old

修士学位論文本審査資料

多フォント文書認識に関する基礎的研究

東北大学大学院工学研究科

情報工学専攻

池田 啓明

目 次

図目次

図 目 次

表目次

第 1 章

序論

1.1 本研究の背景及び目的

各種の文字認識システムの研究は30年余り昔から始まり、十数年におよぶ基礎研究が積まれた後に、文字読取り装置OCR(Optical charactor Reader)が開発された。開発当初は、読取り可能な字種は数字10種程度が限度で、郵便番号の自動読取り区分などにしか実用化できなかった。またOCRの価格も非常に高価で普及には至らなかった。しかし半導体技術の発達、とりわけLSI技術を基盤としたマイクロプロセッサの発明で、廉価なOCRが登場、以来OCRは急速に普及していき、それにともない、認識技術も高度化し、現在では数千種にのぼる手書き漢字をも認識の対象とできるようにまで到達した。

さらに、近年ではOA化により、ワープロやファックス、コピー機等の事務機器が産業から家庭にまで広がっており、大量の文書が出回っている。このような大量の文書を保存するために文書そのものではなく、データとして保存するというデータベースは、現在でも百科事典などで実用化されているが、さらに幅広い分野にわたって普及と発達が見込まれている。紙面上の文字として印字してある情報を、この様なデータベースに登録するにはコンピュータに入力しなければならない。この入力の際に、人間がキーボードから入力するのではなく、OCRを用いて自動的に入力させることできれば作業の効率化をはかれる。今後の情報化社会においてOCRを用いた文書入力は、人間とコンピュータとを結ぶマン・マシーンインターフェースとして、今まで以上に重要な役割を担うことが期待されている。[1][2][3]

しかし、現状では認識率100%の文書認識システムは達成されておらず、どうしても誤認識を生じてしまう。実際の文書認識においては誤認識があるかどうかは、人間が確かめなければならず、さらに誤認識がある場合には、修正に大変な労力をさかなければならない。

このため、文書認識の自動化が大変に期待されている。

自動化を実現する際、現在の文書認識で問題になるのは、その文書のフォントが、既に登録されている辞書のパターンからかけはなれている場合である。認識方法として一般的に用いられているパターンマッチング法では、辞書 (標準パターン・代表ベクトル) に依存する部分が大きい。このため、このような場合には単純にパターンマッチング法だけでは正解を選出することができず、認識結果の上位の数候補に関して詳細に識別しなければ正確な認識を行うことができない。また、詳細識別を行うことでプロセス等は複雑になり、単純パターンマッチング法のみで自動化するよりも難しくなる。そこで、認識方法で高精度化をはかるのではなく、その文書のフォントを抽出し、辞書を作成するという辞書の高精度が考えられる。辞書を新しく作成する方法は同じフォントの文書を大量に認識する、というような場合に有効であると考えられる。[8]

本研究では既存辞書を基に、認識対象となる文書からフォントを抽出し、新辞書を作成すること、その辞書を用いてフォントを抽出した文書を再認識させ誤りを自動的に修正することなど、抽出したフォントを用いて文書認識の高精度化をはかる方法について実験し検討する。

1.2 本論文の構成

本論文の構成は次の通りである。

第 1 章 序論であり、本研究の背景と目的を述べる。

第 2 章 本研究で用いる特徴量や、文書認識システムについて述べる。

第 3 章 文書中から文字のフォントを抽出するアルゴリズムについて説明し、フォントの抽出実験を行い、結果について検討する。

第 4 章 3 章で抽出したフォントより辞書を作成して認識実験を行い、再認識での誤りの自動修正、新辞書の有効性などについて検討する。

第 5 章 本研究の結論・今後の課題について述べる。

第 2 章

文書認識システム

2.1 まえがき

本章では本研究で提案する文書認識システムの概略、システムで用いる文字認識の手法について説明する。

2.2 文書認識システムの構成

認識率を向上させるためには、パターンマッチング法だけでなく、他の認識方法と併用するなど、認識手法を複雑にすることは有効である。しかし、認識対象からフォントを抽出することが可能で、そのフォントを基に標準パターンを新しく作成できれば、1 つの認識手法のみ (パターンマッチング法) を用いても良い認識結果が得られるはずである。つまり、認識手法を改善・改良するのではなく、辞書を改善することによってパターンマッチング法だけを用いても認識結果は向上できるのではないか、ということである。図 2.2に示す本研究で提案する文書認識システムでは、この点に着目し対象文書からのフォントの抽出を中心に据えている。この認識システムでは辞書を抽出した文書を再認識し、誤認識を自動修正することを目的にしているが、この抽出した辞書を、他の文書の認識に用いることで、パターンマッチング法のみで従来法よりも認識精度が向上することも期待している。

- 認識
  認識では予め用意しておいた辞書を用いて、認識対象文書を認識させる。このとき認識結果としては 1 位の候補だけでなく、2 位・3 位 ··· と複数の結果を出力する。

- フォント抽出
  文書中の全ての文字について認識結果を基に、フォントを抽出する。フォント抽出の

アルゴリズムは3章で詳しく説明する。

- 新辞書作成
  抽出されたフォントから、新たに標準パターンを作成する。本研究では、この新標準パターンを既存辞書に加える方法のマルチテンプレート法と、もともとの標準パターンと新標準パターンとを置換する置換法の 2 つの方法で新しい辞書を作成している。

- 再認識
  再認識では新辞書を用いる。

2.3　文字認識の方法

文字認識のアルゴリズムには、大きく分けて次の二つの方法がある。

- パターンマッチング法
- 構造解析法

パターンマッチング法は、パターン同士の重なりぐあいで評価し、認識を行う。このため、文字の多少の変形やノイズには強いが、類似文字の識別は難しい。また、計算機上で容易に高速で実現できる等の特徴をもつ。

これに対し後者の構造解析法は、線分の接続関係や位置関係などの文字構造に着目し、構造の類似性で認識を行う。この手法は類似文字の多い漢字の認識や、個人のくせが顕著である手書き文字の認識に有効であるとされているが、特徴量の定義や抽出が難しく、また計算機上での処理に時間がかかる等の問題も多い。

以後、本研究で用いるパターンマッチング法について説明する。図 2.1に文字認識のアルゴリズムを示す。

2.3.1　文字入力・前処理

- 文字入力
  認識される文字データは文書の形でイメージスキャナによって入力される。出力されたイメージデータは、切り出し処理により、各文字毎に切り出される。

- 前処理
  切り出された文字データには、実際にマッチングに使われる特徴量を求めるために前

処理が施される。前処理は、ノイズ除去・スムージング、正規化、細線化、線素化の4つの処理によって構成されている。

- ノイズ除去・スムージング
  入力画像 (切り出されたデータ) は印刷時のプリンタの状態やイメージスキャナなどの性能によって、ノイズや線分上の凹凸が少なからず発生している。これらは認識精度に悪い影響を及ぼすため、まず入力画像に対してノイズ除去を行う。本研究では2×2ドット以下の孤立点をノイズと見なし除去することにする。スムージングは、周囲8近傍を参照し、中心のドットを決定する。実際には3×3ドットのマスクを用いて行う。

- 正規化
  正規化は、もとの入力イメージ画像を一定の大きさ (本研究では64×64ドット) に拡大又は縮小する処理であり、これによって文字の大きさや領域の位置ずれによる影響を吸収する。正規化の手法には大きく分けると、線形正規化と非線形正規化の二つの手法に分けられるが、本研究では線形正規化を用いている。

- 細線化
  細線化は、通常数ドットの太さを持った文字の線分を1ドットの幅に変換する処理である。この処理によって、文字の線分幅の違いを吸収する。本研究では Hilditch の細線化アルゴリズム[5] を用いる。この方法は、各画素においてその周りの3×3のマスクを用いて、着目するドットの近傍の8ドットの連結状態を見ながら細線化していく方法である。

- 線素化
  線素化は、細線化された図形の各画素について、黒画素であればその周囲3×3の小領域を参照し最も自然な方向と考えられる4つの縦「｜」、横「―」、斜め45°「／」、斜め135°「＼」、のうちの1つの線素に対応させる。

前処理の各処理による効果を図 2.3に示す。

2.3.2　特徴抽出

パターンマッチング法では、処理の高速化、パターン分離の効率化などのために、文字パターンを特徴量という数値ベクトルに変換する。この課程を特徴抽出という。以降は本研究で使用する方向線素特徴量[4]、について説明しておく。

- 方向線素特徴量
  方向線素特徴量[4]の抽出法は図 2.4のように、まず、64 × 64 ドットの線素化文字を 8 ドット間隔に縦横を分割する。次に、左上から 16 × 16 ドットを半分ずつ重複させて 49 領域 (左上を 0 とし、左から右、上から下へ順に並ぶ) ができる。1 領域のベクトルは 4 方向のカウンタからなる 4 次元ベクトルでできてあり、1 領域内の重みは図 2.4 の下図のようになっている。よって 1 文字当りの次元数は N =196(49×7) 次元のベクトルとなる。

2.3.3　認識

ここでは最も一般的な手法である全数整合法を説明する。全数整合法は、各文字の標準パターンと特徴抽出で得られた未知入力文字の特徴量で評価値 (距離) を求め、近いものから順に識別する方法である。アルゴリズムの簡潔性やある程度の高い認識率を得られることから、一般的に用いられている。認識で用いられる辞書・評価値について以下で説明する。

- 辞書の作成方法
  パターンマッチング法では、入力されたパターンと標準パターンとを比較し、一番似ている (距離的に近い) 標準パターンのコードを結果として出力する。この標準パターンは辞書と呼ばれ、予め全ての登録文字に対して用意しておかなければならない。辞書作成法としては、

  - 文字 kのサンプルパターンの集合 S(k) の重心ベクトル

  が一般的であり、本研究ではサンプルパターンの集合として 42 種のプリンター出力を用いて作成した重心ベクトルを辞書 (既存辞書) として使用している。

- 評価値
  未知入力文字の特徴量と標準パターンとの類似性や整合性の評価は、評価値により行われる。評価値の例としてよく使用されるものとして、ユークリッド距離、重み付きユークリッド距離[6] がある。ここで、$\boldsymbol{x}$、$\boldsymbol{u}$ をそれぞれ未知入力ベクトルと、ある文字の標準パターンベクトルとすると、それぞれ次のように定義される。

  ユークリッド距離：

$$d_e(x,u) = \sum_{i=1}^{n}(x_i - u_i)^2 \tag{2.1}$$

重み付きユークリッド距離：

$$d_w(x,u) = \sum_{i=1}^{n} w_i(x_i - u_i)^2 \tag{2.2}$$

ここでw_iは重み付きユークリッド距離の重み[7]で、各次元ごとに次のように定義する。

$$w_i = A\frac{1\backslash v_i}{\sum_{j=1}^{n} 1\backslash v_j} \tag{2.3}$$

但し、A は定数、v_iは標準パターン作成用データセットにおける各次元の分散、n は次元数である。本研究ではこの 2 つを使用する。この 2 つの他にシテイブロック距離

$$d_c(x,u) = \sum_{i=1}^{n} |(x_i - u_i)| \tag{2.4}$$

があり、計算時間が速い等の特徴をもつ。

距離以外の評価値としては、2 つの特徴量の内積から定まる類似度[2]が使われている。

$$E(\mathrm{u},\mathrm{x}) = \frac{\mathrm{u}\cdot\mathrm{x}}{|\mathrm{u}||\mathrm{x}|} = \cos\theta \tag{2.5}$$

である。

2.3.4　候補出力

評価値計算をした後評価値の小さい順に, 第 1 位候補、第 2 位候補、$\cdots$、第 N候補を出力する。

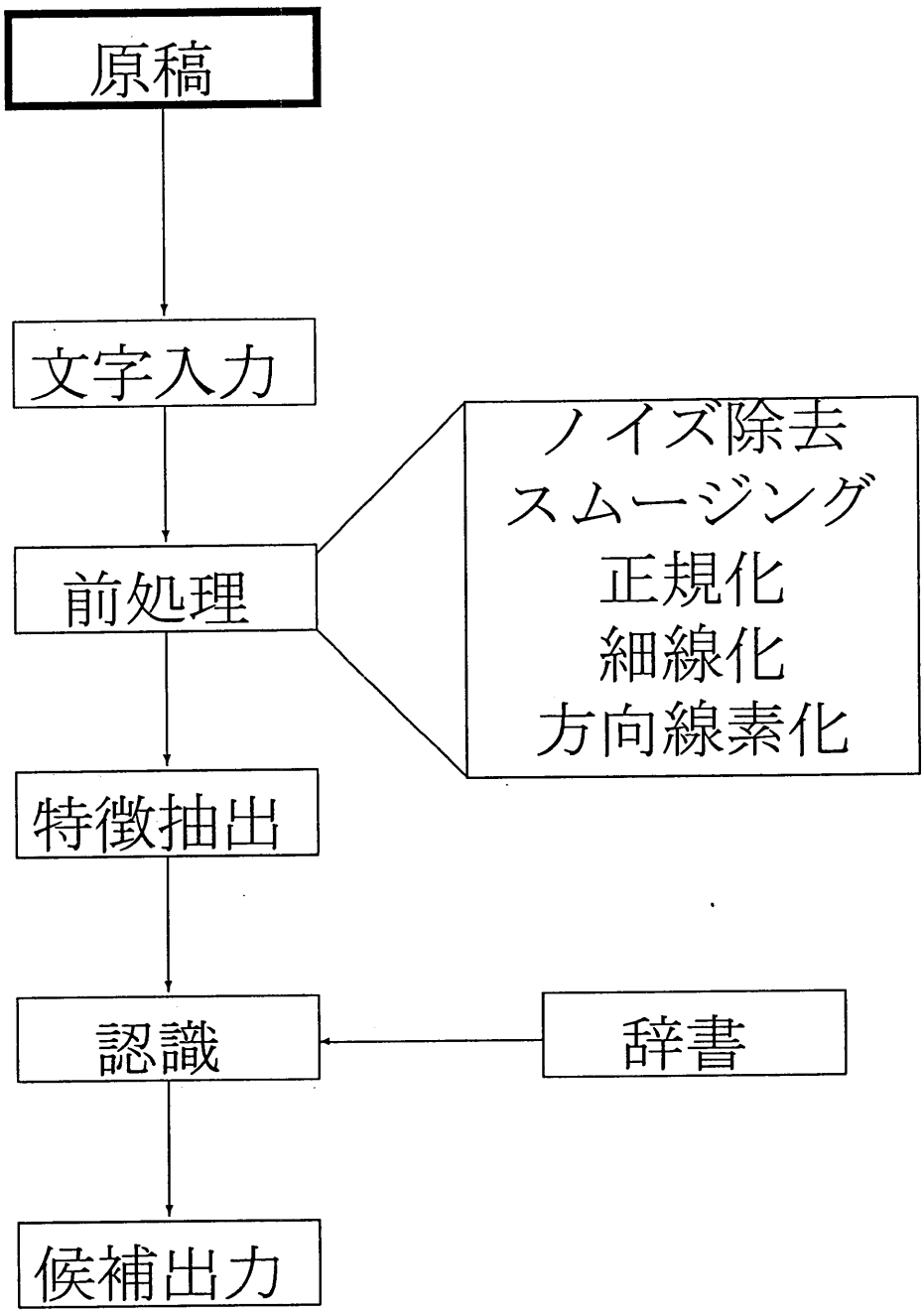


図 2.1: 文字認識アルゴリズム

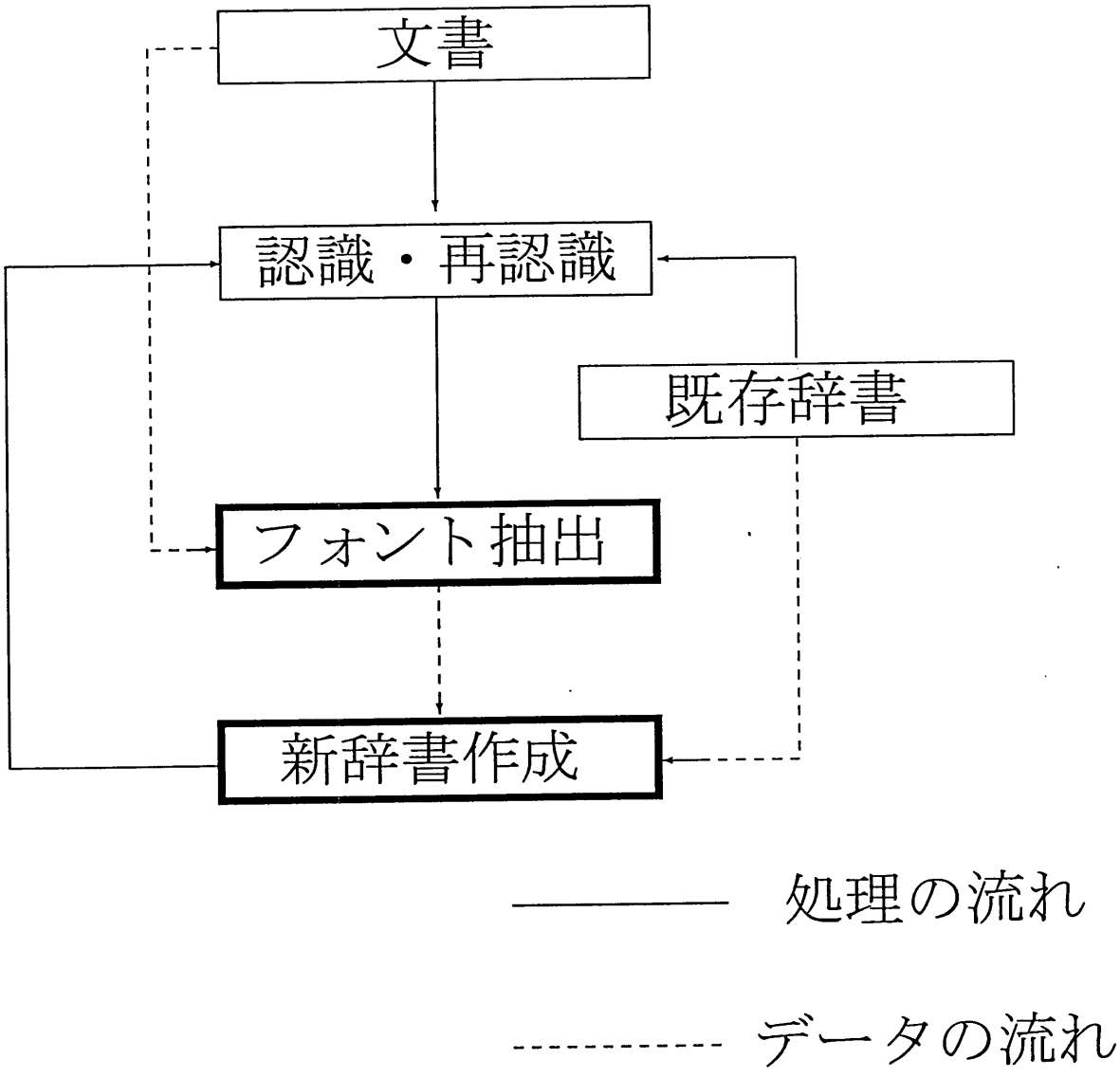
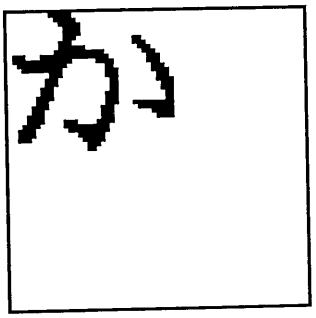
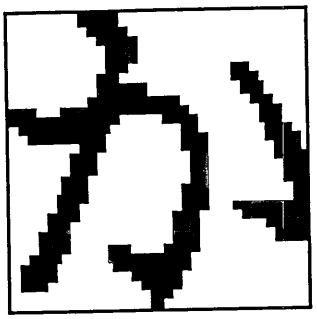


