2003－CH－58 (3)
2003／5／30

n-gram と OCR による定型表現がある古文書の文字の推定

山田　奨治 *1, 柴山　守 *2

*1 国際日本文化研究センター・研究部
*2 大阪市立大学・学術情報総合センター

古文書の翻刻作業中に遭遇する不可読文字について，江戸時代の借金証文類に対象を限定して，前後の文字の n-gram 情報と不可読文字画像の OCR 結果を併用して正解候補を求める手法を提案する．n-gram は $n = 2$ と $n = 3$ を併用する方法を，OCR は改良型方向線素特徴量とユークリッド距離最短法による認識を採用した．250,000 文字を超える古文書文字データと 27,000 文字を超える古文書文字画像データを電子化して手法の検証をおこない，提案手法を翻刻支援システムに適用した場合の性能と有用性について考察した．提案手法により 3,509 文字の試験データの 81.93% について，正解の平均順位が 4.69 位，20 位以内に正解が得られる割合が 79.77% という結果が得られた．古文書の全自動読み取りではなく，あくまで人間の作業を支援するシステムのための方法として提案手法は有効であり，歴史学研究に対する情報処理学のあらたな適用分野としての発展が期待される．

An Estimation Method of Unreadable Historical Character for Manuscripts in Fixed Forms using n-gram and OCR

YAMADA Shoji[*1] and SHIBAYAMA Mamoru[*2]

*1 International Research Center for Japanese Studies
*2 Media Center, Osaka City University

We propose a method of getting some candidates for unreadable characters, which appear in reading historical manuscripts, using the n-gram information of the character sequences and the OCR results of the characters, with restricting the object to written acknowledgements of debts in Edo period. We used $n = 2$ and $n = 3$ together for the n-gram and the improved directional element feature and the minimized Euclid distance for the OCR. We examined the performance and the effectiveness of the proposed method using over 250,000 characters and over 27,000 character image data, which are digitized by our project. Using the proposed method, the correct answers appear at 4.69th of average ranking and at 79.77 percents within top 20 candidates for 81.39 percents of 3,509 test samples. The proposed method is effective not for the automatic reading of historical manuscripts but for the supporting system for human's work, and has possibility to extend the application area of information processing to historical studies.

1　はじめに

歴史資料のディジタル・アーカイブ化が進展をみせるなか，古文書（明治初期ころまでに筆記された文字記録）をたんに毛筆・くずし字の画像データとして蓄積するのではなく，文字データとして全文データベース化する方法が模索されている．古文書の文字データ化（翻刻）には，いまは使われていないくずし字の読解力だけでなく，各種文書の様式や定型的な表現，その文書の背景にある歴史学の知識も必要で，高度に専門的な作業である．しかしながら，翻刻を必要としている無数の古文書の量と比較

して，翻刻作業に従事できる能力を備えた作業者の数は圧倒的に少なく，コンピュータによる翻刻作業の支援環境を作ることが望まれている．

そうしたなか，古文書の翻刻の作業を効率よく進めるために，手書き文字認識 (OCR) 技術などを応用する研究が進められている [1][2]．先行研究として，古文書から採字した小規模な実験データを用いた文字認識実験 [3] や，文字認識処理を援用しながら古文書のつづけ字から文字切り出し処理をおこなう研究 [4] などが，すでになされている．

古文書の翻刻支援に目的を絞った場合，高精度の文字認識は差し当たり絶対に必要な条件とはならない．なぜならば，古文書の完全自動読み取りとは違って，翻刻支援の場合は人間が介在する作業を効率化するような情報をシステムが提供できればよい．つまり，たとえ不完全であっても，人間による推論の助けになる情報を提示することが重要なのである．したがって，翻刻支援システムとして利用価値のあるものにするためには，システムが出力する第 1 候補文字が正解である率を 100% に近いレベルで競うことよりも，たとえば正解が候補文字の上位 20 位に入る割合を 80% 程度にすることが，現段階での目標になる．

