日本史史料読解支援のための候補文字検索

山田太造† 井上聡‡ 遠藤珠紀‡ 久留島典子‡
†人間文化研究機構本部 ‡東京大学史料編纂所

史料を読解して記述内容を活字にする翻刻は歴史学や史料学の研究を進める上で重要な作業の 1 つであるが，非常に高度な知識が必要とされる．本研究では，史料読解を支援するため，入力してあるテキストに応じて，次に入力される文字の候補を提示する n-gram モデルを用いた候補文字検索手法を提案する．文字推奨機能の有効性を評価するため実験を行った．その結果，検索結果の上位 5 件で 0.696，上位 20 件で 0.822 のヒット率であった．

A Candidate Character Search for Reading Support of Japanese Historical Documents

Taizo Yamada† Satoshi Inoue‡ Tamaki Endo‡ Noriko Kurushima‡
†Head Office, National Institutes for the Humanities
‡Historiographical Institute, The University of Tokyo

A decoding to reprint is one of important factors to advance studies of history and historical document, however, its work is needed very high skill and knowledge of history. In this paper, we propose a method of a candidate character search for assisting decode. The search method is based on characteristic n-gram model. Using the method, a user can obtain a set of a candidate character which appears immediately after an entered string. For evaluating the effectiveness of the method, we experimented to hit ranking. As experimental results, hit ratio was 0.696 in case of top-rank 5, and hit ratio is 0.822 in case of top-rank 20.

１．はじめに

史料を読解して記述内容を活字にする翻刻は史料編纂，歴史学・史料学の研究などを進める上で重要な作業の 1 つである．翻刻は史料の内容を正確に読解して活字にする作業およびその作業の結果としての出力である．史料に記述されている内容を確認し，より深く史料を調査するためには翻刻は不可欠である．

また，日本史史料には，図 1 のように難読文字，擦れ，虫食いによる欠損等により読めない部分が少なからず存在する．例として，図 1 左東京大学史料編纂所蔵『大内義隆書状二月廿八日』の矩形で囲った文字は”候”である．字形から判断することは極めて困難であるが，この文字の前後の文字列から判断することで文字を確定することができる．図 1 中同所蔵影写本『吾妻鏡紙背文書』および図 1 右同所蔵影写本『大外記中原師生母記紙背文書』の楕円で囲った部分は虫食い・破損，文字の擦れにより文字を確定することが極めて困難である．史料を読解するためには，言葉の用法，史料の正確な読解，史料の性格，歴史的背景などのさまざまな史料学的知見が必要であり，その習得には長期にわたる修練が必要とされる．

本研究では，日本史史料読解を支援するため，既存の翻刻を用いたテキスト特徴に基づく候補文字検索機能を提案する．本機能を用いると，システムは，ユーザの入力した文字列（翻刻）に応じて，確定したい部分の文字の候補を提示する．提示した候補文字の中からユーザが文字を選択することで文字を確定する．翻刻の作業において実際に文字を確定していくとき，我々は次の 2 つのシナリオを想定した．

- シナリオ 1: 文字を出現順に読解し，確定していく．
- シナリオ 2: 読解困難な文字を飛ばし，それ以降の文字を確定し，再度読解困難な文字を確定していく．

シナリオ 1 での読解支援では，確定した文字列（読解困難な文字よりも前の文字列）を用いて読解困難な文字を推測し，ユーザにその候補を提示する．シナリオ 2 では，読解困難な文字の前後の文字列を用いて推測し，その候補文字を提示する．本論文ではこの 2 つのシナリオに従った候補文字検索の手法を提案する．

また，翻刻を支援するために，研究対象とする史料に関する情報収集，翻刻の記述・データ構造，翻刻管理，および翻刻や史料に関連する情報の検索を行う仕組みが必要であると考えている．われわれがこれまで構築してきた，候補文字検索の機能を用いて翻刻を行うことができる翻刻支援システムの概要を示す．本システムは，ユーザと対話しながら，入力された史料画像に対して翻刻を行い，確定された翻刻データを格納することが可能である．

