n-gramによる古文書証文類翻刻支援の検討

○山田 奨治†, 柴山 守‡

†国際日本文化研究センター・研究部
‡大阪市立大学・学術情報総合センター

われわれは古文書証文類を対象に，翻刻時に遭遇する読めない文字（不明文字）の前後文字からn-gramの情報を使って不明文字の正解候補を提示する可能性について検討した．用例データとして大阪市立大学所蔵『伏見屋善兵衛文書』の全文約243,000文字を翻刻し，翻刻支援手法の検討と検証をおこなった．その結果，前後の既知文字から3-gramおよび2-gramの情報を使って不明文字の正解を検索する実験により，第10候補までで72.70％の正解率を得られると推定できた．本手法をMicrosoft Wordのマクロとして実装した古文書翻刻支援システムの利用試験をおこなったところ，翻刻経験のない初心者が辞書なしで翻刻した結果の正解文字数が有意に増加することがわかり，システムの有効性が確かめられた．

A study of a historical document research supporting system using n-gram

YAMADA Shoji† and SHIBAYAMA Mamoru‡

†International Research Center for Japanese Studies
‡Media Center, Osaka City University

In this article, we study a historical document research supporting system. We examined a n-gram application for candidate selection of unknown characters appearing in business transactions documents of Edo period. We converted the full text of "Fushimiya Zenbei Monjyo" (approximately 243,000 characters) into text file and used it as the corpus of our study. We found that the correct candidates of unknown characters selected from 3-gram and 2-gram were included in the top ten candidates for 72.70 percent. We implemented this method on Microsoft Word and conducted an usability test, which indicated a significant increase in readability for novices without using dictionaries.

1 はじめに

古文書の翻刻（文書を解読して記述内容を活字にすること）は，歴史研究の基礎的

作業である．国内には未翻刻の古文書が膨大な数あり，多くは未だ手つかずのまま文書館などに眠っている．古文書翻刻作業を推進することは，歴史研究を前進させるために必要不可欠のことである．現在のところ，古文書翻刻作業は人手に頼らざるをえないにもかかわらず，翻刻しなければならない残存古文書数に比して翻刻作業にたずさわっている人間の数は非常にすくないといわざるをえない．また古文書翻刻作業が高度に専門的な作業で，一人前の作業者となるまでに相当の訓練を経なければならないことが，問題を困難なものにしている．

われわれは，古文書翻刻作業を支援するシステムの開発をめざして研究を進めている [1][2]．プロジェクトの主眼は，(1) 古文書文字認識システムの研究 [3]，(2) 古文書文字認識システム研究のための古文書文字データベースの作成，(3) 古文書からの文字切り出しの研究 [4][5]，(4) 古文書に関する知識を用いた翻刻支援の研究，(5) 電子化古文書文字辞典の研究においている．本報告は，これらのうちの (4) 古文書に関する知識を用いた翻刻支援の研究に関するものである．

古文書には多くの種類があるが，近世の借金証文類は様式が比較的一定しており，使用されている用語には定型がある．たとえば，「依而如件」「実正也」などの用語は必ずといってよいほど文書のなかに登場する．その他の用語についても，借金証文のなかでよく使われるものがみられる．

借金証文のように使用される用語に定型がみられる種類の文書については，多くの用例を集めてそこから用語に関する知識を抽出し，知識にしたがって翻刻者を支援する方法が考えられる．具体的な方法としては，n-gram を利用することの有効性が予想される．

われわれは古文書証文類を対象に，翻刻時に遭遇する読めない文字（不明文字）の前後文字から n-gram の情報を使って不明文字の正解候補を提示する可能性について検討した．証文類の用例データとするために，大阪市立大学所蔵『伏見屋善兵衛文書』の全文を翻刻した．さらに，本手法を実装した翻刻支援ユーザインタフェースを作成し，被験者を用いた利用試験を実施し，その結果，システムの有効性を確認することができた．

2 n-gram による不明文字候補検索実験

2.1 検索手法

n-gram による不明文字の正解候補検索手法は，つぎのとおりである．

検索対象である不明文字を c_i とすると，その前後の文字のつながりは，

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

と表現される．

一方，用例データから得られる n-gram テーブルはつぎのように定義される．

$$t_{j,1}t_{j,2}\cdots t_{j,n},\quad f_j$$

ここで $t_{j,1}$ は用例中に登場する n 文字のつながりの 1 文字目，$t_{j,2}$ は n 文字のつながりの 2 文字目，f_j はその n 文字のつながりの頻度である．

n-gram テーブルからの不明文字の正解検索は，前方一致の場合と後方一致の場合にわけられる．前方一致は $c_{i-n+1}\cdots c_{i-1}$ と $t_{j,1}\cdots t_{j,n-1}$ のマッチングをとることであり，後方一致は $c_{i+1}\cdots c_{i+n-1}$ と $t_{j,2}\cdots t_{j,n}$ のマッチングをとることになる．

