，新たに定義するスキップタイプのマルコフ連鎖モデル及び、２重と３重の連続タイプのマルコフ連鎖モデルを併用して誤り種別を識別することにより、検出精度の向上めざした誤り検出･訂正方法を提案する。

2 混合誤りの検出訂正法

2.1 諸定義

【１】混合誤りの定義

これまでに、FAX と OCR を通して計算機に入力された日本語文には、置換型、挿入型、脱落型の誤り以外に、これらの混合された誤りが１割程度存在していることが知られている [2]。ここで、これらの誤りの種別を次のように定義する。

「単独誤り」とは，日本語文に含まれる誤り文字列が，単一の誤りタイプ（置換，挿入，脱落誤りのいずれかのタイプ）から構成される場合で，誤り文字数は任意の数を取りうるものとする。これに対し,「混合誤り」とは，置換，挿入，脱落の異なるタイプの誤りが連続して生じる場合を言い，置換と挿入誤りが連続しておこったものを,「置換＋挿入誤り」と表し，また，置換と脱落が連続しておこったものを,「置換＋脱落誤り」と表す。なお，挿入と脱落が連続して起こった場合は，単独の置換誤りと考える。また，本論文では,「置換＋挿入」（又は「置換＋脱落」）の順序は区別せずに扱うこととする。

【２】スキップタイプのマルコフ連鎖モデルの定義

m 重マルコフ連鎖モデル（連続タイプのマルコフ連鎖モデルと呼ぶ）において，文字 x_i の連鎖確率は，条件付き確率の定義から，次のように表される [1].

$$P(x_i \mid x_{i-m}, x_{i-m+1}, \cdots, x_{i-1}) \equiv \frac{P(x_{i-m}, x_{i-m+1}, \cdots, x_{i-1}, x_i)}{P(x_{i-m}, x_{i-m+1}, \cdots, x_{i-1})} \quad (1)$$

これに対して，離れた位置にある文字間の結合力を評価するのに，文字 x_i の連鎖確率が以下のように表されるマルコフ連鎖モデルを新たに定義し，これらをそれぞれ順方向及び逆方向のスキップタイプのm重マルコフ連鎖モデルと呼ぶ.

$$P^{n-SK}(x_i \mid x_{i-m-n}, x_{i-m-n+1}, \cdots, x_{i-n-1}) \equiv \frac{P(x_{i-m-n}, x_{i-m-n+1}, \cdots, x_{i-n-1}, x_i)}{P(x_{i-m-n}, x_{i-m-n+1}, \cdots, x_{i-n-1})}$$

$$P^{n-RSK}(x_i \mid x_{i+m+n}, x_{i+n+m-1}, \cdots, x_{i+n+1}) \equiv \frac{P(x_{i+m+n}, x_{i+n+m-1}, \cdots, x_{i+n+1}, x_i)}{P(x_{i+m+n}, x_{i+n+m-1}, \cdots, x_{i+n+1}))}$$

例として，順方向および逆方向の１文字スキップタイプの２重マルコフ連鎖モデルをそれぞれ図１ (ii),(iii) に示す．また，実験においては誤り文字数の出現頻度から見て，0〜2 文字スキップまでを用いる.

2.2 従来の単独誤りの検出方法

漢字かな交じり文を，$\alpha = x_1 \cdots x_i \cdots x_h$ と表すとき、この中に含まれる「単独誤り」の誤り種別と、誤り文字数は次のように決定される [1].
(1) 文字位置 i から $i+m+q-1$ までの $m+q$ 個のm重マルコフ連鎖確率 ($i \le j \le i+m+q-1$) が、誤り検出用のm重マルコフ連鎖確率のしきい値 T_D より小さいとき、連続落ち込み回数は $m+q$ 回であり、誤り文字列は $x_i\, x_{i+1} \cdots x_{i+q-1}$ で，誤り種別は置換または挿入誤り，誤り文字数は q と判定される．また，
(2) 文字位置 $j (i \le j \le i+m-1)$ の m 個のm重マルコフ連鎖確率 が、しきい値 T_D より小さいとき、連続落ち込み回数は m 回であり、文字位置 i

