

マルコフモデルを用いたOCRからの誤り文字列の訂正効果

荒木 哲郎[†] 池原 悟^{††} 塚原 信幸^{†††} 小松 康則[†][†]福井大学 ^{††}NTTコミュニケーション科学研究所 ^{†††}NEC

Abstract

21世紀の知的通信サービスの1つにファックスを用いた翻訳通信があり、一般の家庭から利用可能で、将来的に大いに期待される。ファックスによって送信された文書は、光学式文字読み取り装置(OCR)を通して入力されるが、このような文書には、一般に置換誤り、脱落誤りおよび挿入誤りの文字列が含まれる。自然言語処理技術を用いて、これらの誤りを自動検出および訂正する技術が期待されている。従来、日本語漢字かな混じり文に対し、 m 重マルコフ連鎖モデルを用いて、これら3つの誤りタイプの判定および訂正を行う「選択的誤り訂正法」が提案されており、疑似的に設定されたランダム誤りに対し、その有効性が示されている。

本論文では、選択的誤り訂正法を、実際に、ファックスを通して送信された文書をOCRにより読み込む場合に含まれる日本語文の誤り文字列の検出・訂正に適用し、本手法が、ファックスとOCRによる複合誤りの検出および訂正に効果的であることを確認する。

フォントサイズとして8、10、12ポイントの3種類を用いた実験により、次の知見を得た。

1. FAX通信された文書のOCR誤りの特徴として、
 - (a) 置換誤りおよび混合誤りタイプ、誤り位置が先頭および内部、誤り文字列長が1または2、文節内の誤り文字が連続したもの、誤り文字種が漢字であるものが多数を占めること。
 - (b) 文字の大きさに比例して、複雑な誤りタイプが減少すること。
2. 従来のランダム誤りと比較して、FAX-OCR複合誤りの適合率および再現率が低下する理由として、
 - (a) 複数の異なる誤りタイプから構成される混合誤りが存在すること。
 - (b) 文節の先頭および末尾にも誤りが存在すること。
 - (c) 文節内で誤り位置が分離している誤りが存在すること。
 - (d) 文節内の誤り文字列長が3以上の誤りが存在すること。

An Evaluation of a Method to Detect and Correct Erroneous Characters in Japanese input through an OCR using Markov models

Tetsuo ARAKI[†] Satoru IKEHARA^{††}Nobuyuki TSUKAHARA^{†††} Yasunori KOMATSU[†][†]Fukui University^{††}NTT Communication Science Laboratories^{†††}NEC

The communication service of machine translation of documents sent using a FAX is expected to develop an intelligent communication service of the 21st century.

Text that has been input through an optical character reader (OCR) using a FAX usually contain erroneous character strings (wrongly substituted, wrongly deleted and wrongly inserted). The techniques of natural language processing can be used to detect and correct these errors automatically. These "Selective Error Correction Method" to judge these three types of errors, and correct them, using m -th order Markov chain model for Japanese 'kannji-kana' characters, has been proposed and shown to be useful to detect and correct errors generated randomly.

In this paper, the Selective Error Correction Method is applied to detect and correct erroneous characters in Japanese text input through a FAX and an OCR. The method is confirmed to be effective to detect and correct the compound errors introduced by a FAX and an OCR.

1. はじめに

21世紀の知的通信サービスの1つにファックスを用いた翻訳通信があり、一般の家庭から利用可能で、将来的に大いに期待される。ファックスによって送信された文書は、OCRによって読み取られて計算機に入力されるが、このような文書には、一般に誤りの文字列が含まれており、これらの誤りは従来のOCR誤り[1]に加えて、FAX通信による誤りが組み合わされ、FAXとOCRの複合誤りの形態をとる。これらのFAX-OCRの複合誤りを自動的に検出し、訂正するために、自然言語処理技術が期待されている。しかし、現在の自然言語解析の技術は正しい文に対して開発されているため、上記の問題に直接適用することができない。

