認識誤りを含む和文テキストにおける全文検索手法

太田 学[†] 高須淳宏[††] 安達 淳[††]

本稿では，OCRによる認識誤りを含む和文テキストに対する3つの確率的な全文検索手法（CMR法，ECMR法，BMR法）を提案する．提案手法は認識誤りの存在を考慮に入れたうえで検索を行うため，OCRによる文書認識後に必要だった手作業による修正編集を行う必要がない．CMR法では検索時に認識誤りを吸収するために，置換誤りの可能性のある文字とその確率を保持した類似文字テーブルを用いる．ECMR法およびBMR法ではそれぞれさらに欠落，挿入，結合，分解誤りの情報を保持した拡張類似文字テーブル，文字のbigram統計に基づいた文字の接続確率を保持したbigramテーブルを用いる．検索時には類似文字テーブルおよび拡張類似文字テーブルを参照することで，1つの入力検索語に対して複数の検索文字列を生成し，それぞれの検索文字列を用いて全文検索を行う．検索された文字列の適否は，OCRの誤りやすさに基づいた確率や文字の接続確率によって判断する．提案手法を用いて検索効率の評価実験を行った結果，test setにおいて認識誤りを考慮しなければ96.01%であった再現率を99.26%に改善できることを示した．

Full-text Search Methods for OCR-recognized Japanese Text with Misrecognized Characters

MANABU OHTA,[†] ATSUHIRO TAKASU[††] and JUN ADACHI[††]

This paper presents three probabilistic full-text search methods, CMR (Confusion Matrix Retrieval), ECMR (Expanded Confusion Matrix Retrieval), and BMR (Bigram Matrix Retrieval), designed for Japanese language documents containing OCR (Optical Character Reader) errors. These methods use search techniques, which assume that errors can exist in the recognized text. Therefore, manual post-editing is not required after OCR scanning. The CMR method uses a confusion matrix which stores all characters likely to be substituted and the respective probability of each substitution. In addition to the single-character confusion matrix, the ECMR and BMR methods also use an expanded confusion matrix for all characters likely to be missed, inserted, combined or decomposed, and a bigram matrix that stores probabilities of character connection, based on bigram statistics. Multiple search terms are generated for an input query term by referring to the confusion matrices, after which a full-text search is tried for each search term. The validity of retrieved terms is determined, based on error-occurrence and character-connection probabilities. The methods' performances were experimentally evaluated by determining retrieval effectiveness. Results indicated that while the recall rate when searching a test set by exact matching was 96.01%, use of the proposed methods improved the rate to 99.26%.

1. はじめに

これまでに印刷文書の形で蓄積された大量の情報を遡及的にデータベース化（以下，DB化）する場合は，OCR（光学的文字読取装置）を用いて文書画像[1]をテキストコードに変換することが普通になっている．また大量の印刷文書を画像で入力し，OCRを使って全文DBを構築する試みが最近行われるようになってきた[2]．その場合OCRによる文字認識の結果にはわずかながら認識誤りが含まれるため，この誤りへの対処が必要不可欠である．この問題に対しては，自然言語処理などのOCRによる文字認識の後処理を強化することで少しでもOCRの認識率を上げようとする試みが多く，認識誤りを少なくするというのが主流であった[3]．しかし認識率を100%にすることは難しく，実用的にはOCRによって得られた結果を効率良く修正するエディタを組み合わせる必要があるが，OCR認識後の人手による修正作業のコストは非常に高い．そこで本稿ではこの認識誤りに対処するために，認識誤

† 東京大学大学院工学系研究科
Graduate School of Engineering, University of Tokyo
†† 学術情報センター研究開発部
Research and Development Department, National Center for Science Information Systems

りを含むことを前提としたテキストの全文検索手法の提案を行い，その検索能力を評価する[4),5)]．提案手法は，OCR認識後のテキストに含まれる認識誤りを訂正するのではなく検索段階で吸収するため，DB化のコストを抑えることができる．

認識誤りを許容する検索手法としては，たとえば類似文字の共通コード化がある[6)]．これはOCRが混同しやすい文字（類似文字）をあらかじめ何らかの方法で共通のテキストコードに変換することによって，検索時に検索文字列と照合するようにする方法で，共通コードに変換するための類似文字のクラスタリングをうまく行うと検索洩れを減らすことができる．しかし類似文字を完全に同一のものと見なすため，検索ノイズの問題が無視できなくなる．また丸川らの提案する複数認識候補型検索手法[7)]は，OCRが出力する複数の認識候補文字（第1位候補文字以外）中に正解文字が含まれている率が高いことを考慮して，これらの第2位以下の認識候補文字を検索に利用している．具体的には，まず不要な検索ノイズを低減するために類似度を用いて各候補文字の絞り込みを行い，この精選した候補文字を認識結果のテキストに持たせることで，検索時にこの複数の候補文字を利用して検索洩れを防いでいる．しかし文字の切り出し誤りに対処していない点や，検索語ではなく検索対象テキストの文字の方に複数の候補を持たせている点が提案手法と異なる．

本稿で新たに提案する3つの全文検索手法を簡単に説明する．

1) Confusion Matrix Retrieval Method（CMR法）
この手法は，OCRの認識誤りのうち置換誤りの可能性のある文字と，その誤りやすさに基づく確率を格納した類似文字テーブル（Confusion Matrix）を利用した全文検索手法である．検索時にこの類似文字テーブルを参照して，認識誤りを含むテキストにおいて検索洩れを起こさないように，1つの入力検索語からOCRの置換誤りを考慮した複数の検索文字列を生成する．生成された複数の検索文字列を用いて単語部分照合も考慮した検索を行い，検索終了後に各生成文字列によって検索された文字列に，OCRの誤る確率を元に計算した得点を与える．その得点が閾値を超えていればその部分を検索結果として出力し，さもなくば棄却する．また類似文字テーブルは，training setとして与えられたOCRの出力した認識誤りを含むテキストと，それに対応する誤りのないテキストを比較することで検索前にあらかじめ作成しておく．

2) Expanded Confusion Matrix Retrieval Method（ECMR法）
この手法はCMR法とほぼ同じアルゴリズムを採用しているが，検索時に置換誤りを扱う類似文字テーブルだけでなく，欠落，挿入，結合，分解誤りを扱う拡張類似文字テーブル（Expanded Confusion Matrix）をも考慮する点が異なる．これによって，認識誤りを含むテキストと誤りのない元のテキストとが1対1に対応しない，文字切り出しの誤りを含むような文字列も検索できる．

