C4－3　スキップタイプのマルコフ連鎖モデルを用いた日本語文の脱落誤り文字列の検出と訂正法

橋本　憲久（福井大学）　荒木哲郎（福井大学）　池原　悟（鳥取大学）

1はじめに

コンピューターとのマンマシンインターフェイスを改善するためにＯＣＲ（光学式文字読みとり装置）や音声認識装置などの発達が期待されている。しかしながら、日本語の文章は通常、多種な文字（特に漢字かな混じり文）によって書かれるため、その入力が容易ではなく、ＯＣＲや音声認識装置を通して入力された文には、通常、誤りが含まれる。これらの誤りを自動的に検出し、訂正する自然言語処理技術の発達が期待され、従来、形態素解析技術や１重、および２重マルコフモデルを応用した誤りの検出・訂正法が知られている。誤り文字は一般に、誤字、誤挿入および脱落誤りの３つのタイプに分類されるが、脱落誤りにおいては、検出段階ではその誤り長が特定できないこと、および検出・訂正精度が他の誤りに比べ低いという問題点があった。そこで本論文では従来までの連続型のマルコフモデルに加え、新たに離散型マルコフモデル（スキップマルコフモデル）を導入してこれらの問題を解決する方法を提案し、その効果を実験により定量的に評価する。

2脱落誤り検出法と従来までの問題点

2.1マルコフモデル

今回脱落誤り検出に用いる２つのマルコフモデルは次のようなものである。

(1)連続タイプ：従来までのマルコフモデルで連続する文字間の結合力を調べるものであり、位置X_iの文字に対するｍ重マルコフ連鎖確率値は、次のような条件付き確率

$$P(X_i \mid X_{i-m} \cdots X_{i-1})$$

によって表される。例として２重マルコフの場合を図１(i)に示す。

(2)スキップタイプ：今回新たに導入するマルコフモデルで、離散的な文字間の結合力を調べるタイプであり、位置X_iの文字に対するｍ重ｎ文字スキップマルコフ連鎖確率値は

$$P(X_i \mid X_{i-n-m} \cdots X_{i-n-1})$$

のように表される。例として２重１文字、２文字スキップマルコフの場合を図１(ii)(iii)に示す。

2.2従来までの脱落誤り検出法

誤りを含む文字列に対する各マルコフ連鎖確率値は、正しい文字列に対するマルコフ連鎖確率値に比べ、小さいという性質がある。これを利用し文字列中の位置 i-1 と i の間に誤り長ｋの脱落誤りを含む文字列$\cdots X_{i-2}X_{i-1}X_iX_{i+1}X_{i+2}\cdots$におけるｍ重マルコフモデルの脱落誤り検出はしきい値をＴとして、次のように行う。

(1) $h = i-1$ かつ $h = i+m$ に対し　$P(X_h \mid X_{h-m} \cdots X_{h-1}) \geqq T$

(2) $i \leqq h \leqq i+m-1$ に対し　$P(X_h \mid X_{h-m} \cdots X_{h-1}) < T$

すなわち、脱落誤りにおいては（２）にしめすように連続ｍ回しきい値を下回る（落ち込む）。２重マルコフの場合は２回である。これによって脱落誤りの位置が特定できる。例として２重マルコフによる２文字脱落誤りの場合を図２(i)に示す。

2.3 従来までの問題点とその改善法

(i)脱落誤り長の検出：従来までの方法で脱落誤りの位置は特定できる。しかし、脱落誤り長によらず常にm回しきい値を下回り、脱落した文字数が検出段階では決定できない。このため訂正精度が下がるという問題点があった。そこで、脱落した文字数を決定するために j 文字スキップマルコフ（$P^{j\text{-}skip}$）を併用し、次のように行う。（$T^{j\text{-}skip}$ は j 文字スキップマルコフのしきい値）。

(3)$j \neq k$ に対して $P^{j\text{-}skip}(X_i | X_{i-m} \cdots X_{i-1}) < T^{j\text{-}skip}$

(4)$j = k$ に対して $P^{j\text{-}skip}(X_i | X_{i-m} \cdots X_{i-1}) \geqq T^{j\text{-}skip}$

（４）の条件を満たすとき、脱落誤り長をｋと検出する。例として２重マルコフによる２文字脱落の場合を図２(ii),(iii)に示す。なお、複数がしきい値を上回った場合は、確率の大きい方を選ぶ。