本論文はこれ以降，2 つのシナリオに応じた読解困難な文字に対する候補文字検索の手法，お

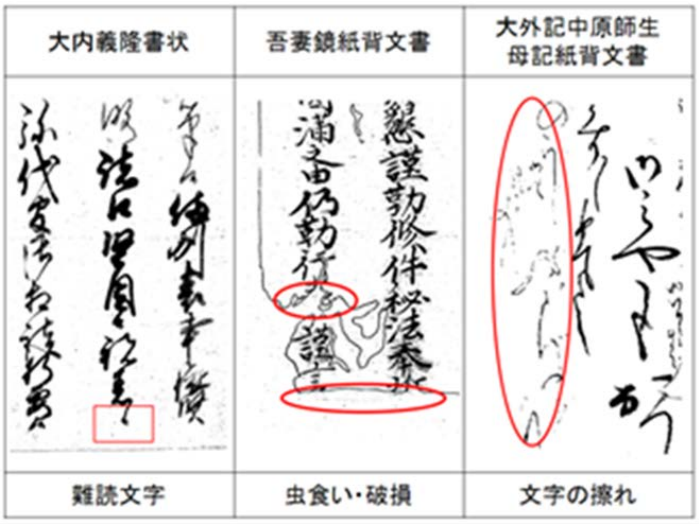

図 1　翻刻を困難にする要因

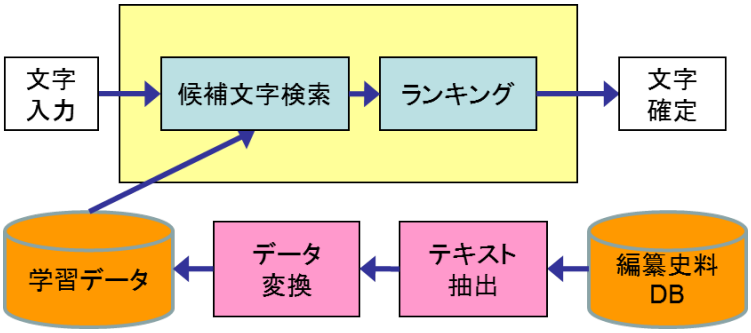

図 2　候補文字検索の手順.

よびその性能を評価した実験結果を示について述べ，これまで構築してきた翻刻支援システムの概要を示す.

2．候補文字検索機能

本研究における候補文字検索機能は，入力されている文字列に応じて，確定したい部分の文字の候補を検索し，候補文字のスコアに応じて上位 r 件をユーザに提示する．最後に，ユーザが候補文字の中から 1 文字を選択することにより文字を確定する．このとき，上位r件の候補文字をユーザに提示する．この手順を図 2 に示す.

文字列$c_1,\ldots,c_{n-1},c_n,c_{n+1},\ldots,c_{n+m}$において，ユーザにより入力された文字列が$c_1^{n-1}=c_1,\ldots,c_{n-1}$および$c_{n+1}^{n+m}=c_{n+1},\ldots,c_{n+m}$であるとき，本候補文字検索は文字$c_n$の候補を検索することになる．シナリオ 1 では$c_{n+1}^{n+m}$は確定していないため$c_1^{n-1}$のみを用いて候補文字検索を行う．シナリオ 2 ではc_1^{n-1}とc_{n+1}^{n+m}はユーザにより確定されているため，両者を用いて候補文字検索を行う．例えば，”駿河国入江庄内”に続く文字を推測するときが，シナリオ 1 に相当する．ある確定できない文字が”駿河国入江庄内”と”沢小次郎妻”の間にあり，それを推測するのがシナリオ 2 に相当する（この間に入る文字は”三”である）．直感としては前後の文字列から推測するシナリオ 2 の方が精度を高くできると考えられる.

c_nの候補文字を検索する手法として，文字 n-gram モデルを用いることにした．一般的に n-gram モデルを用いた場合，スパーネス問題（学習データ中に出現しなかった文字の出現確率が 0 になってしまう問題）がある．例えば，文字列” 庄内”と”沢小”の間の文字を推測するとき，これらの n-gram が学習データ内に存在しない場合，推測することが困難である．これを解決するためスムージングを適用する．多くのスムージング手法では，低次元の n-gram を用いて高次元の n-gram の出現確率を補間することが多い．本研究では Modified Kneser-Ney Smoothing（MKNS）[2,3]によるスムージングを行い，それを用いた候補文字検索も提案する．MKNS は完全ディスカウンティング法（absolute discounting smoothing）とバックオフスムージング法（back-off smoothing）を組み合わせた非線形スムージング手法である.

本章では以下，文字 n-gram モデルおよび MKNS を用いた候補文字検索手法を示し，学習用テキストデータの対象と抽出方法について述べる.

DB名	件数	概要
平安遺文FT	10,842	『平安遺文』,『新訂増補国史大系』など
鎌倉遺文FT	34,291	『鎌倉遺文』など
古文書FT	37,800	『大日本古文書　家わけ文書 東大寺文書』など
古記録FT	64,499	『大日古記録　九暦』など
大日本史料総合	5,580	『大日本史料　6編44』など
合計	153,012	

