# OCR認識誤りの学習方法について

1D－1

太田 学
東京大学大学院工学系研究科

高須 淳宏　安達 淳
学術情報センター研究開発部

## 1 背景と目的

OCRなどの認識系の誤りパタンを学習することは、認識後の誤り訂正や曖昧検索での活用を考えると大変重要である。著者らは、OCR認識誤りを置換・欠落・挿入・結合・分解の5つの誤りに分類し、この5種類の誤りからなる類似文字テーブルを作成する方法を検討している。著者らはこれを曖昧検索に活用する[1, 2]ことを考えており、類似文字テーブルに蓄える情報として、正しい長さmの文字(列)$A^m$、そのOCR認識結果である長さnの文字(列)$B^n$、及びそのような認識結果となる確率$P(A^m \cap B^n)$を求めることを目的とする。但し、定義した5種類の誤りと$\{m,n\}$の関係は、置換誤り$=\{1,1\}$、欠落誤り$=\{1,0\}$、挿入誤り$=\{0,1\}$、結合誤り$=\{2,1\}$、分解誤り$=\{1,2\}$と定めている。

## 2 認識誤りの抽出とその解釈

### 2.1 認識誤りの抽出

まず、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して、異なる部分、即ち認識誤りの部分を抽出する。具体的には図1に示すように、両テキストを最初から1文字ずつ比較して不一致が起こる部分まで進み、そこからは比較元を順に変えながら、連続した2文字(△▲)が一致するまでこの操作を繰り返す。このような2文字が見つかった時点で、不一致が起きた時点からそこまでの間の文字列を認識誤りとして抽出する。この比較を両テキストの最後まで行なうことで、training setの認識誤りを全て抽出する。

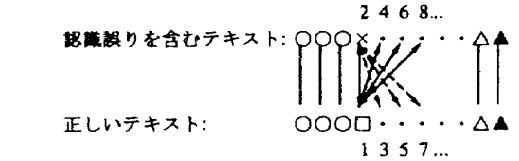


図 1: 認識誤りの抽出

### 2.2 認識誤りの種類判別ヒューリスティクス

前節で抽出された認識誤りは、一般に$\{A^m, B^n\}$の形をしており、これをそのまま類似文字テーブルに蓄えることも考えられるが、$m,n$が大きくなると学習量の観点から$P(A^m \cap B^n)$を正しく推定することが困難となる。そこで$\{A^m, B^n\}$の誤りを、既に定義した5種類の誤りのいずれか、あるいはそれらの複合誤りと解釈するために、以下のヒューリスティクスを適用する。

ヒューリスティクス1　まず以下のヒューリスティクス1を適用して、$\{A^m, B^n\}$の解釈を一意に定めることを試みる。

1. $\{m,n\}$が$\{1,1\}$、$\{1,0\}$、$\{0,1\}$、$\{2,1\}$、$\{1,2\}$のときは、それぞれ無条件に置換誤り、欠落誤り、挿入誤り、結合誤り、分解誤りと解釈する(図2参照)。
2. $\{m,n\}$が$\{m,0\}$及び$\{0,n\}$の場合、それぞれ$m$個の欠落誤り、$n$個の挿入誤りと解釈する。
3. $m=n$の場合、$m$個の置換誤りと解釈する。

ヒューリスティクス2　ヒューリスティクス1を用いて認識誤りを解釈できない場合、即ち$(m,n \geq 2) \cap (m \neq n)$の場合、考えられる全ての誤りの組合せを求める。但し、爆発的な組合せの増加と類似文字テーブルのノイズを防ぐために、以下のヒューリスティクス2を適用する。

1. $A^m$及び$B^n$中に同じ文字が存在し、かつその文字の文字列中の相対位置が近い場合は、その文字は正しく認識されていると考えてその文字で$A^m$及び$B^n$をそれぞれ分割する。分割の結果$\{A^{m_1}, B^{n_1}\}...\{A^{m_k}, B^{n_k}\}$が得られた場合[3]、各$\{A^{m_i}, B^{n_i}\}$にヒューリスティクス1を適用して一意な解釈を試み、それが不可能な場合はヒューリスティクス2の2以降の処理に移る。
2. $A^m$及び$B^n$中にヒューリスティクス1によって得られた既知の置換・結合・分解誤りが存在し、かつその誤りを構成する双方の文字(列)の文字列中

---

How to Learn OCR's Misrecognitions
Manabu OHTA[1], Atsuhiro TAKASU[2], Jun ADACHI[2]
[1]Graduate School of Engineering, The University of Tokyo
[2]Research & Development Department, National Center for Science Information Systems

[3] $(k-1)+\sum_{i=1}^{k} m_i = m, (k-1)+\sum_{i=1}^{k} n_i = n$.

