情 報 学 基 礎 50－2
（1998．7．17）

日本語文の相互干渉誤り文字列の検出・訂正方法

荒木 哲郎　池原 悟　小松 康則
（福井大学）（鳥取大学）（日立情報制御ｼｽﾃﾑ）

本論文では，日本語文の中で近くに位置して相互に干渉し合う誤り文字列を検出・訂正する方法を提案する．ＯＣＲや音声認識装置を通して入力された文書は，通常，置換誤り，誤挿入，脱落誤りの３種類の誤りを含んでいる．最近，選択的誤り訂正法がこれらの３種の誤りの検出・訂正に対して提案され，誤りが孤立（誤りが相互に干渉し合わないように，離れた位置に存在）する場合に有効であることが示されている．しかし，この方法は誤りが互いに近くにあって干渉し合う場合には，誤りのタイプを識別し，誤りの位置と誤り文字列長を決定する事が難しいために有効でない．

この問題を解決するために，本論文では次のような２段階の方法を提案する．最初の第１ステップでは，お互いに近くにある１個以上の誤り文字を含む誤りの区間を検出する．次の第二ステップでは，誤り文字列を検出・訂正するために，今検出された誤りの区間に対して，スキップタイプのマルコフモデルを新しく導入して詳細に解析する．

また，提案した方法が日本語文内で近くに位置して相互に干渉する誤り文字列の検出・訂正に対して有効か否かを評価するための実験結果を示す．

Detection and Correction of Mutually Interfered Erroneous Characters in Japanese Texts

Tetsuo Araki　Satoru Ikehara　Yasunori Komatsu

This paper proposes a method to detect and correct erroneous characters which locate so closely in Japanese text that they interfere each other.

Texts input through OCR or speech recognition devices usually contain three kind of erroneous characters such as wrongly substituted, wrongly deleted and wrongly inserted characters. Recently, selective error correction method has been proposed for these three kind of errors and it was shown that the method is very effective for the errors of isolated type which are located at so different places from each other that there is no mutual interference. However, the method does not work well for the error characters located closely each other, In such a case, it has been very difficult not only to distinguish the types of errors but also to determine the location and the length of erroneous character parts.

In order to resolve this problem, this paper proposed a two tiered methods. First it finds an error section which includes one or more error characters located closely each other. Second the error section is analyzed in detail to find and correct error characters. To perform these processes, skip type Markov Chain model was newly introduced. The experimental results showed that the proposed method was very effective for finding and correcting error characters closely located each other in Japanese texts.

1 はじめに

コンピュータのマンマシンインターフェースを改良するするために，ＯＣＲや音声認識装置が開発されてきた。しかし，日本語文書が多くの字種，特に数千種の漢字などを用いて書かれているため，これらの装置で日本語文書を入力することは容易ではなく、通常，誤りが含まれる。これらの誤りを自動的に検出し，訂正するための自然言語処理の技術が期待されているが、これまで主として正しい文に対して開発されてきたために，これらの問題に直接適用することはできない。これまでに、統計的な手法が（例えば，マルコフモデル）がこの問題の解決に用いられてきた [1]-[7]。

誤りのタイプは次の３つに分類することができる．すなわち、第１は，正しい文字が間違って他の文字に認識された場合であり，置換誤りと呼ばれる。また、第２と第３は，挿入誤りと脱落誤りである。最近，ｎ重マルコフモデルを用いて、誤りのタイプを識別し，各誤り文字列の性質に従ってそれらを訂正する選択的誤り訂正手法が提案されている [8]。FAX とOCR を通して入力された日本語文書の誤り検出・訂正実験の結果によれば、この方法は誤りが互いに離れた位置にあって相互に干渉しない場合 (このタイプの誤りを「孤立誤り」と呼ぶ）に対して，有効である事が示されている。しかし，２つ以上の誤り文字列が互いに近く位置していて それらが干渉し合う場合（このタイプの誤りを，「干渉誤り」と呼ぶ）には，これらの誤りの位置と長さを正しく検出することが難しく，全体の誤り訂正精度が低下するという問題がある。