前方一致のケースと後方一致のケースにおける候補文字の確率を総合して，つぎの

ような第1候補文字 $t_{k,n}$ の選択基準を定義する．

前方一致した n-gram の集合を $\{t_{k*}\}$，後方一致した集合を $\{t_{l*}\}$ とすると，

$$\max_{t_{k,n}} F(f_k, f_l) = \max(f_k, f_l; t_{kn} = t_{l1}).$$

以下，$F(f_k, f_l)$ の降順に，$t_{k,n}$ を第2候補，第3候補…とする．

2.2 用例データベース

n-gram による古文書翻刻支援のための用例データとするために，大阪市立大学所蔵の『伏見屋善兵衛文書』（図1）の全文を翻刻した．『伏見屋善兵衛文書』は，大阪の元伏見坂町（現在の大阪市南区坂町）の茶屋，伏見屋善兵衛家に伝わった文書である．伏見屋善兵衛は，遊興の地である伏見坂町のなかでも最大の茶屋として栄えた．また町年寄をつとめ，芝居興業にも関係し，何軒かの貸家をもち，また金融業を営んだ．

図 1: 『伏見屋善兵衛文書』

本文書は，文化から慶応年間にいたる各種の証文類である．芝居関係では，天保年間を中心に歌舞伎役者の芝翫，我童らの手附証文がある．伏見屋の金融・借家，同家内部の親族関係に関する諸証文・議定等も含まれている．文書の総数は，証文類が約1,300である．『伏見屋善兵衛文書』の全画像および目録情報は，大阪市立大学学術情報総合センターのホームページから公開されている [6]．

『伏見屋善兵衛文書』を全文翻刻した結果，用例データ量は約243,000文字となった．

2.3 不明文字検索実験結果

古文書翻刻中に遭遇する不明文字の正解候補を，用例データから作成した n-gram を用いて検索することの有効性を試験した．『伏見屋善兵衛文書』全文データから無作為に10文書を選択して，それらの日付と署名部分を除く表題と本文部分を試験データとし，残りの文書の全文データから5-gram までを作成して教師データとした．n-gram の作成は，長尾らの方法 [7] によった．

試験データの全1,553文字を1文字づつ取りだし，それらを不明文字と仮定して教師データから作成した n-gram をもとに不明文字の正解候補を出した．$n = 2$ から5までについて n-gram から不明文字の正解候補を出し，候補文字中の累積正解出現率を第50候補まで求めたものが，図2である．