3) Bigram Matrix Retrieval Method（BMR法）
この手法もCMR法とほぼ同じアルゴリズムを採用しているが，生成文字列によって検索された文字列の得点を計算する際に，OCRの誤る確率を保持した類似文字テーブルだけでなく文字のbigram統計に基づいた接続確率を保持したbigramテーブル（Bigram Matrix）をも参照する点が異なる．

本稿ではまず2章で，実際に日本語活字OCRの認識誤りを分類しその分類に基づいて類似文字テーブルおよび拡張類似文字テーブルを構築する方法について述べる．3章では，BMR法で用いるbigramテーブルと，そこに蓄えられた統計的に得られる文字の接続確率をどのようにOCRの誤りやすさに基づいた確率と結び付けて得点計算に利用するかを説明する．4章では，2，3章で説明したテーブルを用いた検索アルゴリズム，特に得点付けのアルゴリズムについて述べる．5章では，提案した3つの検索手法を用いて実際に検索実験を行い，検索精度の評価を行う．6章をもってまとめ，今後の課題について述べる．

2. 類似文字テーブルの構築

2.1 OCRの認識誤り

日本語活字OCRの認識誤りを具体的に把握するために予備実験を行った．具体的には学術論文誌から得たテキスト41,854文字について，OCRに認識させるための元の誤りのないテキストと認識結果のテキストの比較を行って認識誤りを以下のように分類した．この実験で使用した学術論文誌は，情報処理学会論文誌1994年度のNo.1～No.5で，そこから抽出した約80 KBのテキストを実験に使用した．このテキストは，各論文の標題，著者名，および内容梗概を含むもので，本予備実験における誤りの統計的性質の評価だけでなく，類似文字テーブルや拡張類似文字テーブルの学習用テキストとしても用いている．また使用したOCRは，(株）リコーのIMAZONE日本語活字OCRで認識率は98.3%であったが，以下の分類は一般のOCR

表1 日本語活字OCRの認識誤り†
Table 1 Types of Japanese-text OCR errors.

誤りの種類	出現頻度	出現頻度（%）
置換（誤字）	586	80.7
欠落（脱字）	10	1.4
挿入	60	8.3
結合	49	6.8
分解	8	1.1
その他	13	1.8

†41,854文字に対する予備実験の結果

についても成り立つ．

- 置換（誤字）
  1文字がOCRによって別の1文字に置換される誤り．
  例：“な” → “は” “ボ” → “ポ” “問” → “間”
- 欠落（脱字）
  1文字以上がOCRによって抜け落ちる誤り．
  例：“E-R” → “ER”
- 挿入
  1文字以上がOCRによって挿入される誤り．
  例：“グラフィックス” → “グラ，フィックス”
- 結合
  2文字以上がOCRの文字の切り出し誤りによって1文字として認識される誤り．
  例：“し,” → “い” “(こ” → “に” “Pr” → “R”
- 分解
  1文字がOCRの文字の切り出し誤りによって2文字以上として認識される誤り．
  例：“利” → “夫U” “指” → “ま旨” “ル” → “Jし”

またその出現頻度を調べた結果を**表1**に示す．ここで“その他”は上記の方法で分類できなかった誤りで，たとえば“Programs”が“P1び匁ams”と認識されるような誤りがこの中には含まれる．なお認識誤りの原因の約4割は英数字（半角文字）の認識に絡むものであったが，これは本研究で使用したOCRが日本語活字対応のものであったことによる．

表1から，OCRの認識誤りのうち文字列の長さが変わらない置換誤りが約81%と最も多く，続いて文字数が増加する挿入，分解誤りが約9%，文字数が減少する欠落，結合誤りが約8%とほぼ同程度の割合で起こることが分かる．また，調べたテキストにおいては，連続した2文字以上の欠落，挿入，3文字以上への分解はなく，3文字以上の結合は4件だけ（しかも英数字）であった．つまりこの実験結果から，置換誤りと1文字の欠落，挿入誤り，2文字の結合，分解誤りを考慮すれば，OCRの認識誤りの約98%をカバーできることが分かる．

B_y= 認識結果の文字 C_y

A_x= 元の文字 C_x	B_1	B_2	B_3	.....
A_1	$P(A_1\|B_1)$	$P(A_1\|B_2)$	0	
A_2	0	$P(A_2\|B_2)$	0	
A_3	0	$P(A_3\|B_2)$	$P(A_3\|B_3)$	
.....				

図1 類似文字テーブル
Fig. 1 Structure/contents of the confusion matrix.

2.2 文字の確信度

使用するすべての文字集合を $\{C_1, C_2, ..., C_{all}\}$ とする．このとき**文字の確信度**とは，正しいテキストにおける文字が C_x である事象を A_x，OCRのテキストにおける文字が C_y である事象を B_y とするとき，OCRのテキストにおける文字 C_y が正しいテキストにおいて文字 C_x である（と確信できる）確率 $P(A_x|B_y)$ のことである．この $P(A_x|B_y)$ は，Bayesの定理から式(1)で計算される．この文字の確信度は，得点付けの際にすべての提案手法で利用する．

$$P(A_x|B_y) = \frac{P(A_x)P(B_y|A_x)}{\sum_{z=1}^{n} P(A_z)P(B_y|A_z)}. \qquad (1)$$

2.3 類似文字テーブル

類似文字テーブルは，OCRの出力した認識誤りを含むテキストとそれに対応する元の誤りのないテキストを比較することで得られるもので，日本語活字OCRにおいて最も頻度の高い置換誤りに対処するために構築する．このテーブルには，学習用テキストに現れた置換誤りの可能性のある文字がすべてその確信度とともに格納されている（**図1**参照）．またCMR法は予備実験の誤り分類（表1）で約8割を占めた置換誤りに対処するための手法であるため，使用するのはこの類似文字テーブルのみである．

2.4 拡張類似文字テーブル

欠落，挿入，結合，分解誤りは，結果としてテキストの文字数を変化させるため，CMR法の単語部分照合によっても検索できない場合が多い．そこでECMR法では，図1に示す類似文字テーブルと同様な構成を持つ欠落文字テーブル，挿入文字テーブル，結合文字

(a) 欠落文字テーブル

	A_1	A_2	A_3	.....
V_m	$P(A_1 \mid V_m)$	$P(A_2 \mid V_m)$	$P(A_3 \mid V_m)$	

V_m = 欠落文字A_xに対応する仮想的な文字

(b) 挿入文字テーブル

	B_1	B_2	B_3	.....
V_i	$P(V_i \mid B_1)$	$P(V_i \mid B_2)$	$P(V_i \mid B_3)$	

V_i = 挿入文字B_yに対応する仮想的な文字

図2 (a) 欠落文字テーブルと (b) 挿入文字テーブル
Fig. 2 Structure/contents of matrices for (a) missing and (b) inserted characters.

