手書き住所認識の後処理法

1K-1

鈴木　章　　宮原末治　　小橋史彦

ＮＴＴヒューマンインタフェース研究所

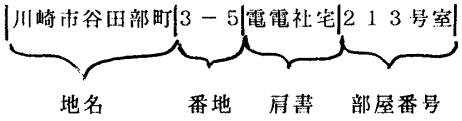
１．はじめに

文字認識技術の応用の一つに，銀行などの窓口で来客によって書かれる帳票類を直接読み取って処理する認識装置が考えられる．だがこのようなシステムに現在の手書き漢字認識装置をそのまま用いようとすると認識精度に問題が残る．そこで我々は認識結果に知識処理を施して読み取り精度を向上する方法について検討した．

今回は読み取り対象として住所を選び，従来あまり扱われていなかった〔1〕べた書きの住所の知識処理について報告する．

２．処理アルゴリズム

日本語の住所は下に示すように，地名，番地，それにアパート名や社宅名など（ここではこれを肩書と呼ぶ）とその部屋番号から構成される．



今回開発した住所データ後処理方式の処理手順を以下に述べる．

（１）地名処理

最初に地名の訂正処理を行う．これは認識結果の先頭から地名の単語辞書と照合する処理である．複数の候補が抽出された場合には，単語長および距離値の合計値などから最尤候補を決定する．

（２）番地，部屋番号処理

次に，番地および部屋番号の処理を行う．番地については（１）で抽出された地名以降のデータに対して末尾方向に番地テンプレート（２２パターン）と照合し，部屋番号についてはデータの最後部から先頭方向に部屋番号テンプレート（３９パターン）と照合する．これについても複数の候補が抽出された場合には，単語長および距離値の合計値などから最尤候補を決定する．

［例］＊＊－＊＊＊－＊＊＊
　　＊丁目＊＊番地
　　＊＊号室
　　（＊は数字を示すワイルドカード）

（３）肩書処理

最後に肩書処理を行う．肩書については，大規模で正確な単語辞書を得ることが困難なこともあり，今回は単語パターンとの照合法を用いた．肩書を複合語として見ると，「地名＋接尾語」，「企業名＋地名＋接尾語」のような単語パターンで表すことができる．

肩書処理は，まず地名処理で抽出した領域を除く認識結果の全ての箇所を開始点として企業名，地名，接尾語，姓名等の単語辞書と照合する．次に抽出された単語を接続し，その組合せが単語パターン（１６パターン）として登録されているものを抽出する．

３．評価実験ならびに考察

住所データ９４件を用いて評価実験を行った．使用したＯＣＲは，ＮＴＴの手書き漢字認識装置ＯＣＲ６０である〔2〕．処理前と処理後の文字単位と帳票単位の正解率の比較を表１に示す．

（１）地名部分の訂正誤りは帳票単位でわずか３件で，誤認識率を９１．１％減少できた．訂正誤りの原因はいずれも地名辞書に記載されていない字が地名中に含まれていたことが原因であった．

（２）番地，部屋番号の部分は帳票単位でみると正解率を約７倍に向上できた．訂正誤りの主な原因は肩書の訂正誤りの波及であった．肩書の処理に用いられる単語パターンには地名も含まれており，肩書に近い部分の文字の認識結果の中に地名を示す組合せが存在するとそれが肩書の一部として抽出されてしまうため，誤りの原因となった．

（３）肩書部分の訂正性能は他に比較して低かったものの，帳票単位で誤認識率を３３．５％減少できた．この部分の訂正誤りの原因は，字や企業名などの単語が辞書に含まれていないことによるものが大部分であった．

表１　自動訂正による正解率の変化（％）

	文字単位		帳票単位	
	処理前	処理後	処理前	処理後
地名	94.7	99.6	63.9	96.8
番地，部屋番号	71.1	91.3	10.6	71.3
肩書	83.0	87.9	48.9	66.0

４．まとめ

べた書きの手書き住所データの認識誤りを自動訂正するシステムを開発し，帳票単位の誤認識率を地名の部分で９１．１％，番地および部屋番号で６７．９％，肩書で３３．５％減少させることができた．今後は字を含めた地名およびその他固有名詞の辞書の拡張，肩書処理アルゴリズムの改良，認識結果の距離値による第１位候補文字の正誤検定などにより，一層の訂正性能の向上を図る予定である．

［謝辞］処理システムの作成を手伝っていただいた北海道大学の堀田祐介君に感謝いたします．

［参考文献］

〔1〕蕪山，菅原，山本，中西「手書漢字認識における単語情報の利用」昭和57年度電子通信学会総合全国大会NO.1341

〔2〕赤松，川谷，塩，飯田「手書き漢字用文字読み取り装置」研究実用化報告 Vol.36,no.4,pp.579～587

Correction of address recognition errors

Akira Suzuki,Sueharu Miyahara,Fumihiko Obashi

NTT Human Interface Laboratories