N-013

日本史史料における翻刻データの管理と編集支援
A Management and an Editing Support Method for Reprint Texts of Japanese Historical Documents

山田 太造† 井上 聡† 遠藤 珠紀† 久留島 典子†
Taizo Yamada Satoshi Inoue Tamaki Endo Noriko Kurushima

1. はじめに

本論文では，歴史学・史料学研究の支援を目的とした日本史史料の翻刻を支援するために，翻刻のデータ構造とそのデータ管理，および翻刻記述支援について述べる．史料の内容を正確に読解して活字にする作業，もしくはその成果を翻刻という．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り，歴史学・史料学の研究や史料集編纂などを行う上で，翻刻は重要な要因である．翻刻を支援することは研究推進を行う上で不可欠であると考えられる．翻刻を支援する上で，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．

近年，計算機の処理能力の急速な向上や低コスト化に伴い，歴史学・史料学研究に有用な資源のデジタル化が推進されている．文化遺産オンライン[19]，人間文化研究機構研究資源共有化データベース[16]，PORTA（国立国会図書館デジタルアーカイブポータル）[14]，SHIPSDB（東京大学史料編纂所データベース）[18][20]などの史料に関連するポータルサイトにおいて，多様かつ多量なデータが蓄積されている．これらのサイトでは，名称や形態などの史料に関するメタデータ，および関連する画像などを提供するサービスを行っており，歴史学・史料学研究を進める上で欠かすことができない研究資源基盤として認識されつつある．

国文学研究資料館における『吾妻鏡データベース』[13]や SHIPSDB における『大日本史料総合 DB』・『古記録フルテキスト DB』などの編纂史料データベースでは，翻刻もしくは編纂成果や関連史料に関する検索サービスを提供している．

しかしながら，多くの史料ポータルサイトにおいて，史料の目録データや史料画像に比べ，翻刻の量は圧倒的に少ない．また研究過程で現れる人名・地名や史料上での体裁に関するアノテーションをシステム的に取り扱うためのデータ構造はあまり考慮されていない．さらに，すべての日本史史料に関する調査が終わっているわけではないため，今後も多くの史料に対して翻刻を行う必要がある．

われわれはこれまでに，翻刻支援システムを構築してきた．本システムでは，ユーザと対話しながら，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できる，翻刻データの管理と翻刻記述の支援を行う．翻刻データはその作成や研究の目的に応じて内容が異なることがある．歴史学・史料学研究では，史料をどのように解釈したかということを翻刻として表現する．そこで，本システムでは，1 つの史料に対して，記述者ごとに翻刻データを格納し，史料名や年代などの史料のメタデータと関連させている．

翻刻を実際に行う際，言葉の用法，史料の正確な読解，

† 東京大学史料編纂所, Historiographical Institute The University of Tokyo

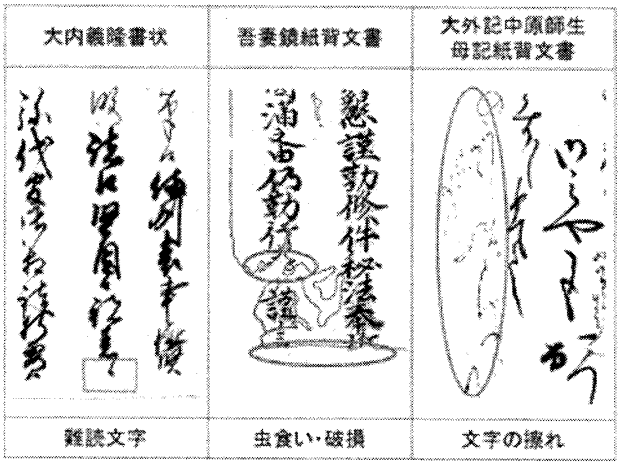

図 1 翻刻を困難にする要因

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．また，史料には図 1 で見られるように，難読文字（矩形で示した部分は“候”と読む），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．そこで本システムでは読解支援を目的とした候補文字検索機能を提供している．文字の擦れの場合は OCR 処理など画像処理による読解支援が可能かもしれないが，難読文字や破損箇所は処理すべき画像が無いため画像処理は困難である．そこで，本研究では，既存の翻刻を用いたテキスト特徴に基づく検索手法を取り入れている．

本論文の構成は次の通りである．2 章で翻刻支援システムでの翻刻データの構造やシステム機能・ユーザインターフェースについて述べる．3 章では候補文字検索機能について述べる．ここでは本研究で扱ったデータ，検索手法，およびその性能を評価した実験結果を示す．4 章では関連研究について述べる．

2. 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，ユーザごとに格納する．また，翻刻は，版管理に基づいており，他ユーザが作成した翻刻の閲覧や利用が可能である．本論文では，翻刻および対象となる史料のメタデータおよびアノテーションから構成される翻刻に必要な情報を翻刻データと呼ぶ．