(a) 結合文字テーブル

	B_1	B_2	B_3	.....
$a_1 = (A^{i-1}A^i)_1$	$P(a_1 \mid B_1)$	$P(a_1 \mid B_2)$	$P(a_1 \mid B_3)$	
$a_2 = (A^{i-1}A^i)_2$	$P(a_2 \mid B_1)$	$P(a_2 \mid B_2)$	$P(a_2 \mid B_3)$	
$a_3 = (A^{i-1}A^i)_3$	$P(a_3 \mid B_1)$	$P(a_3 \mid B_2)$	$P(a_3 \mid B_3)$	
⋮				

(b) 分解文字テーブル

	$b_1 = (B^{i-1}B^i)_1$	$b_2 = (B^{i-1}B^i)_2$	$b_3 = (B^{i-1}B^i)_3$	.....
A_1	$P(A_1 \mid b_1)$	$P(A_1 \mid b_2)$	$P(A_1 \mid b_3)$	
A_2	$P(A_2 \mid b_1)$	$P(A_2 \mid b_2)$	$P(A_2 \mid b_3)$	
A_3	$P(A_3 \mid b_1)$	$P(A_3 \mid b_2)$	$P(A_3 \mid b_3)$	
⋮				

a_x = 元の文字列 $A^{i-1}A^i$ の x 番目の組合せ
b_y = 認識結果の文字列 $B^{i-1}B^i$ の y 番目の組合せ

図3 (a) 結合文字テーブルと (b) 分解文字テーブル
Fig. 3 Structure/contents of matrices for (a) combined and (b) decomposed characters.

テーブル，分解文字テーブルを作成してこれらの誤りに対処する（**図2**，**図3** 参照）．またこれら4つの文字テーブルを総称して拡張類似文字テーブルと呼ぶことにする．

図2 (a)，(b) はそれぞれ，欠落文字テーブルと挿入文字テーブルを表しており，ここで仮想文字 V_m と V_i はそれぞれ認識結果の文字と元の文字を表す．このような仮想文字を考えることで，これらのテーブルを類似文字テーブルと同様に扱うことができ，ECMR 法でも CMR 法と同じ得点付けのアルゴリズムを用いることが可能になる．また図3 (a)，(b) はそれぞれ結合文字テーブルと分解文字テーブルを表し，同様の理由からここでも $a_x = (A^{i-1}A^i)_x$，$b_y = (B^{i-1}B^i)_y$ という元のテキストと認識結果のテキストにおける2つの連続する文字を定義する．さらに認識結果の文字 B_y が元のテキストにおいて a_x という文字列と見なせる確率（確信度）$P(a_x|B_y)$ が結合文字テーブルには蓄えられており，同様に認識結果の文字列 b_y が元のテキストにおいて A_x という文字と見なせる確信度 $P(A_x|b_y)$ が分解文字テーブルに蓄えられている．

3. BMR 法

3.1 bigram 統計に基づいた文字の接続確率

BMR 法では，類似文字テーブルに保持する OCR の誤り確率とは別の観点からの確信度を導くために，bigram 統計から得られる文字の連接情報（接続確率）を利用している．通常我々が使用している日本語の文字種は約 3,000 種類☆と英語に比べて桁違いに多いため，十分な学習量のテキストデータに対して n-gram 統計（$n \geq 3$）を求めるのは困難である☆☆．そこで文字の連接情報として BMR 法では，比較的頻度統計がとりやすく，自然言語の統計モデルとして様々な研究で利用されている bigram[9] を用いている．

本稿でいう文字の接続確率とは，元のテキスト中の文字 A^i が A^{i-1} の次に続く確率の推定値で式 (2) で求められる．

$$P(A^i|A^{i-1}) = \frac{f(A^{i-1}A^i)}{f(A^{i-1})}. \qquad (2)$$

式 (2) において，$f(A^{i-1}A^i)$ と $f(A^{i-1})$ はそれぞれ文字列 $A^{i-1}A^i$（bigram）の出現頻度と文字 A^{i-1}（unigram）の出現頻度を表す．そしてこのようにして求めた文字の接続確率を**図4** に示す bigram テーブルに保持し検索された文字列の得点付けの際に参照する．

3.2 BMR 法における文字の確信度

認識結果の文字 B_y^i が元のテキストにおいて A_z^i と見なせる確率は，類似文字テーブル（図1）によると

☆ たとえば JIS 第1水準の漢字は 2,965 文字で，これに仮名，数字，アルファベットなどを加えれば一般的な文書に含まれる文字は大体まかなえる．

☆☆ 最近では大容量テキストから n-gram 統計（$n \geq 3$）をとることも可能になりつつある[8]．

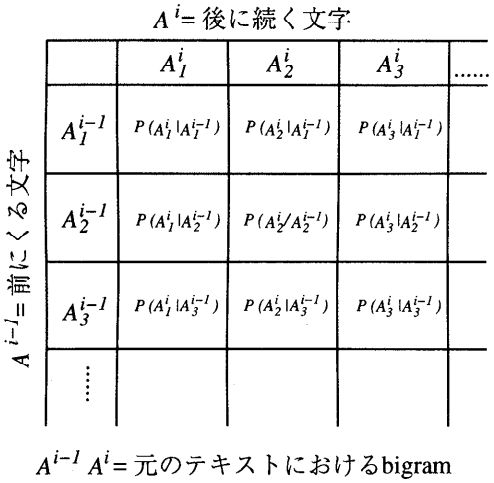

図4 bigram テーブル
Fig. 4 Structure/contents of the bigram matrix.