図 2.2: 文書認識システムのアルゴリズム

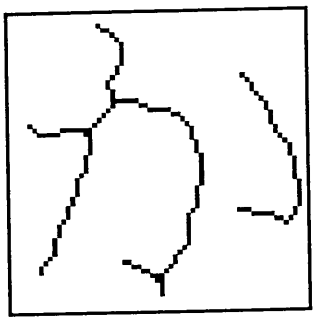
入力画像



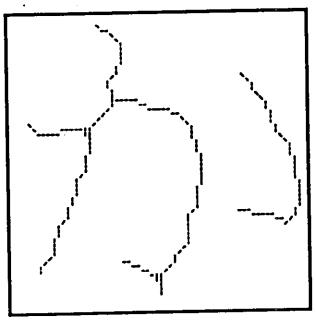
ノイズ除去・スムージング



正規化



細線化



線素化

図 2.3: 各処理後のイメージ

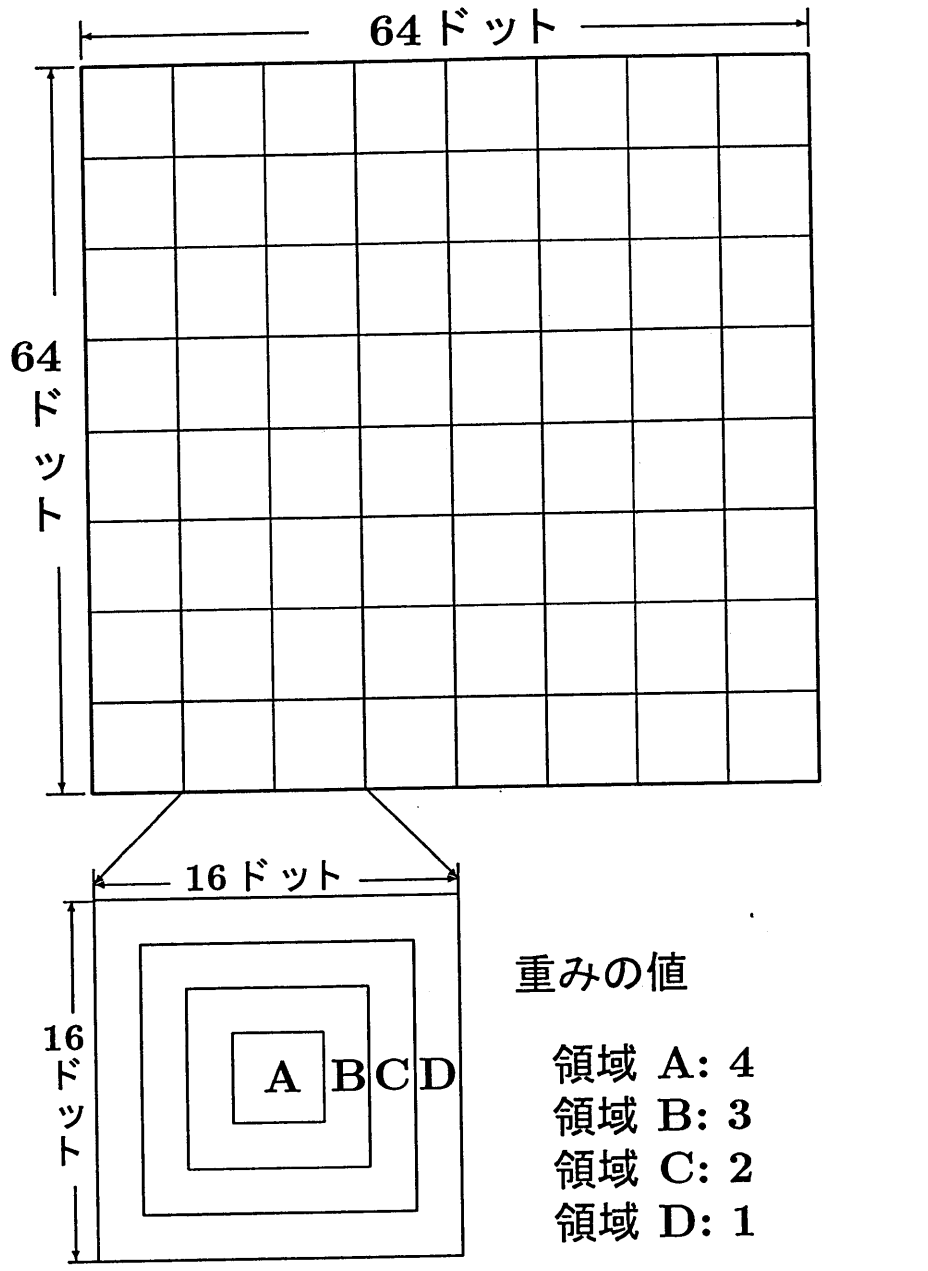

64 ドット
64 ドット
16 ドット
16 ドット
A
B
C
D
重みの値
領域 A: 4
領域 B: 3
領域 C: 2
領域 D: 1
方向線素:
方向線素特徴量: $\mathbf{V}_m = (V_{m_1}, V_{m_2}, V_{m_3}, V_{m_4})$
where $m = 1 \sim 49$
小領域の数: 49(7×7)
方向線素特徴量の次元数: 196(4×49)

図 2.4: 方向線素特徴量

第 3 章

フォントの抽出

3.1 前書き

この章以降で、抽出すべきコードと同じコードのフォントのことを**正解のフォント**と表し、抽出すべきコードとは違うコードのフォントを**不正解のフォント**と表すことにする。文書中から未知のフォントを正確に抽出することは容易ではない[9]。特に次に挙げる 2 つの点が問題になる。

1. 正解 (抽出すべきコード) のフォントを集める。
2. 正解のフォントだけを抽出する。

- 前者の場合は、認識結果を用いて正解のフォントを集める際に、認識結果が 1 位のコードに対応しているフォントだけを集めた場合には、現在誤認識しているものからは、コードとフォントが正しく対応していないフォントが正解のフォントとして集められてしまい、2 位以降にある真の正解を集めることができない。このため入力されたフォントに、2 位以降の複数の候補のコードと対応させたからもフォントを集めてこなければ、全ての正解のフォントを集めることができない。
- 次に上記の様に 1 つの未知フォントに複数のコードを割当てて、フォントを集めた場合、集まったフォントは正解のフォントだけとなることはほとんどない。正解のフォントは含まれていても、その他には不正解のフォントも含まれているはずである。正解のフォントがその集合の大多数を占める場合であれば、集めた全てのフォントのベクトルの重心をとり、重心から離れているようなベクトルを持つフォントを取り除くことで、正解のフォントのみを抽出することは可能である。(図 3.1の成功例) しかし、

出現頻度が低い文字や、類似している文字が多い文字の場合は、正解のフォントがごく少数の場合や、正解のフォントと、類似している文字のコードのフォントが同数ほどの場合には、この方法では正解のフォントが全て取り除かれたり、不正解のフォントも同時に抽出されたりする可能性がある。図 3.1での失敗の例 1 では、抽出したいフォントは「か」であるが、その母集団の中に「か」に類似している「が」のフォントも多く含まれるため、「か」と「が」が抽出されてしまっている。また、失敗例の 2 では「お」が母集団の中の少数なため、誤って「も」が抽出されてしまう場合である。この様にフォントを抽出する母集団の構成により、抽出結果が変ってしまうため、この重心を用いた手法だけではフォントを正確には抽出できない。

そこで、本研究では図 3.2で示すアルゴリズムのフォント抽出法を提案する。

また、今回フォントを抽出する対象としては、事前に新聞社説 (1〜20) の構成する字種について調査し (A章参考)、100 文字以下で、社説中の全ての文字の 50%弱を占める、ひらがなとし、その中から 3.2.5節の除去候補文字の学習用に用いた新聞社説 (1〜10) の 10 部中に出現した 65 種とした。(表 3.3) 実験で用いた新聞社説は図 3.5に例示する。

3.2　フォントの抽出

図 3.2の各処理について以下に詳しく説明する。

3.2.1　認識・結果出力

認識の処理では、予め準備しておいた既存辞書を用いて、未知文書の各文字を全数整合法により認識させる。誤認識する文字を救済するため 1 位の候補だけでなく下位候補まで複数の候補を正解として出力する。ここで何位まで正解として出力するかを**候補数**と定義する。通常の認識と同様に、1 位だけを正解とする場合は候補数が 1 であり、5 位候補まで正解とするならば候補数は 5 となる。

認識に用いた辞書の文字数は、ＪＩＳ第一水準の漢字 2965 文字と、ひらがな・カタカナのうち小文字「っ」「ぁ」等を取り除き、さらに人間でも字体からだけでは、ひらがなかカタカナか判断がつきかねる「へ」「べ」「ぺ」の 3 文字に関しては、同一のものとして扱い、ひらがな 73 文字、カタカナ 71 文字の合わせて 3109 文字である。辞書に含まれるひらがなとカタカナを表 3.4に示す。

この処理を文書中の全ての文字に関して行い、各文字に対して、候補数の複数の正解を与える。通常の認識では入力されたフォント (文字、パターン) と、その入力フォントに一

番近い辞書中の標準パターンのコードとを 1 対 1 で対応させているが、ここでは入力フォント 1 つに対して、複数のコードを対応付けることになる。実際に図 3.6と図 3.7(a) の例で説明する。まず、図 3.6中の「A」、「B」、「C」はそれぞれ字種の異なる文字の入力フォントを表し、「A」と「A'」は同じ字種の文字の異なる入力フォントを表している。右側の 1 位～4 位までの欄は、そのときの上位の認識結果である。通常の認識では入力フォント (「A」) に対し、1 位認識結果である「は」というコードを 1 対 1 で対応させるが、今回用いる手法では入力フォント (「A」) に対し、「は」だけでなく、「け」、「ほ」等複数のコードを対応付ける。(図 3.7(a)) この対応付けた複数のコードの中に、フォントのコードと同じコードがある (フォントとコードが正しく対応している) と考える。

3.2.2　リストの作成・候補集合の作成

認識・結果出力は文書中の各文字毎に処理を行ったが、以降の処理では、結果出力で対応付けした各コード毎に行う。リストの作成では、まず出力したコードそれぞれについて候補数までの認識順位で、そのコードを出力した入力フォント (文字) と対応付ける。次に出力したコード毎にクラスタリングし、リストを作成する。図 3.6から対応関係を表したのが図 3.7(b) になる。「は」というコードを認識時に出力したのは、入力フォント「A」と「C」、「A'」であるので、これをコード「は」と対応させる。同様に、コード「ば」に対しては、入力フォント「B」と「C」を対応させる。

候補集合の作成では、リストに従って各コード毎に実際の入力フォントを集めてくる。この各コードに集まったデータを**候補集合**と定義する。この候補集合が正解のフォントを抽出するための母集団になり、認識する際の候補数が、この候補集合の大きさを決定する。図 3.8に、実際に候補集合として集められた入力フォントを示す。これは、コード「け」に対して候補数を 3 にしたときの例である。この候補集合中には、「は」、「け」、「ゆ」の 3 文字のフォントが集められているが、不正解のコードである「は」、「ゆ」のフォントを除去するためには次の除去候補文字の処理を行う。

3.2.3　除去候補文字の処理

認識において複数の正解を与えるため、候補集合中には抽出すべき正解のフォントのみではなく、不正解のフォントも含まれている。特に文書中で出現頻度が低いコードの場合、候補集合中のデータのほとんどは不正解で、正解のフォントは少数しか含まれていない場合もある。この様に他種のフォントが混在している中から正解のフォントのみを抽出するには、候補集合中に存在する不正解のフォントを除去しなければならない。この候補集合

中に存在する除去しなければならないフォントのコードを**除去候補文字**と定義する。除去候補文字は各コード毎に固有であり、その数もコードによって異る。

除去候補文字の大部分は、フォントによって変わるのではなく、その文字の字体に依存しているため統計的に決定できる。しかし、一部はフォントの違いで大きく異るため事前に、そのフォントでの除去候補文字を学習しておくことが必要になる。除去候補文字の学習については3.2.5節で述べる。

この除去候補文字には、「け」と「ゆ」の様に明らかに異なる文字とわかるものと、「か」と「が」の様に類似度が大きく識別しがたい文字との 2 つがある。除去候補文字の処理もこの 2 つに合せて 2 段階で行う。

まず、明らかに異なる文字の場合は、パターンマッチング法で誤認識した場合の、正解と誤認識との評価値の性質を用いて削除することができる。

いま、「A」というコードの候補集合中に「A」、「B」の 2 種のフォントが集められたとする。ここで「A」、「B」のフォントの入力ベクトルと、それぞれの標準パターンとの評価値の関係は次式の様になると考えられる。

$$E(u_A, v_A) \ < \ E(u_B, v_A) \tag{3.1}$$

$$E(u_A, v_B) \ > \ E(u_B, v_B) \tag{3.2}$$

ただし、v_A、v_B、u_A、u_Bはそれぞれ、A、B の入力ベクトル、標準パターンを表し、$E(u_A, v_A)$は、「A」の入力ベクトルと「A」の標準パターンとの評価値を表すとする。

ここで、不正解である「B」の入力ベクトルは、「A」の標準パターンより、「B」の標準パターンに距離が近いと考えられる。この性質を用いて、予め各コード毎に誤って候補集合中に選出されるような文字 (除去候補文字) を学習させておき、除去候補文字のコードの標準パターンと、正解のコードの標準パターンとを、候補集合中の未知フォント (ベクトル) とマッチングさせ、未知フォントが、正解のコードの標準パターンよりも、除去候補文字のコードの標準パターンに近ければ、その未知フォントは、正解のコードのフォントではないとみなして、除去を行う。

正解のコードを「T」、除去候補文字のコードを「x_n」、候補集合中のあるフォントをY_mとすると、

$$E(u_T, v_{Y_m}) \ > \ E(v_{x_n}, v_{Y_m}) \tag{3.3}$$

のとき、フォントY_mを候補集合から除去する。

この処理で候補集合中の不正解のフォントを除去し、候補集合中のフォント数を、それに含まれる正解のフォント数に近づける。

3.2.3.1　詳細識別

次に類似度が大きく、識別しがたい文字の処理について説明する。「こ」と「ご」のように清音と濁音がある文字や、「し」と「リ」のように全体の字形がベクトル的に似ている文字の識別の際には、文字間の距離は、通常の異なる 2 文字間のものよりも差異が小さいため、上記の方法では、正解かどうか識別が難しくなる。この様な文字は**詳細識別文字**として新しく定義する。詳細識別文字と正解のコードとの識別は文字の画像領域全体ではなく、差異の大きい部分だけを用いて行う。

除去候補文字の処理、詳細識別については、3.2.5節の除去候補文字の学習で具体例を示しながら説明する。

3.2.4　細分類

候補集合中に、各コードの除去候補文字以外の除去すべきコードのフォントが存在していた場合には、上記の除去候補文字の処理だけでは除去できない場合が考えられる。その様なフォントを除去するために次の示す細分類の処理を行う。

1. 候補集合中の全てのフォント (ベクトル) から重心ベクトルを作成する。
2. 全てのフォントベクトルとこの重心ベクトルとマッチングを行い、ユークリッド距離が、個々のコード毎に予備実験で定めた閾値よりも大きくなれば、そのフォントを候補集合から削除する。
3. 候補集合中のフォントで再び重心ベクトルを作成し、マッチングを行う。
4. これを収束するまで繰り返す。

なお、各コード毎の閾値は次のように定めた。あるコードの標準パターンを $\boldsymbol{U}$、その学習サンプルパターンを $\boldsymbol{P}$ とし、$\boldsymbol{U}$ と $\boldsymbol{P}$ のユークリッド距離 ($\boldsymbol{dis}$) の平均 ($\boldsymbol{x}$) と分散 ($\boldsymbol{\sigma^2}$) を求める。このときの平均と偏差の和をそのコードの閾値 ($\boldsymbol{threshold}$) とした。

$$\boldsymbol{U} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{P}_j \tag{3.4}$$

$$\boldsymbol{dis}_n = \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{P}_j - \boldsymbol{U})^2 \tag{3.5}$$

$$\boldsymbol{x} = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{dis}_j \tag{3.6}$$

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(dis_j - x)^2 \tag{3.7}$$

$$threshold = x + \sigma \tag{3.8}$$

この細分類の処理を終えて候補集合中に残っていたフォントは、そのコードのフォントとして抽出する。

3.2.5　除去候補文字の学習

本研究で用いた除去候補文字は表 3.1、表 3.2に示すが、これらは次のようにして学習させた。

- データとして新聞社説 (1〜10) の 10 部を使用し、候補数を 10 として今回提案したアルゴリズムを用いて、各コード毎に候補集合を作成する。図 3.9は「み」のコードの場合で、候補集合中には「ん」、「み」、「入」、「ふ」、「ぶ」の 5 字種が合せて 112 個集まった。