従来、統計的な手法(マルコフ連鎖モデル)がこのような問題に対して用いられている[2][3]。誤り文字は一般に、置換誤り(正しい文字の代わりに、誤って認識された文字)、挿入誤り、脱落誤り(誤って文字をスキップ)の3タイプに分類されるが、これまでのマルコフ連鎖モデルの適用は、最初の置換タイプの誤り検出および訂正に制限されており、第2、第3の挿入誤り、および脱落誤りの訂正に対しては、提案されていなかった。その理由は、削除誤りと挿入誤りを区別し、それらの誤り位置を検出することが難しいと考えられていたためであった。

最近になって、日本語漢字かな混じり文の文節における置換誤り、脱落誤りおよび挿入誤りの3タイプの誤りを判別し、これらの誤りをm重マルコフ連鎖モデルを用いて訂正する、選択的誤り訂正法が提

案されており、この手法が、ランダムに設定された誤りの検出・訂正に有効であることが示されている[4][5]。

本論文では、FAX-OCR複合誤りを含む日本語の誤り文字列の検出・訂正に、選択的誤り訂正法を適用し、本手法が、FAX-OCR複合誤りの検出および訂正に対して効果的であることを確かめる。はじめに、8、10、12ポイントフォントを使用し、FAX-OCR複合誤りの特徴を解析する。つぎに、FAX-OCR複合誤りの検出および訂正結果から、「適合率P」および「再現率R」が得られ、FAX-OCR複合誤りの特徴とPおよびRの関係について述べる。

2. 基礎的な定義および選択的誤り訂正法

本論文で研究の対象としている、ファックスを用いた翻訳通信サービスの構成図を図1に示す。

2.1 基礎的な定義

日本語文は、文節と呼ばれる単位に分割される。

文書中の日本語文節は「正解文節」と「誤り文節」の二つに分けることができる。日本語正解文節は Γ_C を用いて表す。誤り文節は Γ_E で表し、さらに、(1)置換誤りを含む文節のタイプ(これを Γ_S で表す)(2)脱落誤りを含む文節のタイプ(Γ_D)(3)挿入誤りを含む文節のタイプ(Γ_I)の3つに分類される。誤りの検出および訂正の正解率を、適合率Pおよび再現率Rで表す。ここで、Pは検出あるいは訂正された誤りの全体のうち、 Γ_E の誤りが占める割合を示

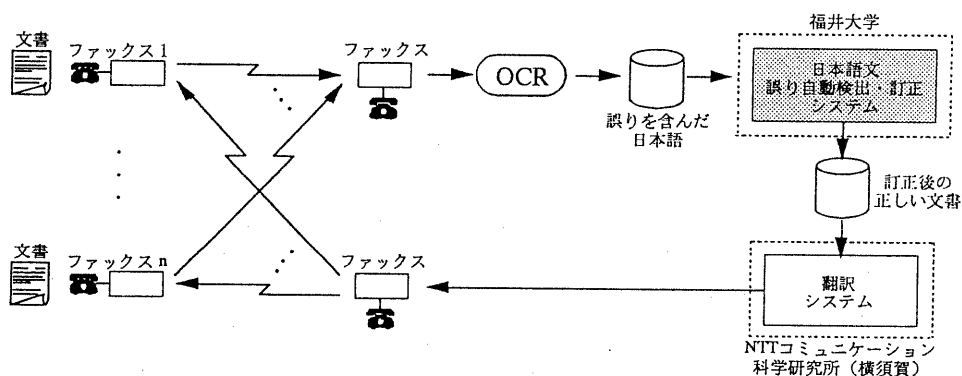


図1. ファックスを用いた翻訳通信サービスの構成図

し、 R は、 Γ_E の要素のうち、どのくらいの誤りが検出あるいは訂正されたかの割合を示す。また、 Γ_S 、 Γ_D および Γ_I に対する適合率を $P_S^{(D)}$ 、 $P_D^{(D)}$ および $P_I^{(D)}$ とそれぞれ表す。