本論文では，２つのタイプのｎ重マルコフモデルを用いて，「干渉誤り」を検出し訂正する方法を提案する。「干渉誤り」を検出し訂正するためには，これらの誤りの近傍にある全ての文字が正しい事が前提となる。最初に，ｎ重マルコフモデルを用いて，互いに近く位置している１個以上の誤り文字を含む区間を検出する。このステップは，信頼できる文字と信頼出来ない文字との境界を決定することであり，信頼できない文字列を，「誤り区間」と呼ぶことにする。次に，「誤り区間」に対して誤り文字を検出し，訂正するために、それらの文字間の結合力を詳細に解析する。この解析は，「誤り区間」の外に位置している信頼出来る正しい文字列だけを用いて行われるために、「誤り区間」の内部に位置している誤り文字列をスキップする新しいタイプのｎ重マルコフモデルを導入する（これをスキップタイプのマルコフモデルと呼ぶ）。

本手法はｎ重マルコフモデルに適用できるが，簡単のために，１重および２重マルコフモデルを用いて実験を行ない、検出・訂正された誤り文字列の適合率Ｐおよび再現率Ｒを評価する。

2 基本的な定義と二つのタイプのマルコフモデルを用いた相互干渉の誤り文字列の検出・訂正方法

2.1 基本的な定義

日本語文を $\Gamma = X_1X_2X_3\cdots X_n$ によって表す。但し，X_i $(1 \leq i \leq n)$ は漢字またはかな文字を表す．OCR を通して入力された文は，正しい文字の部分 $X_iX_{i+1}X_{i+2}\cdots X_{i+r-1}$ と、誤り文字の部分 $\bar{X}_i\bar{X}_{i+1}\bar{X}_{i+2}\cdots\bar{X}_{i+s-1}$ に分れる．ここで，$\bar{X}_i$ は誤り文字 $(1 \leq i,, s \leq n)$ を表す．

n 重マルコフ連鎖確率 $P(\bar{X}_i \mid X_{i-n}\cdots X_{i-1})$ を用いて，誤り文字列の先頭 $\bar{X}_i$ を正しく検出するためには，文字 $X_{i-m}\cdots X_{i-1}$ (これらを X_i の近傍と呼ぶ) が正しいことが前提となる．そのとき，次の二つの誤り文字列 $\bar{X}_i\bar{X}_{i+1}\bar{X}_{i+2}\cdots\bar{X}_{i+s-1}$ と $\bar{X}_j\bar{X}_{j+1}\bar{X}_{j+2}\cdots\bar{X}_{j+u-1}$ の間の関係は次の二つのタイプに分けられる．１つのタイプは，お互いに相互干渉しない誤り文字列すなわち，誤り文字列 $\bar{X}_i\bar{X}_{i+1}\bar{X}_{i+2}\cdots\bar{X}_{i+s-1}$ が，$\bar{X}_j$ の近傍に含まれない場合で，これらの誤りを，「孤立誤り」と呼ぶ．もう一つは，誤りの文字列が相互に干渉する場合で、これらの誤りを「干渉誤り」と呼ぶ．これらの「孤立誤り」と「干渉誤り」の例を，図１に示す．例えば，２重マルコフモデルの場合，２つの誤り文字列 $\bar{X}_i\bar{X}_{i+1}\bar{X}_{i+2}\cdots\bar{X}_{i+s-1}$ と $\bar{X}_j\bar{X}_{j+1}\bar{X}_{j+2}\cdots\bar{X}_{j+u-1}$ の間の距離が３より小さいときには，２つの誤り文字列は相互に干渉している呼ばれ、これらの２つの誤り文字列が「誤り区間」に含まれる（図2）．「干渉誤り」に対して、互いに近く位置している１個以上誤り文字を含んでいる区間を、「誤り区間」と呼ぶ．

「誤り区間」は，明らかに正しい信じられる文字（「誤り区間」の外側）を，正しいと信じられない文字（「誤り区間」の内部）から区別する．正しいと信じられる文字を用いて，「誤り区間」の内部解析するために，新しいタイプの m 重マルコフ連鎖確率が定義される．すなわち、

$P(X_i \mid X_{i-n-k}\cdots X_{i-k-1})$ は文字 X_i のマルコフ連鎖確率値が、k 文字分 $(X_1X_2\cdots X_k)$ をスキップ