(ii)規定回数以下の落ち込みの検出：さらに、脱落誤りにおいては上述のように２重マルコフでは２回の落ち込みが検出されるはずであるが実際の脱落誤りを調べてみると落ち込み回数が２回となるものは７割近くであり、実際は、１回しか落ち込まないものが２割以上ある。そこで、$P(X_i | X_{i-2}X_{i-1}) < T$ が１回（落ち込みが１回）であった場合でもスキップマルコフモデルを用いてこれを検出する。この場合脱落誤りは $X_{i-1}X_i$ の間か X_iX_{i+1} の間にあると考えられるのでこの両方の位置にスキップマルコフモデルを(i)の場合と同様にして用い脱落誤り位置および脱落誤り長を検出する。これを図３に示す。

3 脱落誤り訂正法

上記の手順によって位置 i-1 と i の間に長さ k の文字列が脱落していることが検出された場合、以下の脱落誤りの訂正手順[1) 2)]を行う。最初に脱落誤りを含む文Ｂに長さｋの全ての文字列候補 $X_iX_{i+1} \cdots X_{i+k-1}$ を付加して訂正文字列候補 $X_{i-2}X_{i-1}X_iX_{i+1} \cdots X_{i+k-1}X_iX_{i+1}$ を順次生成する。次に、ｍ重マルコフ連鎖確率のしきい値Ｔに対して、各候補における位置ｐ（$i \leqq p \leqq i+k+m-1$）でｍ重マルコフ連鎖確率が全てが次の条件

$$P(X_p | X_{p-m} \cdots X_{p-1}) \geqq T$$

を満足することを確認する。最後にこの様な条件を満たす文字列候補の中でマルコフ連鎖確率値の積の最も大きい文字列候補を正解文字列とする。２重マルコフ連鎖確率を用いた２文字脱落誤りに対する訂正方法の例を従来の場合と比較して図４に示す。

4 実験

本論文では従来、有効性が知られている２重マルコフモデルを考え、また実際に３文字以上の脱落誤りが少ないために、１文字または２文字の脱落誤りの場合を考える。

4.1 実験条件

(1)マルコフ連鎖確率辞書：漢字かな混じり表記された日経新聞記事７７日分を標本統計データ（文総数は 28,547 文、平均文長は 37.4 文字）として用い、マルコフ連鎖確率辞書の作成を行った。辞書の種類として、連続型およびスキップ型（１文字・２文字）を用いた。

(2)実験用入力データ：日経新聞記事の標本内データから文長１５文字以上（平均 39.8 文字）の文を対象として、１文中に１箇所、ランダムで内部に１文字または連続２文字の脱落誤りを設定したもの各 1000 文について実験を行った。

4.2 実験結果

従来までの連続マルコフモデルのみの方法と今回のスキップタイプのものを併用した方法の検出・訂正精度の結果を表１に示す。同表より検出段階においては、従来の方法より適合率は約５％ほど低下する。これは、１回落ち込みの時に間違った位置を検出してしまうためである。しかし再現率が約１８％ほど向上し両者の調和平均では、約９～１０％向上することが分かった。最終的な検出・訂正を行った結果では適合率が１文字脱落の場合約３％向上、２文字脱落の場合約６％低下であるが、再現率において約１０～２０％向上し、両者の調和平均では約４～１３％向上することが分かった。

5 まとめ

マルコフモデルを用いた日本語文の誤り自動検出・訂正において、脱落誤りの場合には、誤り位置でマルコフ連鎖確率値が必ずしも一定回数連続して減少するとは限らず、誤りを見逃したり、また誤り長を正しく識別できないために訂正精度が他の誤り種別（誤字、誤挿入誤り）に比べ低いという問題点があった。

本論文では、従来の連続タイプのマルコフモデルに加えて、新たにｎ文字スキップのマルコフモデルを導入することにより、検出段階での誤り長の特定および連鎖確率値の減少回数が少ない場合でも誤りを検出する新しい方法を提案した。また、実験では、漢字かな混じり表記された新聞記事（2,000 文）を対象に、それらが脱落文字を含む場合の誤り検出訂正精度を従来の方法と比較評価した。その結果、提案した検出・訂正手順により、１文字、２文字の脱落誤りの検出精度が、適合率 94.3-95.1 %、再現率 85.1-86.5 %が得られ、従来の方法に比べ、両者の調和平均で、約９～１０％向上する事が分かった。また、最終的な検出・訂正精度は適合率 67.9-79.8 %、再現率 59.1-72.3 %となり、従来の方法よりも両者の調和平均で、約４～１３％向上する。以上のことが分かり本手法が有効であることが分かった。