- この候補集合中に存在した正解以外のコードを、仮の除去候補文字として決定する。例では「ん」、「入」、「ふ」、「ぶ」の 4 字が仮の除去候補文字となる。
  この仮の除去候補文字を用いて、式 (3.3) により候補集合中の不正解のフォントを取り除いていくが、仮の除去候補文字 x_n が以下の式 (3.9)、式 (3.10) の条件を満たすときは、その x_n を除去候補文字から削除し、除去候補文字の削減・最適化を行った。式 (3.3) 中の評価値は、各コード毎にユークリッド距離 (式 (2.1))、重み付きユークリッド距離 (式 (2.2)) のどちらかを使用した。(図 3.3)

$$\begin{cases} T(x_1, x_2, \cdots, x_{n-1}) = T(x_1, x_2, \cdots, x_n) = T_{in} \\ D(x_1, x_2, \cdots, x_{n-1}) \geq D(x_1, x_2, \cdots, x_n) \end{cases} \tag{3.9}$$

$$\begin{cases} T(x_1, x_2, \cdots, x_{n-1}) = T_{in} \\ T(x_1, x_2, \cdots, x_n) < T_{in} \end{cases} \tag{3.10}$$

  ただし、T_{in} は最初の候補集合中の正解のフォント数、x_j は除去候補文字とし、$T(x_1, x_2, \cdots, x_{n-1})$ $D(x_1, x_2, \cdots, x_n)$ はそれぞれ除去候補文字が x_1〜x_n のときの除去候補文字の処理を行った後の候補集合中の正解フォント数と、候補集合中のフォント数を表すものとする。式 (3.9) は x_n を除去候補文字から外しても、残りの $n-1$ 字の除去候補文字により不正解のフォントを除去できる場合である。図 3.9の場合は仮の除去候補文字の

うち「ぶ」が、除去候補文字では外れている。これは、「ぶ」を外しても残っている「ふ」によって不正解のコードである「ふ」と「ぶ」のフォントが除去できることを示している。

式 (3.10) は x_n を除去候補文字としたことで、不正解のフォントだけでなく、抽出すべき正解のフォントまで除去している場合を示している。図 3.10では、仮の除去候補文字の「ざ」を用いて除去候補文字の処理をした場合に正解である「ぎ」も除去されてしまったことを示す。このような場合には、「ざ」を除去候補文字から外す。

ここで最適化された除去候補文字を x_1〜x_N とすると、

$$T(x_1, x_2, \cdots, x_N) \;=\; T_{in} \tag{3.11}$$

となることがフォントを正確に抽出するための条件となるが、式 (3.10) の条件で除去候補文字から外された x_n のフォントが候補集合に残った場合や、清音と濁音の文字の様な類似文字では、不正解のフォントを全て除去できない場合がある。そこで、この段階で候補集合中に不正解のフォントがある場合は、その不正解のフォントのコードを詳細識別文字として新たに詳細識別を行うものとする。(図 3.10の「ざ」)

- 詳細識別文字と正解のコードとの識別・除去でも、式 (3.3) を用いるが、マッチングを行う領域は図 3.14で示される 64×64 ドットの文字の画像領域全体ではなく、差異が大きい部分のみでマッチングを行う。この部分を識別領域とする。2.3.2節で説明した方向線素特徴量の画像領域での 49 個の 16×16 ドットの小領域 (マス) のうちの 2 マスを識別領域とする。通常のマッチングでは 49 マス全てを用いるため、ベクトル (方向線素特徴量) は 196 次元であるが、詳細識別では 2(マス)×4(方向) の 8 次元となる。各コード毎に、2 マスを次のように決定した。

  候補集合中に正解「T」のサンプルが N 個、不正解「F」のものが M 個あり、それぞれ t_1〜t_N、f_1〜f_M とする。このとき、

  $$E_m(u_T, v_{t_i}) \;>(<)\; E_m(u_f, v_{f_j}) \tag{3.12}$$

  ただし、$E_m(u_T, v_{t_i})$ は $m(0 \leq m \leq 48)$ マスにおける「T」の標準パターンと、「t_i」の入力フォントとのユークリッド距離を表す。式 (3.12) が $(0 \leq i \leq N, 0 \leq j \leq M)$ の全ての (i, j) で成り立つようなマス m を識別領域とした。m が 2 以上あるときは、式 (3.12) の右辺と左辺の差が大きい順に 2 つを識別領域とした。式 3.12を満たす m が 1 つしかない場合はこの 1 つを識別領域としている。図 3.11では、$N = 13$、$M = 10$ であり、識別領域としては、マス (18, 19) となった。(図 3.13)。今回用いた全ての詳細識別文字と識別領域は表 3.12に示す。

- 以上の除去候補文字の処理と、詳細識別文字の除去を行い、候補集合中には抽出すべき正解のデータのみが残るようになり、除去候補文字の学習を終了した。(学習データ新聞社説 (1〜10) の 10 部において)

3.3　抽出実験

本研究で提案するフォント抽出のアルゴリズムによってフォントが正確に抽出できるかどうかを確認するためにフォントの抽出実験を行った。

3.3.1　実験方法

- フォントを抽出する対象文書として、除去候補文字の学習に用いなかった新聞社説 (11〜20) の 10 部を対象文書とし、1 部ずつに対して実験を行う。また抽出する文字は表 3.3で示したひらがな 65 文字とする。
- フォント抽出のための認識で使用する既存辞書は各種プリンタのフォントセット 42 種から作成した重心ベクトルを用いる。辞書の文字数は 3109 字である。
- 図 3.2で示したアルゴリズムでフォントの抽出を行う。ここで、候補数をどれだけとれば全ての正解を候補集合に出力できるかを調べるため、候補数を 1 から 10 まで変えて実験を行った。

3.3.2　実験結果

候補数と候補集合に集められる正解について表 3.4に示し、抽出の結果については表 3.5に示す。また、候補集合中の正解の割合と、真の正解数までに不足している正解数をグラフにし、図 3.15に示す。実際に抽出できなかった文字は表 3.7に示す。

3.4　考察

表 3.4の A は、除去候補文字の処理を一切行わなかった作成直後の候補集合のデータ数である。実際の実験では社説 1 部毎に行ったが、ここでは 10 部の結果を和で示してある。B は同様に候補集合に取り入れられた正解数を表している。C は、B / 10328 で、D は、B/A で候補集合中の正解数の割合、つまり母集団中の有効なデータの割合を示している。B の欄に注目すると候補数 6 で全ての正解を候補集合に取り込むことができており、この

結果、3.1節で示した条件の 1 つである、文書中の全ての正解を集めることは達成できた。候補数 7 以降では新しく候補集合に取り込まれるデータは全て誤ったものである。候補数 6 以下で抽出できなかった文字の詳細は表 3.7に示す。次に D 欄に注目すると、有効なデータの割合は候補数 6 では 32.12%なのが候補数 10 では 23.80%と 10 ポイント近くさがっており、候補数 7 以上のデータは全体の処理の実行時間を無駄に使っていることになり、候補数の上限として 6 が適当であると考えられる。

表 3.5で、真の出力数 575 は 10 部中に存在したひらがな 65 文字の字種数の和であり、フォントの抽出を期待される数になる。A の欄の正解の出力数であるが、このカウントは社説毎に、1 つでも抽出のあった文字を 1 として数えている。このため、1 つの社説中に同じ抽出すべき文字が複数存在していた場合でも、その全てを抽出しなくても、1 文字を抽出すれば抽出としている。A 欄の () の内に数字は、存在した抽出すべき文字を全て抽出した文字数をカウントした。この正解全てを抽出した文字数は全体の 8 割であり、候補集合中に集められた全ての正解の抽出には成功しなかった。次に、正解以外を抽出した B と C について述べる。B は正解と違う誤った文字を出力した場合で、C は複数字種のフォントを 1 つの字種として出力した場合であるが、今回の実験では C の出力は見られなかった。しかし、候補数 2〜10 で B の誤った抽出が確認された。表 3.6に実際に誤って抽出された文字を示すが、誤抽出が生じると抽出結果から作成した辞書を用いる際に、誤抽出したコード (例「液」) と、実際に抽出すべきコード (例「ゆ」) の両方のコードを持つ文字の認識に影響があると考えられる。実際の影響については、4章での実験で確認する。

誤って抽出された文字は実質「液」、「点」、「ザ」の 3 文字であるが、これらは全て除去候補文字の学習に用いた社説 10 部中には出現しなかった文字である。3 文字とも出力数が 1 であり、候補集合中から除去候補文字を除去する際に、1 文字だけが除去されなかったため、細分類を行っても除去できなかったものである。

また、「ザ」を除く 2 文字は候補数 10 にしたときのみに抽出されている。図 3.16では、「液」の場合について示している。候補数が 9 のときには、候補集合に「液」のフォントは選出されてないため、除去候補文字の除去の処理後に候補集合中にデータは残らない。しかし、候補数を 10 にすると「液」のフォントが、最初の候補集合に選出され、除去候補文字の除去後も候補集合中にあり、そのまま誤抽出されてしまった。

このことより学習したデータにない文字が候補集合中にある場合でも、候補数を抑えることによって誤抽出を避けることも可能であるが、「ザ」の様に候補数が 2 から抽出されてしまう場合には除去候補文字を更新する等の処置を施さなければ、誤抽出を改めることは難しい。

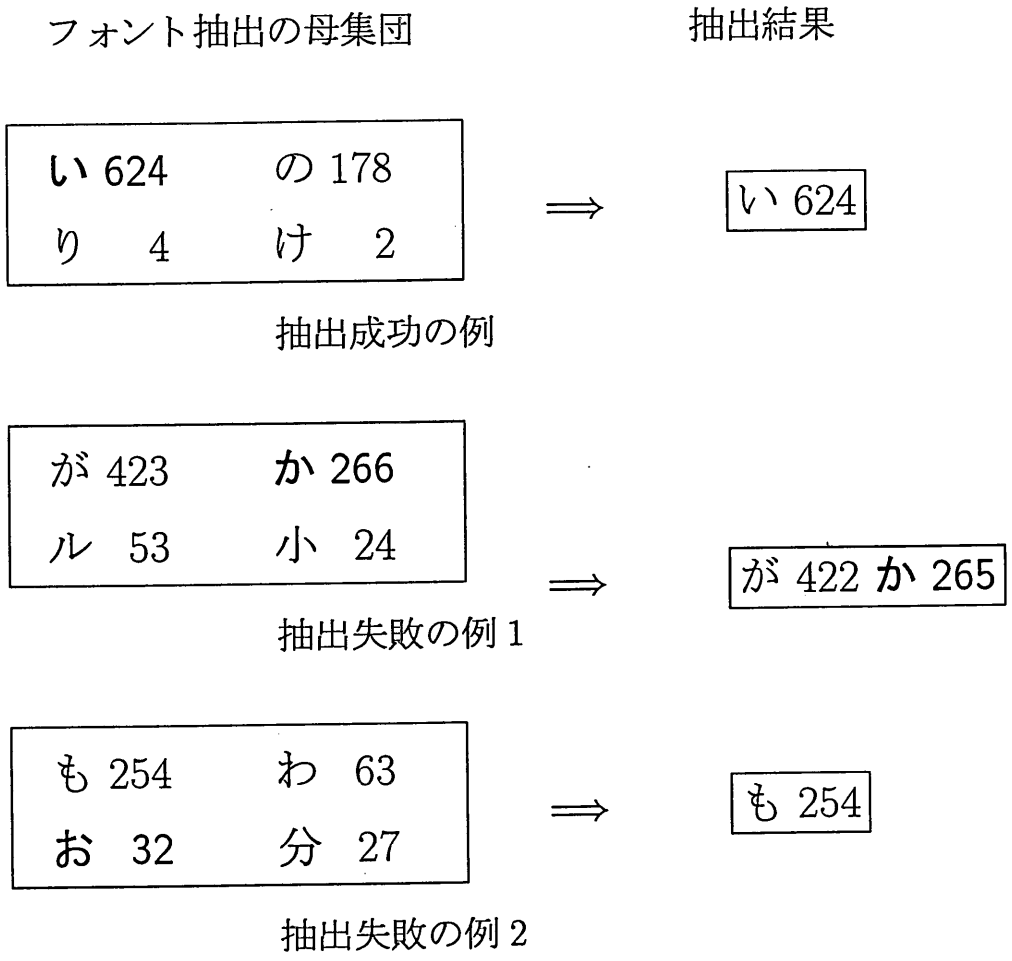


図 3.1: 重心を用いたフォントの抽出法

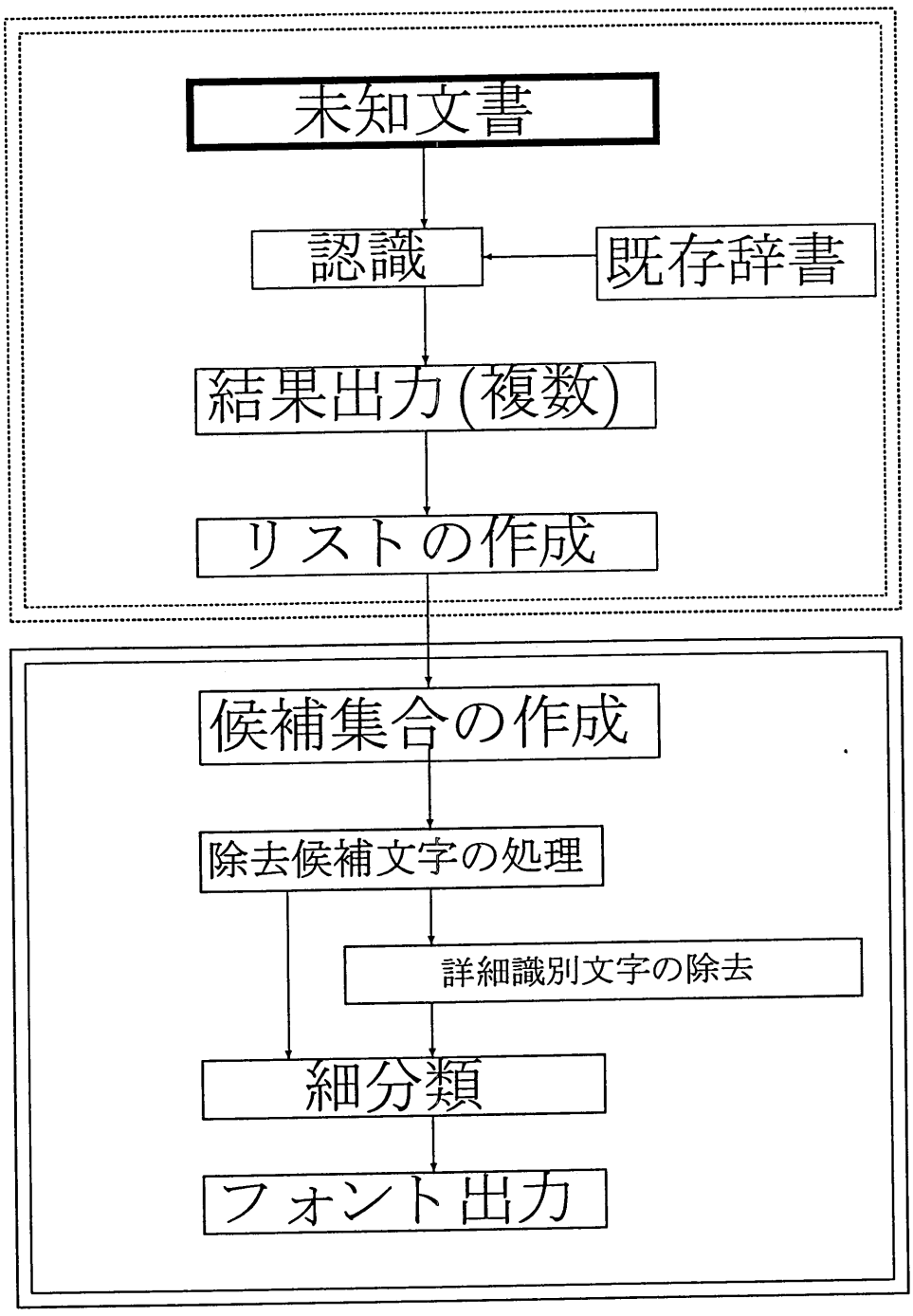


図 3.2: 未知文書からのフォント抽出アルゴリズム

あ	い	う	え	お	か	が	き	ぎ	く	ぐ	け	げ	こ	ご
さ	ざ	し	じ	す	ず	せ	ぜ	そ	ぞ	た	だ	ち	つ	づ
て	で	と	ど	な	に	ぬ	ね	の	は	ば	ひ	び	ふ	ぶ
へ	べ	ほ	ぼ	ま	み	む	め	も	や	ゆ	よ	ら	り	る
れ	ろ	わ	を	ん										

図 3.3: フォント抽出の対象とした文字

あ	い	う	え	お	か	が	き	ぎ	く	ぐ	け	げ	こ	ご
	ざ	し	じ	す	ず	せ	ぜ	そ	ぞ	た	だ	ち	ぢ	つ
づ	て	で	と	ど	な	に	ぬ	ね	の	は	ば	ぱ	ひ	び
ぴ	ふ	ぶ	ぷ	へ	べ	ぺ	ほ	ぼ	ぽ	ま	み	む	め	も
や	ゆ	よ	ら	り	る	れ	ろ	わ	ゐ	ゑ	を	ん		
ア	イ	ウ	エ	オ	カ	ガ	キ	ギ	ク	グ	ケ	ゲ	コ	ゴ
サ	ザ	シ	ジ	ス	ズ	セ	ゼ	ソ	ゾ	タ	ダ	チ	ヂ	ツ
ヅ	テ	デ	ト	ド	ナ	ニ	ヌ	ネ	ノ	ハ	バ	パ	ヒ	ビ
ピ	フ	ブ	プ	ホ	ボ	ポ	マ	ミ	ム	メ	モ	ヤ	ユ	ヨ
ラ	リ	ル	レ	ロ	ワ	ヰ	ヱ	ヲ	ン	ヴ				

図 3.4: 辞書中の仮名文字

社説

なぜ、いま四年制高校なのか

中央教育審議会が、五年半ぶりに再開された。高校教育の改革と生涯学習の基盤整備を中心課題にしている。

諮問された内容を一読してまず気づかされるのは、高校の修業年限に踏み込んだ学制改革案が含まれていることだ。

それに沿って、極めて具体的な提案を盛り込み、それぞれの当否を問う諮問の仕方にもなっている。ごく抽象的な課題を投げかけるにとどめてきたこれまでの諮問方式に比べ、大きく異なるところだ。