した文字列 $X_{i-n-k}\cdots X_{i-k-1}$ によって決定されるもので，スキップタイプのマルコフモデルと呼ばれる．図3に、従来の非スキップタイプと、1文字スキップの2重マルコフモデルの例をに示す．

2.2 2重マルコフモデルを用いて誤り区間を検出する方法

最初に，1文内の「誤り区間」先頭位置と長さを決定する方法を述べる．

[手順 1] (「誤り区間」の先頭と長さを決定する方法)

次の条件を満たす長さkの文字を検出する．そのとき，「誤り区間」の先頭位置が i で長さが k であると判定される．

(1) $P(X_h \mid X_{h-m}\cdots X_{h-1}) > T_L$, 但し，$h = i-1$ または $h = i+k+m$ 及び

(2) 任意なj($i \leq j \leq i+k+m-1$) に対して、$P(X_j \mid X_{j-m}\cdots X_{j-1}) < T_L$

ここで，$P(X_j \mid X_{j-m}\cdots X_{j-1})$ は通常の非スキップタイプの m 重マルコフ連鎖確率であり，X_u は $u < 0$ のとき，空白記号を表す。また．T_L は「誤り区間」を検出するのに用いられる m 重マルコフ連鎖確率のしきい値である．

手順1によって得られた「誤り区間」は，その中に含まれる誤り文字の位置によっていくつかのタイプに分類される（200文を用いたFAX-OCR誤りにおいて、2重マルコフモデルにより検出する場合に生じる誤り区間のタイプ(誤り区間の長さが3，4，5の場合)を図4に示す，）．次に，k文字スキップするマルコフモデルを用いて「誤り区間」のタイプを判定する手順を述べる．これは、スキップタイプのマルコフ連鎖確率を用いて，「誤り区間」に含まれる正しい文字の位置 h を求め、それらの正しい文字の位置から、「誤り区間」のタイプを決定するものである。

[手順 2] (「誤り区間」のタイプを判定する方法)

(1)「誤り区間」における任意の位置 h において、$P(X_h \mid X_{h-m-k-1}\cdots X_{h-k-1}) > T_t$ を調べる。この条件式を満たす文字 X_h を正しい文字としてみなし、その文字位置 h をすべて列挙する。(2) 上記(1)で求めた「誤り区間」における正しい文字個数およびそれらの位置によって、「誤り区間」のタイプを決定する。

ここで，T_t は正しい文字の候補を選択するために用いられるスキップタイプのn重マルコフ連鎖確率のしきい値を表す．

2.3 スキップタイプのn重マルコフモデルを用いた誤り文字候補のラテイスの構成法

手順2によって判定された「誤り区間」内の全ての誤り文字に対して，次の条件を満たす正しい文字候補を，スキップタイプのm重マルコフモデルを用いて選出する手順を述べる．

[手順 3] (誤り文字に対する正しい文字候補を選出する方法)

次の条件を満たす文字の集合を見つける．

(1) 誤り区間内の全ての誤り文字位置を示すjに対して，$P(X_j \mid X_{j-m}\cdots X_{j-1}) > T_s$ となる文字候補の集合 $\{X_j \mid X_j$は漢字またはかな文字$\}$ を選ぶ。を満たす文字 X_j を正しい候補としてすべて求めれる。

(2) 上記(1)で求めた文字候補を，それぞれ確率値の大きい順に並べ、文字候補の配列を構成する(これを、1つの「誤り区間」に対する「候補ラテイス」と呼ぶ)。

スキップタイプの1重マルコフ連鎖確率を用いて構成される候補ラテイス L の例を図5に示す．

2.4 非スキップタイプのm重マルコフモデルを用いて候補ラテイスから正しい文字列を見つける方法

非スキップタイプのm重マルコフモデルを用いて候補ラテイスから正しい文字列を見つける方法を述べる．

[手順 4] (正しい文字列を見つける方法)