誤りタイプや適合率および再現率の正確な定義は、文献[5]を参照のこと。

2.2 2重マルコフモデルを用いた誤り検出法

経験に従って、次のような仮定を導入する。

【仮定】誤りの音節列または漢字かな混じり文に対する各マルコフ連鎖確率の値は、正しい音節列または漢字かな列に対するマルコフ連鎖確率の値に比べて小さい。

この仮定に従うと、誤り位置 i 及び誤り長 k の誤り文字列の検出の手続きは、次のように定義される。

【手続き 1】(置換誤り ($\Gamma_S^{(k)}$ の要素) 及び挿入誤り ($\Gamma_I^{(k)}$ の要素) の誤り位置と誤り長を検出する法)

次の条件を満たす長さ k の部分列を見つけたとき、この部分列は位置 i に誤って挿入または置換されたものであると判定する。

- (1) $h = i - 1$ または $h = i + k + m$ に対して、 $P(X_h | X_{h-m} \cdots X_{h-1}) > T$ 及び、
- (2) $i \leq j \leq i + k + m - 1$ となるような全ての j に対して、 $P(X_j | X_{j-m} \cdots X_{j-1}) < T$ 、

ここで $P(X_j | X_{j-m} \cdots X_{j-1})$ は、列 $X_{j-m} \cdots X_{j-1}$ が起きた時、文字 X_j が起きる確率を表す m 重マルコフ連鎖確率である。また X_u は、 $u < 0$ のとき、ブランク文字を表す。さらに、 T は誤り検出に用いる m 重マルコフ連鎖確率の足切り値を示す。

この手続きは文節に対する m 重マルコフ連鎖確率位置 i から $i + k + m - 1$ において、ちょうど $(k + m)$ 回足切り値 T より小さい値を取り続けるならば、その位置に k 個の文字列が誤って置換、または挿入されていることを検出する。

【手続き 2】(脱落誤りの列 ($\Gamma_D^{(k)}$ の要素) に対する誤り位置の検出法)

次の条件を満たす長さ k の部分列を求め、それが位置 i に脱落していると判定する。

- (1) $h = i - 1$ または $h = i + k + m$ に対して、 $P(X_h | X_{h-m} \cdots X_{h-1}) > T$ 、及び
- (2) $i \leq j \leq i + m - 1$ となるような全ての j に対して、 $P(X_j | X_{j-m} \cdots X_{j-1}) < T$

ここで、 T は誤り検出に用いられる m 重マルコフ連鎖確率の足切り値を表す。

列に対する m 重マルコフ連鎖確率が、位置 i から $i + m - 1$ でちょうど m 回足切り値 T より小さい値を取り続けるとき、ある文字列がその位置に誤って脱落していると判定される。しかしながら、その位置 i で誤って脱落した文字列の長さはこの手続きによっては決定できず、2. 3 節の手続き 4 によって始めて求められる。

2 重および 3 重マルコフ連鎖モデルの場合に、マルコフ連鎖確率が足切り値 T より小さい値を取り続ける回数の関係を、表 1 に示す。

誤り列は、次の 2 つのクラスに分類される。一つは、誤って置換されたり、誤挿入された文字列のクラスであり、もう一つは誤って脱落された文字列のクラスである。しかしながら、この方法は前者のクラスに対しては、誤って置換された文字列と、誤って挿入された文字列を識別することができない。また、脱落の誤り列に対するマルコフ連鎖確率は長さ k に対してちょうど同じ回数小さい値を取り続けるから、 $\Gamma_D^{(k)}$ のタイプに対しては、長さ k を決定する事ができない。これらの問題は、2. 3 節の手続き 3 及び 4 によって解決される。

2.3 2重マルコフモデルを用いた誤り訂正法

マルコフモデルを用いて、誤り列を正しい列に置き換える手続きが、次のように表される。

【手続き 3】(置換誤り ($\Gamma_S^{(k)}$) または挿入誤り ($\Gamma_I^{(k)}$) の列を訂正する方法)