いまの三年制高校のほかに四年制の高校についても検討を求めている。

これらはいずれも、画一的な学校教育に一石を投じる要素を持っている。個々に見れば悪くないと思われるものもある。

だが、多くの国民の受け止め方は、なぜいま四年制高校なのか、というものではないだろうか。職業高校にとどまるにせよ、普通科にも広げるにせよ、どうしてもいま必要なものなのかどうか。このあたりはこれからの検討に待つほかない。

ただ、そうした違和感を覚えるのは、今回の制度改革案が、高校と、大学の一部だけでなく、そうした環境の中で対象に検討を進め、教育を含めた全面改革としているように思えるからだ。

だが、学校の区切りや制度全体をにらんだうえで筋というものだろう。

今回の諮問がわかりにくい大きな背景を持ち、限のみに手をつけるだけでも知れない。それにに乗り出す必要性にという問題もあろう。

とは言っても、高

元秘書

リクルート事件で

図 3.5: 新聞社説の例

フォント	1 位　2 位　3 位　4 位　・・・
「A」	「は」「け」「ほ」「吋」・・・
「B」	「ぱ」「ぽ」「ば」「ぼ」・・・
「C」	「は」「ほ」「ば」「け」・・・
「A'」	「け」「は」「ほ」「げ」・・・

図 3.6: 認識結果の例

フォント		コード
「A」	←	「は」
「A」	←	「け」
「A」	←	「ほ」
⋮	⋮	⋮

(a)

コード		フォント
「は」	←	「A」「C」「A'」
「け」	←	「A」「C」「A'」
「ほ」	←	「A」「C」「A'」
「ば」	←	「B」「C」

(b)

図 3.7: コードとフォントの対応

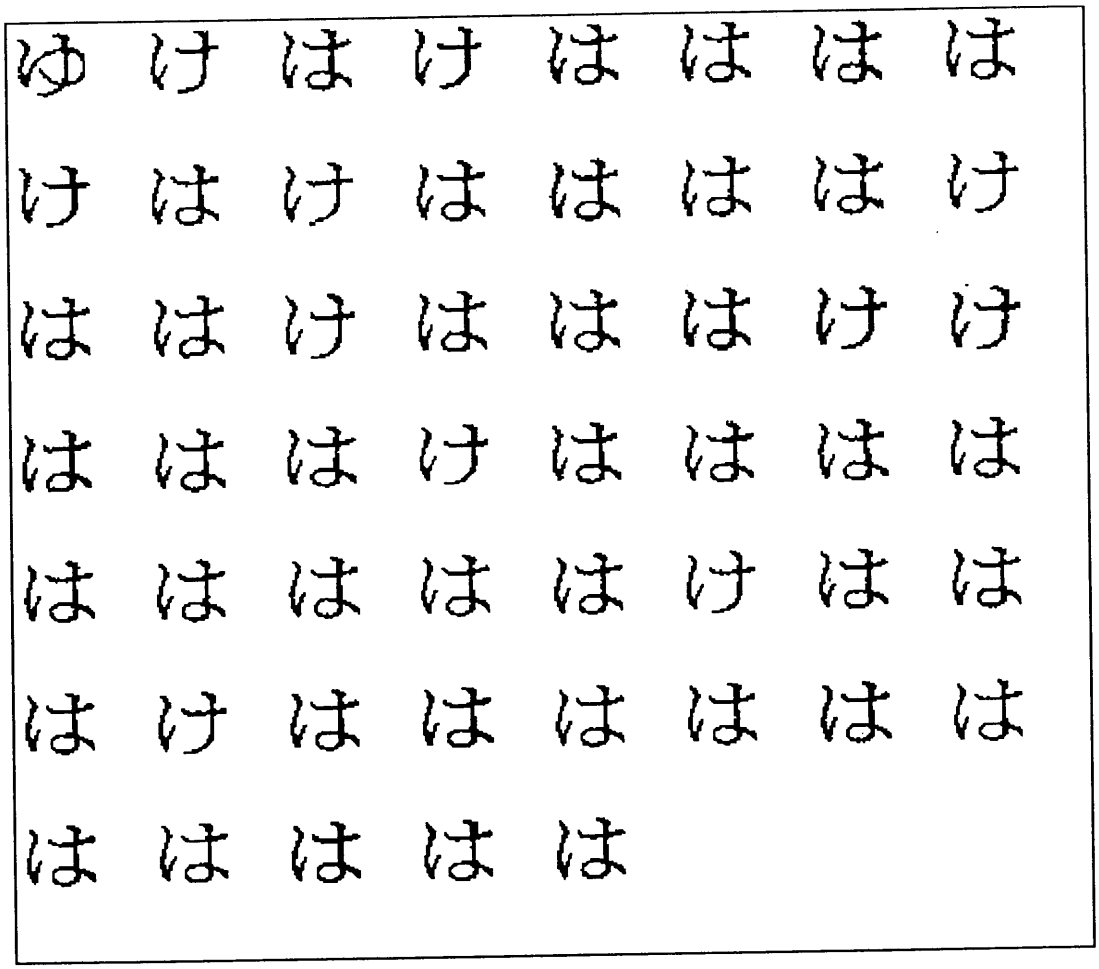

ゆ け は け は は は は
け は け は は は は け
は は け は は は け け
は は は け は は は は
は は は は は け は は
は け は は は は は は
は は は は は

図 3.8: 候補集合の例

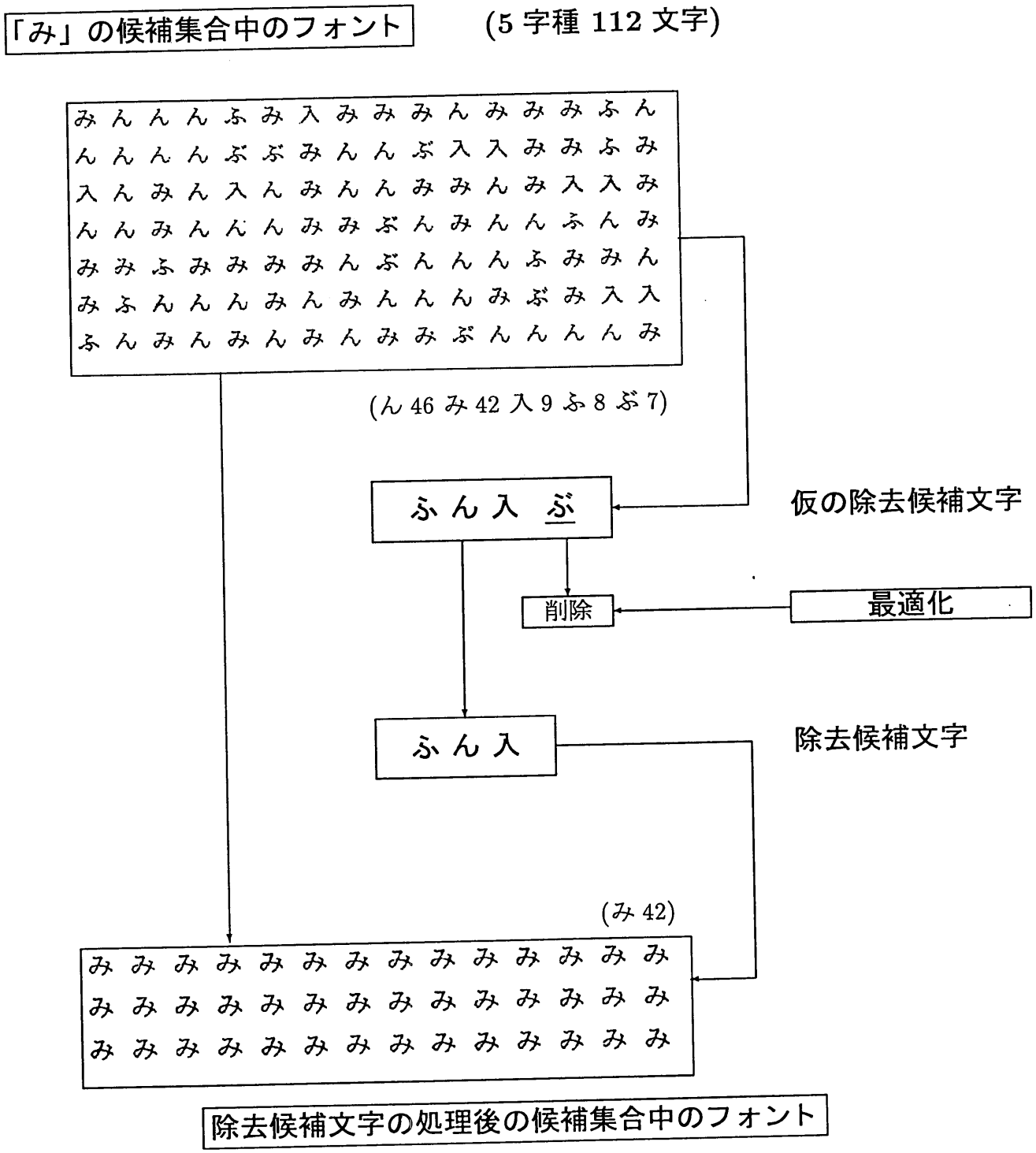


図 3.9: 除去候補文字の学習例 1

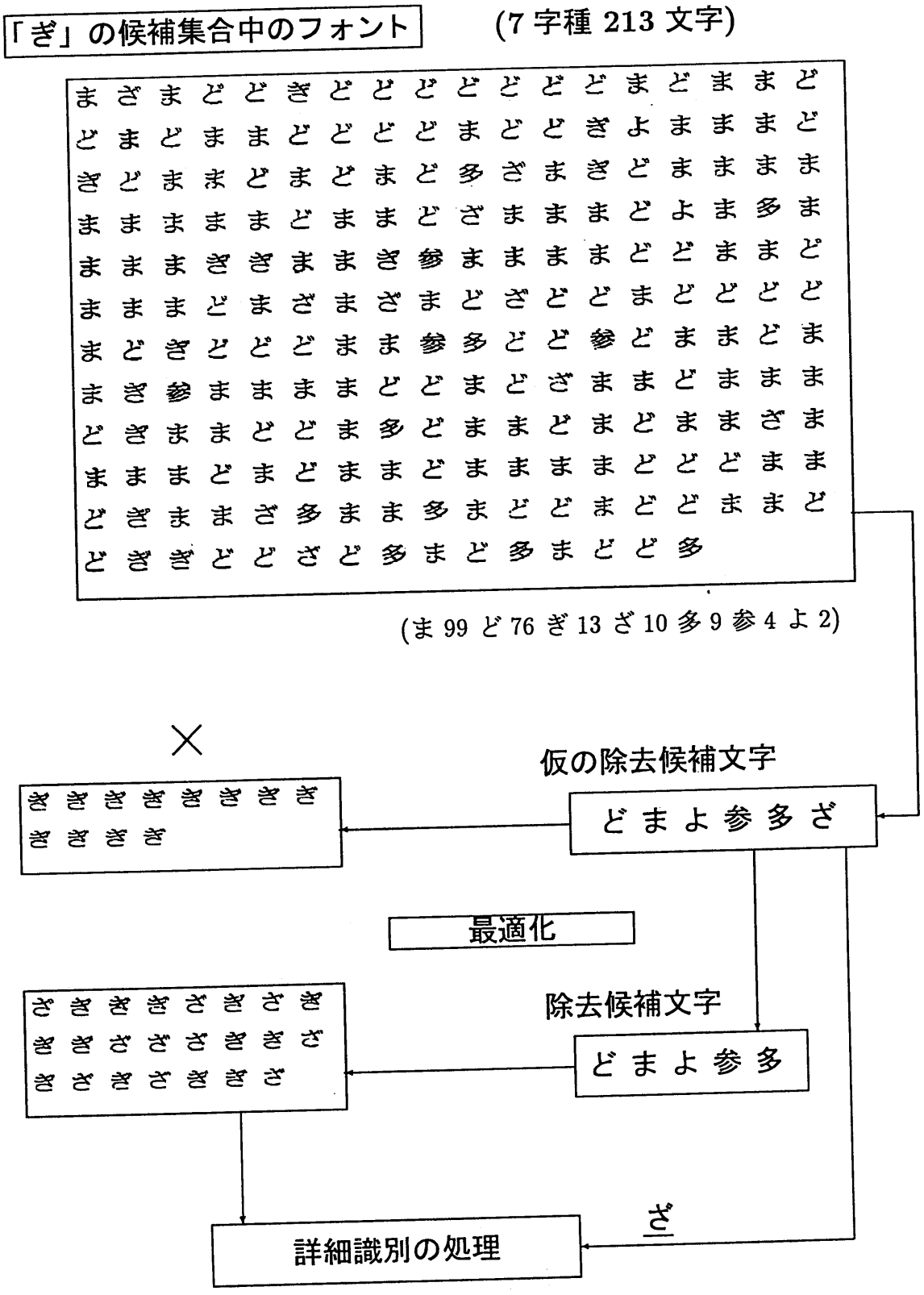


図 3.10: 除去候補文字の学習例 2–1

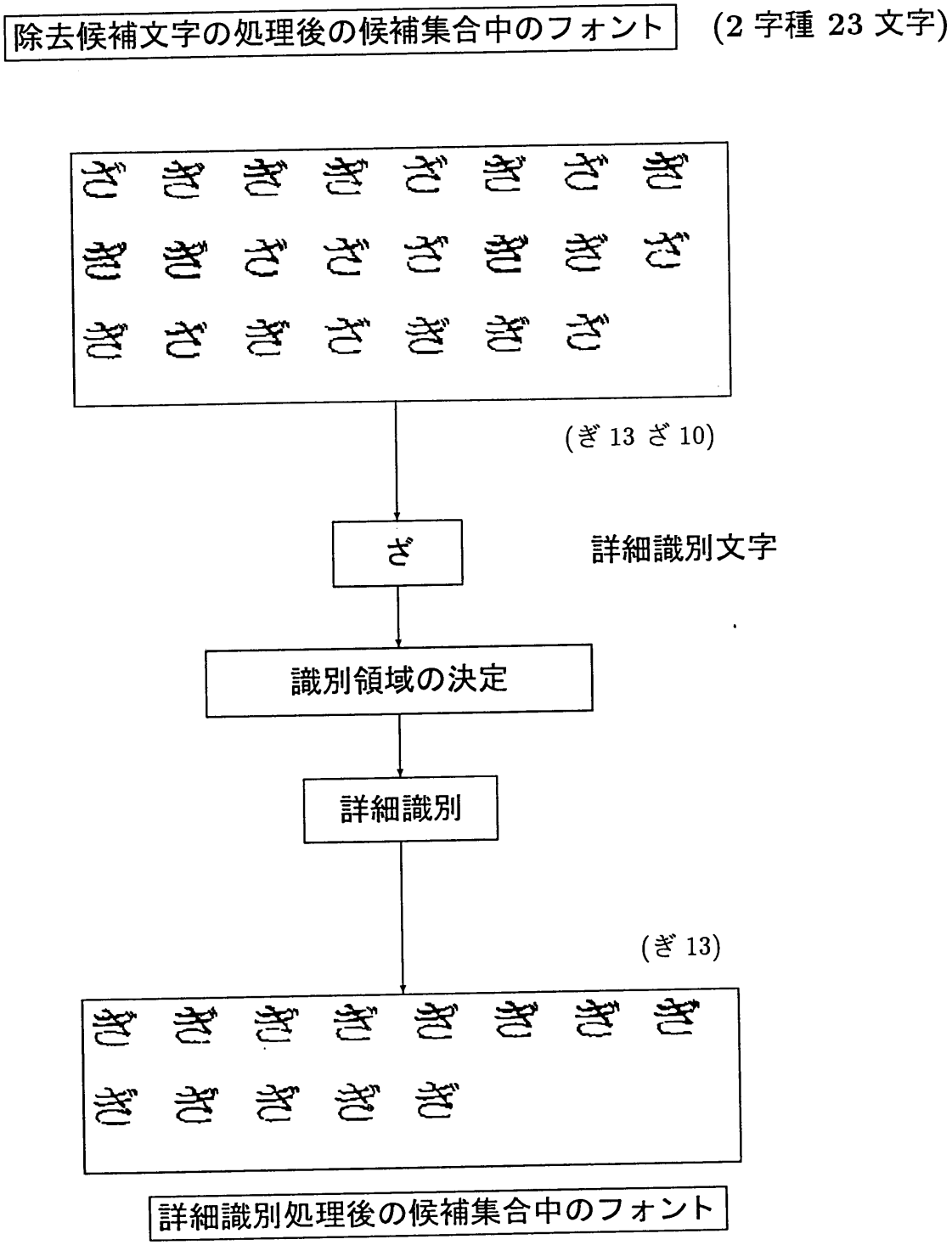


図 3.11: 除去候補文字の学習例 2–2

表 3.1: 除去候補文字 1

コード	除去候補文字	コード	除去候補文字
あ	お ぬ み も ん 占 方 命 力	し	じ に ら ニ レ
い	け の り	じ	け し に ら り ニ ビ リ 亡
う	つ ら る	す	ず オ キ ケ ナ 求 水 本 六
え	く そ た て ミ	ず	す オ ナ プ ボ 求 歩
お	あ ね も 分	せ	ぜ サ
か	が な や ケ ル	ぜ	せ だ ど ゼ
が	か だ べ や ケ ト ボ	そ	ぞ る ろ を
き	さ ま を 吉 参 表	ぞ	ぐ そ で み ろ ゼ デ 守 分
ぎ	ど ま よ 参 多	た	え だ な に わ セ
く	え ぐ 今	だ	が ぜ た づ て な に も わ ゼ
ぐ	く へ イ ウ ジ 一	ち	お た も ら を
け	げ に び ぼ ゆ	つ	う ぐ づ で の コ
げ	は び 庁	づ	う が ぐ つ ウ ガ ゴ ゼ 分
こ	ご と に る エ 二	て	く ぞ つ エ マ 下
ご	ざ じ だ ど ゴ ビ ユ ン 二	で	ぞ つ て の や フ 及
さ	う き ざ と も を 台		
ざ	ぎ す ど ぶ を ゴ		