図３　学習用編纂史料データベース

対象	処理内容	パターン(正規表現)	処理回数 平安遺文FT	鎌倉遺文FT	古文書FT	古記録FT	大日本史料総合
文字以外の表現	削除	"〔参考〕","〔附録〕", "(朱印)","(花押影)","(花押)", "[…—ーー]+","(墨引)","(紙繊目)", "¥(未詳)","※","∴","〇[^本]*本文書[、]*検討を要す[る]*[。、]","〇下略","〇中略","◎[^¥n]*¥n","(〇[^)]*)","{下略[^}]*}","{中略[^}]*}", "(){2,}","( )+","¥()","¥{¥}","()", "「」〈〉∧]","[[](]","[[]]","『』","【[^】]*】"	0 4589 295 143 173953 113883	0 23125 1281 45 443287 34324	0 23592 2957 584 600534 78497	0 524 1396 54 116643 340403	36 1665 1837 3427 38778 4430
欠損文字，難読文字	文字列"xxx"に置換 "？"に置換	"<NOTE CNTS =xxx>□</NOTE>□□","(xxx)□□□","[xxx]□□□" "[=□■@]","△","〃"	65 23806	4644 78735	4406 48871	2518 40897	719 3305
傍注，脚注	削除	"<(/*)NOTE[^>]*>","<NOTE[^>]*>@</NOTE>","<NOTE[^>]*>『</NOTE>","¥$([^¥$]*)¥$","¥([^¥)]*¥)","¥([^¥)]+¥)¥(〃+¥)","¥[([^¥]]*)¥]","note","[[^]]*]","([^)]*)","[[^]]]"	62105	180003	321620	209865	45818
細字双行	文字列"xxx"に置換	"{xxx}"	11689	6826	5588	7164	6002
大漢和コード	"？"に置換	"¥大漢和{¥d*}","<NOTE CNTS=xxx>&</NOTE>#m¥d*;","&#m¥d*;	9921	5801	5520	7148	2184
出力用途の文字列(HTMLやTeXのタグなど)	削除	"-.5zw","0mm","</*TI[^>]*>","<<"," =( )?","[<]?BR/[>]?","【¥d*】"	19	2	4607	167750	41722
文書番号，ページ番号	削除	"</PG>","<PG PID="[^"]*">","#[^¥n]*¥n"	18	0	1223	2	1979
上記以外	削除	"[¥{|¥}]","[x×]","[〇●]{2,}","[〇●o]","[＊－／]","？[～～]？","TE"	173953	443287	600534	116643	38778

図４　テキスト抽出のルール

２．１．n-gram 手法

この手法はc_nの推定を文字 n-gram モデルを用いて行う．文字c_nの生起は先行する N-1 文字にのみ依存する（N-1）重マルコフ過程として仮定し，$P(c_n|c_1^{n-1})$~$P(c_n|c_{n-N+1}^{n-1})$と表す．$P(c_n|c_{n-N+1}^{n-1})$は，学習データ中に出現する文字 n-gram から最尤推定を行うと，

$$P_{ML}(c_n|c_{n-N+1}^{n-1}) = \frac{freq(c_{n-N+1}^{n})}{freq(c_{n-N+1}^{n-1})} \quad (1)$$

となる．ここで $freq(c_1^n)$は学習データでの文字列c_1^nの出現回数を示す．

シナリオ１では$P_{ML}(c_n|c_{n-N+1}^{n-1})$を求め，この値に応じてランキングし，上位 r 件を検索結果としてユーザに提示する．シナリオ２では，c_{n+1}^{n+m}の逆方向の文字 n-gram c_{n+N-1}^{n+1}に対して（1）式を計算し$P_{ML}(c_n|c_{n+N-1}^{n+1})$を求め，

$$P_{ML}(c_n|c_{n-N+1}^{n-1}, c_{n+N-1}^{n+1}) = P_{ML}(c_n|c_{n-N+1}^{n-1}) \cdot P_{ML}(c_n|c_{n+N-1}^{n+1}) \quad (2)$$

を計算することでc_nのスコアを求める．

２．２．MKNS 手法

この手法では n-gram スムージングである Modified Kneser-Ney Smoothing（MKNS）を用いた手法である．シナリオ１では次式で n-gram の確率を計算する．

$$P_{KN}(c_n|c_{n-N+1}^{n-1}) = \frac{freq(c_{n-N+1}^{n}) - D(freq(c_{n-N+1}^{n}))}{\sum_{c_n} freq(c_{n-N+1}^{n})} + \gamma(c_{n-N+1}^{n-1}) P_{KM}(c_n|c_{n-N+2}^{n-1}) \quad (3)$$

$$D(freq) = \begin{cases} 0, if\ freq = 0 \\ D_1, if\ freq = 1 \\ D_2, if\ freq = 2 \\ D_{3+}, if\ freq \geq 3 \end{cases} \quad (4)$$

$$\gamma(c_{n-N+1}^{n-1}) = \frac{D_1(N_1(s\cdot)) + D_2(N_2(s\cdot)) + D_{3+}(N_{3+}(s\cdot))}{\sum_{c_n} c_{i-n+1}^{n}} \quad (5)$$