文節 $\alpha = s_1 s_2 \cdots s_{i-1} \hat{s}_i \cdots s_{i+k-1} s_{i+k}$

または $\alpha = s_1 s_2 \cdots s_{i-1} \hat{s}_i \cdots s_{i+k-1} s_{i+k} \cdots s_m$ が、“ (i, k) -EBWS” または “ (i, k) -EBWI” であり、部分列 $\beta = t_1 t_2 \cdots t_k$ がそれぞれ α の位置に i に誤って置換、または誤って挿入されていると仮定する。そのとき誤り列 α は、条件 (1) が満たされるとき、次の正しい列 γ (Γ_C の要素) に置き換えられる。

表 1: 誤り列のマルコフ連鎖確率が足切り値Tを下回る回数

種別	2重マルコフ	3重マルコフ
$\Gamma_S^{(1)}$	3回	4回
$\Gamma_S^{(2)}$	4回	5回
$\Gamma_S^{(k)}$	(k+2)回	(k+3)回
$\Gamma_D^{(1)}$	2回	3回
$\Gamma_D^{(2)}$	2回	3回
$\Gamma_D^{(k)}$	2回	3回
$\Gamma_I^{(1)}$	3回	4回
$\Gamma_I^{(2)}$	4回	5回
$\Gamma_I^{(k)}$	(k+2)回	(k+3)回

$\gamma = \alpha^{(i)} \parallel \beta \equiv s_1 s_2 \cdots s_{i-1} t_1 t_2 \cdots t_k s_{i+k} \cdots s_m$,
但し、 $t_1 \leftarrow s_i, \dots, t_k \leftarrow s_{i+k-1}$ または
 $\gamma = \alpha^{(i)} \gg \beta \equiv s_1 s_2 \cdots s_{i-1} s_{i+k} \cdots s_m$, 但し
 $t_1 = s_i, \dots, t_k = s_{i+k-1}$

- (1) $i+k \leq j \leq i+k+m-1$ となるような全ての j に対して、
 $P(X_j | X_{j-m} \cdots X_{j-1}) > T$

上記の二つの場合の正しい列に対するマルコフ連鎖確率値を比較し、大きい方のマルコフ連鎖確率を持つ正しい列を選ぶ。

【手続き 4】(脱落誤り ($\Gamma_D^{(k)}$) の列を訂正する方法)

文節 $\alpha = s_1 s_2 \cdots s_{i-1} s_i \cdots s_m$ が、“(i, k)-EBWD”であり、また部分列 $\beta = t_1 t_2 \cdots t_k$ が α の位置 i に誤って脱落されているとする。誤り列 α を条件 (1) が満たされるとき、次のような Γ_C の正しい列 γ に置き換える。

$\gamma = \alpha^{(i)} \ll \beta \equiv s_1 s_2 \cdots s_{i-1} t_1 t_2 \cdots t_k s_i \cdots s_m$

- (1) $i+k \leq j \leq i+k+m-1$ となるような全ての j に対して、
 $P(X_j | X_{j-m} \cdots X_{j-1}) > T$

3. FAX-OCR 複合誤りの検出・訂正実験

本論文では、上記の選択的誤り訂正法を、ファックスを通して入力された文書をOCRで読み込んだときに含まれる誤り（以後FAX-OCR複合誤りと呼ぶ）の検出・訂正に用いる。

3.1 実験条件

- (1) 日本語新聞記事 70 日分の文節数：283,963
- (2) マルコフ連鎖確率辞書：2重の漢字マルコフ連鎖確率辞書
- (3) FAXを通してOCRに入力された文節数：1000
 - (a) 平均文節長（漢字かな混じり文節）：6
 - (b) 文字の大きさ：8,10,12 ポイント

3.2 実験結果

誤り検出および訂正は、足切り値Tに依存し、検出および訂正の「適合率P」と「再現率R」は、Tの値を変化させることによって得られる。

3.2.1 FAX-OCR複合誤りの特徴

従来提案されていた手法を用いた日本語文誤り検出・訂正実験では、誤りをランダムに設定していた。ファックスを通してOCRに入力された文書の誤りは、ランダム誤りとは異なる特徴を持つ。以下に、FAX-OCR複合誤りの特徴について述べる。