図2によると，$n = 2$ から5までの間の累積正解出現率は $n = 3$ で最大となることがわかる．したがって，古文書翻刻支援の

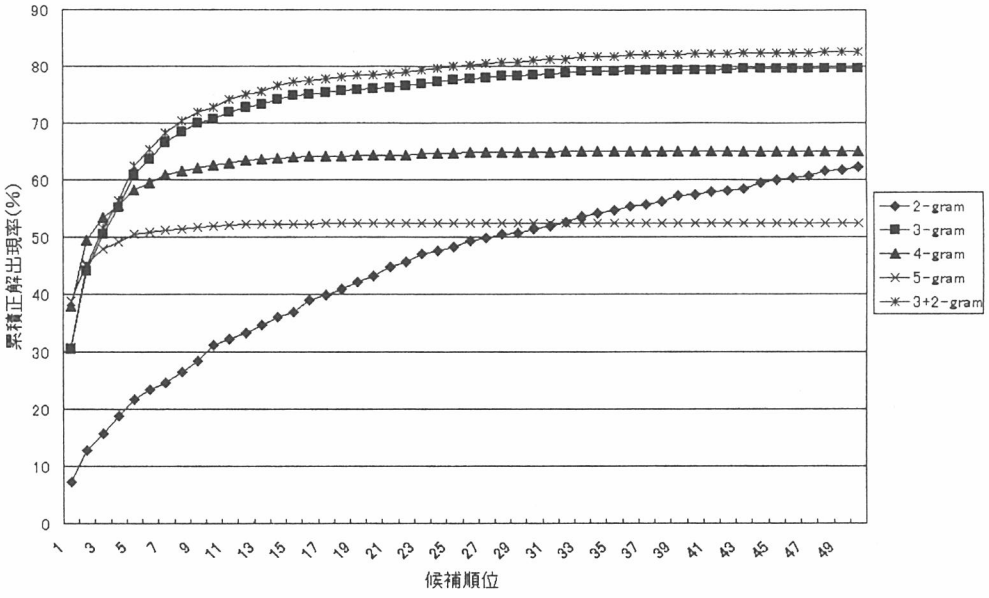

図 2: $n = 2$ から 5 までの候補順位別累積正解出現率（第 50 位まで）

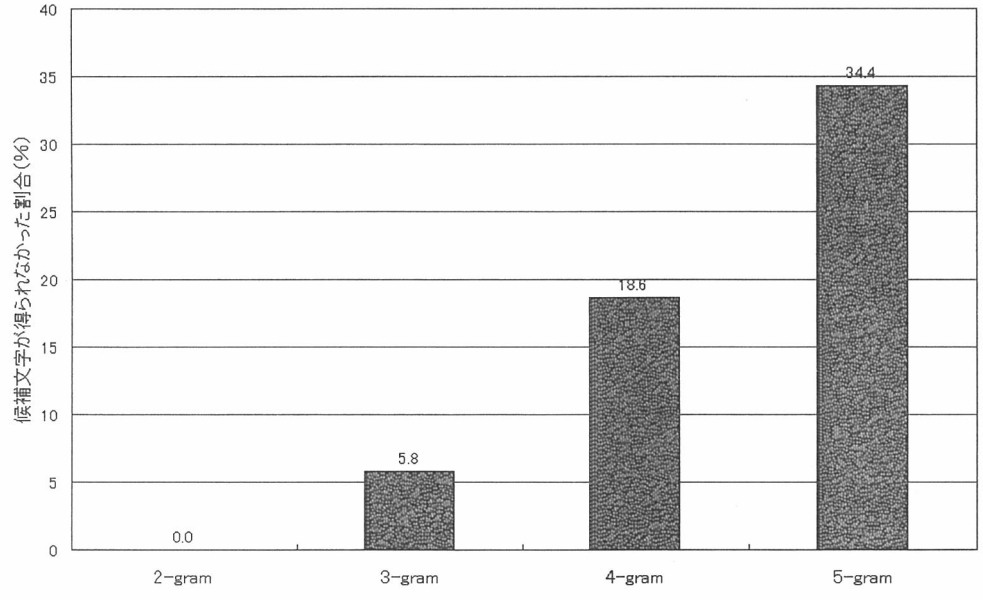

図 3: 候補文字が得られなかった割合

ためには，用例データの3-gramを知識として用いることが適当であると考えられる．

2-gramから5-gramまでで候補文字が得られたなかった割合を示したものが，図3である．3-gramでは候補文字が得られなかった割合が5.8％であるのに対して，2-gramではすべての不明文字に対して候補文字が得られた．2-gramは図2にみられるように正解出現率の点で3-gramに劣るものの，候補文字を提示する能力においては3-gramよりも優れている．したがって古文書翻刻支援のためには，3-gramで正解候補を示し得ない不明文字に対しては2-gramを適用することが有効であると考えられる．実際に3-gramで正解候補が得られなかった場合に2-gramを適用する手法（以降3+2-gramとする）を用いて，おなじ試験をしてみた結果が，図2中の3+2-gramのグラフである．

図4は，3+2-gramで得られた正解候補数の頻度分布である．正解候補数の平均値は18.47候補，最頻値は1候補，最大値は286候補であった．

システムとしての実用性を考慮した場合，正解が第10候補までに入ることをひとつの目安としうる．表1は，3+2-gramを用いた場合の第10候補までに正解があらわれた累積割合である．第10候補までに正解があらわれた割合は，72.70％であった．また正解があらわれた最高は第250候補で，その累積正解出現率は83.77％あった．

3 古文書翻刻支援システムの利用試験

3.1 ユーザインタフェースの実装

『伏見屋善兵衛文書』の全文用例データから3+2-gramを用いて不明文字の正解候補

表1: 第10候補までに正解があらわれた累積割合 (3+2-gram)

候補	累積割合（％）
1	30.97
2	44.95
3	51.44
4	56.34
5	62.40
6	65.23
7	68.26
8	70.32
9	71.93
10	72.70

を提示する機能を持った，翻刻支援のためのユーザインタフェースを試作した．ユーザインタフェースは，Microsoft Word 2000のマクロ言語であるVisual Basic for Applicationを利用して作成した．Wordの操作画面から本手法による不明文字検索機能を呼び出し，正解候補をWord入力画面に反映できるようになっている．画面例を図5に示した．