この論文では，江戸時代の借金証文類に対象を限定して，翻刻作業中に遭遇する判読不能な文字（不可読文字）を，その前後の文字の n-gram 情報と不可読文字の画像データの OCR 結果から不可読文字の正解を推定する方法を検討し，当手法を古文書翻刻支援システムに応用した場合の有効性を，大量の実データを使って検証する．

江戸時代の借金証文類に対象を限定する理由は，この種の文書には「預り申所実正也」「急度返済可仕候」「仍而如件」といった定型表現が頻出するため，n-gram のような統計情報で不可読文字を推定できる可能性がたかいからである．さらに，借金証文のような江戸時代の公文書は「御家流」という書体で筆記されているため，文字のくずし方にある程度の法則性がみられ，毛筆・くずし字という OCR に不利な条件が緩和される．また，借金証文類の翻刻は江戸時代の経済史研究にとって重要な作業であるにもかかわらず，未翻刻の文書数は，各地の文書館や個人の蔵で眠っているものも含めると，それこそ無数にある．したがって，借金証文類に対象を限定した研究であっても，実用への期待と可能性はたかいといえる．

われわれは，実証性を重視した検証を進めるために，江戸時代の借金証文類 231,161 文字を翻刻して用例データを作成し，それらのうちの 3,509 文字についてくずし字のなかから 1 文字を切り出した文字画像データを作成した．さらに，標準的な古文書文字辞典から 24,244 文字を採字して，その文字画像データと文字データを電子化し，OCR のための学習データにした．これらのデータを使って，n-gram 情報と OCR のそれぞれによる不可読文字の推定と，両者を総合した推定結果を示し，翻刻支援システムにこの手法を適用した場合の性能と有用性について考察して，情報処理学のあらたな適用分野の開拓を試みる．

2 n-gram 情報による不可読文字の推定

2.1 方法

用例データから作成する n-gram[5] は $n = 2$ と $n = 3$ を併用し，不可読文字の推定に当たって $n = 3$ では候補が得られなかった場合に $n = 2$ の情報を使用する方法を採用した．この方法は，本論文の実験で使用するものと同種の古文書データを使って，有効性がすでに検証されている [6]．方法の概略は，以下のとおりである．

推定対象である不可読文字を c_i とすると，その前後の文字のつながりは，

$$\cdots c_{i-1}c_ic_{i+1}\cdots$$

と表現され，一方 n-gram テーブルは，

$$t_{j1}t_{j2}\cdots t_{jn}, f_j$$

と表現される．ここで t_{j1} は用例データ中に登場する n 文字のつながりの 1 文字目，t_{j2} は 2 文字目，f_j はその n 文字のつながりの出現頻度である．

n-gram 情報を使って不可読文字を推定する方法は，文献 [6] では前方一致と後方一致が取られているが，本論文ではそれらに加えて $n = 3$ の場合の中間一致も考慮することにする．すなわち，不可読文字 c_i に対して，

- 前方一致した集合:

$$F_f(c_i) = \{(t_{k3}, f_k)|t_{k1} = c_{i-2}, t_{k2} = c_{i-1}\}$$

- 中間一致した集合:

$$F_m(c_i) = \{(t_{l2}, f_l) | t_{l1} = c_{i-1}, t_{l3} = c_{i+1}\}$$

- 後方一致した集合:

$$F_b(c_i) = \{(t_{m1}, f_m) | t_{m2} = c_{i+1}, t_{m3} = c_{i+2}\}$$

となり，不可読文字 c_i の正解候補の集合 $G(c_i)$ には，前方・中間・後方一致のうち頻度が最大となるつぎのような要素を与える．

$$\begin{aligned} G(c_i) &= \{(t_{**}, f_*)\} \\ &= \{\max_{f_*}(F_f(c_i), F_m(c_i), F_b(c_i)) \\ &\quad | t_{k3} = t_{l2} = t_{m1})\} \end{aligned}$$

