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