表 3.2: 除去候補文字 2

コード	除去候補文字	コード	除去候補文字
と	ど ら 公 上 占	ぼ	け に ば ほ 庁
ど	じ と ら ビ	ま	す 安 拳 主 全 歩
な	す た を ウ 女 珍 立	み	ふ ん 入
に	い け し た 広 仁	む	だ な も れ わ 以 位 心 打
ぬ	ね め 応 山 内 約	め	あ い ぬ の み わ カ 小 心
ね	お か む め も れ 付 料	も	ち む ら 合
の	い つ み め り 山	や	み ネ ヤ 一
は	な に ば ほ	ゆ	が の や 中 仲
ば	げ な は び ぼ ゆ ビ 体	よ	ま エ 少 上 歩
ひ	い け し な び ユ レ 以	ら	し ち ふ も る 一 各
び	が げ ひ ゆ ブ	り	い う し つ の む ゆ ウ リ ロ
ふ	ぶ よ ん ム ン	る	ち ろ ミ 各 石
ぶ	ふ よ ん ム 小	れ	か な ん 介 九
へ	な べ ん レ 外 人	ろ	あ う そ ち み る
べ	ぐ な へ ん イ ド バ	わ	あ お か み も れ ん 心 分
ほ	に は ぼ 行 低	を	あ さ ち と ら る 去 合
		ん	た ふ み ル

表 3.3: 除去候補文字数と評価値

文字数	用いる評価値	
	ユークリッド距離	重み付き
2	せ	
3	い く み	う げ
4	ぜ は も や ん	お そ ど
5	え か ぎ し ち ふ ぶ ほ ぼ ゆ よ れ	け と び る
6	ぐ ざ た つ に ぬ の へ ま	き こ て ろ
7	が ず な べ ら	さ で
8	ば を	ね ひ
9	あ じ す ぞ づ む め	ご わ
10	だ り	

識別文字	識別領域	識別文字	識別領域
う　ら	6　13	う　ラ	0　0
ぎ　ざ	18　19	こ　ご	13　48
ざ　さ	13　48	し　リ	13　20
そ　モ	2　3	だ　た	18　18
て　で	19　20	ね　れ	39　46
ね　わ	38　39	は　け	39　40
は　ゆ	27　34	ば　は	6　6
ひ　か	34　41	ぶ　ふ	6　13
ほ　ぼ	6　13	わ　れ	46　47

図 3.12: 詳細識別文字

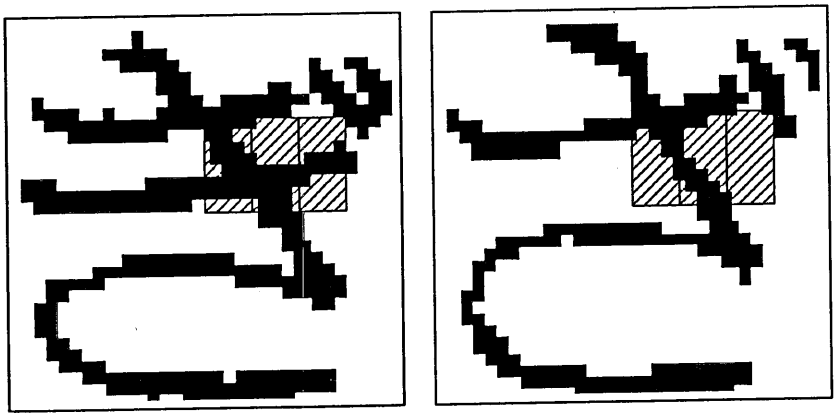


図 3.13: 識別領域の例

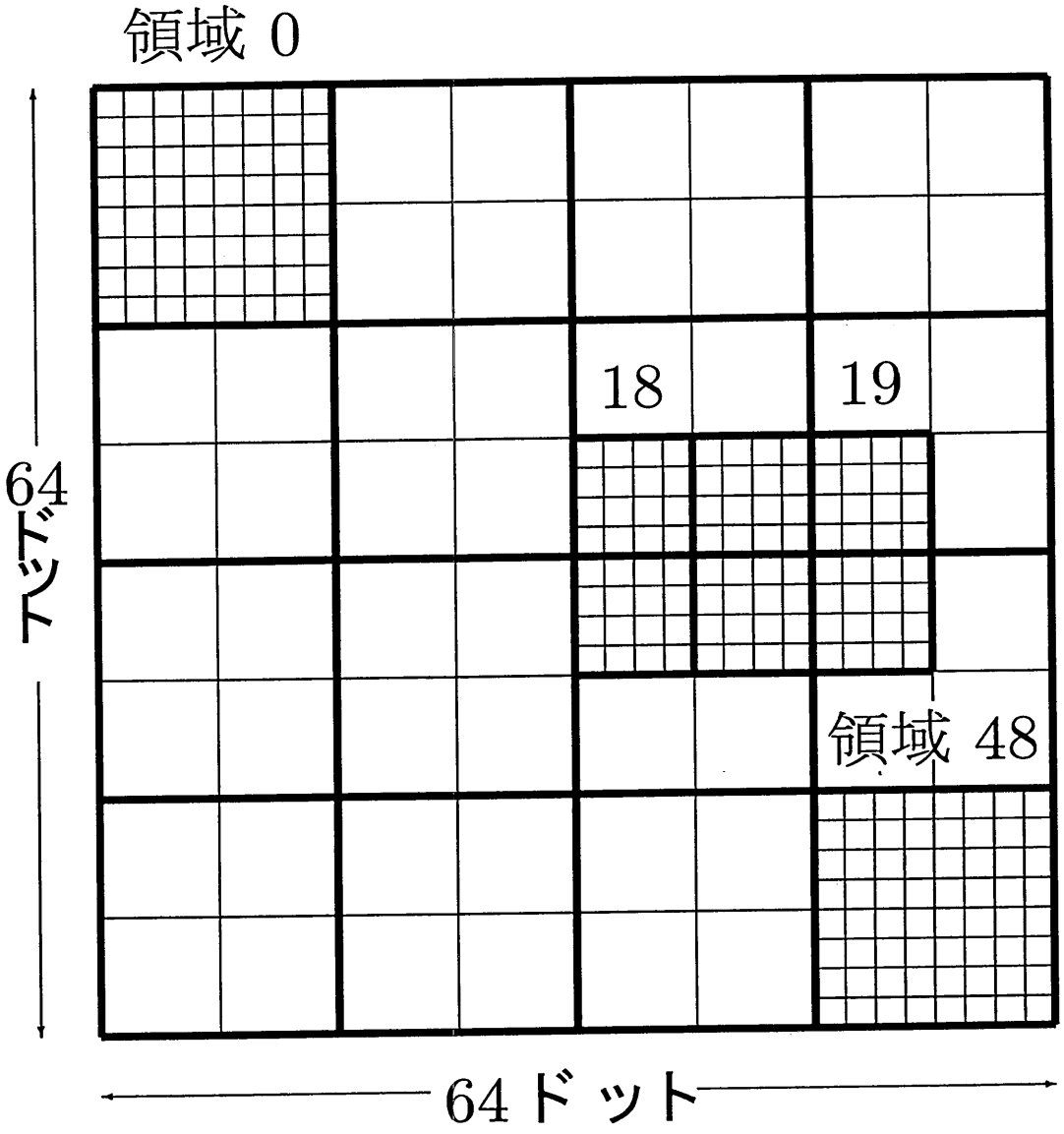


図 3.14: 識別領域

表 3.4: 候補数と候補集合

候補数	A	B	C(%)	D(%)
1	10131	9958	**96.42**	98.29
2	16391	10264	**99.38**	62.62
3	21898	10316	**99.88**	47.11
4	25904	10323	**99.95**	39.85
5	29296	10325	**99.97**	35.24
6	32150	10328	**100.00**	32.12
7	34984	10328	**100.00**	29.52
8	38006	10328	**100.00**	27.17
9	40905	10328	**100.00**	25.25
10	43390	10328	**100.00**	23.80

A:全候補集合のデータ数 B:正解数

C: B/全正解数 (10328) D: B/A

表 3.5: 候補数と出力数

候補数	出力数			
	A	B	C	合計
1	533(389)	**0**	0	533
2	568(449)	**1**	0	569
3	572(465)	**1**	0	573
4	574(469)	**1**	0	575
5	574(472)	**1**	0	575
6	575(473)	**1**	0	576
7	575(473)	**1**	0	576
8	575(471)	**1**	0	576
9	575(471)	**1**	0	576
10	575(471)	**3**	0	578

A:全て正解が出力 B:全て不正解を出力
C:正解と不正解の混在を出力
() は文書中の全正解を出力した数
※期待される真の出力数は 575

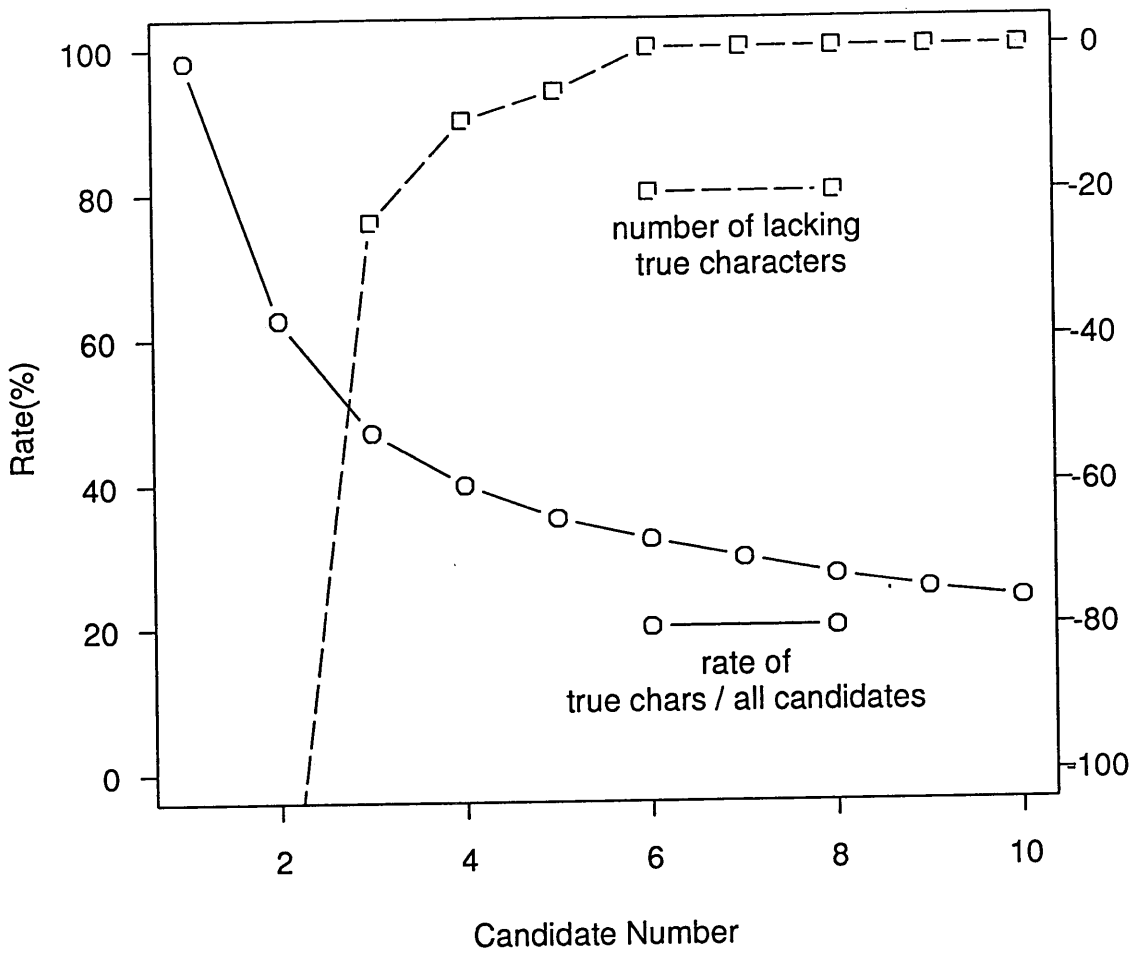
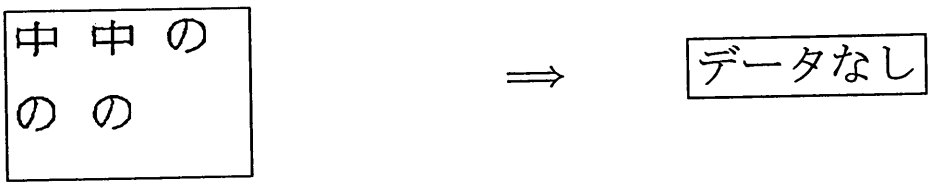


図 3.15: 候補数と出力数

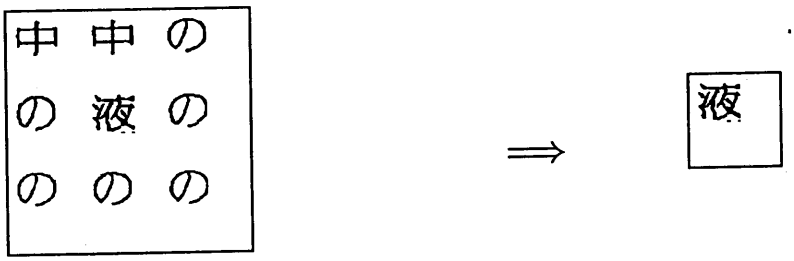
表 3.6: 誤抽出された文字

文書	候補数	抽出すべき文字	出力数	誤抽出した文字
11	10	ゆ	1	液 (1)
13	10	ぬ	1	点 (1)
14	2	ぜ	1	ザ (1)
14	3	ぜ	1	ザ (1)
14	4	ぜ	1	ザ (1)
14	5	ぜ	1	ザ (1)
14	6	ぜ	1	ザ (1)
14	7	ぜ	1	ザ (1)
14	8	ぜ	1	ザ (1)
14	9	ぜ	1	ザ (1)
14	10	ぜ	1	ザ (1)

候補集合　　除去候補文字の処理後



抽出成功の例 (候補数 9)



誤抽出の例 (候補数 10)

図 3.16: 誤抽出の例

表 3.7: 抽出できなかった文字

文書	候補数	抽出できなかった文字
11	1	ぎ　ば　び　わ
12	1	ぎ　ば　び　ゆ　わ
12	2	ゆ
12	3	ゆ
13	1	さ　ば　び　わ
14	1	ぎ　ご　ば　び　わ
14	2	ば
15	1	ぎ　ば　わ
16	1	ぎ　ご　ざ　ば　び　わ
16	2	ぎ　ば
17	1	ぎ　ざ　わ
18	1	ぎ　ご　ば　ば　わ
18	2	ご
18	3	ご
19	1	ば　ゆ　わ
19	2	ゆ
19	3	ゆ
19	4	ゆ
19	5	ゆ
20	1	ぎ　ば　び　わ
20	2	ぎ

第 4 章

認識実験

4.1 前書き

抽出されたフォントを用いて認識実験を行う。この認識実験の結果により、本研究で提案した文書認識システムの有効性を確認する。

4.2 実験方法

3章で抽出されたフォントの重心ベクトルと既存辞書から新辞書を以下に示す2つの方法で作成し、認識実験を行う。実験は次の3種類について行った。

1. クローズ実験
2. オープン実験
3. 更新した辞書による認識実験

また比較のために3章でフォントの抽出に用いた辞書(既存辞書)を用いても認識実験を行う。認識方法は評価値としてユークリッド距離を用いた全数整合法によって行う。

4.3 新辞書の作成方法

社説からフォントを抽出し作成した辞書を以降は抽出辞書と呼ぶことにする。新辞書は、抽出辞書と既存辞書から作成されることになる。

4.3.1　マルチテンプレート法

既存辞書に、抽出辞書を付加して新辞書を作成する。フォントが抽出されなかった文字に関しては、全て 0 の値を持つベクトルを代用する。この新辞書は、最大でひらがな 65 文字に対しテンプレート数 2 のマルチテンプレートの辞書となる。辞書の字数は 3109+65 で 3174 文字になる。

4.3.2　置換法

フォントを抽出したコードの抽出辞書を既存辞書のものと置換して新辞書を作成する。抽出されなかった文字は既存辞書のベクトルをそのまま用いる。辞書の字数は 3109 文字のままである。

4.4　クローズ実験

フォントの抽出に用いた新聞社説 (11〜20) の 10 部について、抽出したフォントから作成した新辞書を用いて認識実験を行う。フォントの抽出の際には候補数を 1〜10 まで変えたものを用いる。

フォントを抽出した文書を、その抽出したフォントから作成した新辞書を用いて認識するので、再認識することになり、誤認識の自動修正の有効性を確認する実験となる。

表 4.1に既存辞書を用いた認識結果を、表 4.2にクローズ実験の 1 位の認識結果を示す。また、表 4.3、表 4.4に従来法と比較して、クローズ実験でどれだけ認識結果が改善されたかを示す。マルチテンプレート法、置換法ともに候補数 6 以上ではフォント抽出の対象であるひらがなの 65 種に対して、従来法では誤認識してしまった 370 文字を全て正しく認識している。また、表 4.2を見ると候補数 6 でひらがな 65 種に対する認識では、マルチテンプレート法で 100%、置換法でも 99.99%の 1 位認識率が得られている。3章の結果とも合わせて、以降の実験では候補数を 6 に固定して実験を行うことにした。

4.5　オープン実験

フォントの抽出に用いた新聞社説 (11〜20) の 10 部について、認識する文書以外の 9 つの文書からそれぞれ作成した 9 つの新辞書を用いて認識実験を行う。

この実験で 1 部の文書からフォントを抽出して作成した新辞書に、他の文書の認識に対する有効性を確認する。この有効性が確認できれば、認識する文書全てにフォント抽出・

再認識 (クローズ実験) を行わなくても、1 部の文書でフォントの抽出を行い、それから作成した新辞書を用いてパターンマッチング法のみで、クローズ実験程度の認識結果が得られることになる。

以降の実験では、抽出辞書を抽出する文書と、認識を行う文書が異なる。そこで表 4.5に本研究で認識の対象とした社説 (11〜30) の 20 部の文書に存在しないひらがな 65 種を示しておく。

表 4.6、表 4.7にマルチテンプレート法と置換法による 1 位認識結果を示す。表中の A の欄は前節のクローズ実験の結果である。B 欄では、9 つの新辞書を用いた認識結果のうち一番悪かったものと一番良かったものを、C 欄では 9 つの新辞書の平均認識率を示した。また表 4.8、表 4.9には、フォントを抽出した文書毎の 1 位認識率を示した。

表 4.1と表 4.6、表 4.7とを比較すると、従来法の 3 位認識率と、オープン実験の平均の 1 位認識率がほぼ同じになる。一番悪いもので比較すると従来法の 2 位認識率よりも劣るものもある。これは認識文書中の対象文字と、辞書を抽出した文書とが一致していないためで、クローズ実験と同程度の認識結果が得られたのはいずれも、認識文書中の対象文字を全て含んだ抽出辞書を用いた場合であった。この様に単に社説 1 部から作成した新辞書を用いてパターンマッチング法のみで他の文書を認識した際には、従来法よりは良い結果が得られるが、クローズ実験にまでは及ばなかった。

