I-75

アイルランド向け手書住所認識システム
Address Recognition System for Irish Handwritten Mail

秋山 達勇† 西脇 大輔‡ 林 祐人† 山内 俊史†
Tatsuo AKIYAMA Daisuke NISHIWAKI Masato HAYASHI Toshifumi YAMAUCHI

1. はじめに

郵便書状に記載された住所の認識は，郵便番号認識の時代を含めると歴史は深い．近年においてはその対象を郵便番号から住所へと移しながら，現在でも興味深い研究開発が行われている．その研究の多くは，郵便番号と住所の双方が記載されていることが前提となっている[1]．

一方，アイルランドのように郵便番号制度がない国の手書き書状に対しては，単に住所のみを読み取る必要があり，高度な知識処理技術と，知識処理と密接に結びついたパタン認識技術が要求される．本論文では，アイルランド手書き書状の処理を前提とした手書住所認識システムについて論じる．

2. アイルランド住所の特徴

アイルランドの住所は，County, City, Locality, Street, House No. の 5 つの階層で構成される．ただし，すべての住所にこれら 5 つの階層が存在するとは限らず，さらに，住所の同一性を失わない範囲で，いくつかの階層の記載が省略されることがある(図 1 参照)．また，City の中で "Dublin" だけは，"Dublin 2"というように，地域を示す番号とともに記載される．番号は 1 から 26 及び "6W" からなる．

さて，郵便番号の記載がない住所に対する最も単純な手法として，County, City,...と一つずつ上位階層から順に確定していく方法が考えられる．単語認識方式として[2]の方法を用いることとし，Lexicon には直上の階層の住所要素に属する名称を登録することとして認識実験を行うと，認識率及び誤読率は表 1(改善前)のようになる．

ここで，OUTWARD とは，配達局が決定されるレベル(City レベル)であり，INWARD とはさらに細かい地域を決定するレベル(一部の Locality、一部の Street name 及び House No)である．

3. 改良手法

3.1 問題点の分析と改良

単純な手法の問題点の第一は，住所要素の記載省略があった場合に，それより下位の階層の読取が不可能であったことである．従って，

（１Ａ）記載省略に対応した読取フロー

とする必要がある．

第二に，入力画像に対する正解文字列が Lexicon 中に登録されていない場合の誤読の問題である．本来，Lexicon 駆動型の単語認識は，Lexicon 中に正解文字列が含まれていることが前提である．ところが，記載省略が生じる影響で，必ずしも Lexicon に正解が含まれる条件で認識が行えるとは限らない．従って，正しい認識結果が得られるように

（２Ａ）Lexicon に正しい文字列が含まれる処理

（２Ｂ）不正な文字列に対する棄却能力の向上

が求められる．また，

（２Ｃ）住所知識を用いた単語認識結果の検証

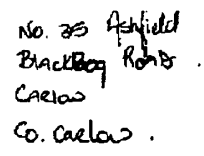
することで，単語認識誤りがあっても住所としての出力を補正できるような仕組みも求められる．

第三に，さらなる住所認識性能の向上のために，

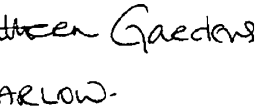
（３Ａ）単語認識の精度向上

（３Ｂ）住所知識を利用した知識処理

も求められる．下記に述べる改良手法では，これらの課題に関して改良を行う．



(a) 省略のない場合, COUNTY-CITY-LOCALITY-STREET-House No



(b) 省略のある場合, CITY-STREET-House No

図 1. 記載省略の例

表 1. 認識率,誤読率の比較

評価項目	改善前	改善後
OUTWARD 認識率	12%	48%
OUTWARD 誤読率	12%	1%
INWARD 認識率	3%	24%
INWARD 誤読率	1%	1.2%

3.2 改良処理手順

改良した処理手順の概略を図 2 に示す．国内住所としての認識を行う場合は，以降の 3.3, 3.4 に述べる処理は住所階層毎のループ処理となる．すなわち，単語認識の結果得られた評価値の最大のものを各住所階層の認識結果とする．ただし County 及び City は同一ループ内での識別を行う．これは、County と City の双方が記載されている場合に，それぞれの上位 3 候補づつの認識候補を比較し矛盾のない候補を選ぶことによって精度向上（３Ｂ）を目的としている．