時代区分	西暦		平安遺文FT		鎌倉遺文FT		古文書FT		古記録FT		大日本史料総合		合計	
			異なり文字数	延べ文字数	異なり文字数	延べ文字数	異なり文字数	延べ文字数	異なり文字数	延べ文字数	異なり文字数	延べ文字数	異なり文字数	延べ文字数
0	～	887	3,834	609,970	0	0	2,092	127,378	135	313	0	0	3,936	737,661
1	887 ～	1086	3,738	464,737	0	0	1,814	77,465	2,401	251,220	0	0	3,942	793,422
2	986 ～	1185	3,848	747,400	0	0	2,414	206,397	3,252	1,681,281	1,946	108,394	4,279	2,743,472
3	1086 ～	1185	3,953	1,305,701	699	8,394	2,800	406,005	3,364	3,355,053	2,275	120,039	4,444	5,195,192
4	1185 ～	1221	1,158	13,098	3,527	945,148	2,486	195,299	2,746	569,550	0	0	3,963	1,723,095
5	1221 ～	1333	1,285	16,250	4,787	6,817,515	3,479	1,473,340	3,371	1,618,059	0	0	5,002	9,925,164
6	1333 ～	1392	597	2,684	1,999	96,670	3,247	893,586	1,379	32,700	3,607	312,415	4,150	1,338,055
7	1392 ～	1466	517	2,896	0	0	3,168	1,250,186	4,113	1,313,271	2,560	119,201	4,539	2,685,554
8	1467 ～	1508	468	1,209	0	0	2,885	623,909	4,008	892,426	2,496	147,879	4,401	1,665,423
9	1508 ～	1568	621	6,520	0	0	3,317	1,088,514	1,588	42,773	2,604	101,746	3,716	1,239,553
10	1568 ～	1582	282	1,865	0	0	2,525	284,610	1,376	71,144	3,228	249,825	3,581	607,444
11	1582 ～	1603	0	0	0	0	2,820	573,231	2,215	328,852	2,272	117,190	3,342	1,019,273
12	1603 ～	1651	586	1,696	0	0	3,135	690,373	1,277	49,307	0	0	3,227	741,376
13	1651 ～		1,784	17,118	45	72	3,475	1,055,563	1,895	192,854	0	0	3,695	1,265,607
		合計	4,438	3,191,144	4,352	7,867,799	4,441	8,945,856	4,986	10,398,803	4,014	1,276,689	5,866	31,680,291

図 5　抽出したテキスト

$$
\begin{aligned}
\mathrm{Y} &= \frac{n_1}{n_1 + 2n_2} \\
\mathrm{D}_1 &= 1 - 2\mathrm{Y}\frac{n_2}{n_1} \\
D_2 &= 2 - 3\mathrm{Y}\frac{n_3}{n_2} \\
D_{3+} &= 3 - 4Y\frac{n_4}{n_3}
\end{aligned}
\tag{6}
$$

ここでsはc_{n-N+1}^{n-1}であり，s·は文字列sの直後に任意の文字が出現するすべての文字列である．$N_1(\mathrm{s}\cdot) = |\{c_n : freq(s)\}|$であり，$N_2(\mathrm{s}\cdot)$および$N_{3+}(\mathrm{s}\cdot)$も同様に定義される．$\mathrm{n}_1 = |t_i : freq(s)|$であり，同様に$\mathrm{n}_2$，$\mathrm{n}_3$および$\mathrm{n}_4$も定義される．

シナリオ 2 では n-gram 手法と同様に文字 n-gram c_{n+N-1}^{n+1}から（3）式で $\mathrm{P_{KN}}\left(\mathrm{c_n}\middle|\mathrm{c_{n+N-1}^{n+1}}\right)$を求め，

$$
\begin{aligned}
&P_{\mathrm{KN}}(c_n | c_{n-N+1}^{n-1}, c_{n+N-1}^{n+1}) \\
&= \mathrm{P_{KN}}(c_n | c_{n-N+1}^{n-1}) \cdot \mathrm{P_{KN}}(c_n | c_{n+N-1}^{n+1})
\end{aligned}
\tag{7}
$$

を計算することで$\mathrm{c_n}$を求める．

２．３．学習用テキストデータの抽出

本研究では，SHIPSDB（東京大学史料編纂所データベース）にある『大日本史料総合 DB』，『平安遺文フルテキスト DB』，『鎌倉遺文フルテキスト DB』，『古文書フルテキスト DB』，『古記録フルテキスト DB』から抽出したテキストデータを学習データとして扱う．図 3 は本研究で対象としたデータベースのデータ件数と登録されているデータの出典について示しており，『大日本史料総合 DB』，『平安遺文フルテキスト DB』，『鎌倉遺文フルテキスト DB』，および『古文書フルテキスト DB』では 1 史料を，『古記録フルテキスト DB』では古記録の 1 段落を 1 件としている．