2.1 翻刻データ

本論文での翻刻データについて定義する．翻刻データ doc は，翻刻対象の史料の識別子doc_idと 1 枚以上の史料画像{$image_1$,$image_2$,…,$image_n$}，および史料自体のアノテーション{$docmemo_1$,$docmemo_2$,…,$docmemo_n$}で構成される

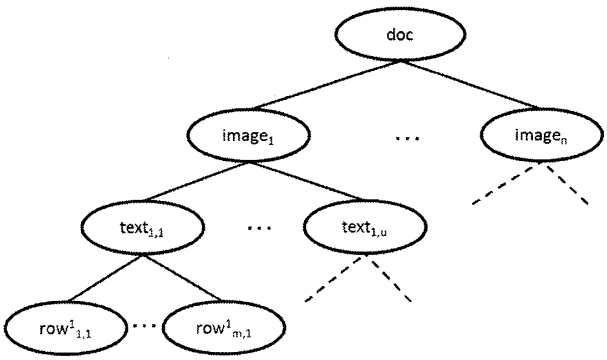

図 2　翻刻データの構造

表 1　アノテーションの種類

アノテーションの種類	詳細
用語・表記に関するアノテーション	人名，地名，用語説明，校訂
体裁に関するアノテーション	抹消，細字双行，傍書，補入，補書

データとする．

$$doc=(doc\_id,\{image_1,image_2,\ldots,image_n\},\{docmemo_1,docmemo_2,\ldots,docmemo_n\})$$

史料画像imageは，その画像の識別子image_idと，翻刻テキスト$\{text_1,text_2,\ldots,text_u\}$，および画像に対するアノテーション$\{imgmemo_1,imgmemo_2,\ldots,imgmemo_n\}$で構成される．

$$image=(image\_id,\{text_1,text_2,\ldots,text_u\},\{imgmemo_1,imgmemo_2,\ldots,imgmemo_n\})$$

ここで，$text_u$はユーザ u が作成した翻刻テキストを示す．翻刻テキストtextは，記述内容を示す行$\{row_1,row_2,\ldots,row_m\}$の集合とする．

$$text=\{row_1,row_2,\ldots,row_m\}$$

行rowは，行番号 row_{num}，文字列 row_{str}，および行に対するアノテーション rowmemo で構成される．

$$row=\{row_{num}, row_{str}, rowmemo\}$$

docmemo およびimagememo は，それぞれ史料，史料画像に対するメモ書きに相当する．rowmemo は行に対する注記であり，ウハ書，裏書，行間補書，追筆，頭書のような行の性質を示すアノテーションである．また，各行の文字列である row_{str} は単なる文字列と文字列の用語・表記もしくは史料上での体裁を示すアノテーションで構成される．アノテーションの種類を表 1 に示す．用語・表記に関するアノテーションでは，文字列に対する説明を記述し，それとともに格納する．体裁に関するアノテーションは文字列がどのような性質であるかを記す．

2.2 システム機能

本システムでは，史料画像の入力や管理を支援するための翻刻テキストに関する検索機能と史料画像に対する翻刻編集の 2 つの機能を有する．

検索機能では，史料の目録階層，翻刻テキスト，ユーザの各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（図 4）．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

データ定義（DTD）

```
<!DOCTYPE root[
<!ELEMENT root(doc*)>
<!ELEMENT doc(date*,image*,docmemo*)>
<!ATTRLIST doc path CDATA>
<!ELEMENT date(#PCDATA)>
<!ATTRLIST date type CDATA>
<!ELEMENT image(text*,imagememo*)>
<!ATTRLIST image src CDATA>
<!ELEMENT docmemo(row*)>
<!ATTRLIST docmemo userid CDATA>
<!ATTRLIST docmemo modified CDATA>
<!ELEMENT text(row*)>
<!ATTRLIST text userid CDATA>
<!ATTRLIST text modified CDATA>
<!ELEMENT row(str*,note*)>
<!ATTRLIST row num CDATA>
<!ATTRLIST row x CDATA>
<!ATTRLIST row y CDATA>
<!ATTRLIST row height CDATA>
<!ATTRLIST row size CDATA>
<!ELEMENT str(#PCDATA)>
<!ELEMENT note(comment,str)>
<!ATTRLIST note type CDATA>
<!ELEMENT comment(#PCDATA)>
<!ELEMENT textmemo(row*)>
<!ATTRLIST textmemo userid CDATA>
<!ATTRLIST textmemo 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">1186008030</date>
  <image src="00000011.tif">
   <text userid="t_yamada" modified="20100401101123">
    <row num="1" x="805" y="109" height="200" size="14">
     <str>下 島津御領住官等</str>
    </row>
    <row num="2" x="756" y="121" height="344" size="14">
     <str>可令停止手継介</str><note type="入名注">
<comment>千葉常胤</comment><str>常胤</str></note><str>代官平経友</str>
    </row>
    <row num="3" x="713" y="126" height="194" size="14">
     <str>清遠非道雑務事</str>
    </row>
    ・・・
   </text>
   <imagememo userid="t_yamada" modified="20100401101123">
画像のメモはここ</imagememo>
  </image>
  <docmemo userid="t_yamada" modified="20100401095541">
史料のメモはここ</docmemo>
 </doc>
</root>
```