n-gram 情報からの推定による正解候補のスコア $NScore$ は，頻度の合計からの比例配分値の逆数であるつぎのような値を与える．

if $f_* > 0$
$$NScore(t_{**}) = \Sigma f_* / f_*$$
else
$$NScore(t_{**}) = 1$$

$NScore$ は $(0 < NScore \leq 1)$ の値をとる．ただし，不可読文字が n-gram テーブルに対して前方・中間・後方のいずれにも一致しない場合は，$NScore$ は不定とする．すなわち，

if $\Sigma f_* = 0$
$$NScore(t_{**}) = NONE$$

計算順序は，まず $n = 3$ で $NScore$ を求め，それが不定になる場合に限って $n = 2$ で同様の操作をおこなう．$NScore$ は小さいほど良好な推定となる．

2.2 実験

実験には，大阪市立大学が所蔵する近世の借金証文類である「伏見屋文書」の全文を翻刻して用いた．「伏見屋文書」は金融・借家・親族関係に関する議定書などからなる総数 1,300 の文書群で，翻刻後の総文字数は 231,161 文字である．そこから，後述する OCR の実験にも用いる 30 文書 3,509 文字の試験データを除いたものを用例データとして n-gram を作成した．すなわち，用例データと試験データは重複しない．

翻刻にあたっては，古文書の文字を MS 明朝フォントが表示する SJIS コードの範囲内でもっとも近い字形を取る文字コードを選択した．したがって，たとえば「返済」と「返濟」がおなじ意味であっても，それぞれ異なる用例として扱われている．

実験では，試験データの 3,509 文字のすべての文字を 1 文字づつ順に取り出して仮想の不可読文字として，正解候補の何番目に正解が出現するかを調べる方法をとった．

2.3 結果と考察

提案手法によって，試験データ全体の 79.62% にあたる 2,794 文字について正解候補が得られた．正解候補が得られながら，そのなかに正解が含まれなかった事例は，この試験データ中にはなかった．正解順位の平均値は 5.42 位 ($\sigma = 8.76$)，最頻値は 2 位，最大値は 129 位であった．正解が候補の 1 位となった割合は 8.49%，10 位以内に入った割合は 71.19%，20 位以内では 76.63% であった．一方，正解候補が得られたものの候補数の平均値は 18.10 個 ($\sigma = 22.33$)，最頻値は 1 個，最大値は 290 個であった．

システムとしての実用性を考えると，正解候補として出力される候補数は 20 個程度以下，もし可能ならば 10 個以内であることが望ましいと思われる．あまりにおびたただしい数の正解候補を示されても，人間の作業の助けにならないからである．提案手法で得られた正解候補数の平均値は 18.10 個で，20 個以下に収まっている．

しかしながら，試験データの 20.38% にあたる 715 文字について，提案手法では正解候補が得られなかった．すなわちこれらの 715 文字は，その前後の文字列が用例データにマッチしなかった文字である．

正解候補が得られた仮想不可読文字について，平均値で 5.42 位に正解が位置するという結果は，翻刻作業の支援システムとして実用可能な水準であろう．一方で試験データの 20.38%，すなわち平均して 5 文字に 1 文字は，正解候補が出力されないという結果は，翻刻作業支援システムとしての実用化に向けて障害となる．したがって，n-gram 情報になんらかの補助的な情報を加えて，正解順位を向上させると同時に，候補を出力しない例を削減しなければならない．

3 OCR による不可読文字の推定

3.1 方法

n-gram 情報に加える補助的な情報として，不可読文字の画像情報を与えて，その OCR 結果を加味して総合的な順位を求める方法を試みる．その前にまず，古文書文字の場合に OCR でどの程度の認識率が出るかを検討する．

OCR にはさまざまな文字特徴量の求め方があるが，われわれは日本語手書き文字認識研究で ETL9B データベースに対してたかい認識率を出している改良型方向線素特徴量 [7] をそのまま適用してみることにした．改良型方向線特徴量は，文字を非線形正規化した後に文字の輪郭線を構成する線分の方向の分布を小領域ごとに重み付けをして抽出する方法で，特徴量は 196 次元のベクトルとして得られる．