図 4 は『大日本史料総合 DB』におけるテキストの表現とタグ処理などを施して抽出したテキストデータの例を示す．データベース内でのテキストデータ（「本文」データ）は史料のメタデータ，テキスト，およびテキストのメタデータで構成されている．史料のメタデータには，その史料の ID，日付，史料名，刊本データ（この例では『大日本史料』における編冊）などが記述されている．テキストデータには，史料メタデータの ID，刊本での掲載ページ，テキストなどが記述されている．上記に掲げた他のデータベースにおけるテキストデータはいずれも大日本史料総合 DB と同様の形式で格納されている．この例で示しているように，データベース内に格納されているテキストには，その史料を読むために付与された注記，実物の史料の状態などを示すアノテーション，刊本やデータベースシステムで表示するために必要とされるタグなどが付与されていることが多い．例えば「爲<NOTE CNTS=``報恩寺">新寺</NOTE$>」では``新寺"は``報恩寺"を示す，と意味する．本研究では，上記のような実際の史料に記述されていない情報を排除したテキストを学習に用いることにした．図 4 は学習する上で必要とするテキストのみを抽出するために施したルールと，各データベースでそれぞれのルールを適用した回数を示す．データベースに格納されているテキストの記述方法はいずれも同じ形式である．

図 5 は各データベースから抽出したテキスト内の異なり文字数と延べ文字数を時代区分ごと，および対象となるデータベースごとに示している．各時代区分は『大日本史料』における各編の範囲[4]を元に設定した．また，時代区分0は『大日本史料』第 1 編よりも前の時代区分を，時代区分 13 は『大日本史料』第 12 編よりも後の時代区分を示す．

３．実験

３．１．実験準備

ここでは前章で示した候補文字検索の各手法の有効性を示す．この指標としては，推奨結果内に正解データが含まれる確率（ヒット率）と再現率とした．候補文字リストの上位r件内に正解文字が含まれる確率
(正解が含まれていた件数)/(テストデータ件数)
をヒット率として求めた．再現率は，r→∞であ

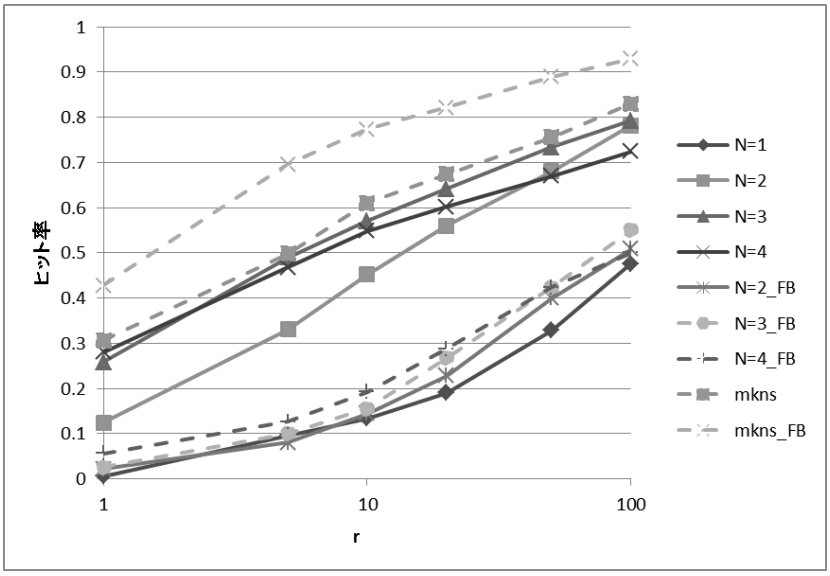

図 6　r を変えたときのヒット率

時代区分	N=1	N=2	N=3	N=4	N=2_FB	N=3_FB	N=4_FB	mkns	mkns_FB
6	1	0.998	0.922	0.852	0.992	0.92	0.782	1	1
11	1	0.992	0.916	0.808	0.99	0.884	0.708	1	1

図 7　再現率

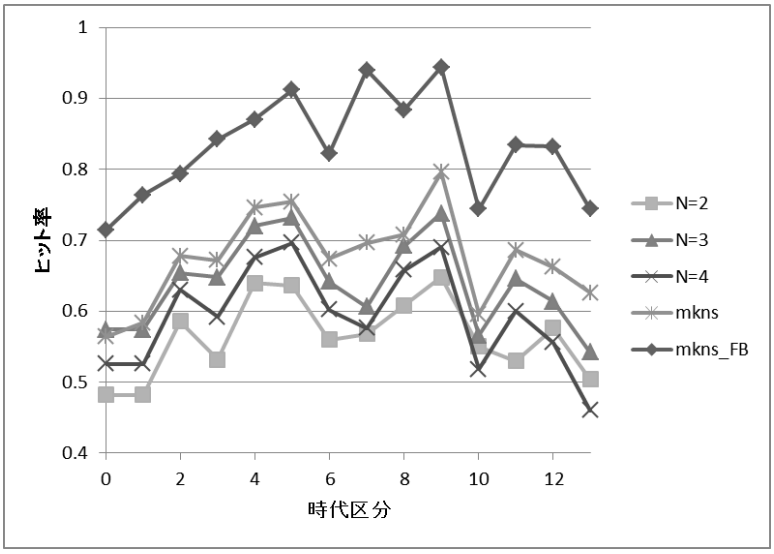

図 8　r=20 のときの全時代区分でのヒット率

りときのヒット率として求めた．各時代区分のデータのうち，500 件をテストデータ，残りを学習データとして扱うことにした．テストデータから任意の位置にある文字を 500 箇所選択し，これをテストデータとした．また，N=1,...,4 とした．