$P(A_z^i|B_y^i)$ であった．しかしBMR法では認識結果のテキストにおける B_y^i の直前の文字 B^{i-1} に着目して式(3)の確率を計算する．

$$P(A_z^i|B^{i-1}) = \sum_x P(A_z^i|A_x^{i-1})P(A_x^{i-1}|B^{i-1}). \quad (3)$$

つまりこれは認識結果のテキストにおいてその前の文字が B^{i-1} であるとき，次の文字が元のテキストにおいて A_z^i と見なせる確率である．式(3)の右辺の $P(A_z^i|A_x^{i-1})$ がbigramテーブル（図4）を参照することで得られる文字の接続確率で，一方 $P(A_x^{i-1}|B^{i-1})$ が式(1)で定義した，認識結果のテキストにおいて B^{i-1} である文字が元のテキストにおいて A_x^{i-1} と見なせる確率，つまり確信度である．このように類似文字テーブルとbigramテーブルを参照することで，認識結果のテキストにおける直前の文字 B^{i-1} から，正しいテキストにおける次の文字が A_z^i である確率を求めることができる．その結果，類似文字テーブルのみから得られる確信度 $P(A_z^i|B_y^i)$ とは別に，認識結果のテキストにおけるその直前の文字から，$P(A_z^i|B^{i-1})$ という異なる観点からの確率が得られる．

さてここで類似文字テーブルのみによって定まる $P(A_z^i|B_y^i)$ を $m_1(A_z^i)$，式(3)で求めた $P(A_z^i|B^{i-1})$ を $m_2(A_z^i)$ とおく．するとこの2つの確率を統合した新たな確率 $m(A_z^i)$ は，Dempsterの結合規則[10),11)]を用いると式(4)のように計算される．

$$m(A_z^i) = \frac{m_1(A_z^i)m_2(A_z^i)}{1 - \sum_x m_1(A_x^i)(1 - m_2(A_x^i))}. \quad (4)$$

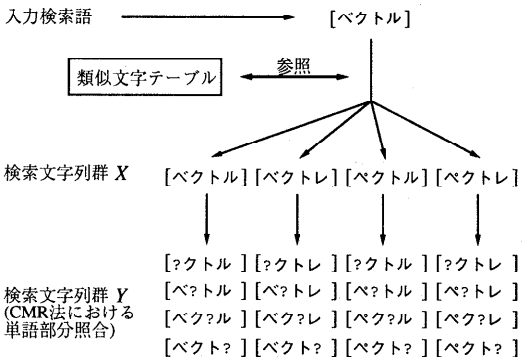

図5 検索文字列群の作成
Fig. 5 Diagram of the search term expansion process.

BMR法では，式(1)で定義した元の確信度 $m_1(A_z^i) = P(A_z^i|B_y^i)$ の代わりに，文字の接続確率も考慮したこの $m(A_z^i)$ を新たな文字の確信度として使用する．

4. 検索アルゴリズム

4.1 検索の手順

本稿でいう検索は全文検索を念頭においており，テキストは原則として1つの長大な和文の文字列として記憶されている．認識結果のテキストにおいてユーザの入力した検索語で検索を行う場合，提案手法はまず類似文字テーブルを参照し，以下の手順で複数の検索文字列を生成する（**図5**参照）．拡張類似文字テーブルを用いた場合も同様であるのでここでは類似文字テーブルによる検索文字列の生成のみを示す．

(1) ユーザから検索語が入力されると，類似文字テーブルを参照してその入力文字列中の各文字について，置換誤りの可能性のあるすべての文字を置き換えて新たに検索文字列群 X を作成する．図5では入力検索語は“ベクトル”で，類似文字テーブルを参照した結果，“べ”→“ぺ”，“ル”→“レ”という置換を行っている．

(2) CMR法においては，検索文字列群 X の各検索文字列中の1文字を前から順にワイルドカード（長さ1）に置き換えた単語部分照合用の検索文字列群 Y を作成する．これはCMR法が置換誤りにしか対応していないためで，単語部分照合を用いるとそれ以外の誤りにもある程度対処できる．

(3) CMR法では検索文字列群 X と Y の，ECMR法およびBMR法では検索文字列群 X のすべ

ての検索文字列を用いて全文検索を行う．検索された各文字列には次節で詳説する得点が付けられており，閾値を用いてそれらの文字列の適否を判断する．

4.2 文字列の確信度と部分確信度

検索された文字列の得点として，CMR 法は文字列の確信度と部分確信度の 2 つを用い，ECMR 法と BMR 法は文字列の確信度のみを用いる．

4.2.1 文字列の確信度

CMR 法では，認識結果のテキストを検索して得られる文字列が正しい確率を，文字列中の各文字が正しい確率の積と仮定する．つまり，認識結果の文字列 $B^{012\ldots n}$ が元の文字列 $A^{012\ldots n}$ に対応する確率を，

$$P(A^{012\ldots n}|B^{012\ldots n}) = P(A^0|B^0)P(A^1|B^1)\ldots P(A^n|B^n) \quad (5)$$

で表す．ここで $P(A^{012\ldots n}|B^{012\ldots n})$ のことを**文字列の確信度**と定義する．

BMR 法でも文字列の確信度は式 (5) で表されるが，式 (4) で定義した $m(A^i)$ を式 (1) で定義した $P(A^i|B^i)$ の代わりに用いて計算する．

しかし ECMR 法では，文字列の確信度の計算方法が若干異なる．これは拡張類似文字テーブルを用いて入力検索語から複数の検索文字列を生成すると，入力文字列と生成文字列との間に 1 対 1 の文字の対応関係が成り立たなくなるためである．よって ECMR 法では文字列の確信度は以下のようにして算出する（**図 6** 参照）．