また，国外宛としての認識は，3.3 に述べる処理で国名のみの認識を行う．国外宛を仮定した場合の単語認識信頼度と国内宛を仮定した場合の各住所階層における単語認識信頼度の最大値を比較し，信頼度の大きいほうの結果を最終的な出力とする．

3.3 単語認識部, 単語検証部, Lexicon 作成部

Lexicon 作成に関わる階層間の依存関係は，グラフ表現を用いた住所関係モデルにて規定する (図 3 参照)．処理対象階層を現す接点へ入る終点を持つノードの始点が絞り込み合に，その結果に応じてその名称に属す処理対象名称文字列を Lexicon として登録する．*Root* が始点の場合は全名称の読み取りであることを表す．また，上位階層の読み取り結果が存在する始点が複数存在する場合には，（２Ａ）に従い，上位の階層を優先とする．以上は（１Ａ）に従い記載の省略に対応したフローとするための規定である．

また，（３ A）を目的として，特徴抽出として 64 次元加重方向指数ヒストグラム[2]，識別器として疑似ベイズ識別[2]による方式をマルチテンプレートに拡張した方式を単語認識部とする．さらに，（２ B）を目的とし，特徴抽出として 392 次元加重方向指数ヒストグラム，識別器として GLV

† NEC 社会情報ソリューション事業部
‡ NEC マルチメディア研究所

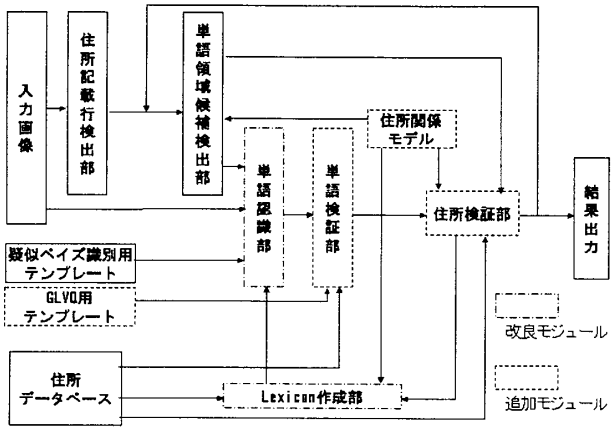


図 2. 改良処理手順

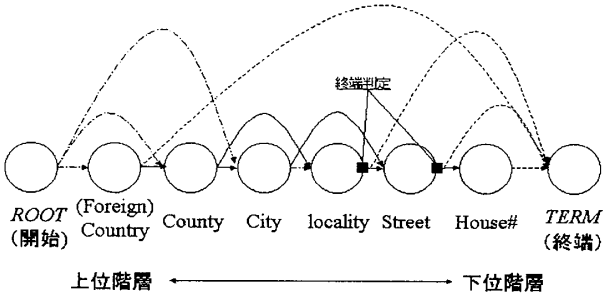


図 3. 住所関係モデル

Q[3]による方式を単語検証部とする．疑似ベイズ識別で得られる尤度をベースとする信頼度により『文字列らしさ』を判定し，結果として得られた個別文字領域に対して，他カテゴリの距離が学習に反映されている GLVQ による識別を検証として利用する．

さらに，棄却能力の向上（２Ｃ）を目的として，"Dublin 2" などの頻出文字列に対する個別文字領域の位置を利用した検証を行う．

3.4 住所検証部

認識結果が当該階層の住所として妥当かどうかを判定するために，上位階層の読み取り結果が，住所関係モデルとデータベースに矛盾していないかどうかの検証を行う．これは，記載省略に対応するために，階層の跳び越しを許容している (3.3 節で述べた Lexicon 作成ルールに従うと，City が認識されていても，County が認識されていれば,County に属する Locality の認識を行う) ためである．

また，上位階層の読み取り結果との相対位置による検証ルールを規定し検証を行う．ルールの例としては，『City の記載位置は County のすぐ上方に記載される』，『Locality は，City の記載がある場合は City のすぐ上方に記載され，City の記載が無い場合には County のすぐ上方に記載される』などがある．