３．２．実験準備

時代区分 6 におけるの r を 1 から 100 まで変動させたときのヒット率を図 6 に示す．x 軸はランクを，y 軸はヒット率を示している．N=1,...,4 はシナリオ 1 における n-gram 手法，N=1_FB,...,4_FB はシナリオ 2 における n-gram 手法，mkns はシナリオ 1 における MKNS 手法，mkns_FB はシナリオ 2 における MKNS 手法の結果を示す．図 7 は再現率を示している．

シナリオ 1 での n-ram 手法において，5≤r≤50 のとき，N=3 がもっともヒット率が高く，r=5 のとき 0.49，r=20 のとき 0.642 だった．r≥100 であれば N=2 のときがもっともヒット率が高くなったが，N=3 とあまり変わらない．また r=1 のとき，N=4 では 0.28 であり，N=3 は 0.258 であったためもっともヒット率が高かったが r=5 で 0.468，r=20 で 0.602 であり，r≥5

データ定義（DTD）

```
<!DOCTYPE root[
<!ELEMENT root(doc*)>
<!ELEMENT
doc(date*,image*,docnote*)>
<!ATTRLIST doc path CDATA>
<!ELEMENT date(#PCDATA)>
<!ATTRLIST date type CDATA>
<!ELEMENT image(text*,imagenote*)>
<!ATTRLIST image src CDATA>
<!ELEMENT docnote(line*)>
<!ATTRLIST docnote userid CDATA>
<!ATTRLIST docnote modified
CDATA>
<!ELEMENT text(line*)>
<!ATTRLIST text userid CDATA>
<!ATTRLIST text modified CDATA>
<!ELEMENT line(str*,note*)>
<!ATTRLIST line num CDATA>
<!ATTRLIST line x CDATA>
<!ATTRLIST line y CDATA>
<!ATTRLIST line height CDATA>
<!ATTRLIST line size CDATA>
<!ELEMENT str(#PCDATA)>
<!ELEMENT note(comment,str)>
<!ATTRLIST note type CDATA>
<!ELEMENT comment(#PCDATA)>
<!ELEMENT textnote(line*)>
<!ATTRLIST textnote userid CDATA>
<!ATTRLIST textnote modified
CDATA>
]>
```

出力例（島津家文書源頼朝下文文治二年八月三日条）

```
<?xml version="1.0" encoding="UTF-8" ?>
<root>
 <doc path="/M00/M/23/1">
  <date type="wareki">文治2年8月3日</date>
  <date type="seireki">11860080030</date>
  <image src="00000011.tif">
   <text userid="t_yamada" modified="20100401101123">
    <line num="1" x="805" y="109" height="200" size="14">
     <str>下島津御庄官等</str>
    </line>
    <line num="2" x="756" y="121" height="344" size="14">
     <str>可令早停止千葉介</str><note type="人名注">
<comment>千葉常胤</comment><str>常胤
</str></note><str>代官字紀太</str>
    </line>
    <line num="3" x="713" y="126" height="194" size="14">
     <str>清遠非道狼藉事</str>
    </line>
    ・・・
   </text>
   <imagenote userid="t_yamada"
modified="20100401101123">画像のメモはここ
</imagenote>
</image>
  <docnote userid="t_yamada"
modified="20100401095541">
史料のメモはここ</docnote>
 </doc>
</root>
```

図 9　翻刻データの定義と出力例

図 10　翻刻支援システムの画面．（左）史料検索機能（右）翻刻編集機能

以上では N=3 でのヒット率には及ばなかった．図 7 から N が高くなるほど再現率が低くなっていることがわかる．本実験での再現率はヒット率の最大値であるため N=4 ではこれ以上のヒット率を示すことはできず，r≥5 以上でもヒット率が高くならない要因となっている．他方，N=1，および N=2 では r の値が低いときのヒット率は低い．これは検索条件から推定される候補文字の選択が困難となるためである．N が大きいほど上位に正解文字が含まれやすくなるが，大きすぎると正解データが含まれにくくなることがわかる．