(1)「誤り区間」に対する「候補ラテイス」 L の先頭列 $i = 1$ に対して，各文字候補 j の中から次の条件を満たすものを選ぶ。

$P(X_{i,j} \mid X_{j-m}\cdots X_{j-1}) > T$

(2)「候補ラテイス」 L のすべての列位置 i に対して、上記(1)を繰り返す。

(3) 上記(1)(2)より求められた候補文字列の中で、確率のもっとも大きい文字列を正しい文字列とみなす。

ここで，P は通常のm重マルコフ連鎖確率を示す．

3 実験結果

3.1 実験条件

1. ７０日分の日本語新聞記事の総文節数: 283,963
2. 誤りのタイプ：相互干渉誤り
3. 誤り区間の総数: 1000
4. 誤り区間の平均長：5.7
5. 日本語漢字かな文字のマルコフモデルの次数: 1重 および2重
6. マルコフモデルのスキップタイプ：非スキップタイプとスキップタイプ

3.2 実験結果と議論

[1] 手順1による誤り区間検出の適合率 P と再現率 R の関係

手順1による誤り区間の位置に対する適合率 P と再現率 R の関係を図6に示す．この図から，次の結果がフォントサイズが8ポイントの場合に得られる．
(1) 誤り区間の検出に対する P と R の最大値は，P= 85 % と R = 85 %
(2) この図におけるこれらの最大値を比較すると，手順1によって得られる P と R の積の最大値は，文献(8)で示されている方法によって求められた積の最大値より約40%（R の場合）および約22%（P の場合）大きな値となっている．

この結果から，手順1は文内の相互干渉の誤り文字列を検出に対する P と R の最大値を改善するのに有効であることがわかる．

[2] 手順2と手順3を用いた誤り区間タイプを判定する正解率

手順2を用いた誤り区間のタイプ判定の平均正解率は，約90%である．手順3によって得られる候補ラテイス L に含まれる正しい候補累積正解率を図7に示す．これらの図から，スキップタイプのマルコフモデルは誤り区間タイプを判定するのに有用であること，また誤り区間内の誤り文字の正しい候補を見つけるのに有効である事がわかった．

[3] 手順4によって決定された正しい文字列候補の正解率

手順4によって決定された正しい文字列候補の累積正解率を図8に示す．この図から，２重マルコフモデルを用いた第１位候補に含まれる正解文字列の正解率は80%であり，その値は１重マルコフモデルの場合に比べて20%大きい．

非スキップタイプに加えてスキップタイプのマルコフモデルを用いて，相互干渉誤りを検出・訂正する方法は有効である事が結論づけられる．

[4] 処理時間の比較

２章で述べられた手順を実行するのに必要な処理時間の値 T を図９に示す．これらの手順は、誤り区間に対して正しい文字位置が存在するか否かを判定し、その位置を高い精度で特定化することができるために、誤り区間全体を誤りと判定して訂正を行なう場合に比べて、訂正を行なうために文字候補数が大幅に削減でき、相互干渉誤りを検出し訂正するのに非常に有効である．

文字列候補の全ての組み合わせを解析するのに要する時間 $\bar{T}$ と比べると，T は $\bar{T}$ より数倍小さい事がわかる．例えば，誤り区間の長さが４に等しいとき，T は $\bar{T}$ より６桁小さくなっている．

4 むすび

本論文は，スキップマルコフモデルおよび非スキップモデルを用いて，互いに干渉し合う誤り文字列を検出し訂正する方法を提案した．本手法の効果は，スキップタイプの２重マルコフモデルに対して実験的に評価された．実験結果は次のように要約される．

1. 誤り区間の検出に対する再現率 R と適合率 P は，それぞれ P= 85 % and R = 85 % であった．この値は共に，文献(8)で述べられた方法による値よりも約40%大きな値である．

2. ２重マルコフモデルを用いて得られる第１位候補に含まれる正しい文字列の正解率は，誤り区間の長さ３の場合には80%である．この値は文献(8)の方法による値よりも約20%大きい．

3. 本手法を用いるのに要する処理時間 T は，網羅的な選択的方法の値よりも数桁小さい．

これらの結果から，非スキップタイプに加えて，スキップタイプのマルコフモデルを用いて相互干渉の誤り文字列を検出・訂正する方法が有効であることが分かった．

参考文献

[1] C.E.Shannon, "Prediction and Entropy of Printed English", *Bell System Technical Journal*, Vol.30, pp.50-64, January (1951)