と $i+1$ の間に脱落誤りがあると判定される。

2.3 混合誤りの検出及び訂正手順

【1】誤り種別の判定法

2重マルコフ連鎖モデルを用いた時に、連続4回落ち込みがある場合は、「置換または挿入2文字誤り」の「単独誤り」と「置換1文字+挿入1文字誤り」の「混合誤り」の3つの場合が存在する(図2)．これらの誤りに対して、連続タイプのマルコフ連鎖確率とスキップタイプのマルコフ連鎖確率の落ち込み回数を調べ、「単独誤り」及び「混合誤り」の種別を決定する方法を以下に示す(図3参照)。但し、スキップタイプの連鎖確率を適用する場合は、連続タイプの連鎖確率が落ち込む位置と落ち込み回数より決定された誤り文字列候補を、元の文字列から削除した位置に対して行なうものとする。

＜1＞「0文字スキップタイプ」の連鎖確率が、落ち込まず、他の「1文字及び2文字スキップタイプ」の連鎖確率が落ち込む場合は、挿入1文字の「単独誤り」とする。
＜2＞「1文字スキップタイプ」の連鎖確率が落ち込まず、他の「0文字及び2文字スキップタイプ」の連鎖確率が落ち込む場合は、置換1文字の「単独誤り」とする。
＜3＞　「2文字スキップタイプ」の連鎖確率が、落ち込まず、他の「0文字及び1文字スキップタイプ」の連鎖確率が落ち込む場合は、置換1文字と脱落1文字の「混合誤り」とする。

【2】「混合誤り」の検出手順

このような誤り種別の判定法に基づき、「単独誤り」及び「混合誤り」を検出する方法を以下に示す。なお、ここでは誤り文字数は最大で2文字(すなわち、落ち込み回数は最大4回まで)とする。また、誤りの中で推定どおりの位置で、推定どおりの回数より一回少ない落ち込みとなる誤り(全体の15%程度存在)も検出の対象としている。
＜ステップ1＞しきい値 T_D を用いて文字連鎖確率を評価し、確率値が落ち込む文字 X_i と連続落ち込み回数を調べ、その落ち込み回数が、(i) 2回の時はステップ2へ、(ii) 3回の時はステップ4へ、(iii) 4回の時はステップ6へ進む。
＜ステップ2＞(2回落ち込みの場合)
(1) スキップタイプのマルコフ連鎖確率 $P^{n-SK}(X_i|X_{i-2}X_{i-1})$ と $P^{n-RSK}(X_{i-1}|X_{i+1}X_i)$ ($n=1$ 又は $n=2$) を調べる。
(2) 1文字または2文字のいずれかの連鎖確率がしきい値より高い値となれば、X_{i-1}, X_i の間の脱落型誤りとして検出する。
(3) いずれのスキップタイプの連鎖確率も低い値となった場合は、推定どおりの落ち込み回数より1回少ない場合の誤りか否かを調べるために、ステップ3へ進む。
＜ステップ3＞(推定どおりの落ち込み回数(3回)より1回少ない場合の落ち込みか否かの判定)
(1) 前方向の文字 X_{i-1} における順方向及び、逆方向のスキップタイプのマルコフ連鎖確率 $P^{n-SK}(X_i|X_{i-3}X_{i-2})$ と $P^{n-RSK}(X_{i-2}|X_{i+1}X_i)$ ($0 \leqq n \leqq 2$) の最大値(P_A とする)および、後方向の文字 X_i における順方向及び、逆方向のスキップタイプのマルコフ連鎖確率 ($0 \leqq n \leqq 2$) の最大値(P_B とする)を P_B とする。
(2) $P_A \geq P_B$ ときは、X_{i-1} が置換型または挿入型の1文字誤りとして、また、$P_A < P_B$ のときは、X_i が1文字誤りとして検出する。P_A,P_B 共にしきい値より低いときは、その連続した落ち込み箇所には誤りなしとする。
＜ステップ4＞(推定どおりの3回落ち込みの場合)
文字 X_i における順方向及び、逆方向のスキップタイプの連鎖確率を調べ、いずれかのスキップタイプの連鎖確率が大きければ、X_i の1文字誤りとする。また、どのスキップマルコフ連鎖確率も低い値となった場合は、ステップ5へ進む。
＜ステップ5＞(推定どおりの落ち込み回数(4回)より1回少ない場合の落ち込みか否かの判定)
文字 X_{i-1} と X_i の2文字誤りに対する順方向、逆方向のスキップタイプの連鎖確率 $P^{n-SK}(X_{i+1}|X_{i-3}X_{i-2})$ と $P^{n-RSK}(X_{i-2}|X_{i+2}X_{i+1})$ ($0 \leqq n \leqq 2$) の連鎖確率の最大値(P_A とする)及び、文字 X_i と X_{i+1} の2文字誤りに対する順方向、逆方向の連鎖確率 $P^{n-SK}(X_{i+2}|X_{i-2}X_{i-1})$ と $P^{n-RSK}(X_{i-1}|X_{i+3}X_{i+2})$ ($0 \leqq n \leqq 2$) の最大値(P_B とする)を求める。
P_A が P_B より大きいとき、X_{i-1} と X_i の2文字誤りとし、また、P_B が P_A より大きいときは、X_i と X_{i+1} の2文字誤りとして判定する。また、P_A と P_B 共にしきい値より小さいときは、落ち込み箇所には誤りなしとする。
＜ステップ6＞(推定どおりの4回落ち込みの場合)
文字 X_i,X_{i+1} の2文字誤りに対する順方向、逆方向のスキップタイプの連鎖確率 $P^{n-SK}(X_{i+2}|X_{i-2}X_{i-1})$ と $P^{n-RSK}(X_{i-1}|X_{i+3}X_{i+2})$ ($0 \leqq n \leqq 2$) を求める。いずれかのスキップマルコフ連鎖確率が高い値となれば、X_i と X_{i+1} の2文字誤り(置換2文字、挿入2文字又は置換1文字+挿入1文字)として検出する。また、どのスキップマルコフ連鎖確率も低い値となった場合は、その連続した落ち込み箇所には誤りなしとする。