OCR で高認識率を出すためには，文字認識用辞書をどのように作るかが重要である．われわれは，専門の翻刻者の間で標準的な辞書のひとつになっている『くずし字解読辞典』[8] を選択して，その本編ならびに付録に掲載されている文字画像と，それらに対応する非くずし字・読みなどの情報の文字コードを電子化して文字認識用辞書を作成した．文字画像の電子化は，辞書のページを 400dpi2 値でスキャニングし，1 文字づつを手作業で切り出す方法をとった．このようにして電子化した総文字数は 4,795 字種 24,244 文字である．

『くずし字解読辞典』では，ひとつの文字について 2 種類のくずしのパターンが例示され，その非くずし字と読みが示されている．おなじ文字の異なるくずし文字が複数の場所に掲載されている場合もあるので，得られるサンプル数は文字によって異なるが，ひとつの文字に対するサンプル数は非常に少ない．1 文字あたりのサンプル数の平均値は 5.06 個 ($\sigma = 4.43$)，最大値は 55 個，最頻値は 2 個である．

『くずし字解読辞典』では，くずし字に対応する非くずし字は活字ではなく手書きであるため，われわれは手書き非くずし字にもっとも近い字形の SJIS コードを与え，SJIS コードに対応する文字がない場合は今昔文字鏡コード [9] を割り振った．その際，たとえば「済」と「濟」がおなじ文字であるといった字形の包摂概念については考慮せず，与えた文字コードが異なっていればそれらは別の文字として取り扱った．『くずし字解読辞典』のうち SJIS コードを割り振ることができたのは，4,053 字種 22,061 文字である．

このようにして，『くずし字解読辞典』から抽出した文字画像について，改良型方向線素特徴量を算出し，文字認識用辞書とした．文字認識は，試験データの文字画像から改良型方向線素特徴量を求め，文字認識用辞書のなかからユークリッド距離が近い順に正解候補を選択し，そのユークリッド距離を認識スコアとする方法をとった．その際，正解候補中におなじ文字コードを持つ候補が複数出現した場合は，それらのうちのユークリッド距離の最小値をもってその文字の認識スコアとした．

3.2 実験

古文書文字認識の試験データとして，「伏見屋文書」から 30 文書 (3,509 文字) をランダムに選択し，そのすべての文字を手作業で切り出して文字画像データを作成した．試験データの作成は，作業進行上の制約により，つぎのような手法をとった．

1. 原文書をスキャニング
2. 画像をいったんシートにプリント
3. 専門の翻刻者がマーカーで 1 文字を囲むようにシート上に記入
4. マーク済みシートをスキャニング
5. マーキングされた 1 文字を画像から自動切り出し
6. 2 値化してノイズ除去処理
7. 文字コードとの対応づけ

このようにして作成された試験文字画像データについて，前節の方法によって文字認識を施した．正解候補の算出にあたっては，計算時間の短縮のため，文字認識用辞書データ数の 5% にあたる上位 1,212 文字まで候補を求め，それ以下の順位をとる候補は切り捨てた．

3.3 結果と考察

実験の結果，正解が 1,212 位までに入ったものは，試験データ全体の 73.64% にあたる 2,584 文字で，正解順位の平均値は 112.80 位であった．