FAX-OCR複合誤りにおける、誤りタイプ、誤り位置、誤り文字列長、文節内における誤り文字の連結性および誤り文字種についての、各ポイント同士の比較を表2に示す。

1. FAX-OCR複合誤りでは、次に挙げる誤りの割合が多い。
 - (1) 置換誤りタイプと置換誤り、脱落誤りおよび挿入誤りで構成される混合誤りタイプ
 - (2) 誤り位置は文節の先頭文字および内部
 - (3) 誤り文字列長は1および2
 - (4) 文節内で誤り文字が連続した誤り
 - (5) 誤り文字種では漢字誤り
2. 誤りタイプ、誤り文字列長および誤り文字連結性の、各ポイント同士の比較から、文字の大きさに比例して、複雑なタイプの誤りが減少していることが分かる。

表 2: FAX-OCR複合誤りの特徴

項目	分類	8 p	10 p	12 p
誤り個数	文節数	586	263	72
誤りタイプ	置換誤り	89.8	90.3	86.6
	挿入誤り	0.4	2.8	7.3
	脱落誤り	0.0	0.0	0.0
	混合誤り	9.8	6.9	6.1
誤り位置	先頭	39.8	32.9	36.6
	内部	50.1	55.4	50.0
	末尾	10.1	11.8	13.4
誤り文字列長	1	67.3	76.1	85.4
	2	24.4	21.4	8.5
	3以上	8.3	2.4	6.1
誤り文字連結性	連続型	75.1	90.5	94.9
	分離型	24.9	9.5	5.1
誤り文字種	漢字のみ	69.2	71.9	68.4
	ひらがな	25.8	23.5	23.7
	カタカナ	1.1	2.5	2.6
	その他	3.9	2.1	5.3

単位: %

3.2.2 FAX-OCR複合誤りの検出・訂正結果

FAXおよびOCRを通して入力された文節は、次の二つに分けられる。一つは、ファックス通信の際に誤りが生じず、OCRによって正しく認識された文節（正解文節）であり、もう一つは、誤り文字列を含む文節（誤り文節）である。2節で述べた選択的誤り訂正法を、FAX-OCR複合誤りの検出および訂正に適用した。

ファックスおよびOCRを通して入力された正解文節および誤り文節に対する、選択的誤り訂正法の適用結果から、次のことが言える。

[1] 正解文節に対する適用結果

誤りを含む文節に対するPとRの積が最大となるように決定された足切り値Tを用いて、選択的誤り訂正法を適用した結果、全ての正解文節が正解であると判定された。

[2] 誤りを含む文節に対する適用結果

FAX-OCR複合誤りに対する項目別の検出・訂正におけるPおよびRを図2-(a)～図5-(e)に示

表 3: 全体の適合率・再現率の比較

種別	8 p	10 p	12 p
$P^{(D)}$	74.0	82.0	83.2
$R^{(D)}$	43.8	66.8	64.2
$P^{(C)}$	61.7	66.9	61.0
$R^{(C)}$	24.7	51.2	42.4

単位: %

す。それらの図から、次のことが言える。

- (1) 誤り位置が内部・末尾のものに比べ、先頭の検出・訂正のPおよびRは、20～50%低い。
- (2) 誤り文字列長が長くなるにつれて、検出・訂正のPおよびRの値が低下する。
- (3) 各ポイント同士の比較では、特に顕著な違いは見られない。

続いて、表3に、検出・訂正における、全体のPおよびRを示す。この表から、ランダム誤り[5][6]の場合と比較して、 $P^{(D)}$ が15～25%、 $R^{(D)}$ が15～32%、 $P^{(C)}$ が10～30%、 $R^{(C)}$ が20～45%それぞれ低下していることが分かった。

[3] FAX-OCR複合誤りに対する検出・訂正のPおよびRが低下する原因

PおよびRの値が低下する理由として、ランダムに設定された誤りの場合と比べ、FAX-OCR複合誤りが次の4タイプの誤りを多く含むためと考えられる。

- (1) 置換誤りや脱落誤りや挿入誤りから構成される混合誤り
- (2) 文節の先頭および末尾の誤り
- (3) 文節内で誤り位置が分離している誤り
- (4) 文節内の誤り文字列長が3以上の誤り

