名前に対する文字認識後処理手法

2V－5

古和田孝之、吉川隆敏、山本英人、堀井　洋
三洋電機株式会社　ハイパーメディア研究所

1．はじめに

近年、手書き文字認識の研究が進められ、商品化が盛んに行われている。しかし、不特定多数のユーザを対象にしたシステムの場合、文字データなどのパターン情報だけで高認識率を達成するのは困難である。それを解決する方策として、言語情報を用いた認識後処理技術が必要となる。

現在の手書き入力システムの使用シーンを考えると、一般文書の入力よりも、住所、氏名、会社名、商品名など、記入すべき文字列の属性があらかじめ決まっているような形式の入力が主流で、また通常このような入力形式では、一般文書の入力よりも高い認識精度が要求される。

このような状況を踏まえ、住所、氏名のように入力属性が限定されている場合の後処理手法の開発を行った。本報告では、その中から、名前に対する認識後処理として、姓名辞書と単漢字辞書を利用し、フリガナ連動処理と単漢字処理を行う手法について述べる。

2．開発方針

名前データは、(1)種類が多い，(2)漢字に対する読み方が様々である，(3)漢字の使用頻度に偏りがある，などの一般単語とは異なる特徴をもっており、同音異字／同字異音の名前や1文字違いの名前などが多く存在する。

一方、文字認識では、丁寧に書かれた筆記状態の良好な文字に対しては高精度の認識が期待できるが、筆記状態の悪い文字に対しては正解文字が認識候補中に含まれないことも少なくない。

我々は、正解文字が認識候補中に含まれないような筆記状態の悪い文字に対しても、正しい修正が可能で、かつ辞書に正解が登録されていない場合などに起こる未処理や訂正誤りが少ない認識後処理の実現を目的として、単語照合処理を基本にして、フリガナ処理による相互補完、単漢字処理による補正を行う手法を開発した。

3．使用する辞書データ

次の3種類の辞書データを使用している。

(1)姓名辞書 ···· 名前の漢字表記、読みと頻度の情報（姓2万件，名2万件）

(2)単漢字読み辞書 ···· 名前で使用される漢字に対する読みの情報（2．3万種類）

(3)単漢字頻度辞書 ···· 姓、名それぞれの特定位置における漢字使用頻度の情報

4．処理の概要

①単語照合処理

文字認識候補を組み合わせた文字列パターンと姓名辞書に登録されている文字列パターンを比較し、ある基準以上の一致が見られた場合に、辞書の文字列パターンを正解候補として残す（認識候補にない文字を含む名前も許容する）。辞書における頻度情報と認識確信度から正解候補としての確からしさを求める。

A Post-Processing Method for Name Recognition.

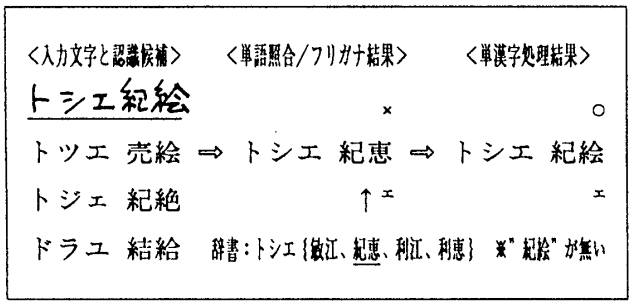
Takayuki Kowada, Takatoshi Yoshikawa, Yamamoto Hideto, Hiroshi Horii
SANYO Electric Co., Ltd. Hypermedia Research Center

②フリガナ処理

漢字部（フリガナ部）の単語照合処理を行う際に、正解候補として得られた漢字（かな）名前の読み（漢字）を姓名辞書を用いて求め、求めた読み（漢字）がフリガナ部（漢字部）の認識候補に含まれる度合いを連動得点として加算する。それぞれの第１正解候補の整合性が取れていない場合は、得点の高いほうを基準にして姓名辞書データで他方を補う。

③単漢字処理

単語照合処理、フリガナ処理で得られた漢字部の第１正解候補の各位置の漢字について読みを調べ、同一位置の認識候補の中に同じ読みを持つ漢字が存在するときは、その漢字で第１正解候補の漢字を置き換える。



④単漢字頻度処理

姓名辞書のデータ不足によって単語照合処理で一致が見られない場合を考慮して、単漢字の位置別頻度情報を用いて処理を行う。これは、ある漢字が、ある長さの名前（姓また名）の、ある位置で使用される頻度情報を利用するもので、各位置について、認識候補の中から最大頻度を持つものを正解候補として選択する。

5．認識実験

実験では、姓と名の漢字とフリガナを分かち書きした１６０人分の筆記データに対して文字認識を行い、それぞれ認識結果の上位５位までを後処理の入力データとした。

表１　実験結果　　単位(%)

	正解率(姓)	正解率(名)	正修正率	誤修正率
単語照合処理 ・・・・①	80.6	78.1	52.7	3.7
①+フリガナ処理 ・・②	87.5	81.9	73.3	7.4
①+②+単漢字処理 ・③	89.4	86.9	76.3	3.7
(文字認識単独)	(55.0)	(63.1)	----	---

後処理として、①単語照合処理のみ、②フリガナ処理を加えたもの、③さらに単漢字処理を加えたものの３種を行い、単語単位の正解率と、文字認識単独の結果に対する正修正率（誤認識したものを正しく修正した率）と誤修正率（正しく認識しているものを誤って修正した率）を算出した。実験結果を表１に示す。

文字認識単独での認識率が低い（本実験では平均認識率６５％）場合でも、フリガナ処理を行うことにより、高い率で正解候補を補完することができた。また、単漢字処理を行うことによって、フリガナ処理だけでは増加していた、辞書データの不足が原因で生じる未処理や訂正誤りを低減させることができた。

6．おわりに

単語照合処理、フリガナ処理に加えて、単漢字処理を行うことにより、比較的ラフな筆記状態の名前入力データに対しても有効に機能する認識後処理が実現できた。

【参考文献】

(1)黒川，吉川，長沢，亀田："認識情報及び言語情報を利用した文字認識後処理"，1990信学秋季全大，D-364