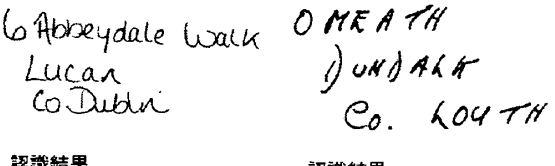
さらにデータベースを参照し，終了判定を行う．区分上必要な階層の認識が行われているかどうかを判定する．終了と判定された場合には処理を終了する．

4. 評価実験

4.1 改善効果

アイルランドの実便から作成した手書き書状の画像テストデータセット 1618 通を用いて評価を行った．Pentium III 1GHz, PC-Linux, 処理時間 2.5 秒以内の条件で，Outward レベル認識率 48%，誤読率 1% ，INWARD レベルでは，認識率 24%，誤読率 1.2%という結果を得た(表 1 参照)．

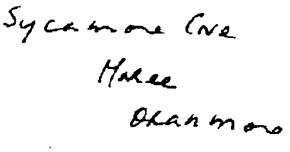
図 1(b)の場合，記載省略があるために単純な手法では単純な手法では，1 行目の "Carlow" が County としてのみ認識され City 以降が正しく認識できなかったのに対し，改良手法では，"Carlow"が City として正しく認識できるようになった．そのほか良好な認識結果が得られた例を図 4 に示す．



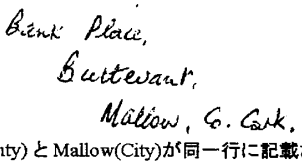
認識結果
County: DUBLIN
City: LUCAN
Locality: (記載なし)
Street: ABBETDALE WALK
House No: 6

認識結果
County: LOUTH
City: DUNDALK
Locality: OMEATH
Street: (記載なし)
House No: (記載なし)

図 4. 良好な結果が得られた例



(a) CountyもCityも記載されていない例



(b) Cork(County) と Mallow(City)が同一行に記載されている例

図 5. 改善を必要とする例

4.2 今後の課題

今後の課題の一つは，County, City いずれの記載も無い場合に誤読する方法の開発である．図 5(a)の例では，2 行目を Traree(City) と誤認識している．検証機能の強化とともに住所記載領域の検出などの前処理情報の活用などを検討する予定である．

また，County と City が同一行に書かれている場合(図 5(b) 参照)のような場合にも精度良く住所を認識するために，単語切り出し性能の向上させる必要がある．

さらに，"Castlerea" と "Castlebar" といった類似単語が存在する場合の単語認識，単語検証の各機能を強化する必要がある．

5. おわりに

本論文では，アイルランド手書き書状を題材として，郵便番号の記載がない場合や住所要素の記載省略がある場合に，単語認識の精度向上，GLVQ を用いた単語検証処理，住所知識を用いた Lexicon 作成や検証処理を導入することにより，アイルランド実便画像テストデータセットに対し実用レベルの性能を実現したことを示した．

今後は，単語認識，単語検証の精度向上とともに，住所記載領域の検出や単語切り出しの精度向上を行い，更なる住所認識性能向上を目指す予定である．

謝辞

英文手書住所システム開発当初に暖かいご指導をいただきました，三重大学の木村文隆先生に深謝致します．

本システムの開発にご協力いただきました NEC 社会情報ソリューション事業部の上谷昌昭氏，NEC ポスタルテクノレクスの近藤克彦氏を始め，関係各位に感謝致します．

参考文献

[1] A. FILATOV, V.NIKITIN, A.VOLGUNIN, P. ZELINSKY: The AddressScript RECOGNITION SYSTEM FOR HANDWRITTEN ENVELOPES, *Proc. of IAPR 1998 Workshop on Document Analysis System* , pp. 222-236, 1998.

[2] Fumitaka KIMURA, Shinji TSURUOKA, Yasuji MIYAKE and Malayappan SHRIDHAR: A Lexicon Directed Algorithm for Recognition of Unconstraind Handwritten Words, *IEICE Transactions on Information and Systems*, Vol. E77-D, No. 9, pp. 785-793, 1998.

[3] A. Sato, K. Yamada: A formulation of learning vector quantization using a new misclassification measure, *Proc. of 14th ICPR*, Brisbane, VOL. I , pp322-325, 1998.