【3】「混合誤り」の訂正手順

2の検出手順より得られた誤り種別に対して、2重及び3重の連続タイプの連鎖モデル[3]及び2重のスキップタイプの連鎖モデルを組み合わせた訂正手順を示す。
＜ステップ1＞
2の検出手順より求めた落ち込み回数(落ち込み開始位置を X_i とする)が、(i) 2回の時はステップ2へ、(ii) 3回の時はステップ3へ、(iii) 4回の

時はステップ４へ進む。
＜ステップ２＞２回落ち込み（脱落型誤りの場合）
文字 X_{i-1} と X_i の間の脱落型誤りとして、順方向及び逆方向のスキップタイプの連鎖確率値の平均（$n=1$ 又は $n=2$）と、１文字または２文字の脱落誤りの訂正訂正候補に対する連続タイプの連鎖確率値の積が最大となる候補を訂正候補する。
＜ステップ３＞３回落ち込み（置換１文字、挿入１文字、置換１文字＋脱落１文字）の場合
文字 X_i の１文字誤りとして、順方向及び逆方向のスキップタイプの連鎖確率値の平均（$n=1$ or 2）と、挿入型１文字誤りの訂正候補、置換型１文字誤りの訂正候補、または、置換１文字＋脱落１文字誤りの訂正候補に対する連続タイプの連鎖確率値の積が最大となる候補を訂正候補する。
＜ステップ４＞４回落ち込み（置換２文字、挿入２文字、置換１文字＋挿入１文字）の場合
文字の X_i と X_{i+1} の２文字誤りとして、順方向及び逆方向のスキップタイプの連鎖確率値の平均（$0 \leqq n \leqq 2$）と、挿入型２文字誤りの訂正候補、置換型２文字誤りの訂正候補、または、置換１文字＋挿入１文字誤りの訂正候補に対する連続タイプの連鎖確率値の積が最大となる候補を訂正候補する。