この結果は，この方式による OCR 単独では古文書翻刻支援のための実用にはほど遠いことを示している．しかしこれは，つぎの理由からじゅうぶんに

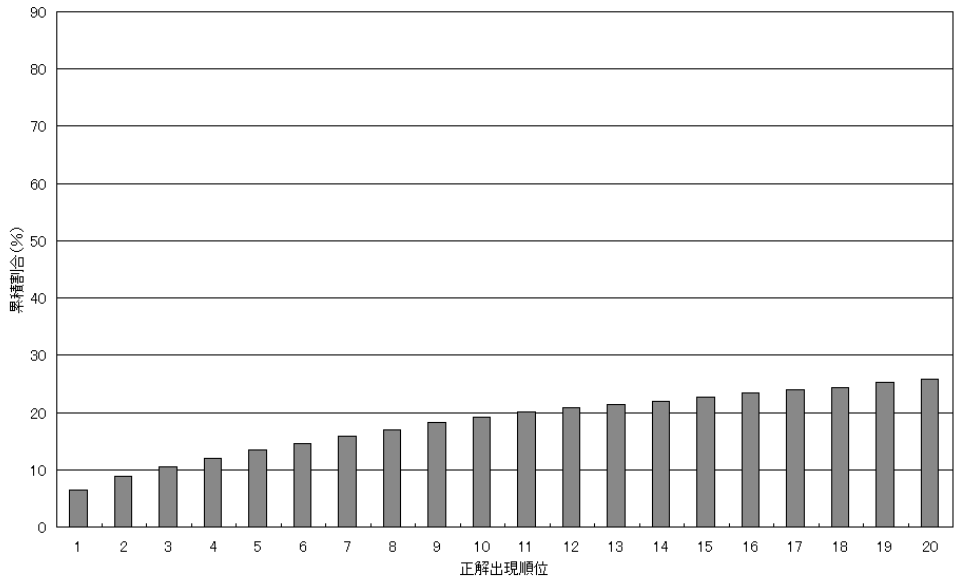

図 1: OCR による正解出現順位の累積割合

予想される結果であった．第 1 に，文字認識用辞書の規模が小さい．すなわち，1 文字あたりのサンプル数が少ない．第 2 に，OCR アルゴリズムは既存の日本語手書き文字認識用のものをそのまま適用しているので，古文書文字に対して最適化がされていない．第 3 に，認識方法としてユークリッド距離法というごく単純な方法を用いている．これらはひとつひとつが大きな研究テーマであるので，本論文の課題からは除外する．

本論文では，OCR としては改良の余地を残す方法ではあっても，そこから得られるあらたな情報を有効に活用する方法を探りたい．本手法は，全体としての認識率の点で劣るとはいえ，なかには良好に文字認識ができている試験データもある．図 1 は，OCR で得られた正解候補のうち，20 位以内に正解があった例の累積割合である．OCR の結果，正解が 1 位にきた例は，試験データ全体の 6.41% にあたる 225 例あり，上位 10 位に入った例は 19.12% にあたる 671 例，上位 20 位では 25.79% にあたる 905 例あった．数は少ないとはいえ，これら OCR から得られた情報を有効に利用し，n-gram 情報による方法と OCR 結果を組み合わせることで，n-gram 情報のみの場合よりも不可読文字の推定結果を向上させうる見込みがある．

4 n-gram と OCR の併用方法の考察

n-gram 情報と OCR 結果を併用した総合スコア ($TScore$) を，つぎのように設定する．

if $NScore(t_{**}) \neq NONE$

$$TScore(t_{**}) = NScore(t_{**}) * OScore1(t_{**})$$

else

$$TScore(t_{**}) = OScore2(t_{**})$$

現状では OCR の信頼性が低いため，$TScore$ の算出にあたっては OCR 結果のなかでとりわけたかいスコアを出した結果のみを選択して使用することが妥当である．すなわち，試験データと学習データの文字特徴量のユークリッド距離を $ED(t_{**})$ とすると，

if ranking of $t_{**} < Threshold1$

$$OScore1(t_{**}) = ED(t_{**})$$

else

$$OScore1(t_{**}) = NONE$$

if ranking of $t_{**} < Threshold2$

$$OScore2(t_{**}) = ED(t_{**})$$

else

$$OScore2(t_{**}) = NONE$$

と定式化され，$TScore(t_{**})$ の昇順で t_{**} を正解候補とする．

この操作はすなわち，n-gram 情報からの推定結果が上位にあっても OCR 結果が悪い場合はスコアを下げ，前者の結果がさほど上位でなくとも，後者の結果がとくに良ければスコアを上げることになる．また，n-gram 情報から正解候補が得られない場合は，OCR 結果のみから正解候補を出す．