の相対位置が近い場合は、その誤りが存在すると考えてそれらの文字(列)で $A^m$ 及び $B^n$ をそれぞれ分割する。分割後は1と同様に、各 $\{A^{m_i}, B^{n_i}\}$ にヒューリスティクス1を適用して一意な解釈を試み、それが不可能な場合はヒューリスティクス2の3または4の処理に移る。

3. $m > n$ の場合、$A^m \rightarrow B^n$ は置換、欠落、結合の3種類の誤りのいずれかによって構成される複合誤りと解釈する。さらに複合誤りを構成する要素誤りは、欠落誤りを除いて互いの文字列中の相対位置が近いもの同士で誤りを構成するものとする。

4. $m < n$ の場合、$A^m \rightarrow B^n$ は置換、挿入、分解の3種類の誤りのいずれかによって構成される複合誤りと解釈する。さらに複合誤りを構成する要素誤りは、挿入誤りを除いて互いの文字列中の相対位置が近いもの同士で誤りを構成するものとする。

ヒューリスティクス2の1及び2において、複数の分割方法が考えられる場合は分割を行なわず、それぞれ次の処理に移る。

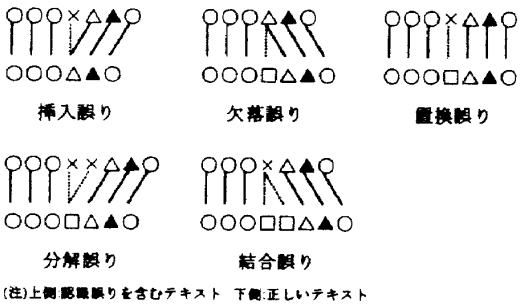


図 2: 認識誤りの抽出例

## 3 OCR 認識モデル

本稿では、OCR は認識対象テキストが与えられた場合、各文字(列)を正しく認識するか、あるいは5種類のいずれかの認識誤りを起こす。よって、$\{A^m, B^n\}$ というシンボルを $P(A^m \cap B^n)$ の確率で出力するものとして OCR をモデル化することができる(図3参照)。

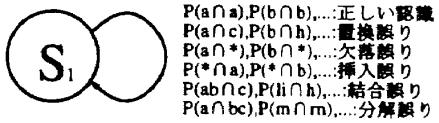


図 3: OCR 認識モデル

このシンボル出力確率は、training set における $\{A^m, B^n\}$ の出現頻度を $C(A^m, B^n)$ とすると、

$$P(A^m \cap B^n) = \frac{C(A^m, B^n)}{\sum C(A^m, B^n)}. \qquad (1)$$

である。ここで誤りの解釈が一意に定まっている場合は、training set 中の $\{A^m, B^n\}$ の出現頻度を単純に数えて $C(A^m, B^n)$ を求めればよい。一方複合誤りにおいて誤りの解釈が $x$ 通り考えられる場合は、その中の1つの解釈の中に現れる $\{A^m, B^n\}$ の出現頻度をそれぞれ $1/x$ と数えることで対処する。

## 4 考察

Elsevier から電子形態で出版されている "Artificial Intelligence" の 1995 年 8 月号〜1996 年 5 月号から得たテキストデータ約 80KB を training set として、認識誤りの抽出・分類を試みた。その結果を表 1,2に示す。尚、文字認識に用いられた OCR の認識率は 98.9%である。

表 1: 抽出された認識誤り 1

| $\{m,n\}$ | $\{1,1\}$ | $\{1,0\}$ | $\{0,1\}$ | $\{2,1\}$ | $\{1,2\}$ |
|---|---|---|---|---|---|
| 頻度 | 138 | 41 | 19 | 176 | 34 |

表 2: 抽出された認識誤り 2†

| $\{m,n\}$ | $\{m,0\}$ | $\{0,n\}$ | $\{m,m\}$ | $\{m,n\} m \neq n$ |
|---|---|---|---|---|
| 頻度 | 8 | 13 | 16 | 33 |

† $m, n > 1$

表 1,2から、合計 478 の認識誤りのうち、33 の複合誤りを除いた 445 の誤りはヒューリスティクス1によって直ちに一意な解釈が可能である。また、33 の複合誤りのうち 25 は、ヒューリスティクス2の1及び2に示した分割によって要素誤りを一意に定めることができ、複合誤りからも確度の高い誤りパタンを抽出できることが示された。

現在複合誤りの解釈が $x$ 通り考えられる場合それぞれの解釈を $1/x$ の重みで頻度に反映させているが、正しい解釈が1の重みでモデルに反映されるのが理想的である。よって、なるべくヒューリスティクスに頼ることなく解釈の曖昧性を扱うことができ、学習によってその理想形に近付けるような OCR 認識モデル及びその学習方法の検討が目下の課題である。

## 参考文献

[1] 太田学, 高須淳宏, 安達淳: 認識誤りを含む和文テキストにおける全文検索手法, 情報処理学会論文誌, Vol. 39, No. 3, pp. 625–635 (1998).

[2] Ohta, M., Takasu, A. and Adachi, J.: Retrieval Methods for English-Text with Missrecognized OCR Characters, *Proc. of ICDAR'97*, Vol. 2, Ulm, Germany, pp. 950–956 (1997).