シナリオ 1 での MKNS 手法では，ヒット率の結果から，いずれの r の値においても，n-gram 手法よりも，格段に高いヒット率を示すことがわかった．また，再現率でも他の方法よりも高く，N=1 と同等であることがわかった．MKNS では，出現しない n-gram を単に線形に補間するのではなく，他の n-gram の出現頻度に応じて n-gram の確率値をディスカウントしている．

図 11　候補文字検索機能

さらに低次元での n-gram の出現頻度も考慮している．そのため，n-gram モデル自体の単純さを強固にサポートすることができていると考えられる

シナリオ 2 での n-ram 手法は，シナリオ 1 に比べ格段にヒット率が低下していることがわかる．この結果だけでは確定したい文字の前方の文字列のみで判断したほうがよい，という直感とは反した結果となった．その理由としては，前方文字列と後方文字列での候補文字検索結果の両方に含まれる文字が少なるためである．後方文字列での候補文字検索の再現率は前方向文字列と同程度であるため（結果は示していない），N=4 のとき，時代区分 6 の再現率は 0.782 程度まで低下してしまった．

シナリオ 2 での MKNS 手法では，n-gram 手法とは異なり，再現率が低下しないことが図 7 より分かった．ヒット率は図 6 から，明らかに n-gram 手法よりも格段に向上していることが分かった． また，シナリオ 1 における MKNS 手法よりもヒット率が高く，図 6 では r=10 のときに約 0.164 も向上した．MKNS 手法においては前方文字列の候補文字検索の結果を，後方文字列での候補文字検索の結果で補正できていることがわかり，読解困難な文字が現れた場合，前方だけではなく後方の文字列も確定した後で候補文字検索を行ったほうがヒットしやすいことがわかった．

図 8 は r=20 のときの全時代区分でのヒット率の結果を示している．この結果より，どの時代区分であってもシナリオ 2 での MKNS 手法がもっとも良い結果を示した．また，図 5 のテキストの延べ文字数と比較した場合，延べ文字数の多い時代区分ほどヒット率が高いことが分かった．これは学習データの量が多いほど，候補文字検索の精度が高くなることを示唆していると考えられる．MKNS 手法を用いた場合，これ以上のヒット率向上を行うためには，翻刻を推し進める必要がある，ということになる．

４．翻刻支援システム

本研究における翻刻支援システムは，ユーザと対話しながら，入力された史料画像に対して翻刻を行い，確定された翻刻データを格納する．以下，本システムにおける翻刻のデータ構造，翻刻の検索・編集機能および本システムでの候補文字検索機能を述べる．

４．１．翻刻データ

本論文では，翻刻および対象となる史料のメタデータから構成される翻刻に必要な情報を翻刻データと呼ぶ．翻刻データは XML 形式で表現する．DTD による翻刻データの定義と『島津家文書源頼朝下文文治二年八月三日条』の一部の翻刻データの例を図 9 に示す．

このデータ定義は以下に示すような階層構造とした．要素``doc"はその子要素``path"で識別される史料を表す．``doc"はさらに要素``image"を持つ．``image"は史料画像を表す．``image"は``text"を要素として持つ．``text"は翻刻を表す．翻刻は，ユーザによって設定される行を単位とし，記述内容を格納する．

４．２．史料検索機能

史料検索機能では，史料の目録階層，テキスト，ユーザ情報に基づいて検索でき，検索結果として史料に関連する史料画像もしくは条件に一致した史料画像が得られる．外部の目録管理システムを利用することで，目録情報および，関連する画像を得ており，本システム独自に史料の目録管理および検索の機能は有していない．翻刻テキストに関する検索では，翻刻もしくは語・表記に関するアノテーション対する解説文に対する全文検索を行う．（図 10（左））．

４．３．翻刻編集機能

翻刻編集機能では，史料画像に対して，画像上の任意の位置へのテキスト配置，画像の拡大表示，翻刻表示などの機能を持つ（図 10（右））．ユーザが画像上の任意の場所をクリックするとその場所にテキストフィールドが配置される．そこにテキストを入力することで翻刻データを編集することができる．テキストフィールドに入力されたテキストにより図 9 で示した``line"要素を構成する． 翻刻は版管理されており，版はユーザ単位に，ユーザによるコミットごとに作成される．履歴表示では，対象史料画像に対する翻刻テキストについて検索・表示・利用できる機能である．自分で作成した過去の版や他ユーザが作成した版を検索し，さらに利用することもできる．他ユーザのテキストを利用・編集し保存した場合，保存したユーザの新たな番として格納される．そのため，あるユーザの操作が他ユーザの翻刻テキストに影響を与えることはない．

４．４．候補文字検索機能

文字推奨機能はユーザの操作によって呼び出され，入力されている文字列に応じて次に入力

される文字の候補を検索し，候補文字の上位 r 件をユーザに提示する．最後に，ユーザによって確定された候補文字が入力対象のテキストフィールドに追記される．本ユーザインターフェースでは，図 11 で示すように，候補文字のリストはセレクトボックス形式で提示し，上位 s(s < r) 件を表示する．また，最大 r 件まで下位の候補文字をスクロールすることで確認することができる．