- 欠落誤りを含む場合
  欠落文字 A_x に対応する仮想的な文字 V_m を認識結果のテキストにおいて考え，欠落文字の確信度 P_m を次式で与える．
  $$P_m = P(A_x|V_m) \cdot P_{m_0}. \quad (6)$$
  式 (6) において $P(A_x|V_m)$ は図 2 に示した欠落文字テーブルより得られる．また P_{m_0} は，欠落の起こる確率で，表 1 から式 (7) で求められる．
  $$P_{m_0} = \frac{10}{41,854} \doteqdot 0.00024. \quad (7)$$
  実際には認識結果のテキストにおける文字 V_m は仮想的なもので観測できないため，P_{m_0} は必ず考慮しなければならない．文字列の確信度の算出は，この欠落文字の確信度を考慮したうえで式 (5) と同様に算出する．
- 挿入誤りを含む場合
  認識結果のテキストに挿入された文字に対応する仮想的な文字 V_i を元のテキストにおいて考え，図 2 に示した挿入文字テーブルから挿入文字の確信度を得る．式 (5) による文字列の確信度の算出の際にその挿入文字の確信度も余分に掛け合わせる．
- 結合，分解誤りを含む場合
  それぞれ図 3 に示した結合文字テーブルと分解文字テーブルを参照して結合文字の確信度，分解文字の確信度を得る．文字列の確信度は，結合文字あるいは分解文字の確信度も置換文字のそれと同様に扱い式 (5) によって算出する．

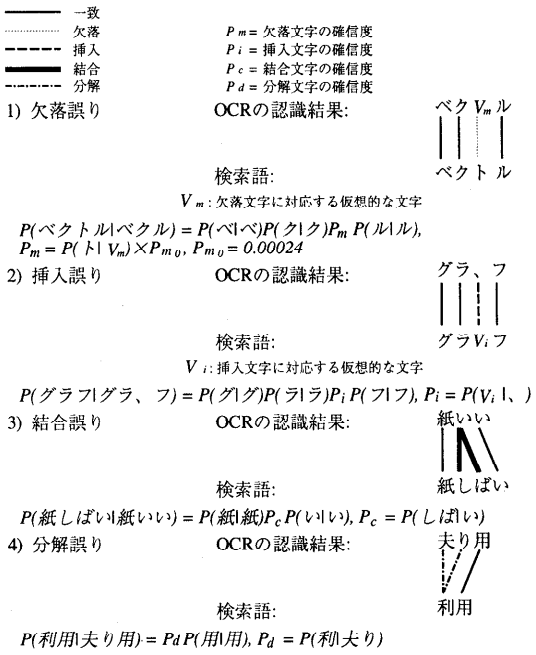

図 6 ECMR 法における文字列の確信度の計算
Fig. 6 Diagram showing how the conviction degree is calculated in the ECMR method.

4.2.2 文字列の部分確信度

CMR 法で用いる単語部分照合は，n 文字の検索文字列のうち任意の $(n-1)$ 文字が完全に一致すれば成り立つというものである．つまり，CMR 法が図 5 の検索文字列群 X と Y の文字列を生成して検索するとき，検索対象となる認識結果のテキストにおいて，検索文字列群 X の文字列には照合しないがワイルドカードのために検索文字列群 Y の文字列には照合する場合に成り立つ．しかしこの場合式 (5) に基づくとこの文字列の確信度 $P(A^{012\ldots n}|B^{012\ldots n})$ は 0 となるため，**文字列の部分確信度** $PCD(A^{012\ldots n}|B^{012\ldots n})$ を式 (8) のように定義する．

$$P(A^k|B^k) = \min\{P(A^0|B^0),...,P(A^n|B^n)\} \text{ のとき,}$$
$$PCD(A^{012...n}|B^{012...n}) = P(A^0|B^0)... P(A^{k-1}|B^{k-1})P(A^{k+1}|B^{k+1})... P(A^n|B^n). \quad (8)$$

換言すれば部分確信度は，文字列中の各文字の確信度を比較してその中で最低のものを除いて掛け合わせた文字列の確信度であるといえる．ゆえに式(5)の確信度が0となる生成文字列によって検索された文字列の適否の判断に利用できる．

5. 実　　験

5.1 実験条件

提案した3つの手法の検索効率について評価実験を行った．そのときの条件を以下に示す．

(1) 使用したOCRは，(株) リコーのIMAZONE日本語活字OCRで認識率は98.3%であった．

(2) 類似文字テーブルおよび拡張類似文字テーブルの構築には，training setとして1994年度の情報処理学会論文誌No.1～No.5から得たテキスト（約80KB）を用いた．

(3) bigramテーブルを構築するために，情報分野の学術論文雑誌から得たテキストデータ約500万文字（約8MB）に対してbigramおよびunigramの頻度統計をとり，3.1節の式(2)を用いて文字の接続確率を計算した．統計用データの内訳は，
- 情報処理学会論文誌1994年1年分の全文データ（training setを含む）
- 電子情報通信学会論文誌1993年1年分のC1～D2分冊の標題，著者名，内容梗概からなるテキスト
- 人工知能学会誌1986～1995年10年分の標題，著者名，内容梗概からなるテキスト

の3つである．また文字の接続確率が0になったものについては，1.0×10^{-5}に底上げすることで補間[12]を行った．

(4) 全文検索の対象としたテキストは，類似文字テーブルや拡張類似文字テーブルの作成に用いたtraining setとそれに含まれないtest setである．またtest setは，1994年度の電子情報通信学会論文誌Vol.J77-A No.1～No.3の標題，著者名，内容梗概からなるテキスト（約55KB）である．

(5) 検索語としては，検索対象テキストに含まれる文字列で，名詞のものを無作為に50☆抽出して用いた．検索効率の値は，それぞれその50単語で検索した結果の平均値である．

(6) 本稿でいう検索効率とは全文文字列検索における再現率・適合率のことでそれぞれ検索洩れの少なさおよび検索ノイズの少なさを表し，式(9)および式(10)を用いて計算される．

(7) CMR法によって検索された文字列にはすべて，得点として確信度Pおよび部分確信度PCDが付けられており，ECMR法およびBMR法によって検索された文字列にはそれぞれ確信度P_EおよびP_Bが付けられている．検索された文字列のうち，この得点がある閾値を超えたものだけが検索結果となる．

(8) CMR法では各文字列に2種類の得点が付与されているので，以下の3つの検索条件を用意し，それぞれCMR法I，II，IIIと命名する．
- (a) CMR法I：$P \geq$ 閾値
- (b) CMR法II：$PCD \geq$ 閾値
- (c) CMR法III：$(P > 0) \cap (PCD \geq$ 閾値$)$