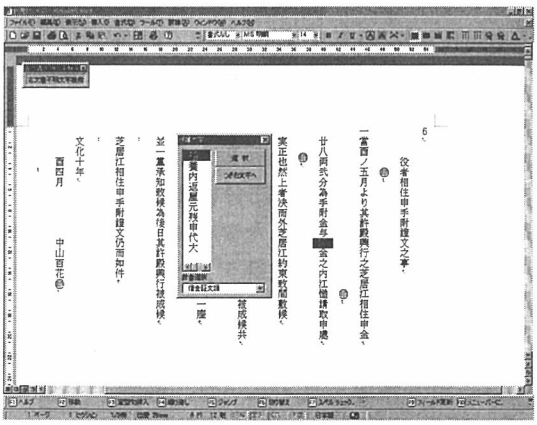

図5: 翻刻支援ユーザインタフェース

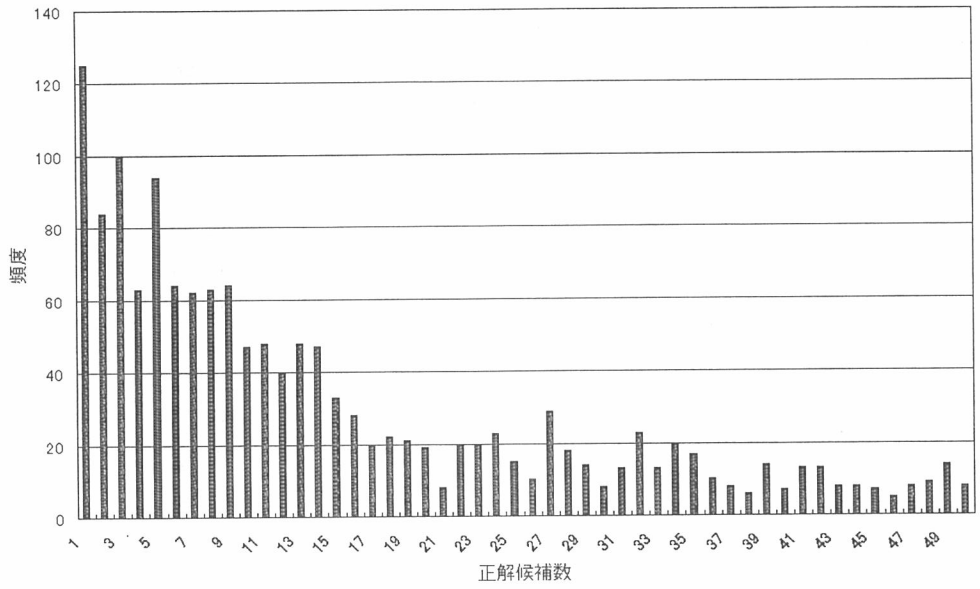

図 4: 正解候補数の頻度分布 (3+2-gram，50 候補まで)

3.2 利用試験

試作した古文書翻刻支援システムの有効性を試験するために，古文書翻刻経験のない被験者を使って利用試験を実施した．被験者に『伏見屋善兵衛文書』のなかの 1 文書（図 1）の紙焼きを示し，その表題と本文部分のみを辞書など参考資料を一切使わずに自分の力で翻刻し，翻刻文を Microsoft Word で入力してもらった．解読できない不明文字は,「□」で入力するよう支持した．その作業が終了した後，Word 上で古文書翻刻支援システムを起動し，システムから提示された「□」の部分の候補文字をみて，被験者が正解と思った文字を「□」と置換した．システムの教師データからは，翻刻対象文書の用例データを除外した．

作業時間の制限は設けず，被験者が納得いくまで作業してもらった．被験者は 30 〜 40 歳代の男女 3 名である．被験者はいずれも古文書翻刻の経験はないが，1 名（被験者 A）は入門程度の古文書読解教育を受けたことがある．

被験者ごとの利用試験結果を，システム使用前と使用後でまとめたものが表 2 である．3 被験者を平均すると，システムの利用によって正解文字数は 9.3 %増加し，不明文字数は 10.8 %減少したが，不正解文字数も 1.5 %増加した．しかし平均の片側 t-検定の結果，正解文字数は有意水準 5 %で，不明文字数は有意水準 10 %でそれぞれ増加しているといえるが，不正解文字数は有意に増加しているとはいえないことがわかった．したがって，本システムは古文書翻刻支援に有効であるといえることが確かめられた．