3 誤り検出訂正実験

3.1 実験条件

（１）連鎖確率辞書
(i) 日本語文の種類：漢字かな混じり表記された日経新聞記事（株式欄，広告欄等を除く記事データの部分），(ii) 連続タイプのマルコフ連鎖モデルの次数：２重及び３重，(iii) スキップタイプのマルコフ連鎖モデル次数：２重、スキップ文字数：0〜2文字 (n=0〜3)，(iv) 標本統計データ量：上記の新聞５年分（1677 日）から得られた統計データ (総文数 8,212,612 文 平均文長 39.3 文字)
（２）試験文用の入力データ
(i) 試験文データ：標本外データ[1]からランダムで擬似的に内部に設定，(ii) 誤り種別と文字数：置換型，挿入型，脱落型，「置換型＋挿入型」，「置換型＋脱落型」（以上連続２文字まで），(iii) データ量：１文中の誤りは１箇所として各 1000 文．計 8,000 文 (平均文長 39.8 文字)

3.2 実験結果

「単独誤り」及び「混合誤り」の検出・訂正実験結果を表１に示す。
【１】誤り検出の結果と考察
今回提案した方法は、従来の誤り検出精度（適合率と再現率の調和平均）比べると、置換型及び挿入型の「単独誤り」の検出結果に対しては、１文字誤りの場合で 10.1%（置換型の場合）11.8%（挿入型の場合）向上し、また２文字誤りの場合で 9.1%（置換型の場合）〜10.4%（挿入型の場合）向上することが分かった。
また、混合誤りに対しては、従来の方法では検出できず、置換１文字と脱落１文字の「混合誤り」の場合で 78.2%、置換１文字と挿入１文字の「混合誤り」の場合で 82.0%向上することがわかり、本方式の効果が確認できる。

【２】誤り訂正の結果と考察
4.2【１】で得られた誤り候補に対して、3．の誤り訂正法を適用した検出訂正の精度を表１に示す。
同表より、従来の誤り訂正精度と比べると、本誤り検出訂正方式はいずれの「単独誤り」及び「混合誤り」種別に対しても有効であり、「単独誤り」の１文字の場合で約 5%（置換型の場合）〜16%（挿入型の場合）向上し、「単独誤り」の２文字の場合で約 4%（脱落型の場合）〜30%（挿入型の場合）向上すること、及び、「混合誤り」の場合で約 38%（「置換１文字と脱落１文字の混合誤り」の場合）〜55%（「置換１文字と挿入１文字の混合誤り」の場合）がわかる。

4 おわりに

本論文では，従来検出・訂正が困難であった「混合誤り」に対して、(i) 誤り種別の判定しに連続タイプ及びスキップタイプのマルコフ連鎖モデルを併用し、(ii) 推定どおりの落ち込み回数より１回少ない落ち込みを誤り訂正候補に加え、(iii) 誤り訂正に２重及び３重マルコフ連鎖モデルを併用することにより、従来の「単独誤り」の検出・訂正精度を低下させることなく、「混合誤り」を新たに検出・訂正する方法を提案し、その有効性を実験により確認した。
今後の課題として，検出・訂正精度の向上を図るために，構文的な情報や単語の意味的な情報などを用いることがあげられる．

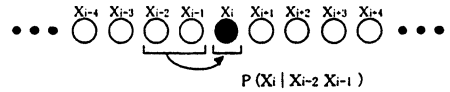
参考文献

[1] 荒木哲郎, 池原　悟, 塚原信幸, 小松康則, 田川崇史, 橋本憲久: “m重マルコフ連鎖モデルを用いた日本語文の誤字，脱落，挿入誤り文字列の検出と訂正法”, 信学論 (D-II)，vol.J83-D-II, no.6，pp.1-13, 2000.

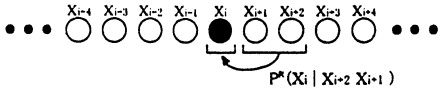
[2] 荒木哲郎, 池原　悟, 橋本憲久：“スキップマルコフ連鎖モデルを用いた日本文の誤り検出・訂正法 ”，信学技報，pp.1-8 , 2000.

[3] 荒木哲郎, 池原　悟, 佐藤　正伸, 榮代　正男：“マルコフ連鎖モデルを用いた日本文の置換型、挿入型及び脱落型誤りの検出・訂正法の改善 ”，信学論 (D-II)，vol.J85-D-II，no.1，pp.66-78, 2002.