一方ECMR法とBMR法ではそれぞれ，$P_E \geq$ 閾値，$P_B \geq$ 閾値 という1つの検索条件を用いる．

$$\text{再現率} = \frac{\text{提案手法で検索された正解文字列数}}{\text{元のテキストで検索された文字列数}}. \quad (9)$$

$$\text{適合率} = \frac{\text{提案手法で検索された正解文字列数}}{\text{提案手法で検索された全文字列数}}. \quad (10)$$

5.2 実験結果

5.2.1 確信度 P，P_E，P_B の閾値と検索効率

まず確信度P（CMR法I）の閾値と検索効率の関係を調べるために，閾値を0から1まで0.1刻みで変えながらtraining setおよびtest setにおいて検索実験を行った．**図7**にその結果を示す．図7から，
- 閾値が0.1付近において再現率・適合率ともに高くなる．
- training setの検索効率の方がtest setのそれよりも全般的に高い．

ことが分かる．またP_E（ECMR法），P_B（BMR法）の閾値と検索効率の関係を調べる実験でもほぼ同様な結果が得られた．

5.2.2 部分確信度 PCD の閾値と検索効率

同様の実験をPCD（CMR法II）の閾値と検索効率の関係を調べるためにも行い，**図8**に示す結果を得

☆ 検索語の具体例は，“幾何図形”，“効率”，“アルゴリズム”などである．

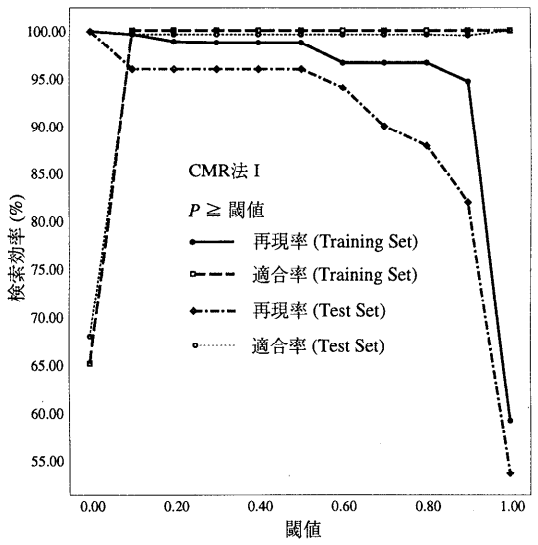

図7 P の閾値と検索効率（CMR 法 I）
Fig. 7 Effect of threshold of P on retrieval effectiveness (CMR method I).

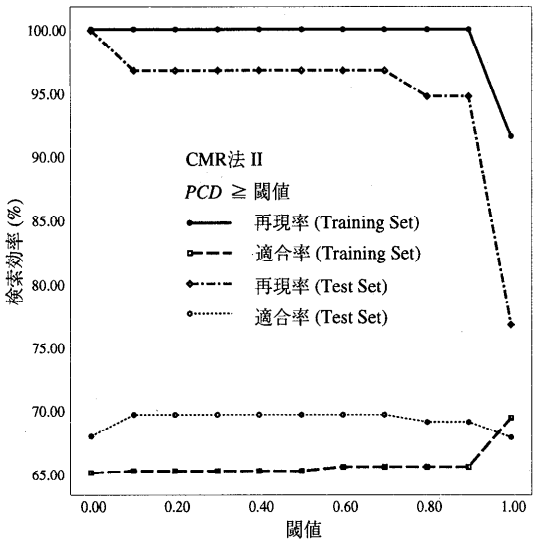

図8 PCD の閾値と検索効率（CMR 法 II）
Fig. 8 Effect of threshold of PCD on retrieval effectiveness (CMR method II).

た．PCD による評価ではワイルドカードによる単語部分照合を許容するため再現率は高いが適合率は全般的に低い．

5.2.3 P と PCD を組み合わせた評価

再現率・適合率ともに高い値を保つために，$(P > 0) \cap (PCD \geq$ 閾値$)$ という条件（CMR 法 III）を用いて評価した結果を**図 9** に示す．これは，検索文字列中の各文字が類似文字テーブルに存在し（$P > 0$），かつ PCD が閾値を超えるもの（適合率を高い値で維持するため）を検索結果として出力したものである．

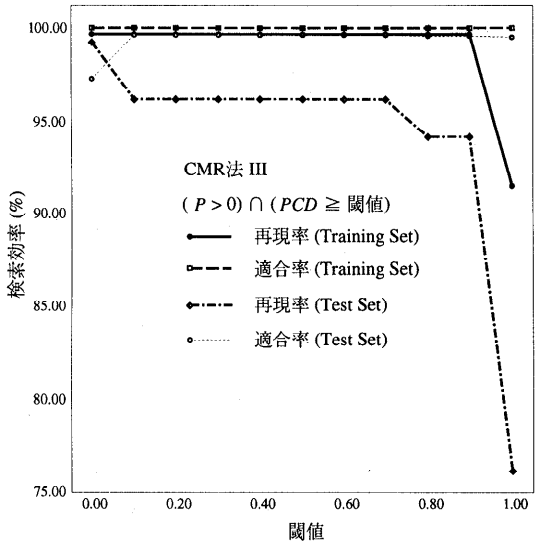

図9 $(P > 0) \cap (PCD \geq$ 閾値$)$ の検索効率（CMR 法 III）
Fig. 9 Retrieval effectiveness vs. $(P > 0) \cap (PCD \geq$ threshold) (CMR method III).

5.2.4 最適閾値における検索効率

training set と test set それぞれの最適閾値における検索効率を**表 2** および**表 3** に示す．ここで最適閾値は，

(1) 再現率が最大となる（ただし適合率が著しく悪くなるため 閾値 $= 0$ は除く）．
(2) 再現率に変化がない場合は，適合率が最大となる．
(3) 再現率，適合率ともに変化がない場合は，閾値自体が最大となる．

という 3 つの条件を (1) から順に適用して定めた．

検索対象テキストである training set および test set を完全照合で検索した場合，再現率はそれぞれ 96.62%（表 2），96.01%（表 3）であった．これに対して CMR 法 I, III を用いると，training set では 99.69%の再現率（表 2），test set では 99.26%の再現率（表 3）を実現している．結局，test set においては若干適合率を下げる（99.60%から 99.28%）ものの，どちらのテキストにおいても完全照合と比べて再現率を約 3%改善できることが分かる．また CMR 法 II は適合率が CMR 法 I および III に比べて著しく悪く，CMR 法 I と CMR 法 III では達成している検索効率が等しいので，以後 CMR 法とは CMR 法 I を指すものとする．

ECMR 法は training set に対しては適合率を下げる

表2 training set における検索効率
Table 2 Retrieval effectiveness when searching training set.