4. 結 論

本論文では、従来提案されていた選択的誤り訂正法を、ファックスを通して入力された文書をOCRを用いて読み込んだ場合に含まれる日本語の誤りの検出・訂正に適用した。FAX-OCR複合誤りの検出・訂正結果は、ランダム誤りに比較して劣るが、ランダム誤りの場合と同様にFAX-OCR複合誤りの場合にも、本手法が有効であることが示された。

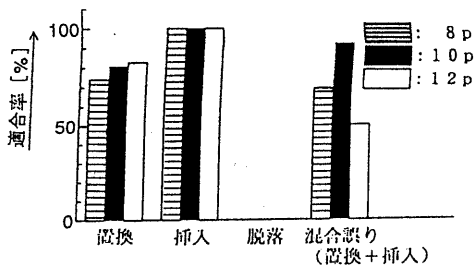


図 2-(a). 検出・適合率の比較 (誤りタイプ)

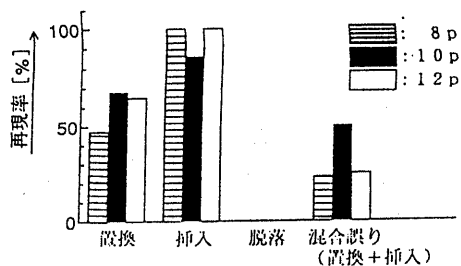


図 3-(a). 検出・再現率の比較 (誤りタイプ)

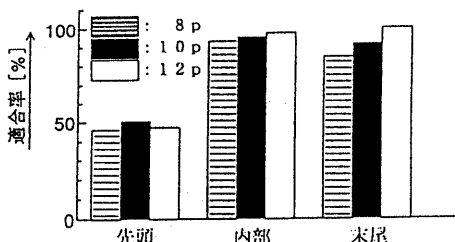


図 2-(b). 検出・適合率の比較 (誤り位置)

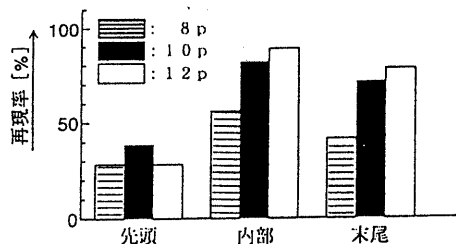


図 3-(b). 検出・再現率の比較 (誤り位置)

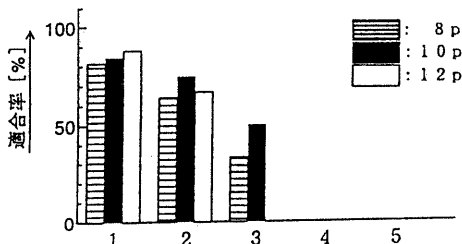


図 2-(c). 検出・適合率の比較 (誤り文字列長)

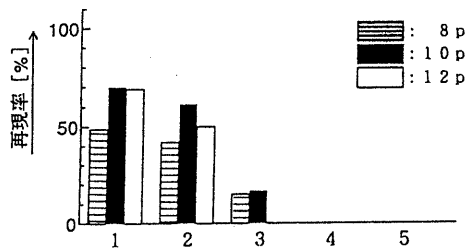


図 3-(c). 検出・再現率の比較 (誤り文字列長)

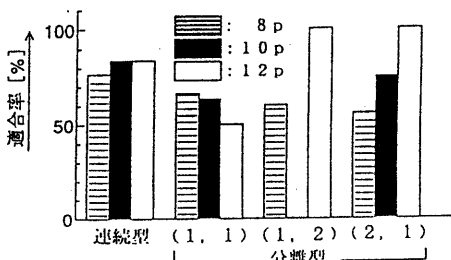


図 2-(d). 検出・適合率の比較 (誤り文字連結性)

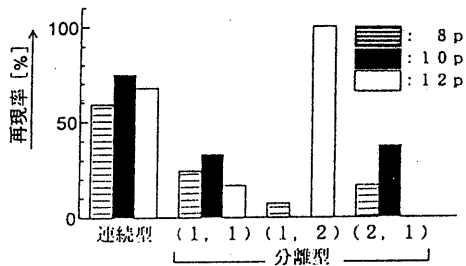


図 3-(d). 検出・再現率の比較 (誤り文字連結性)

