OCRの認識結果に対する文字認識後処理方式の検討 *

2E-3

磯山秀幸　木谷強 †

NTTデータ通信株式会社 ‡

1 はじめに

OCRによる文字認識結果に対して、自然言語処理技術を適用し、その認識精度を向上させる試みがなされている[1][2][3]。一般的に対象分野を限定すると分野特有の知識が利用できるため、認識精度の向上が期待できる。そこで、我々は帳票に多く現れる住所文字列に限定して、文字認識後処理を行う方法を検討した。なお帳票には文字単位にマス目があり、文字の「分離」や「くっつき」はないことを前提とした。

2 本研究の特徴

文字認識候補文字の絞り込み方法　文字認識部からの出力には、各々の文字に対する複数の候補文字と、原パターンからの距離値が存在する。距離値とは文字パターンの適合度合いの逆数であり、値が大きくなるほど確からしさは低くなる。しかし、全ての候補文字を処理対象にすると、単語を照合する段階で組合せ数が増大し、多大な処理時間を要する。そこで本研究では正解文字の含有率を下げずに、候補文字を絞り込むアルゴリズムを提案する。

住所構造知識の定義　住所表記の構造を表現する方法として、図1に示すように住所構造の他に照合範囲切り出しのキーとなる文字(県, 市, 区, 町など)の属性情報として、キー文字の直前に現れる文字種と文字数を住所構造定義ファイルに定義する。これを基に認識結果文字列を解析することで、明らかに無効と思われる候補文字の組合せを、前もって除外することができる。さらに住所構造を外部ファイルとしてまとめて定義できるため、住所構造知識の追加・変更が容易となる。

3 処理方式

文字認識候補文字の絞り込み　候補文字を絞り込む方法として、一定の候補順位による方法と距離値を利用する方法とが考えられる。出力される候補文字の傾向は、文字認識のアルゴリズムによって異なるため、柔軟性のある絞り込み方法が必要である。そこで、距離値を利用して以下に示す数値計算を行い、既定値を越えるものを候補から除外する処理を検討した。

```
((2:2J? 都
  (1:3J? 区
   2:4HJ? 市
   (1:5HJ 町 (1:6HKJ) 1:5HJ 村 (1:6HKJ))
   1:5HJ 町 (1:6HKJ) 1:5HJ 村 (1:6HKJ))
  2:3HJ 村 (0:0J? 大 0:0J? 字 1:6HKJ)
  1:5HJ 郡 (0:0J? 大 0:0J? 字 (1:5HJ 町 (1:6HKJ) 1:5HJ 村
(1:6HKJ))))
```

?: 省略可能　N: 数字　H: ひらがな
K: カタカナ　J: 漢字　A: アルファベット
"2:5HJ? 市" は、'市'の前に入る文字が2文字以上5文字以下、文字種が「ひらがな」または「漢字」であり、この範囲が省略可能であることを意味する。

図 1: 住所構造定義ファイルの内容

(1) **距離差**：第1候補からの距離値の差。第1候補から一定の距離のなかに、正解文字が含まれている場合に有効。

(2) **距離比**：第1候補の距離値に対する第n候補の距離値の増加比率。第1候補の距離値に対して、一定の増加比率のなかに正解文字が含まれている場合に有効。

(3) **項比**：前候補の距離値に対する距離値の増加比率。正解文字の前後で、距離値が一定の増加比率以上で変化する傾向があるときに有効。

(4) **項差**：前候補の距離値に対する距離値の増加量。正解文字の前後で、距離値が一定の値以上で変化する傾向があるときに有効。

(5) **項差比**：前候補の距離値の増加量に対する距離値の増加量の増加比率。正解文字の前後で、距離値の増加量が一定の比率以上で変化する傾向があるときに有効。

上記の絞り込み処理はそれぞれ性質が異なるため、これらを組合せることにより、ある処理で落とされた正解文字を補うことができると考えられる。上記(1)～(5)の値と候補順位とを組合せて、絞り込みを行った結果を図2に示す。今回対象とした文字認識データでは、候補文字数に対する正解文字の含有率を見ると、全体的には距離差と候補順位の組合せによる絞り込みが最も有効であった。

*Post-processing for correcting the character recognition output
†Hideyuki ISOYAMA, Tsuyoshi KITANI
‡NTT DATA COMMUNICATIONS SYSTEMS CORP.