検索条件	再現率（%）	適合率（%）
完全照合	96.62	100.0
$P \geq 0.01$ (CMR 法 I)	99.69	100.0
$PCD \geq 0.9$ (CMR 法 II)	99.95	65.64
$(P > 0) \cap (PCD \geq 0.9)$ (CMR 法 III)	99.69	100.0
$P_E \geq 0.01$ (ECMR 法)	100.0	100.0
$P_B \geq 0.01$ (BMR 法)	99.69	100.0

表3 test set における検索効率
Table 3 Retrieval effectiveness when searching test set.

検索条件	再現率（%）	適合率（%）
完全照合	96.01	99.60
$P \geq 0.00001$ (CMR 法 I)	99.26	99.28
$PCD \geq 0.001$ (CMR 法 II)	99.91	68.84
$(P > 0) \cap (PCD \geq 0.001)$ (CMR 法 III)	99.26	99.28
$P_E \geq 0.00001$ (ECMR 法)	99.26	99.28
$P_B \geq 0.01$ (BMR 法)	99.26	99.28

ことなく再現率を完全に回復することができた（表2）．これは，拡張類似文字テーブル（ECMR 法）によって類似文字テーブルだけ（CMR 法）ではカバーできなかった認識誤りにも対処できたことを示している．一方 test set においては CMR 法と比較して検索効率の差が見られないが，これは拡張類似文字テーブルにも含まれない認識誤りが原因であった．具体的には次のような認識誤りが存在したためである．

(1) “比較” → “上ヒ薮”
“比” → “上ヒ” という分解誤りが拡張類似文字テーブルに存在せず，“較” → “薮” という置換誤りも類似文字テーブルに存在しなかったのでまったく検索することができなかった．

(2) “非線形” → “ヲ E 線形”
“非” → “ヲ E” という分解誤りが拡張類似文字テーブルに存在しなかったため確信度 $P_E = 0$ となってしまい，検索結果として不適と判断された．

これらはどちらもテーブルを構築する際の学習用テキストに存在しなかった誤りであり，当然のことながらテーブルの構築には十分な量の学習用テキストが必要であることがいえる．

BMR 法を CMR 法と比較すると，最適閾値における検索効率に差はなく（表2，表3），適合率にもほとんど差が見られなかったが，再現率については training set および test set においてそれぞれ**図10，図11** に示す結果が得られている．図10では，BMR 法の方が全般的に高い再現率を実現しているが，図11ではそうとはいえない．しかし図11においても，最適閾値付近においては BMR 法の方が高い再現率を実現している．以上の結果から，BMR 法は文字の接続確率を考慮することで検索されるべき文字列の確信度を引き上げることに効果があることが分かる．つまり CMR 法の与える P よりも BMR 法の与える P_B の方がより得点として好ましいといえる．ただし，統計という性質上常に BMR 法が CMR 法よりもすぐれた得点付けを行っている訳ではなく，中には文字の接続確率を考慮することでかえって確信度が改悪され，再現率が下がる場合もあることに注意する必要がある（図11）．また図10，図11の結果は，検索対象となるテキストの分野にふさわしく，かつ十分な量の統計用学習データを用いることの重要性も示唆している．

最後に BMR 法と ECMR 法を比較すると，以下に示す利害得失があげられる．

- 達成できる再現率の点では ECMR 法の方が優れている（表2参照）．
  これが4種類の文字の切り出し誤り（欠落，挿入，結合，分解誤り）それぞれに対応して検索文字列を生成する ECMR 法の長所である．
- ECMR 法との比較においても，BMR 法による検索されるべき文字列の確信度の引き上げは有効である．
  CMR 法によって検索できる文字列の集合は，ECMR 法によるそれのサブセットとなるが両集合の差は大きくない．これは日本語活字 OCR の認識誤りでは置換誤りがその約8割を占めるためで，実験結果にもその差は trainig set にこそ 0.31%の再現率の差（表2）として現れているが，test set においては見られない（表3）．また，ECMR 法について P_E と閾値の関係を調べると図7とほぼ

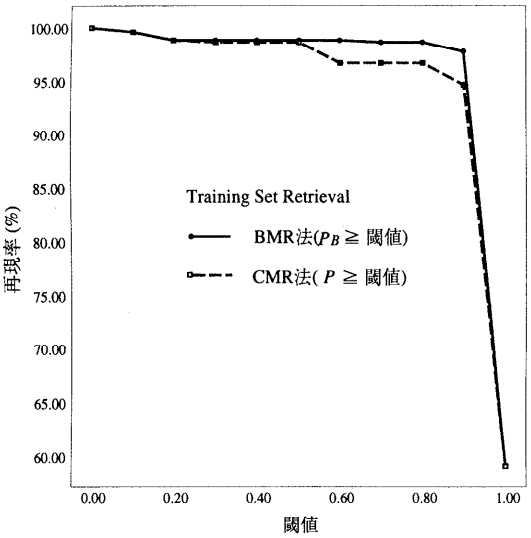

図10 training set における CMR 法と BMR 法の再現率
Fig. 10 Recall rates of the CMR and BMR methods when searching training set.

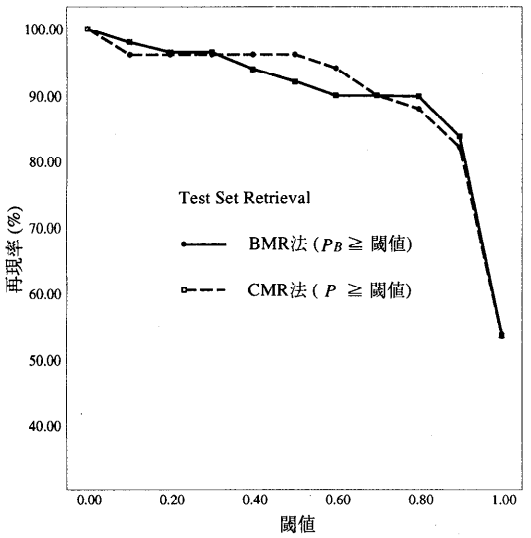

図11 test set における CMR 法と BMR 法の再現率
Fig. 11 Recall rates of the CMR and BMR methods when searching test set.