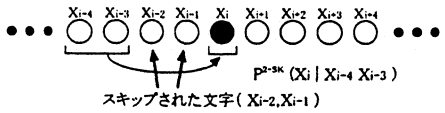
[1]試験文を連鎖確率辞書作成に用いた新聞記事文以外の記事文から選択する場合を言う．



(i) 2重順方向マルコフ連鎖モデル

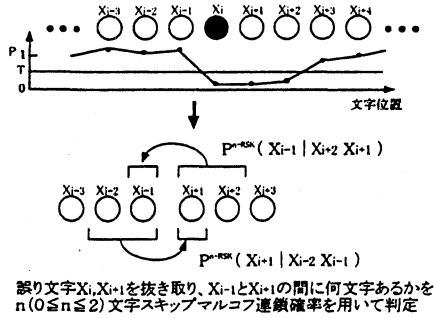


(ii) 2重逆方向マルコフ連鎖モデル



(iii) 2重順方向2文字スキップマルコフ連鎖モデル

図1 マルコフ連鎖モデル



誤り文字Xi,Xi+1を抜き取り、Xi-1とXi+1の間に何文字あるかを
n(0≦n≦2)文字スキップマルコフ連鎖確率を用いて判定

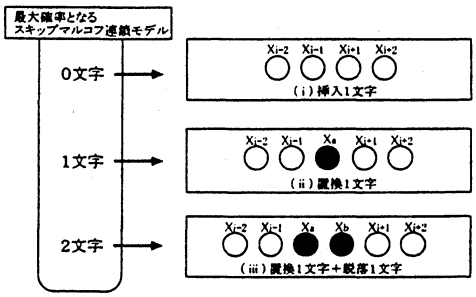


図3 2重マルコフ連鎖モデルにおける3回落ち込みの
スキップマルコフ連鎖モデルを用いた誤り種別判定法

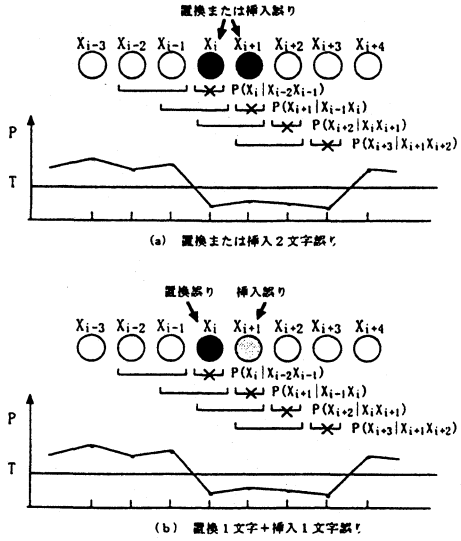


図2 混合誤りの例 (2文字誤りの場合)

表1 混合誤りに対する検出・訂正精度の比較

		従来の方法			新方法		
		再現率(%)	適合率(%)	調和平均(%)	再現率(%)	適合率(%)	調和平均(%)
検出	脱落1文字	33.1	52.0	40.5	30.2	51.5	38.1
	脱落2文字	36.9	66.5	47.5	33.1	64.8	43.8
	置換1文字	76.5	69.8	73.0	83.7	82.5	83.1
	挿入1文字	75.8	71.7	73.7	85.0	85.9	85.5
	置換1文字+脱落1文字	—	—	—	77.8	78.5	78.2
	置換2文字	74.8	66.0	70.1	79.8	78.5	79.2
	挿入2文字	74.1	65.5	69.5	82.5	77.5	79.9
	置換1文字+挿入1文字	—	—	—	83.0	81.0	82.0
訂正	脱落1文字	14.2	22.3	17.4	19.7	33.6	24.8
	脱落2文字	18.0	32.4	23.2	20.7	40.5	27.4
	置換1文字	53.0	48.4	50.6	56.2	55.4	55.8
	挿入1文字	58.1	55.0	56.5	72.5	73.3	72.9
	置換1文字+脱落1文字	—	—	—	38.2	38.5	38.4
	置換2文字	29.4	25.9	27.6	38.8	38.2	38.5
	挿入2文字	39.9	35.3	37.4	69.8	65.5	67.6
	置換1文字+挿入1文字	—	—	—	55.8	54.4	55.1