4.6　更新した辞書による認識実験

4.5節の結果より、社説 1 部だけから作成した新辞書を用いた場合、その認識結果は、抽出に用いた社説中の文字の内容に依存してしまい、他の文書の認識に用いた場合にクローズ実験と同程度の結果が必ずしも得られないことがわかった。そこで、フォントの抽出用の社説 (11〜20) の 10 部を用いて抽出辞書を更新し、抽出辞書を十分に学習させて、他の社説 (21〜30) の 10 部に対して認識実験を行った。次に辞書の更新法を説明する。

4.6.1　辞書の更新方法

抽出辞書を更新させる場合には、学習させる順番が問題になるが、今回は社説の内容は考慮せず、社説の番号が若い順に学習させ更新することにした。更新に用いた社説数を学習文書数とする。例えば、社説 11 と社説 12 で抽出辞書を更新・作成した場合は学習文書数 2 とする。社説 11〜20 まで全て用いた場合は学習文書数 10 となる。学習文書数と、抽出された文字数との関係は、図 4.6、表 4.10に示す。これらから、今回用いた更新法では、

学習文書数 5 で、ひらがな 65 種の全ての抽出辞書を獲得できたことがわかる。

- 社説 (11〜20) からフォントを抽出し、各社説毎に重心ベクトルを作成する。
- 社説の番号の若い順に、フォントが抽出された場合、その重心ベクトルを抽出辞書とする。
- 以後の社説でフォントが抽出された場合、抽出数が現在の抽出辞書のものより多い場合はその社説から抽出した重心ベクトルを抽出辞書とする。

$か_n$が社説 n から抽出されたフォントの重心ベクトルを表すものとして図 4.1で説明する。最初は社説 11 から抽出された 19 個のフォントの重心ベクトルが抽出辞書になるが、次の社説 12 では 34 個が抽出されたので抽出辞書が$か_{12}$に更新される。

ただし、現在の抽出辞書の抽出数が 1 のときに、次の社説の抽出数が 1 の場合は、現在の抽出辞書と、抽出されたベクトルの重心ベクトルを抽出辞書とする。これは 3章での結果より、抽出数が 1 の場合、誤抽出している可能性がある。このために抽出数が少ない場合には抽出辞書に用いない方がより正確な抽出辞書を作成できるが、出現頻度の低い文字が正確に抽出されている場合も抽出辞書が作成できなくなるという弊害が生じる。そこで、抽出数が 1 のときは上記の処理を行う。ただし、この後に 2 以上の抽出数が得られた場合には、通常通りにその重心ベクトルが抽出辞書になる。図 4.2中の $G(A,B)$ は、A と Bの重心ベクトルを示している。文書 12 でフォントの抽出があり、抽出数が 1 なのでこのベクトル ($ぼ_{12}$) がそのまま抽出辞書になる。次に抽出があるのは文書 16 だが、抽出数が 1 であるので、この $ぼ_{16}$ と現在の抽出辞書の $ぼ_{12}$ の重心ベクトル $G(ぼ_{16}, ぼ_{12})$ が新しい抽出辞書になる。

4.6.2　実験結果

1 位認識率は表 4.12に、学習文書数毎の 1 位認識率を表 4.13、表 4.14に示す。比較のために従来法による社説 (21〜30) までの認識結果を表 4.11に示す。

マルチテンプレート法は学習文書数 6 で 99.99%、置換法は学習文書数 4 以上で 100%の 1 位認識率を達成できた。これはクローズ実験と同程度の結果であり、この辞書の更新による学習法が有効であると考えられる。

4.7　考察

初めに、3 つの実験で正解を得られなかった文字について詳しく解析してみる。その理由は大きく次の 3 つに大別できる。

1. 正解のフォントが抽出されない場合
   社説のフォントが抽出できないため、従来法で誤認識した文字をクローズ実験でも誤認識する例や、フォントを誤抽出し、誤って抽出したフォントのコードに誤認識する例がクローズ実験の候補数が少ない場合にみられた。

2. 正解のフォントが全て抽出していない場合
   クローズ実験で誤認識した場合を例として挙げる。誤認識した文字は「ば」で、その社説中には 2 つの「ば」が存在していて、1 つは正しく認識できたが他方は誤認識した。便宜上誤認識した方のベクトルを $\mathbf{ば_f}$、認識できた方のを $\mathbf{ば_t}$ で表す。$\mathbf{ば_f}$ は候補数 1 で候補集合に取り入れられたが詳細識別で除去されている。一方の $\mathbf{ば_t}$ は候補数 3 で候補集合に取り入れられ、そのまま $\mathbf{ば_t}$ だけが抽出されており、候補数 2 までは抽出がなく、候補数 3 以降では $\mathbf{ば_t}$ が抽出辞書として用いられることになる。$\mathbf{ば_f}$ との評価値は

   $$E(ば_f, ば_f) < E(ば_f, ば_s) < E(ば_f, ぱ_s) < E(ば_f, ば_t) \quad (4.1)$$

   の順となる。このため候補数 2 までででは従来の既存の辞書のベクトルで正解が得られているが、候補数 3 以降での置換法では $\mathbf{ば_s}$ が $\mathbf{ば_t}$ に更新されてしまい、誤認識している。また辞書の更新実験で不正解になった文字の場合は、逆にマルチテンプレート法でだけ誤認識している。これは「し」で、抽出辞書を $\mathbf{し_n}$ と表すと、

   $$E(し_f, じ_s) < E(し_f, し_s) < E(し_f, し_n) < E(し_f, じ_n) \quad (4.2)$$

   の順になるため誤認識した例である。この場合でも、抽出辞書には文書中の全ての「し」のフォントが抽出されないため、正しく認識できなかったものである。これらを正しく認識させるには、類似文字と詳細識別類似文字の学習、さらに細分類での閾値の値を十分に検討しなければならない。

3. 類似文字も対象の文字の場合
   「わ」が「ね」に誤認識する場合の例である。「わ」と「ね」は相互も他方の詳細識別類似文字になっている。このため「わ」が抽出されず「ね」だけが抽出されている場合に評価値は

$$E(わ_f, ね_s) < E(わ_f, わ_s) < E(わ_f, ね_n) \tag{4.3}$$

の順になり、置換法の場合は $ね_s$ が $ね_n$ に更新されているので正解を得るが、マルチテンプレート法では不正解となってしまう。

次に、誤って抽出された「液」、「点」、「ザ」からそれぞれ作成された「ゆ」、「ぬ」、「ぜ」の抽出辞書を用いた認識実験での影響について述べる。「ザ」を除く 2 つの場合は候補数 10 のときに誤抽出しているため、クローズ実験でしか評価できず、それぞれの社説中には「ゆ」、「ぬ」が存在していないため、従来法では正しく認識できた「液」、「点」をマルチテンプレート法、置換法とも「ゆ」、「ぬ」に誤認識している。一方、「ザ」の場合は社説 (11〜20) 中に「ザ」は誤抽出されたものだけで、「ぜ」は 9 文字が存在していた。この 9 文字に対する認識結果ではマルチテンプレート法は全てを正解し、置換法では 9 文字中 4 文字を「せ」と誤認識していた。この様に、マルチテンプレート法では誤抽出した場合でも既存辞書のベクトルを認識に使用できるので救済できるが、置換法では既存辞書のベクトルを削除するため、実際は「ザ」であるものを「ぜ」の標準パターンとして用いなければならなくなり、誤認識が生じる。しかし、9 文字の「ぜ」を全て誤認識するのではなく、5 文字がこの条件で正解を得たということは、同じフォントの類似文字 (除去候補文字) が認識に及ぼす影響が大きいことを示している。

続いて、フォント抽出の対象ではない文字に対する影響について考察する。今回の実験では、フォント抽出の対象文字 (ひらがば 65 種) の認識精度を従来法と比較して向上させることが目的であった。実際、1 位認識率で比較すれば従来法よりも認識精度は上がっているが、対象文字以外の認識結果が従来法よりも大きく下回れば問題になる。従来法で正解だったもののうち、今回の実験で不正解になったものは、マルチテンプレート法、置換法ともクローズ実験では 4 字 (社説 10 部中)、オープン実験では 25 字 (社説 10 部×9 セット中) であった。辞書の更新による実験では、どちらも従来法と同じ認識結果になった。誤認識した理由は、フォントの抽出の失敗によるものが一部で、ほとんどはひらがな 65 種の除去候補文字になっているものが、そのひらがなと誤認識してしまっている。カタカナの「リ」がひらがなの「り」に誤認識するのがこの例になる。この結果より、抽出対象とする文字の拡張を検討しなければならない。特に、対象文字の除去候補文字・詳細識別文字となっている文字を抽出の対象とすることで、対象文字以外への影響を緩和できると思われる。

今迄述べてきたように誤認識した文字はフォントの抽出に何らかの問題があると考えられる。今回用いたフォントの抽出法は 3の考察でも触れたが、全てのフォントを正確に抽出するには至っていない。候補集合に選出された全ての正解のフォントを、全て抽出して

いないものが全体の 2 割あり、これらから作成した抽出辞書で誤認識したものが数例確認された。特に「し」と「じ」の様な類似文字の認識の際に、この影響が顕著であり、このままでは抽出辞書を用いてもパターンマッチング法のみでは完全に類似文字の識別を行うことは難しい。正確にフォントの抽出を行うために、除去候補文字・詳細識別文字、細分類での閾値を検討しなければならない。

また、辞書の作成法に関しては、マルチテンプレート法と、置換法で対象文字に関して認識結果に差が出たのもフォントの抽出の影響であると考えられる。上記したフォントの抽出法の問題点が解決し、実際に文書認識を行う際には、誤認識の影響を考え、マルチテンプレート法が有効であると考えられる。

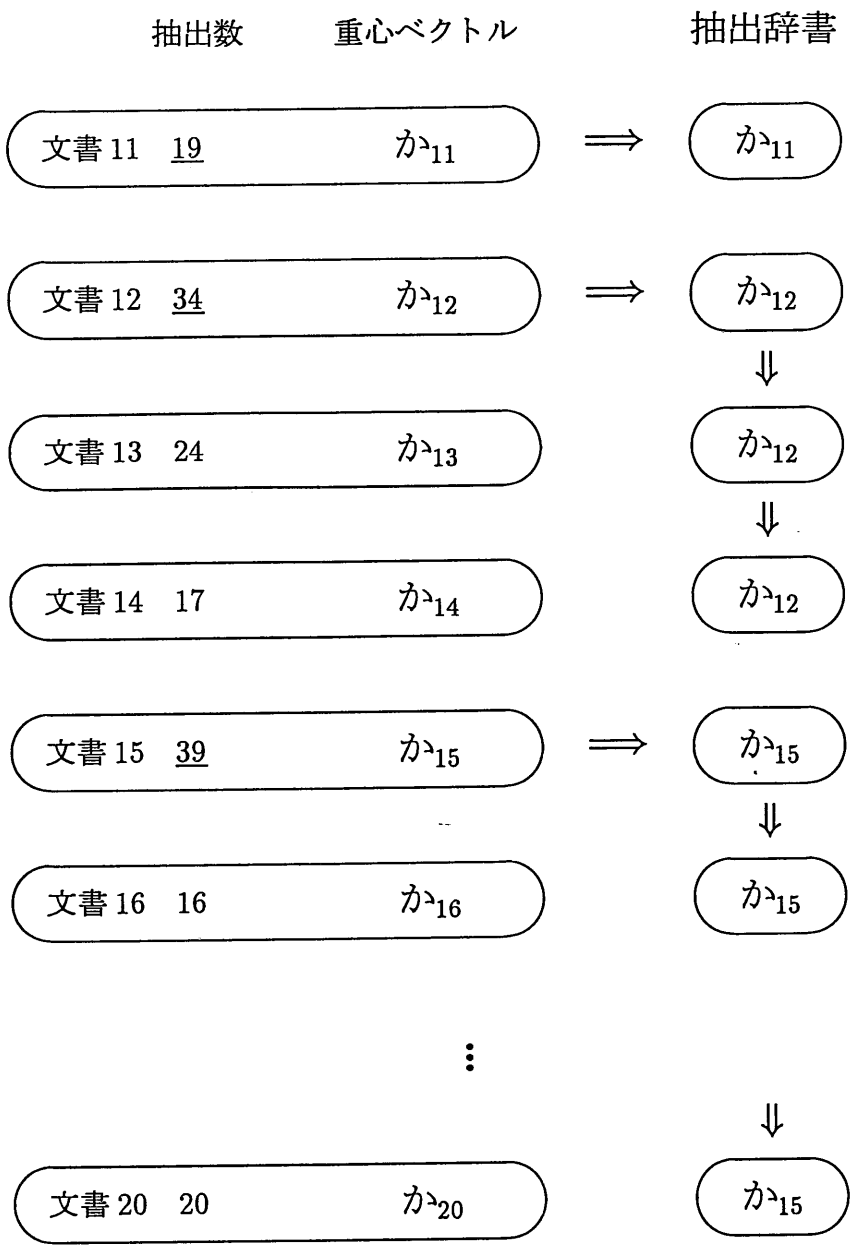


図 4.1: 辞書の更新の例 1

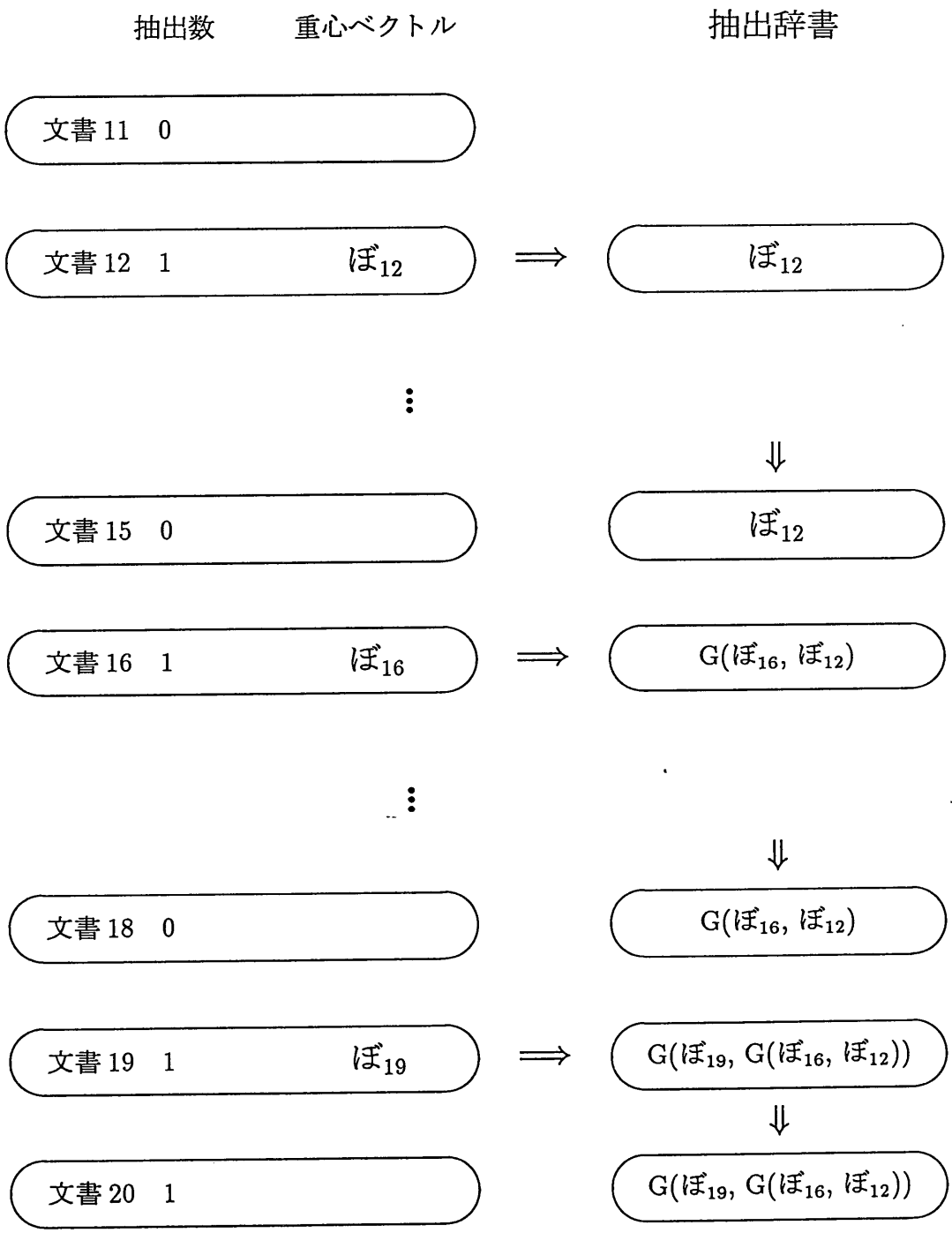


図 4.2: 辞書の更新の例 2

表 4.1: 従来の辞書の認識率 (社説 11〜20)

社説	ひらがな 65 種		
	1 位認識率 (%)	2 位認識率 (%)	3 位認識率 (%)
平均	96.34 (37)	99.37 (6)	99.88 (1)
11	96.77 (34)	99.71 (3)	99.90 (1)
12	95.90 (43)	99.33 (7)	99.81 (2)
13	96.18 (34)	99.10 (8)	99.89 (1)
14	95.80 (41)	99.18 (8)	99.90 (1)
15	96.55 (39)	99.29 (8)	99.91 (1)
16	96.50 (35)	99.50 (5)	100.00 (0)
17	95.85 (45)	99.45 (6)	100.00 (0)
18	97.07 (32)	99.63 (4)	99.82 (2)
19	96.46 (36)	99.51 (5)	99.90 (1)
20	96.44 (31)	98.85 (10)	99.66 (3)

社説	その他		
	1 位認識率 (%)	2 位認識率 (%)	3 位認識率 (%)
平均	93.11 (70)	97.97 (21)	98.86 (12)
11	91.83 (83)	97.93 (21)	98.82 (12)
12	93.42 (68)	97.87 (22)	98.55 (15)
13	93.20 (67)	98.17 (18)	98.78 (12)
14	94.42 (59)	97.73 (24)	99.05 (10)
15	92.23 (74)	97.69 (22)	98.53 (14)
16	92.03 (80)	97.81 (22)	99.30 (7)
17	93.97 (62)	98.93 (11)	99.61 (4)
18	93.62 (64)	97.81 (22)	98.50 (15)
19	92.71 (76)	97.70 (24)	98.66 (14)
20	93.56 (63)	98.06 (19)	98.77 (12)