[2] F.Jelinek, "Continuous Speech Recognition by Statistical Methods", *Proc. of the IEEE*, Vol.64, No.4, pp.532-556 (1976)

[3] 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)

[4] 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)

[5] 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)

[6] 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混じり候補の抽出精度”，信学論,J75-D-II,1,pp11-20 (1992)

[7] 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 Markov Models", *Coling-94*, Vol.1, pp187 - 193 (1994)

[8] T.Araki,S.Ikehara,N.Tsukahara and Y.komatsu, "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)

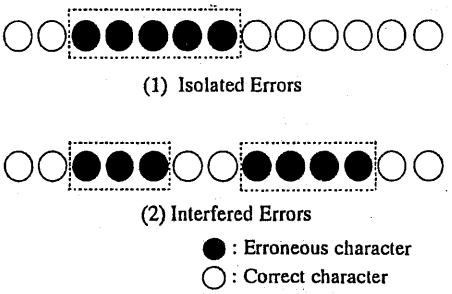

Fig.1 Interference between Erroneous Characters in a Sentence

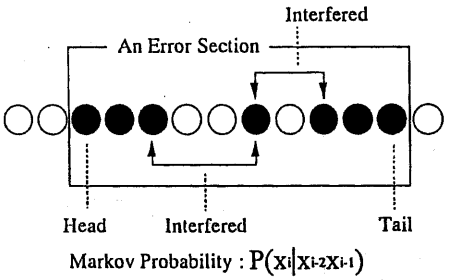

Fig.2 An Example of An Error Section
(The case of 2nd-order Markov Model)

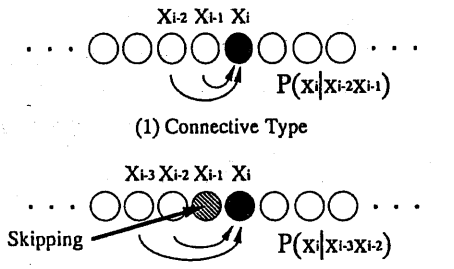

Fig.3 Types of Markov Chain Model
(The case of 2nd-order Markov Model)

	The Lenght of Error Section (n)	Type	FAX-OCR Error 8 point	Accuracy Ratio of Judging the Type
1	n = 3	(A)●○●	40 (17.5%)	100.0%
		(B)●●●	17 (7.5%)	95.9%
2	n = 4	(A)●○○●	32 (14.0%)	100.0%
		(B)●○●●	17 (7.5%)	90.5%
		(C)●●○●	14 (6.1%)	96.7%
		(D)●●●●	8 (3.5%)	77.4%
3	n = 5	(A)●○○●●	9 (3.9%)	98.9%
		(B)●○●○●	6 (2.6%)	97.8%
		(C)●○●●●	1 (0.4%)	80.9%
		(D)●●○○●	8 (3.5%)	98.9%
		(E)●●○●●	7 (3.1%)	85.1%
		(F)●●●○●	2 (0.9%)	91.3%
		(G)●●●●●	0 (0%)	71.3%

* The total number of Error Sections: 228

Fig. 4 Types of the position of Erroneous Character in an Error Section
(The case of 2nd-Order Markov Model)

Sentence : 所得税減税問題に関する首相の発言は次の通り

Error Section

所得稀減俣問題に関する首相の発言は次の通り

Rank	location : 1 Candidates (Cost)	location : 2 Candidates (Cost)	location : 3 Candidates (Cost)
1st rank	な (2.7)	い (2.8)	税 (3.9)
2nd rank	税 (3.4)	れ (3.8)	る (4.1)
3rd rank	る (3.7)	減 (4.0)	の (4.5)
4th rank	ら (3.8)	の (4.3)	な (4.5)
5th rank	の (3.9)	と (4.6)	と (4.6)
6th rank	意 (4.4)	る (5.1)	び (5.1)
7th rank	た (4.5)	伸 (5.2)	い (5.4)
8th rank	に (4.8)	か (5.8)	こ (5.7)
9th rank	を (4.8)	な (5.8)	た (5.8)
10th rank	て (4.8)	く (5.9)	は (5.8)