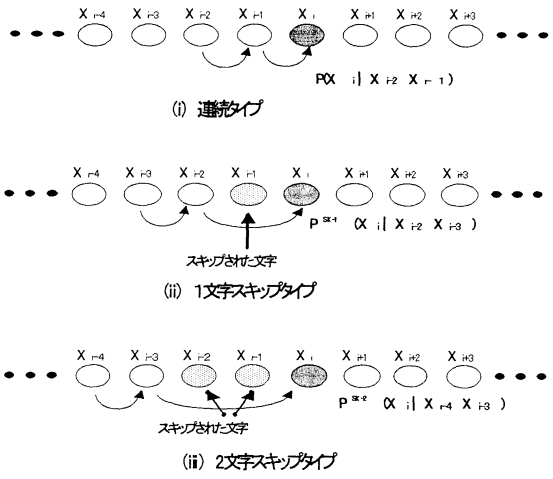

図1 マルコフモデルの概念図
(2重マルコフの例)

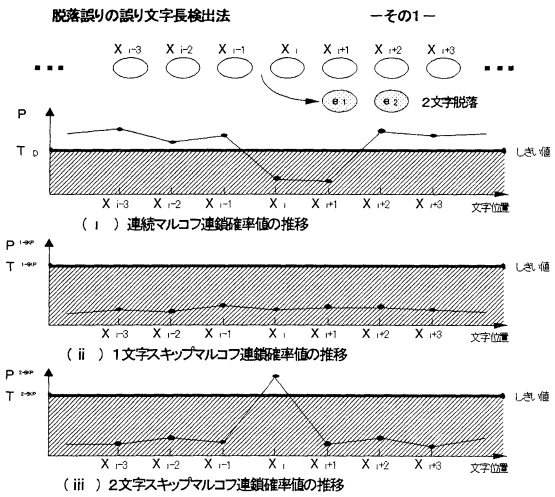

図2 連続2重マルコフ及びスキップ2重マルコフ
による2文字脱落誤りの検出の考え方

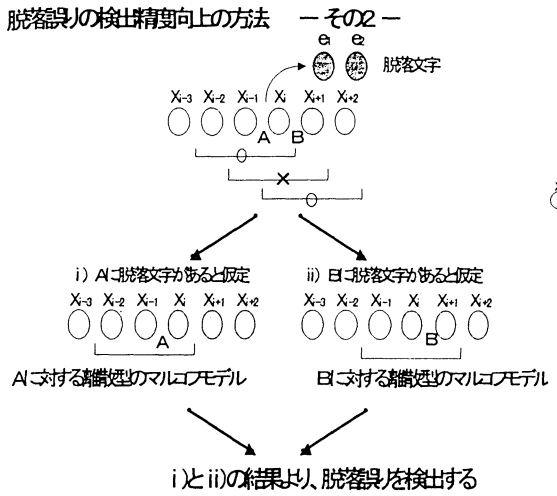

図3 連続2重マルコフが1回落ち込みの時
スキップマルコフを用いて脱落誤りを検出する方法

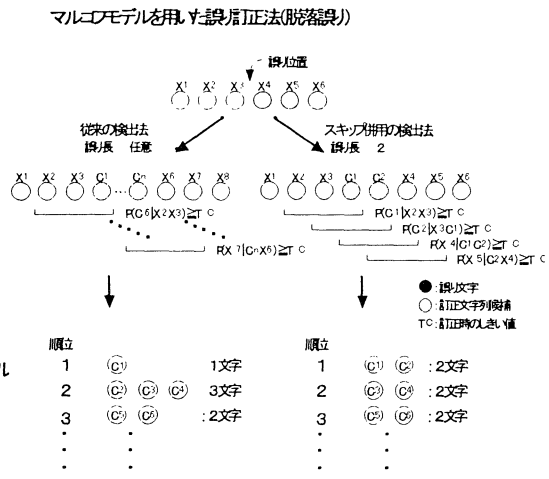

図4 2重マルコフ連鎖モデルによる誤り訂正例
(連続2文字脱落誤りの場合)

表 1 マルコフモデルを用いた脱落誤り実験結果の比較

		従来の方法 (連続型のマルコフモデルのみ)			本論文の方法 (スキップマルコフモデルを併用)		
		再現率 R(%)	適合率 P(%)	調和平均	再現率 R(%)	適合率 P(%)	調和平均
検出	1文字	68.4	100.0	81.2	86.5	94.3	90.2
	2文字	66.7	100.0	80.0	85.1	95.1	89.8
訂正	1文字	52.7	77.0	62.6	72.3	79.8	75.7
	2文字	49.5	74.2	59.4	59.1	67.9	63.2

参考文献

1) T.Araki,S.Ikehara,N.Tsukahara and Y.komatsu," An Evaluation to Detect and Correct Erroneous Characters wrongly Substituted,Deleted and Inserted in Japanese and English Sentences Using Mrkov Models " ,*Coling-94*,Vol.1,pp187-193(1994)

2) T.Araki,S.Ikehara,N.Tsukahara and Y.komatu," An Evaluation of a Method to Detect and Correct Erroneous Characters in Japanese Input Through an OCR using Markov Models" ,*Applied Natural Language Processing*,pp98-199(1994)