() の内は誤認識数

表 4.2: 1 位認識率 (クローズ実験)

候補数	マルチテンプレート法	
	ひらがな 65 種 (%)	その他 (%)
1	94.46 (157)	93.10 (697)
2	99.90 (10)	93.09 (698)
3	99.96 (6)	93.08 (699)
4	99.99 (1)	93.08 (699)
5	99.99 (1)	93.08 (699)
6	100.00 (0)	93.08 (699)
7	100.00 (0)	93.08 (699)
8	100.00 (0)	93.08 (699)
9	100.00 (0)	93.08 (699)
10	100.00 (0)	93.05 (702)

候補数	置換法	
	ひらがな 65 種 (%)	その他 (%)
1	99.09 (92)	93.09 (698)
2	99.90 (10)	93.08 (699)
3	99.95 (5)	93.07 (700)
4	99.98 (2)	93.07 (700)
5	99.98 (2)	93.07 (700)
6	99.99 (1)	93.07 (700)
7	99.99 (1)	93.07 (700)
8	99.99 (1)	93.07 (700)
9	99.99 (1)	93.07 (700)
10	99.99 (1)	93.07 (700)

() の内は誤認識数

表 4.3: 改善の結果 (マルチテンプレート法)

	全体		ひらがな 65 種				その他			
候補数	A	C	A	B	C	D	A	B	C	D
1	213	2	213	1	0	33	0	0	2	0
2	360	3	360	0	0	7	0	1	3	0
3	366	4	366	0	0	4	0	1	4	0
4	369	4	369	0	0	1	0	1	4	0
5	369	4	369	0	0	1	0	1	4	0
6	370	4	370	0	0	0	0	1	4	0
7	370	4	370	0	0	0	0	1	4	0
8	370	4	370	0	0	0	0	1	4	0
9	370	4	370	0	0	0	0	1	4	0
10	370	7	370	0	0	0	0	0	7	0

A:誤認識が正しく認識された数

B:A 以外で認識順位が上がった数

C:正解から誤認識へ変った数

D:C 以外で認識順位が下がった数

表 4.4: 改善の結果 (置換法)

	全体		ひらがな 65 種				その他			
候補数	A	C	A	B	C	D	A	B	C	D
1	278	2	278	5	0	12	0	0	2	0
2	360	4	360	4	0	1	0	1	3	0
3	366	5	366	3	1	0	0	1	4	0
4	369	5	369	1	1	0	0	1	4	0
5	369	5	369	1	1	0	0	1	4	0
6	370	5	370	0	1	0	0	1	4	0
7	370	5	370	0	1	0	0	1	4	0
8	370	5	370	0	1	0	0	1	4	0
9	370	5	370	0	1	0	0	1	4	0
10	370	8	370	0	1	0	0	1	7	0

A:誤認識が正しく認識された数

B:A 以外で認識順位が上がった数

C:正解から誤認識へ変った数

D:C 以外で認識順位が下がった数

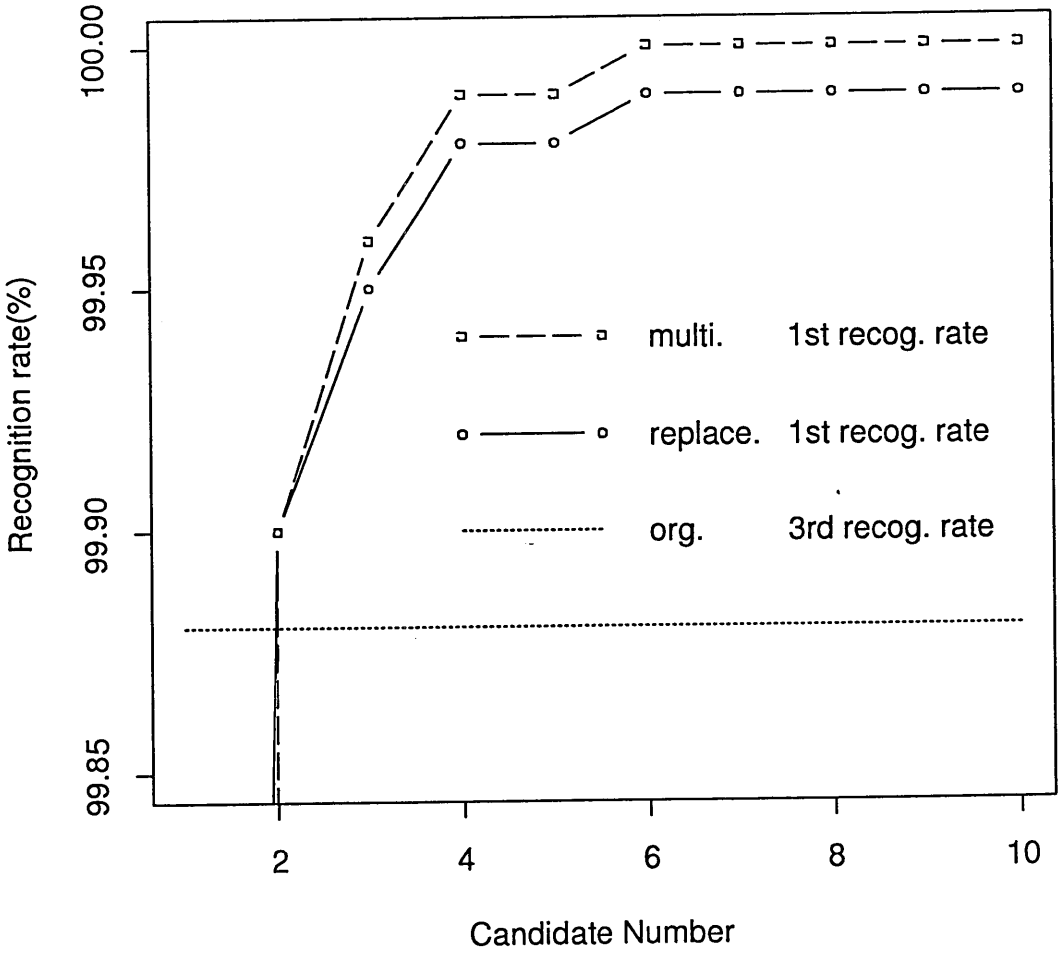


図 4.3: 候補数と認識率 (ひらがな 65 文字)

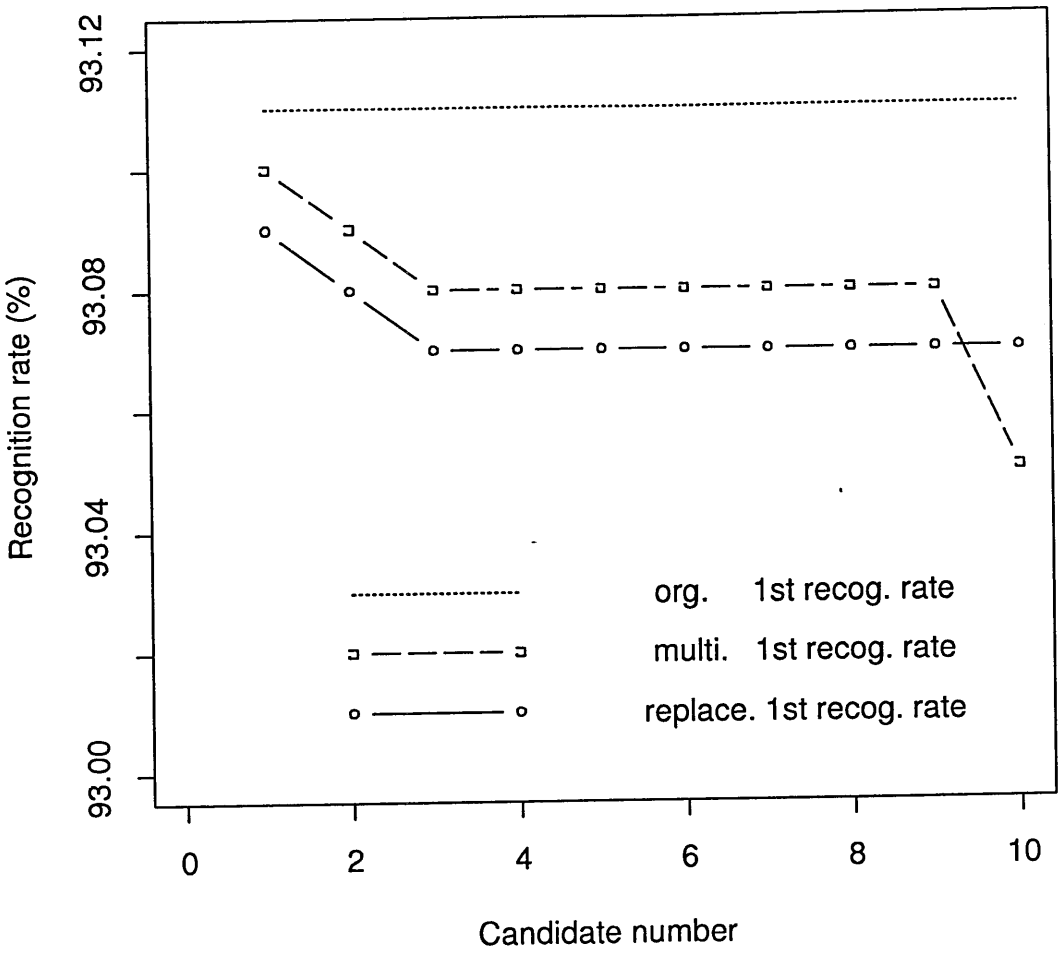


図 4.4: 候補数と認識率 (その他)

表 4.5: 認識文書中に存在しないひらがな 65 種の文字

社説	ひらがな 65 種で文書中に存在しない文字
11	ご ざ ぞ へ ぼ ゆ
12	ご ざ ぜ ぞ づ ね ひ ふ
13	お ぎ げ ご ぜ ぞ づ ぬ ね ふ ぼ ゆ
14	ぐ じ ぜ ぞ ぬ ふ ぶ ぼ
15	ぬ ふ ぼ
16	げ ぜ ぞ ぬ ゆ
17	げ ご ぞ づ び ふ ぶ ぼ ゆ
18	ざ ぜ ぞ ふ ほ ぼ ゆ
19	ぎ ご ぜ ぬ ひ ほ む ん
20	ご ぜ ぞ ぬ ね ほ ぼ む ゆ
21	ご ざ ぜ ぞ ぬ ね ひ ふ ぶ ほ ゆ
22	ぐ ぞ ひ ふ ぶ ほ ぼ
23	お げ ご ぜ ぬ ひ ぼ ゆ
24	ぜ ぞ ぬ ひ び ふ ほ
25	ざ ぜ ぬ ひ ふ ぶ む ろ
26	ぐ ご ぞ ぬ ね び ぶ む ゆ ん
27	ぞ ぬ ゆ
28	ざ せ づ ぬ ね ひ び ふ ぼ ゆ
29	ご ざ ぜ ぞ ね び む ゆ
30	お ぎ ご ぜ ぞ ぬ ひ へ

表 4.6: マルチテンプレート法の 1 位認識率 (オープン実験)

社説	ひらがな 65 種			その他		
	A	B	C	A	B	C
11	100.00	99.43〜100.00	99.79	91.83	91.83	91.83
12	100.00	99.24〜100.00	99.69	93.42	93.32〜93.42	93.41
13	100.00	99.21〜100.00	99.79	93.20	93.20	93.20
14	100.00	99.49〜100.00	99.68	94.32	94.32〜94.42	94.41
15	100.00	99.47〜 99.91	99.71	92.12	92.12	92.12
16	100.00	99.60〜100.00	99.79	92.03	91.93〜92.03	92.02
17	100.00	99.63〜100.00	99.89	93.97	93.97	93.97
18	100.00	99.54〜100.00	99.81	93.72	93.62〜93.72	93.66
19	100.00	99.51〜100.00	99.74	92.71	92.71	92.71
20	100.00	99.77〜100.00	99.87	93.36	93.36〜93.46	93.42

A:認識文書から辞書を抽出 (クローズ)

B:認識文書以外から辞書を抽出したときの min〜max

C:B の平均

表 4.7: 置換法の 1 位認識率 (オープン実験)

社説	ひらがな65種			その他		
	A	B	C	A	B	C
11	100.00	99.24～100.00	99.76	91.83	91.83	91.83
12	100.00	99.24～100.00	99.67	93.42	93.32～93.42	93.41
13	100.00	99.21～100.00	99.79	93.20	93.20	93.20
14	99.90	99.49～ 99.90	99.61	94.32	94.32～94.42	94.41
15	100.00	99.20～ 99.91	99.68	92.12	92.12	92.12
16	100.00	99.65～100.00	99.79	92.03	91.93～92.03	92.02
17	100.00	99.54～100.00	99.87	93.97	93.97	93.97
18	100.00	99.54～100.00	99.82	93.62	93.62～93.72	93.67
19	100.00	99.51～100.00	99.72	92.71	92.71	92.71
20	100.00	99.77～100.00	99.87	93.36	93.36～93.46	93.42

A:認識文書から辞書を抽出 (クローズ)

B:認識文書以外から辞書を抽出したときの min～max

C:B の平均

表 4.8: 抽出フォント毎の 1 位認識率 (マルチテンプレート法)

認識社説	フォント抽出社説									
	11	12	13	14	15	16	17	18	19	20
11	100.00	100.00	99.62	99.43	100.00	99.90	99.81	99.71	99.71	99.90
	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83
12	99.81	100.00	99.24	99.81	100.00	99.62	99.52	99.71	99.71	99.81
	93.42	93.42	93.32	93.42	93.42	93.42	93.42	93.42	93.42	93.42
13	99.89	99.78	100.00	99.21	99.89	100.00	99.89	99.66	99.78	100.00
	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20
14	99.49	99.59	99.59	100.00	100.00	99.90	99.59	99.59	99.59	99.80
	94.32	94.42	94.42	94.32	94.42	94.42	94.42	94.42	94.42	94.42
15	99.73	99.82	99.56	99.47	100.00	99.82	99.47	99.73	99.91	99.82
	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12
16	99.80	99.80	99.80	99.60	100.00	100.00	99.80	99.70	99.70	99.90
	92.03	92.03	92.03	92.03	92.03	92.03	92.03	92.03	91.93	92.03
17	99.91	99.91	99.91	99.63	100.00	100.00	100.00	99.72	99.91	100.00
	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97
18	99.91	99.91	99.63	99.82	100.00	99.82	99.54	100.00	99.82	99.82
	93.62	93.72	93.72	93.62	93.72	93.62	93.62	93.72	93.62	93.72
19	99.51	99.61	99.70	99.70	100.00	99.80	99.70	99.70	100.00	99.90
	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71
20	99.89	99.89	99.77	99.89	100.00	99.89	99.77	100.00	99.77	100.00
	93.36	93.46	93.36	93.46	93.36	93.46	93.36	93.46	93.46	93.36

上段　ひらがな 65 種の 1 位認識率 (%)

下段　ひらがな 65 種以外の 1 位認識率 (%)

表 4.9: 抽出フォント毎の 1 位認識率 (置換法)

認識社説	フォント抽出社説									
	11	12	13	14	15	16	17	18	19	20
11	100.00	100.00	99.62	99.24	100.00	99.81	99.81	99.71	99.71	99.90
	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83	91.83
12	99.81	100.00	99.24	99.71	100.00	99.52	99.52	99.71	99.71	99.81
	93.42	93.42	93.32	93.42	93.42	93.42	93.42	93.42	93.42	93.42
13	99.89	99.78	100.00	99.21	99.89	100.00	99.89	99.66	99.78	100.00
	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20	93.20
14	99.49	99.49	99.59	99.90	99.90	99.80	99.49	99.59	99.49	99.69
	94.32	94.42	94.42	94.32	94.42	94.42	94.42	94.42	94.42	94.42
15	99.73	99.82	99.56	99.20	100.00	99.82	99.47	99.73	99.91	99.82
	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12	92.12
16	99.80	99.80	99.80	99.50	100.00	100.00	99.80	99.80	99.70	99.90
	92.03	92.03	92.03	92.03	92.03	92.03	92.03	92.03	91.93	92.03
17	99.91	99.91	99.91	99.54	100.00	100.00	100.00	99.63	99.91	100.00
	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97	93.97
18	99.91	99.91	99.73	99.82	100.00	99.82	99.54	100.00	99.82	99.82
	93.72	93.62	93.62	93.72	93.62	93.72	93.62	93.62	93.72	93.72
19	99.51	99.61	99.70	99.51	100.00	99.80	99.70	99.70	100.00	99.90
	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71	92.71
20	99.89	99.89	99.77	99.89	100.00	99.89	99.77	100.00	99.77	100.00
	93.36	93.46	93.36	93.46	93.36	93.46	93.36	93.46	93.46	93.36

上段　ひらがな 65 種の 1 位認識率 (%)

下段　ひらがな 65 種以外の 1 位認識率 (%)