被験者の誤りについてさらに詳しく分析してみる．システム使用前は不明文字で，使用後も不明文字のままであるか不正解文字となった文字のうち，システムが提示した候補文字のなかの第 1 候補に正解があった文字数は，被験者 A が 1 文字，被験者 B が 6 文字，被験者 C が 3 文字であった．また

表 2: 「古文書翻刻支援システム」利用試験結果

	被験者 A		被験者 B		被験者 C		平均	
	使用前	使用後	使用前	使用後	使用前	使用後	使用前	使用後
正解文字数	47(69.1)	55(80.9)	23(33.8)	26(38.2)	22(32.4)	30(44.1)	(45.1)	(54.4)**
不正解文字数	5(7.4)	5(7.4)	13(19.1)	12(17.6)	22(32.4)	26(38.2)	(19.6)	(21.1)
不明文字数	16(23.5)	8(11.8)	32(47.1)	30(44.1)	24(35.3)	12(17.6)	(35.3)	(24.5)*

括弧内は%. $*: p < 0.1, **: p < 0.05$.

第10候補文字のなかに正解があった文字数は，それぞれ2文字，8文字，5文字であった．これらの文字は，システムが上位に正解をあげていたにもかかわらず，被験者に古文書読解知識がないために，正解として認知されなかった誤りである．その1例を図6に示した．1文字目の「然」について，被験者Aはシステム使用によって正解を得たが，被験者BとCは正解が全11候補中の第1位に示されていたにもかかわらず，この文字を「然」とは翻刻しなかった．3文字目の「者」も第1候補に正解が示されていたが，すべての被験者が「者」とは翻刻しなかった．

図 6: 被験者が正解を認知できなかった1例（然上者）

図7の部分は，すべての被験者が末尾の「為」は翻刻できたが，他の部分は文字の切り出しができず，文字数すら判断できない結果となった．

図 7: 被験者が正解を認知できなかった1例（急度返済可仕候為）

4 まとめ

以上の結果,『伏見屋善兵衛文書』の全文を対象として，前後の既知文字から3-gramおよび2-gramの情報を使って不明文字を検索する実験により，第10候補までで72.70%の正解率を得られると推定できた．さらに本手法を実装した古文書翻刻支援システムの利用試験をおこなったところ，翻刻経験のない初心者が辞書なしで翻刻した結果の正解文字数が有意に増加することがわかり，システムの有効性が確かめられた．この結果は，辞書を併用した場合や翻刻経験者が使用した場合のさらなる有効性を示唆するものである．

本手法は，不明文字の前後の文字が正しいと仮定して，その情報から不明文字の候補を提示するものである．したがって，前後の文字がそもそも誤っていたり，文字数の推定が誤っていたり，不明文字が連続してしまった場合には，正しい候補文字の提示ができない．本手法の応用として，英文のスペルチェックに対応するような，翻刻済み文字に対する検証システムのようなものも考えられるだろう．また本手法は，証文類という一定の表現が頻出するパターンをとる文字列に対して有効な手法であって，その他の種類の文書対してこの手法がどの程度有効であるかは今後の検討が必要である．

謝辞

本研究は，日本学術振興会科学研究費補助金・基盤研究 (B)(1) 一般研究「古文書解読プロセスの知能情報学的解明」（平成 11 ～ 13 年度，研究代表者：山田奨治），同展開研究「手書き文字ＯＣＲ技術を援用した古文書翻刻支援システムの開発」（平成 11 ～ 13 年度，研究代表者：山田奨治），同一般研究「古文書ＯＣＲの試論的研究」（平成 11 年～ 13 年度，研究代表者：柴山守）の支援を得て実施しているものである．

参考文献

[1] 山田奨治, 加藤寧, 川口洋, 原正一郎, 石谷康人, 柴山守, 笠谷和比古, 小島正美, 梅田三千雄, 山本和彦: 古文書翻刻支援システム開発プロジェクト報告（１）ープロジェクト概要ー, 情報処理学会研究報告, Vol.2000, No.8, pp.1-8, 2000.

[2] 山田奨治, 柴山守: 平成 11 ～ 13 年度日本学術振興会科学研究費補助金研究成果（中間）報告書　古文書翻刻支援システムの研究 (1), 2000.

[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個別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] 尾崎浩司, 柴山守, 荒木義彦: 古文書画像のレイアウト認識と標題抽出, 情報処理学会研究報告, Vol.2000, No.67, pp.47-54, 2000.

[5] 原正一郎: 古典 OCR のための文字切り出しについて, 情報処理学会研究報告, Vol.2000, No.67, pp.55-64, 2000.

[6] http://fdip01.media.osaka-cu.ac.jp/opening/fushimi1.html

[7] 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.