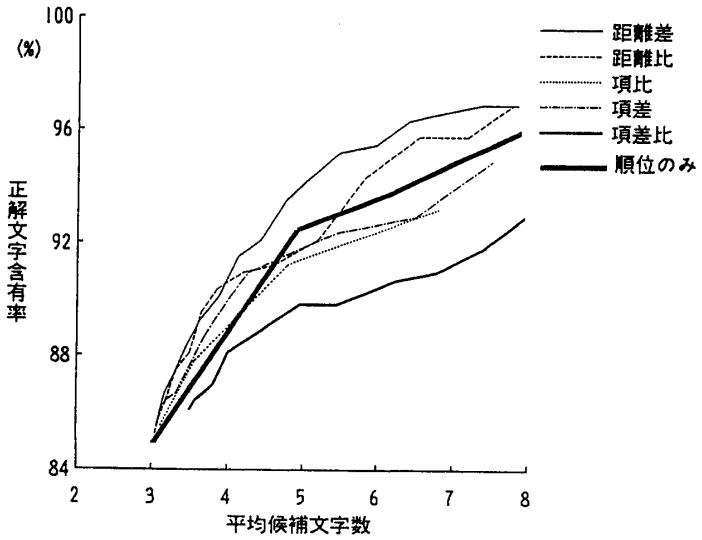

図 2: 候補順位と距離値による絞り込みの組合せ結果

このようにOCRの認識傾向を分析することにより、異なったOCRの出力データに対しても、効果的な候補文字の絞り込みを行うことができる。さらに知識処理の精度・処理時間との関係で、前処理が出力すべき候補文字数が変化することが考えられるが、候補文字数に応じてどの絞り込み方法を適用するのが最適であるかを、上記の手法によって見つけることができる。

4 住所構造の解析

郵便番号を利用した照合 帳票には郵便番号が記入される場合が多い。郵便番号をもとにして住所文字列の照合を行うことは、処理時間の短縮と精度向上のために有効である。郵便番号から住所辞書を検索して得られた住所文字列と、文字認識結果の住所文字列との照合を次のように行う。

(1) 候補文字の順位に対応する点数を設定し、1カラムごとに文字列の比較を行い、一致すれば点数を加算する。

(2) システム既定値以上の得点が得られれば、これを住所文字列として決定する。この場合には、次に述べる住所地名部の解析は不要になる。

住所構造定義に基づいた住所文字列の解析 住所文字列における各部分の名称を図3のように定義する。本稿では、地名部を解析する方法について述べる。

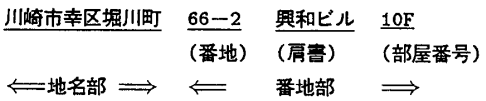

図 3: 住所文字列の各部分の名称

複数の候補文字を含んだ文字認識結果文字列において、文法ルールに定義された単語切り出しキーとなる文字をサーチし、定義された住所構造に一致するようにキー文字の組合せを求める(図4)。こうして得られた組合せを住所構造の候補とする。

次に、住所構造の各レベルごとに住所辞書との照合を行い、一致するものが見つかれば、順に下位レベルへ照合を続けていく。このとき一致する組合せが全くなくとも、文字列の類似度を求め、既定値以上の類似性のある地名単語を住所辞書から検索する。これによって正解文字が含まれていない場合でも、正しい地名を選択することができる。

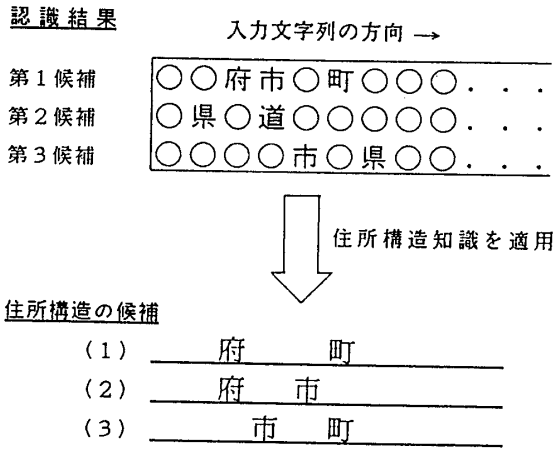

図 4: 住所構造候補の決定

5 おわりに

住所地名部については、住所構造定義による解析と郵便番号の他に、ふりがなを利用することによって、さらに高精度な認識率が期待できる。しかし、番地部については、辞書未登録語が多いため、地名部に比べて精度が低下することは避けられない。したがって帳票全体での認識率を向上させるため、番地部の解析を行うアルゴリズムを強化する必要がある。今後は、実際にOCRから入力したデータをもとに、アルゴリズムの有効性の検証を行う。

参考文献

[1] 清野和司, 柳楽さつき, 中尾和則 「自由記載住所文字列に対する知識処理」 平成元年電子情報通信学会春季全国大会 D-465

[2] 鈴木章, 宮原末治, 小橋史彦 「手書き住所認識の後処理法」情報処理学会第38回(昭和64年前期)全国大会 1K-1

[3] 宮尾滋, 中尾和則, 清野和司 「OCRにおける住所データ読み取りについて」 情報処理学会第34回(昭和62年前期)全国大会 4E-5