Fig.5 An Example of Candidates Lattice for an Error Section

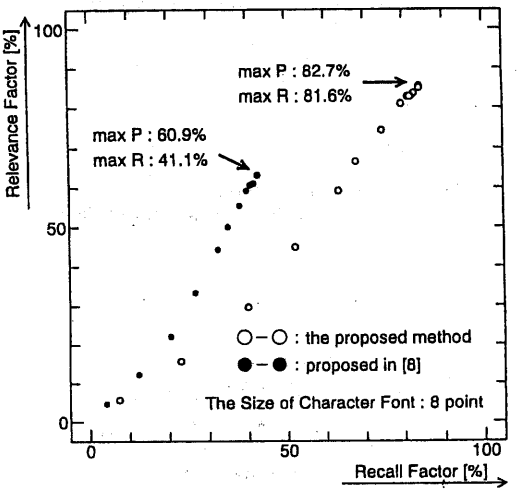

Fig.6 Experimental Results of Detecting the Error Sections
(The case of FAX-OCR Error)

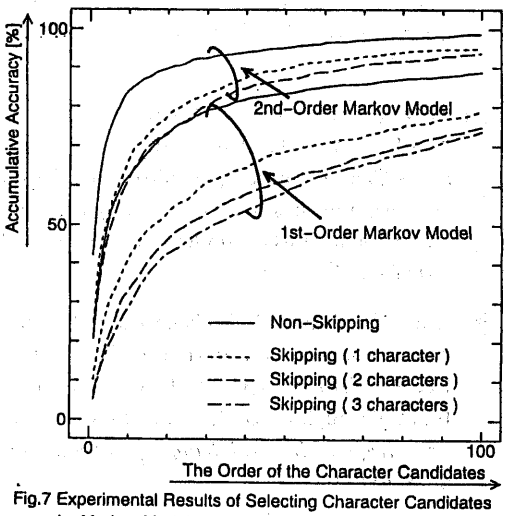

Fig.7 Experimental Results of Selecting Character Candidates by Markov Models

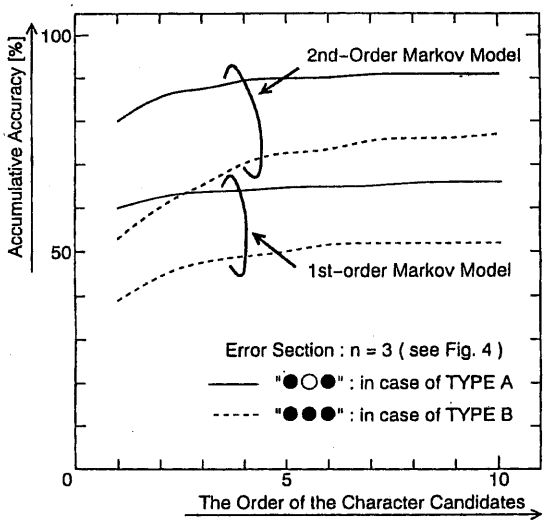

Fig.8 Experimantal Results of Correct String Candidates Determined from Candidates Lattice by Procedure 4

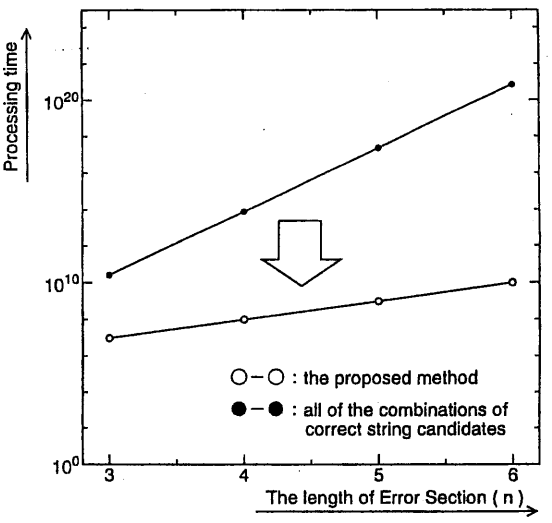

Fig.9 Comparison of the Processing Time required to Detect and Correct Errors