ここで問題になるのは，OCR 結果の正解候補数のしきい値 $Threshold1$ と $Threshold2$ をどのレベルにするかである．

まず，試験データのなかで n-gram 情報から正解候補が得られた 2,794 文字について，OCR 結果の併用を検討する．

図 2 は，n-gram 情報がある場合について，$Threshold1$ の変化による正解の平均順位をみたものである．図 2 によると，$Threshold1 = 15$ 付近で平均順位が 4.77 位となることがわかる．この付近の整数値を調べたところ，実際に $Threshold1 = 15$ で平均順位がもっともたかくなる．したがって，n-gram 情報による候補が得られた試験データについては，$Threshold1$ をこの値に設定するのが妥当であると考えられる．

つぎに n-gram 情報により候補が得られなかった 715 文字について検討する．n-gram 情報で候補が得られない場合，OCR 結果のみを情報として用いる．当然のことながら，$Threshold2$ の値を大きくするにしたがって候補中に正解が出現する割合はたかくなるが，候補数は $Threshold2$ 個得られることになる．候補中に正解が出現する率と比較して正解が出現しない率のほうがたかいため，$Threshold2$ の値を大きくすることは，正解を含まない候補をむやみに多く出力する結果を招く．したがって，これら 715 文字についても，$Threshold2$ の妥当な水準を決定する必要がある．

$Threshold2$ の妥当な水準の決定方法として，$TScore(t_{**})$ を基準にした場合の，候補中の正解順位の平均値を最小化する方法を採用することにする．図 3 は，その結果である．$Threshold2 = 4$ で，正解の平均順位が最小の 4.69 位 ($\sigma = 7.54$) となった．

図 4 は，n-gram 情報のみの場合と OCR 結果を併用した場合とで，不可読文字が候補中の何位にあらわれるかを比較したものである．両者を併用した場合，正解順位の最頻値は 2 位，最大値は 129 位となった．これらのしきい値で正解が 1 位となる文字数は，全体の 17.98% にあたる 631 文字，正解が 10 位以内に入る文字数は，全体の 74.35% にあたる 2,609 文字，20 位以内だと全体の 79.77% にあたる 2,799 文字である．この結果は，n-gram 情報のみの場合に正解の平均順位が 5.42 位，1 位が 8.49%，10 位以内が 71.19%，20 位以内が 76.63% であったのと比較すると，不可読文字の推定性能が上昇していることを示している．とくに，正解が候補の 1 位となる割合について，OCR 結果を併用することの効果が顕著である．同時にこれらのしきい値では，全体の 18.07% で正解を含まない 4 個の候補を出力することになる．

5 おわりに

本論文では，古文書の翻刻作業中に遭遇する不可読文字について，前後の文字の n-gram 情報と不可読文字画像の OCR 結果を併用して正解候補を求める手法を提案し，250,000 文字を超える古文書文字データと 27,000 文字を超える古文書文字画像データを電子化して手法の検証をおこなった．提案手法により 3,509 文字の試験データの 81.93% について，正解の平均順位が 4.69 位，20 位以内に正解が得られる割合が 79.77% という結果が得られた．これらの結果は，提案手法を古文書翻刻支援システムに実装した場合の有効性を示唆するものである．

ただし，本論文で問題にした不可読文字にはいくつかのタイプがあり，提案手法では対応できないものもある．たとえば，n-gram 情報では不可読文字の前後の文字はただしく翻刻されていることが前提になる．前後の文字がそもそも誤って読まれていたり，不可読文字が連続する場合には，提案手法ではよい精度は得られない．また，OCR では背景ノイズが少なく，「にじみ」や「かすれ」の少ない文字画像が必要である．古文書の文字では，紙の虫食いなどによって文字の一部が欠けてしまっていて，OCR がそもそも不可能な例も多い．

これらの限界はあるものの，古文書の全自動読み取りではなく，あくまで人間の作業を支援するシステムのための方法として，提案手法がある程度有効である可能性を示すことができたのではないかと