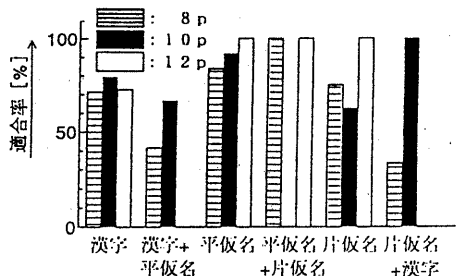


図 2-(e). 検出・適合率の比較 (誤り文字種)

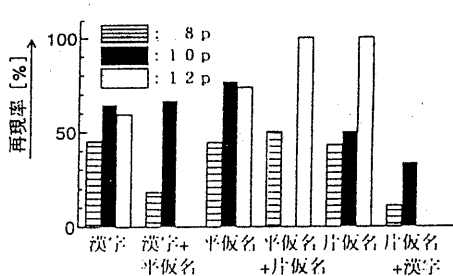


図 3-(e). 検出・再現率の比較 (誤り文字種)

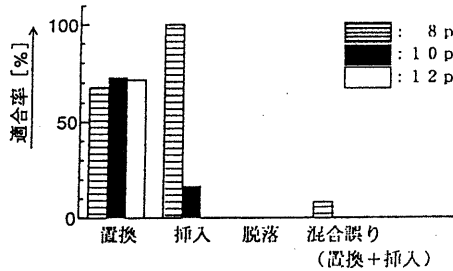


図 4-(a). 訂正・適合率の比較 (誤りタイプ)

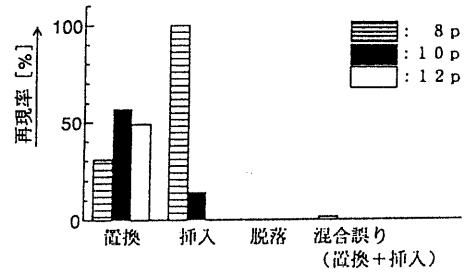


図 5-(a). 訂正・再現率の比較 (誤りタイプ)

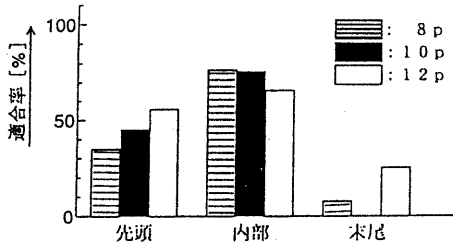


図 4-(b). 訂正・適合率の比較 (誤り位置)

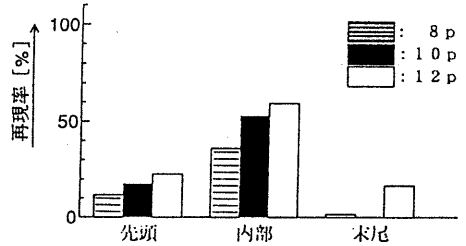


図 5-(b). 訂正・再現率の比較 (誤り位置)

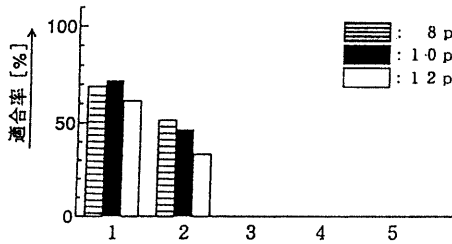


図 4-(c). 訂正・適合率の比較 (誤り文字列長)

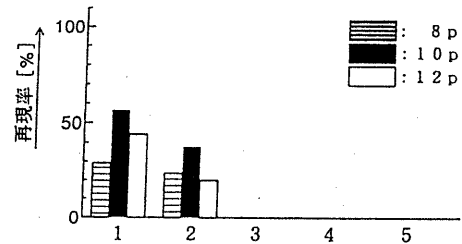


図 5-(c). 訂正・再現率の比較 (誤り文字列長)