５．関連研究

史料の読解支援として以下のシステムが挙げられる．1 つは文字画像を用いたシステムである．トランスメディア[5]，SMART-GS[6]，Mokkanshop[7] などの文字画像検索に基づく支援システムは文字画像をクエリとして同型の文字画像を検索できる機能を有している．電子くずし字字典[8] や文字管理システム[9] などでは，史料に出現する文字画像を 1 文字単位，もしくは，文字列単位で切り出し，それにテキストをつけている．しかしながら，図 1 ような史料に欠損・破損がある場合には用いることができない．他方，本研究の手法と同様にテキスト特徴に基づいて支援を行なうシステムもある．HCR（Historical Character Recognition）プロジェクト[10,11] による古文書翻刻支援システムがあり，本研究での候補文字検索手法における n-gram 手法に近い手法であるが，（2） 式とは異なり，N=3 で前方・後方・中間での文字列一致した n-gram の頻度がもっとも高かったものを用いる．N=3 で不定になる場合に限って N=2 を用いているが，本研究のようにスムージングは行わない．

6. おわりに

本研究では，日本史史料の読解を支援すべく，テキスト特徴に基づく候補文字検索を提案し，2 つのシナリオに応じた検索手法とその効果を示す実験を行った．結果として，Modified Kneser-Ney Smoothing を用いた n-gram 手法であれば，南北朝期のテキストに対して，検索結果の上位 5 件で 0.696，上位 20 件で 0.822 のヒット率であることがわかった．また，史料画像と翻刻を関連づける翻刻データの構造．翻刻データの検索・編集するためのユーザインターフェースを示し，これらの機能を有する翻刻支援システムについて示した．

候補文字検索での実験結果では，テキストが少ない時代ではヒット率が低かった．そこで時代区分に応じた学習データを作成するなどしてあらゆる時代でのヒット率向上を目指す．また，検索結果をテキストとして返すが，電子くずし字字典における代表文字のような文字画像とともに結果を示すことでよりユーザに多くの情報を提示することで，読解の支援を向上させることができると考えている．

本研究での翻刻支援システムでは翻刻のみを編集・検索しているが，史料の状態のような史料自体の情報のような文化財としての情報も扱えるようにするなど，史料に関わるあらゆる情報をマッシュアップさせていくことを考えている．史料学・歴史学を進める上で必要なあらゆる情報を管理・編集していくことで，翻刻支援システムを”史学支援システム”として位置づけていく．

謝辞

研究の一部は，日本学術振興会科学研究費基盤研究（S）「史料デジタル収集の体系化に基づく歴史オントロジー構築の研究」（20222001），および若手研究（B）（21700274）「史料学研究支援のためのアノテーション管理基盤に関する研究」による．

参考文献

[1] Chelba, C. and Jelinek, F.: Self-organized language modeling for speech recognition, Readings in Speech Recognition, Morgan Kaufmann, pp.450–506 (1990).
[2] Chen, S.F. and Goodman, J.: An empirical study of smoothing techniques for language modeling, Proceedings of the 34th annual meeting on Association for Computational Linguistics (ACL-96), pp.310–318 (1996).
[3] James, F.: Modified Kneser-Ney Smoothing of n-gram Models, Technical report, RIACS Technical Report 00.07, http://www.riacs.edu/navroot/Research/TRpdf/TR00.07.pdf. (2000).
[4] 東京大学史料編纂所：大日本史料・史料綜覧, http:// www.hi.u-tokyo.ac.jp/publication/nihonshiryo-shiryosoran-j.html.
[5] 田中知朗，田中譲：トランスメディアシステムによる英文テキスト画像処理，情報処理学会論文誌, Vol.38, No.7, pp.1389‐1398 (1997).
[6] SMART-GS: SMART-GS: a tool for humanistics. http://www.shayashi.jp/SMARTGS/ mainjp.html.
[7] 高倉純，SHERINI, S.，耒代誠仁，石川正敏，中川正樹，馬場　基，渡辺晃宏：木簡解読支援のための情報検索，人文科学とコンピュータシンポジウム論文集, Vol.2008, No.15, pp.75–80 (2008).
[8] 東京大学史料編纂所：電子くずし字字典データベース. http://www.hi.utokyo.ac.jp/ships help/ OSIDE/W34/.
[9] 岡本隆明：古文書・典籍を対象とした文字管理システムとその可能性，情報処理学会研究報告, Vol.2008, No.47, pp.77‐84 (2008).
[10] HCR プロジェクト：古文書翻刻支援システム開発プロジェクト. http://www.nichibun.ac.jp/shoji/hcr/index.html.
[11] 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，人文科学とコンピュータシンポジウム論文集，Vol.2000, No.17, pp.185‐192 (2000).