同様な結果が得られる．よって BMR 法と CMR 法の比較の際とほぼ同様な議論が ECMR 法との比較においても成立つ．

よって実用に供する際には BMR 法と ECMR 法それぞれの長所を活かした，両者を統合した手法が望まれ，そのためには BMR 法において利用する文字の接続確率を，何らかの方法で文字切り出し誤りも考慮に入れたものに拡張する必要がある．

6. おわりに

本稿では，OCR の出力した認識誤りを含むテキストを有効に活用するという観点から，認識誤りを含むテキストにおける検索手法を 3 つ提案し，検索精度の評価を行った．

これまでの文字認識に関する研究，開発は文字の認識率向上に主眼が置かれてきたが，最近では単に文字認識精度の向上だけでなく，文字認識をどのように利用するかという後処理を含めたシステムとして考える動きがある．その一例として，大量の印刷文書を画像で入力し，OCR を使って全文 DB を構築する試みが最近行われるようになってきており，本稿ではこの場合問題になる OCR の認識誤りに対処するため，この誤りを訂正するのではなく検索段階で吸収するための手法を提案した．

いずれの提案手法も OCR の認識誤りを学習したテーブルを検索時に参照することで検索洩れを減らすとともに，検索された各文字列に確率に基づく得点を与え，その得点を閾値によって評価することで検索ノイズの低減を図るものであった．またその得点付けでは，OCR による文字認識においてその誤りやすさを考慮した確率，および統計的に得られる文字の接続確率の 2 つを用いた．実験結果から，どの手法も検索効率の改善に有効であり，特に ECMR 法では training set において再現率，適合率ともに 100%を達成できた．しかし training set に存在しなかった誤りのため，test set における再現率は 100%とはならなかったことから，十分な量のテキストを類似文字テーブルおよび拡張類似文字テーブルの学習に用いなければならないことも示唆された．

今後の課題としては，文字の切り出し誤りを扱う ECMR 法と文字の接続確率を扱う BMR 法を統合することがあげられる．また実験で用いた training set のための検索語の各文字は，類似文字テーブルを参照することで平均 1.26 文字に拡張されており，これはたとえば CMR 法では長さ n の 1 つの検索語から平均 1.26^n の検索文字列が生成されることを意味している．仮に $n = 10$ とするとこの検索語からは $1.26^{10} \approx 10$ の検索文字列が生成されるため，拡張されない場合のおよそ 10 倍のコストがかかることになる．そのため既存の高速検索エンジンを使った実装では，この問題を現実的に回避することが必要となるが，これは次に取り組むべき課題としている．

参考文献

1) 飯沢篤志：文書画像データベースシステム，情報処理，Vol.33, No.5, pp.497–504 (1992).
2) Fujisawa, H. and Marukawa, K.: Full-Text Search and Document Recognition of Japanese Text, *Proc. 4th Symp. DA & IR*, Las Vegas, Nevada, pp.55–80 (1995).
3) 西野文人：文字認識における自然言語処理，情報処理，Vol.34, No.10, pp.1274–1280 (1993).
4) 太田 学，片山紀生，高須淳宏，安達 淳：統計的手法による文字誤りテキスト検索，第52回情報処理学会全国大会論文集 (4)，5P-10, pp.211–212 (1996).
5) Ohta, M., Takasu, A. and Adachi, J.: Probabilistic Retrieval Methods for Text with Miss-Recognized OCR Characters, *Proc. Workshop on Information Retrieval with Oriental Languages*, Taejon, Korea, pp.35–41 (1996).
6) Myka, A. and Güntzer, U.: Fuzzy Full-Text Searches in OCR Databases, *Advances in Digital Libraries (Preliminary Version)*, Chapter7, pp.87–100, Springer-Verlag, New York (1995).
7) 丸川勝美，藤澤浩道，嶋 好博：文書認識と全文検索の融合技術に関する実験的検討，情報処理学会研究報告，95-FI-39, pp.65–72 (1995).
8) 長尾 眞，森 信介：大規模日本語テキストのnグラム統計の作り方と語句の自動抽出，情報処理学会研究報告，93-NL-96, pp.1–8 (1993).
9) Jardino, M.: Multilingual Stochastic N-gram Class Languge models, *Proc. ICASSP'96*, Vol.1, pp.161–163 (1996).
10) 石塚 満：Dempster と Shafer の確率理論，電子通信学会誌，Vol.66, No.9, pp.900–903 (1983).
11) 小林邦勝，鈴木伸明，根元義章，佐藤利三郎：Dempster と Shafer の確率理論に基づく情報量に関する一考察，電子通信学会論文誌，Vol.J68-A, No.8, pp.741–747 (1985).
12) 森 大毅，阿曽弘具，牧野正三：2重マルコフモデルを用いた日本語文書認識後処理，情報処理学会研究報告，94-NL-102, pp.89–96 (1994).

(平成9年 4 月 30 日受付)
(平成9年 12 月 1 日採録)

太田 学（学生会員）

1971年生．1994年東京大学工学部電気工学科卒業．1996年同大学院工学系研究科電気工学専攻修士課程修了．現在，同大学院博士課程に在学中．電子図書館における検索システムの開発研究に従事．

高須 淳宏（正会員）

1984年東京大学工学部航空学科卒業．1989年同大学院工学系研究科博士課程修了．工学博士．同年，学術情報センター研究開発部助手．1993年より同センター助教授．データベースシステム，文書画像理解，機械学習の研究に従事．電子情報通信学会，人工知能学会，ACM各会員．

安達 淳（正会員）

1976年東京大学工学部電気工学科卒業．1981年同大学院博士課程修了．工学博士．同年同大学大型計算機センター助手．1983年同大学文献情報センター講師および助教授．1986年より学術情報センター研究開発部助教授を経て，現在教授．オンライン情報システム，分散処理システム，電子図書館の開発研究に従事．電子情報通信学会，IEEE，ACM，各会員．