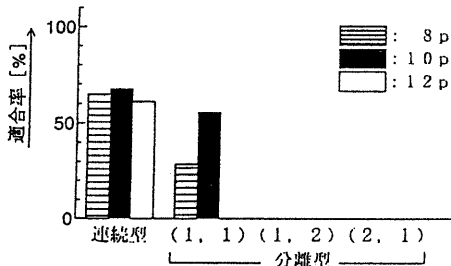


図 4-(d). 訂正・適合率の比較 (誤り文字連結性)

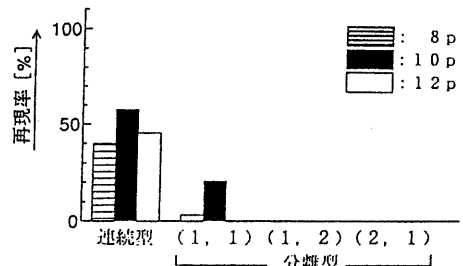


図 5-(d). 訂正・再現率の比較 (誤り文字連結性)

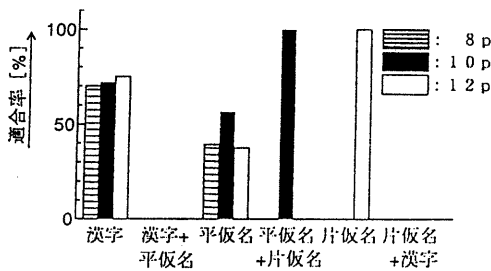


図 4-(e). 訂正・適合率の比較 (誤り文字種)

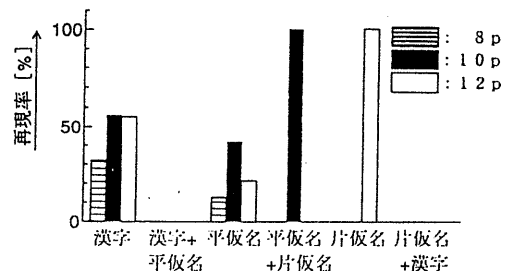


図 5-(e). 訂正・再現率の比較 (誤り文字種)

8、10、12 ポイントの文字サイズを使用したFAX-OCR複合誤りの特徴として、

1. 置換誤りおよび混合誤りタイプ、誤り位置が先頭および内部、誤り文字列長が1または2、文節内の誤り文字が連続したもの、誤り字種が漢字であるものが多数を占め、
2. 文字の大きさに比例して、複雑な誤りタイプが減少する

ことが分かった。

また、FAX-OCR複合誤りの検出および訂正から、「適合率P」および「再現率R」は、ランダムに設定された誤りよりも10-45% 下回ることが示され、PおよびRが低下する要因としては、

1. 複数の異なる誤りタイプから構成される混合誤り
2. 文節の先頭および末尾の誤り
3. 文節内で誤り位置が分離している誤り
4. 文節内の誤り文字列長が3以上の誤り

が考えられる。

今後は、PおよびRの値を減少させている要因を取り除き、選択的誤り訂正法を、FAX-OCR複合誤りの検出および訂正に対してより効果的に適用できるよう、拡張を図っていく予定である。

参考文献

- [1] 西野: “文字認識における自然言語処理”, 情報処理学会論文誌, vol.34, No.10, pp.1274-1280 (1993)
- [2] 荒木、村上、池原: “2重音節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”, 情報処理学会論文誌, 30-4, pp467-477 (1989)
- [3] 村上、荒木、池原: “日本語文節入力に対して2重マルコフ連鎖モデルを用いた漢字かな候補の抽出精度”, 電子情報通信学会論文誌, J75-D-II-1, pp11-20 (1992)
- [4] 荒木、池原、塚原: “べた書き日本語文の脱落・挿入誤りの検出方法”, 情報処理学会自然言語処理研究会, 93-NL-94-7, pp49-54 (1993)
- [5] 荒木、池原、塚原: “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”, 情報処理学会自然言語処理研究会, 93-NL-97-5, pp29-35 (1993)