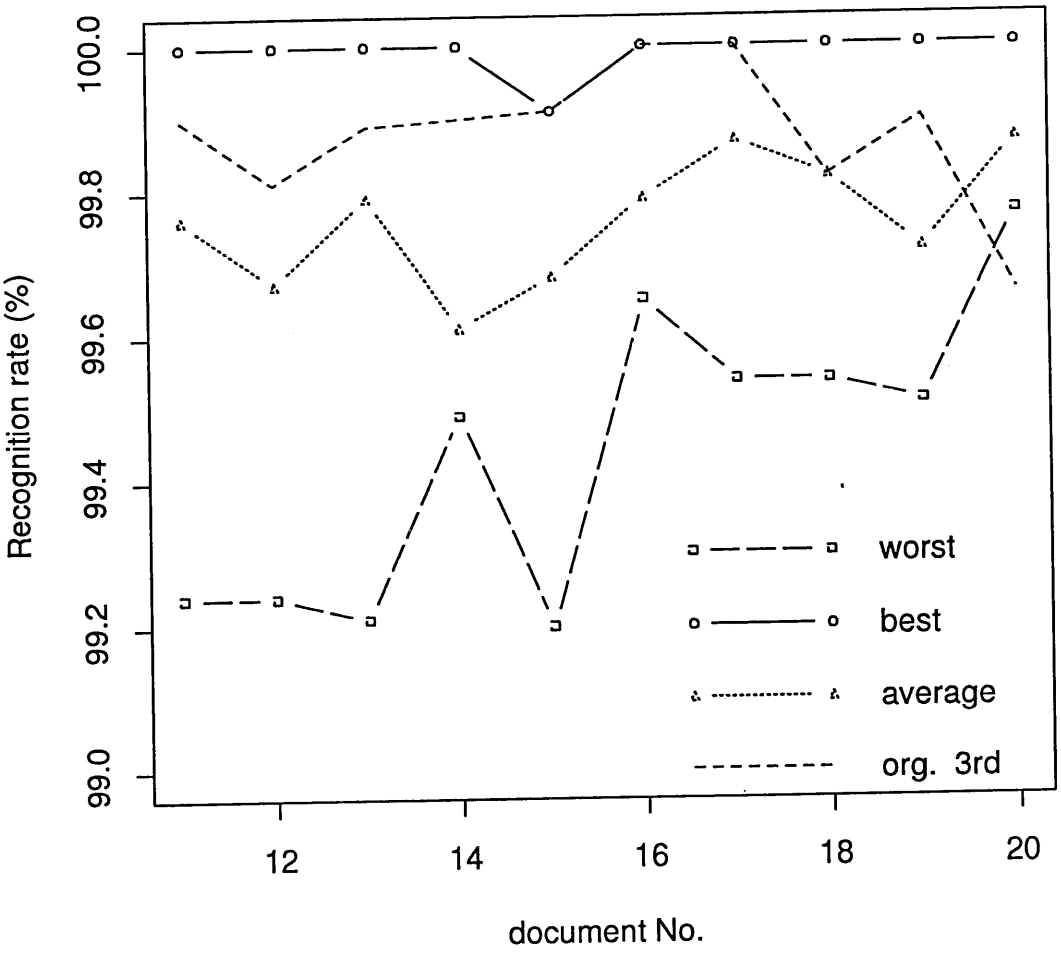


図 4.5: 認識文書毎の 1 位認識率

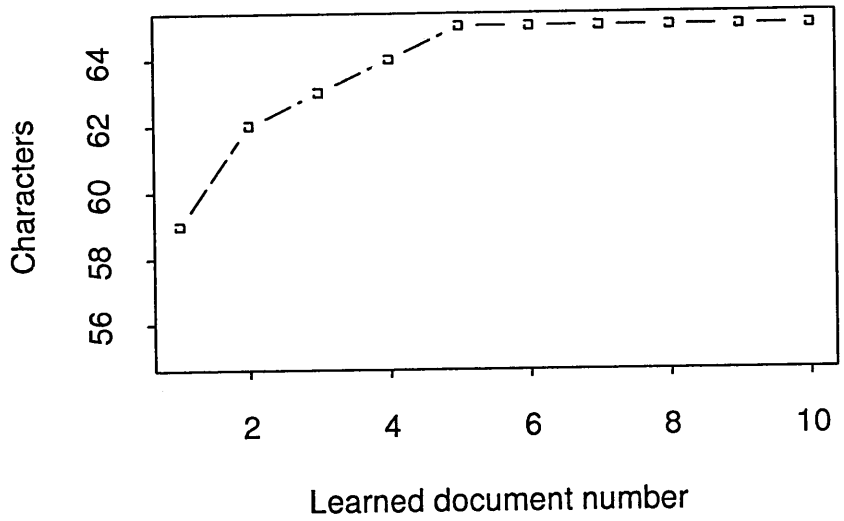


図 4.6: 学習文書数と抽出辞書中の文字

表 4.10: 抽出辞書にない文字

学習文書数	辞書にない文字
1	ご ざ ぞ へ ぼ ゆ
2	ご ざ ぞ
3	ご ぞ
4	ぞ

表 4.11: 従来の辞書の認識率 (社説 21〜30)

社説	ひらがな 65 種		
	1 位認識率 (%)	2 位認識率 (%)	3 位認識率 (%)
平均	96.08 (41)	99.35 (7)	99.88 (3)
21	97.65 (26)	99.82 (2)	100.00 (0)
22	95.84 (39)	99.04 (9)	99.68 (3)
23	95.63 (53)	99.18 (10)	99.67 (4)
24	95.71 (44)	99.80 (2)	99.80 (2)
25	95.30 (47)	99.10 (9)	99.80 (2)
26	96.99 (33)	99.63 (4)	100.00 (0)
27	95.45 (52)	98.95 (12)	99.82 (2)
28	96.56 (35)	99.41 (6)	99.90 (1)
29	96.16 (40)	99.23 (8)	99.81 (2)
30	95.44 (42)	99.35 (6)	99.89 (1)

社説	その他		
	1 位認識率 (%)	2 位認識率 (%)	3 位認識率 (%)
平均	92.50 (78)	97.66 (24)	98.80 (13)
21	93.92 (60)	98.48 (15)	99.09 (9)
22	91.51 (98)	97.14 (33)	98.96 (12)
23	93.77 (53)	98.24 (15)	98.82 (10)
24	90.60 (99)	97.72 (24)	99.34 (19)
25	92.41 (81)	97.28 (29)	98.22 (19)
26	94.00 (63)	98.29 (18)	98.95 (11)
27	91.58 (81)	97.30 (26)	98.65 (13)
28	91.55 (84)	97.08 (29)	98.59 (14)
29	93.64 (71)	97.31 (30)	98.66 (15)
30	92.28 (106)	97.88 (25)	98.73 (15)

() の内は誤認識数

表 4.12: 1 位認識率 (辞書の更新による実験)

学習文書数	マルチテンプレート法	
	ひらがな 65 種 (%)	その他 (%)
1	99.66 (36)	92.50 (781)
2	99.76 (25)	92.50 (781)
3	99.90 (11)	92.50 (781)
4	99.98 (2)	92.50 (781)
5	99.98 (2)	92.50 (781)
6	99.99 (1)	92.50 (781)
7	99.98 (2)	92.50 (781)
8	99.98 (2)	92.50 (781)
9	99.98 (2)	92.50 (781)
10	99.98 (2)	92.50 (781)

学習文書数	置換法	
	ひらがな 65 種 (%)	その他 (%)
1	99.68 (34)	92.50 (781)
2	99.78 (23)	92.50 (781)
3	99.91 (9)	92.50 (781)
4	100.00 (0)	92.50 (781)
5	100.00 (0)	92.50 (781)
6	100.00 (0)	92.50 (781)
7	100.00 (0)	92.50 (781)
8	100.00 (0)	92.50 (781)
9	100.00 (0)	92.50 (781)
10	100.00 (0)	92.50 (781)

() の内は誤認識数

表 4.13: 学習文書数毎の 1 位認識率 (マルチテンプレート法)

社説	学習文書数									
	1	2	3	4	5	6	7	8	9	10
21	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92
22	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51
23	99.75	99.75	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77
24	99.61	99.71	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60
25	99.70	99.80	99.80	99.90	99.90	100.00	99.90	99.90	99.90	99.90
	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41
26	99.63	99.82	99.91	99.91	99.91	99.91	99.91	99.91	99.91	99.91
	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00
27	99.30	99.30	99.82	100.00	100.00	100.00	100.00	99.72	99.91	100.00
	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58
28	99.80	99.80	99.80	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55
29	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64
30	99.57	99.89	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28

上段　ひらがな 65 種の 1 位認識率 (%)

下段　ひらがな 65 種以外の 1 位認識率 (%)

表 4.14: 学習文書数毎の 1 位認識率 (置換法)

社説	学習文書数									
	1	2	3	4	5	6	7	8	9	10
21	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92	93.92
22	99.25	99.57	99.68	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51	91.51
23	99.75	99.75	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77	93.77
24	99.61	99.71	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60	90.60
25	99.80	99.90	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41	92.41
26	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00	94.00
27	99.30	99.30	99.82	100.00	100.00	100.00	100.00	99.72	99.91	100.00
	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58	91.58
28	99.80	99.80	99.80	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55	91.55
29	99.90	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64	93.64
30	99.57	99.89	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28	92.28

上段　ひらがな 65 種の 1 位認識率 (%)

下段　ひらがな 65 種以外の 1 位認識率 (%)

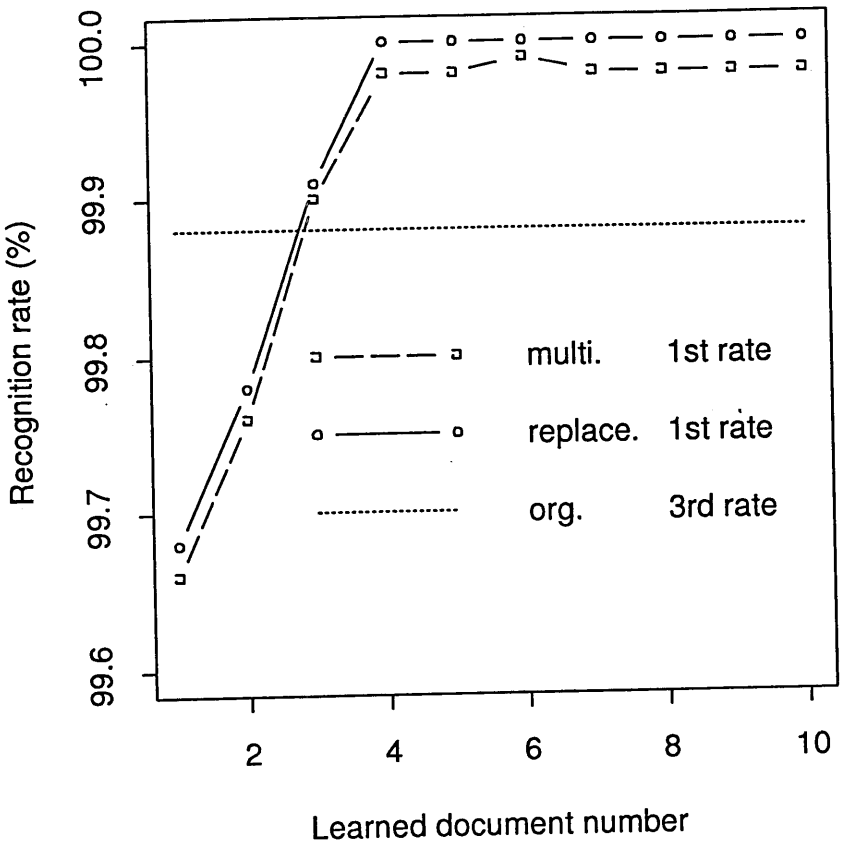


図 4.7: 学習文書数毎の 1 位認識率

第 5 章

結論

本研究では既存辞書を基にして、認識文書からフォントを抽出し、フォントの利用によって文書認識の高精度化を目的とした。その結果について記す。

1. フォントの抽出
   新聞社説 (1〜10) の 10 部を用いて学習させた類似文字の利用により、候補数 6 までのデータの使用で新聞社説 (11〜20) の 10 部から抽出対象となるひらがな 65 種を全て抽出することができ、今回提案したフォントの抽出のアルゴリズムが有効であることがわかった。

2. 文書認識の高精度化

   - 再認識による誤りの自動修正
     抽出したフォントから作成した新辞書を用いて、抽出した辞書を再認識することで、フォントを抽出する際には 1 位で認識できなかった 370 文字 (10 部の合計) を全て正しく認識でき、誤りの自動修正に成功した。

   - 新辞書による認識
     社説 1 部から抽出した抽出辞書を用いて、他の新聞社説の認識を行ったが、辞書が抽出文書中に含まれる対象文字に依存するため認識結果は従来法よりは向上したが誤認識の全てを修正するまでには至らなかった。しかし、新聞社説 (11〜20) を用いて新辞書の更新を行った実験では、新聞社説 (21〜30) の 10 部に対して従来法での誤認識を全て修正できた (置換法)。

   従来法と比較して上記の点で、今回提案したフォントの抽出を基にした文書認識システムの有効性が確認できたと考える。

しかし、本研究で提案した文書認識システムは未完成であり、今後の課題として下記の様な項目が挙げられる。

1. フォントの抽出法の高精度化

   - パターンマッチング法だけでなく、構造解析法を用いる。
   - 除去候補文字を検討する。

   以上の様な手段によりフォントの抽出法を高精度化する。特に、除去候補文字については、本研究中では事前に学習させたものを固定し、実験の結果をフィードバックしていない。これを抽出結果により、除去候補文字を学習・更新するシステムへ発展させていくことが多種フォント文書の自動認識のために有効であると考える。

2. 文字の出現頻度を利用した文書認識システムの構築
   今回のフォントの抽出では、事前の調査から対象をひらがな 65 文字にし、認識文書の約 50%弱の文字をカバーし、認識精度の向上に成功したが、文書認識の自動化を目指すには、文字の出現頻度を考慮し、フォントを抽出する対象を拡大することが必要である。

謝辞

本研究を進めるにあたり、多大な御指導、御助言をいただいた東北大学工学部 阿曽弘具教授、北陸先端技術大学院大学 木村正行教授に心より感謝致します。また、本論文をまとめるにあたり、貴重な御意見をいただいた東北工学部 丸岡章教授、白鳥則郎教授に深く感謝致します。

また、東北大学大型計算機センター 孫寧助手、東北大学大学院博士課程 大町真一郎氏、同修士課程 越後和徳氏、セイコーエプソン株式会社 内山喜照氏には、文字認識の研究全般にわたって数多くの御指導、御助言をいただきました。ここに感謝の意を表します。

さらに、北陸先端技術大学院大学 下平博 助教授、東北大学情報処理教育センター 阿部亨助手、東北大学工学部 瀧本英二助手、東北大学大学院博士課程 成富敬氏、沼田一成氏、後藤英昭氏、同修士課程 福田大氏、中井満氏には、研究全般にわたる有益な御示唆をいただくとともに、計算機環境の整備を計っていただきました。厚く感謝いたします。

最後に御討論、御協力を戴き、また日頃の生活においてお世話になった旧木村研究室、丸岡研究室阿曽グループの皆様に感謝いたします。

参考文献

[1] 飯島泰蔵：「パターン認識理論」
森北出版 (1989 年)

[2] 森健一：「パターン認識」
電子情報通信学会編 (1988 年)

[3] 長尾真：「パターン情報処理」
コロナ社 (1983 年)

[4] 孫寧、田原透、阿曽弘具、木村正行：「方向線素特徴量を用いた高精度文字認識」
電子情報通信学会論文誌 (D-II), **Vol.J74-D-II**, No.3, pp. 330–339 (平成 3 年 3 月)

[5] C. J. Hilditch : "Linear Skeleton from Square Cupboards"
In Machine Intelligence 6, B. Meltzer & D. Michie, Eds., Univ. Press, Edinburgh, pp. 403–420 (1969)

[6] 江島俊朗, 中村洋介, 木村正行：「 構造情報を含む手書き漢字認識のための特徴量」
電子情報通信学会論文誌 (D), Vol.J68-D, No.4, pp.789–796 (1985 年 4 月)

[7] 八代博昭：「手書き漢字認識の高精度化に関する基礎研究」
東北大学大学院工学研究科情報工学専攻 修士学位論文 (1988 年 2 月)

[8] 池田啓明, 阿曽 弘具, 木村 正行: 「文書認識における辞書の半自動作成システム」
電気関係学会東北支部連合大会, 1H-3 (1991 年 8 月).

[9] 池田啓明, 阿曽 弘具: 「未知文書からの文字フォントの自動抽出」
電気関係学会東北支部連合大会, 2C-1 (1992 年 8 月).

研究業績一覧

- 文書認識における辞書の半自動作成システム
  池田 啓明、阿曽 弘具、木村 正行
  平成3年度電気関係学会東北支部連合大会 1H3 (1991-08)

- 未知文書からの文字フォントの自動抽出
  池田 啓明、阿曽 弘具
  平成4年度電気関係学会東北支部連合大会 2C1 (1992-08)

第 A 章

社説の構成字種

今回用いた社説 (1〜20) の 20 部の構成字種について次に示す。なお、辞書に含まれていない字種 (数字、英字、かな文字の一部) もある。

表 A.1: 社説 20 部中の各字種の出現頻度

	字種	文字数 (累計)	割合 (%)
記号	14/ 107	3650	8.02
数字	1/ 10	13	0.03
英字	18/ 52	206	0.45
かな	70/ 83	21138	46.45
カナ	72/ 86	1550	3.41
漢字	1276/ 2965	18937	41.61
合計	1451/ 3303	45504	100.00
その他	7/ -	10	0.03

字種の項はセット中の字種数 /存在した字種数

表 A.2: 各字種毎の高出現頻度の文字

記号	、1698	。1075	ー 231	「191	」189
数字	0 13				
英字	N 46	A 32	T 26	S 22	E 18
かな	の 2010	い 1164	る 1145	に 1124	を 931
カナ	ン 105	ル 91	ト 88	ア 79	ス 77
漢字	国 336	政 233	日 201	一 186	大 155
全体	の 2010	、1698	い 1164	る 1145	に 1124
	。1075	を 931	は 887	な 862	が 850

() 内は 20 部中での出現数

表 A.3: 高出現頻度の文字の全体に占める割合

文字数	5	10	20	30	40	50	100	150	200	250
割合 (%)	15.7	25.8	39.5	46.6	51.2	54.6	64.7	71.0	75.9	79.6

表 A.4: 上位 30 位までに含まれる文字

	ひらがな		記号		漢字	
	字種	文字数	字種	文字数	字種	文字数
10 位	8	8973	2	2773	—	—
20 位	18	15213	2	2773	—	—
30 位	27	20366	2	2773	1	336

表 A.5: 各字種の出現頻度

社説	記号	数字	英字	ひらがな	カタカナ	漢字	その他	計
1	158	0	48	1134	58	877	0	2275
2	167	0	0	1057	59	974	1	2257
3	179	0	2	1126	66	876	1	2249
4	189	0	3	1073	90	948	1	2303
5	168	3	0	1062	190	885	0	2308
6	176	0	0	1050	66	1026	0	2318
7	186	0	9	1086	79	950	2	2310
8	168	4	0	1029	103	995	0	2299
9	189	0	4	1020	135	928	2	2276
10	211	3	0	1034	90	1005	0	2343
11	183	1	8	1077	40	973	1	2282
12	201	1	24	1078	40	992	3	2236
13	200	0	22	930	75	917	0	2144
14	178	0	8	1032	63	997	0	2278
15	161	0	0	1149	57	898	0	2265
16	216	0	3	1028	47	956	0	2250
17	173	0	68	1109	64	957	0	2371
18	176	0	0	1126	96	894	0	2292
19	207	0	0	1044	74	965	0	2290
20	164	1	7	899	53	924	0	2048