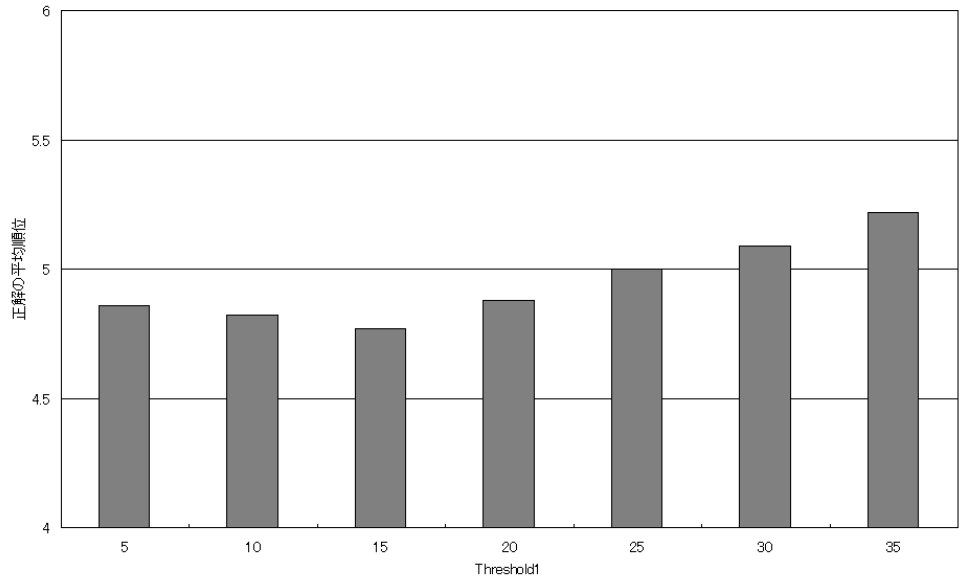

図 2: OCR 結果を加味することによる正解の平均順位（n-gram 情報がある場合）

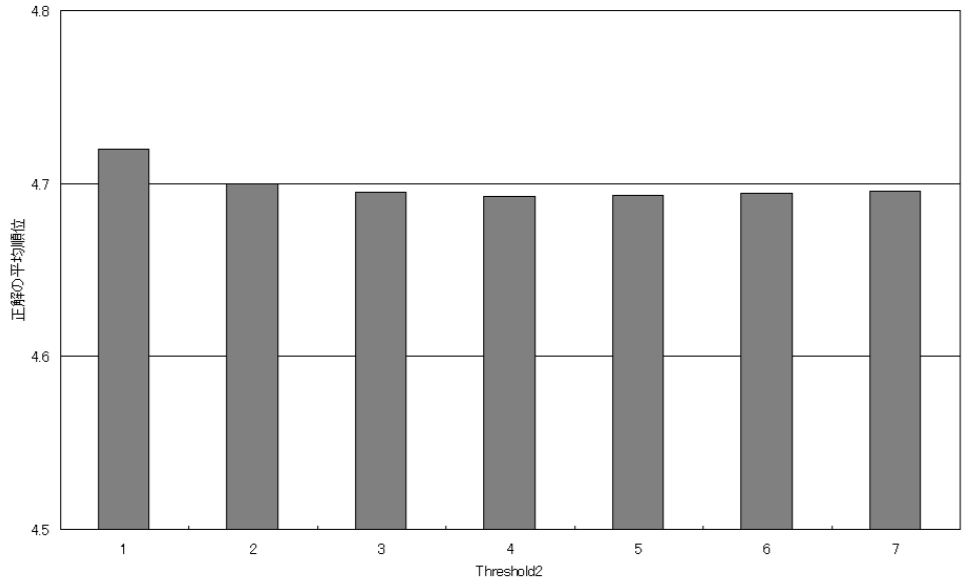

図 3: 正解の平均順位 ($Threshold1 = 15$)

考えられる．本論文では OCR の認識方法については，ごく初歩的な方法をとった．今後 OCR を古文書のために最適化することにより，不可読文字の推定精度がさらに向上することが，じゅうぶんに期待できる．

謝辞

OCR についてご指導をいただいた東北大学の加藤寧，和泉勇治の両氏，『くずし字解読辞典』の電子化と利用を許諾いただいた（株）東京堂出版，ならびに「伏見屋文書」のデータ作成にご尽力いただいた岡屋純子さんと常田律子さんに謹んで感謝の意を表する．

参考文献

[1] 山田奨治: 古文書 OCR 研究の現在, 人文学と情報処理, No. 18, pp. 2–5 (1998).

[2] 山田奨治, 柴山守: 古文書を対象にした文字認

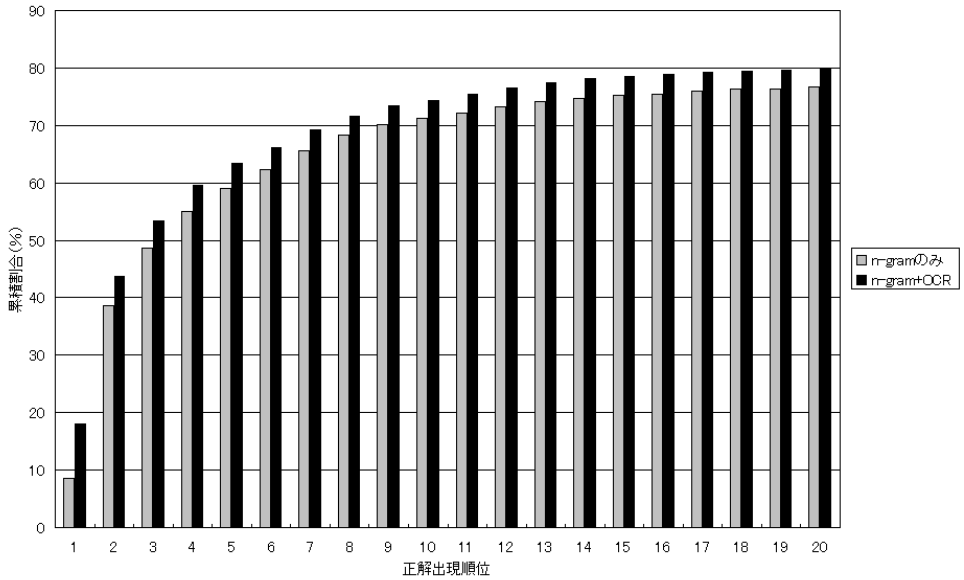

図 4: n-gram と OCR の併用による不可読文字の正解順位の分布 ($Threshold1 = 15, Threshold2 = 4$)

識の研究, 情報処理, Vol. 43, No. 9, pp. 950–955 (2002).

[3] 和泉勇治, 加藤寧, 根元義章, 山田奨治, 柴山守, 川口洋: ニューラルネットワークを用いた古文書文字認識に関する一検討, 情報処理学会研究報告, Vol. 2000, No. 8, pp. 9–15 (2000).

[4] 梅田三千雄, 橋本智広: 認識処理を援用した文字切り出しによる古文書のキャラクタスポッティング, 電気学会論文誌 C, Vol. 122, No. 11, pp. 1876–1884 (2002).

[5] Nagao, M. and Mori, S.: A New Method of N-gram Statistics for Large Number n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, *COLIN 94 : The 15th International Conference on Computational Linguistics : Proceedings,* pp. 611–615 (1994).

[6] 山田奨治, 柴山守: n-gram による古文書証文類翻刻支援の検討, 情報処理学会シンポジウムシリーズ, Vol. 2000, No. 17, pp. 185–192 (2000).

[7] 孫寧, 安部正人, 根元義章: 改良型方向線素特徴量および部分空間法を用いた高精度な手書き文字認識システム, 電子情報通信学会論文誌 D-II, Vol. J78-D-II, No. 6, pp. 922–930 (1995).

[8] 児玉幸多編: 毛筆版くずし字解読辞典, 東京堂出版 (1999).

[9] http://www.mojikyo.org/