図 3　翻刻データの定義と出力例

図 4　検索機能

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（部分文字列完全一致での検索）を行う．このとき，各全ユーザを通して最新版である翻刻テキストが検索対象となる．ユーザを指定した検索では，指定したユーザが作成した翻刻テキストが検索対象となる．このとき，翻刻テキストに検索条件を指定した場合，指定したユーザが作成した最新版が検索対象となる．また，史料に対するアノテーションの付与も行うことができる．

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・編集，画像の拡大表示，史料画像・行・文字列へのアノテーション付与，翻刻テキスト表示・出力，履歴表示，翻刻データ保存の機能を持つ（図 5）．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テキストを編集することができる．翻刻データを本システムで取り扱うため，XML 形式で記述し，格納する．DTD と『島津家文書源頼朝下文文治二年八月三日条』の一部の XML の記述例を図 3 に示す．文字コードは，Shift-JIS や EUC では表現できない文字が多く出現するため，UTF-8 を使用した．また，各行の画像上での

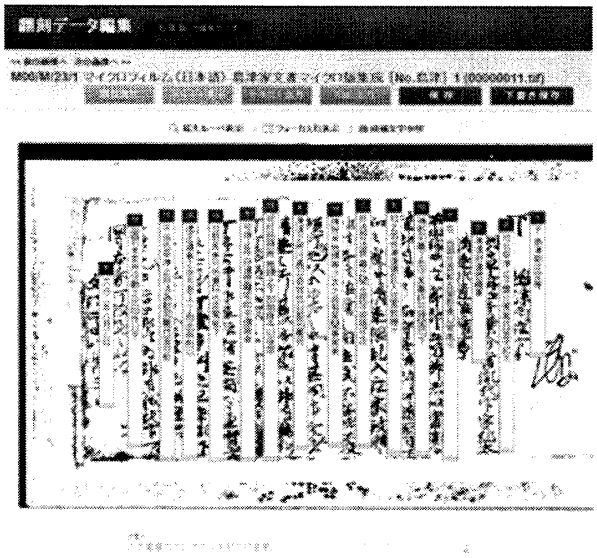

図 5　翻刻編集機能

配置位置を保持するため，史料画像上での記述位置，記述されている文字列の組，作成日時，およびユーザ ID の組で行を表現することにした．記述位置は史料画像上での 2 次元座標で示す．本研究で扱う画像は，ラスタイメージであるため，ラスタイメージ上での座標(x,y) を記述位置として扱う．翻刻テキストの表示機能では，現在対象としている翻刻テキストをプレーンテキスト，もしくは XML 形式で出力することができる．またこの機能では行番号の修正や史料画像なしで翻刻できるインターフェースも備えている．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．翻刻データを保存すると，翻刻テキストの版が新たに生成される．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．利用した翻刻テキストを保存した場合，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．

3. 候補文字検索機能

テキスト配置・編集機能では，記述されている文字の候補をユーザに提示する候補文字検索機能を実装している（図 6）．この候補文字検索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．本システムではでは，図 6 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．本節では，候補文字検索機能として文字 N グラムモデルを利用した手法，およびその評価結果を提示する．

3.1 候補文字検索

本研究における候補文字検索は，ユーザが入力している

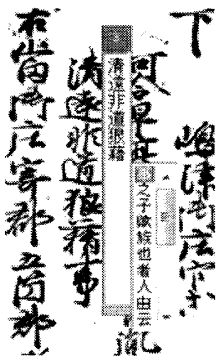

図 6　候補文字検索の例

表 2　学習データ

DB 名	件数	異なり文字数	延べ文字数
平安遺文 FT	13,500	4,765	2,861,330
鎌倉遺文 FT	34,495	4,817	7,293,059
古文書 FT	39,719	4,838	7,956,168
古記録 FT	60,805	5,149	8,698,724
大日本史料総合	5,797	4,485	975,755
合計	154,316	5,997	27,785,036

文字列が$c_1^{n-1}=c_1,\ldots,c_{n-1}$であるとき，この文字列の直後に出現する文字 c_n を確率的に推定し，c_n の候補をユーザへ提示する方法とした．c_n の推定では，文字 N グラムモデルを用いる．文字列c_1^nから確率$P(c_n|c_1^{n-1})$を求めるとき，文字 N グラムモデルでは，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するため，$P(c_n|c_1^{n-1})\approx P(c_n|c_{n-N+1}^{n-1})$と表せる[6]．$P(c_n|c_{n-N+1}^{n-1})$は，学習データ中に出現する文字 N グラムから最尤推定を行うと，

$$P(c_n|c_{n-N+1}^{n-1})=\frac{freq(c_{n-N+1}^{n})}{freq(c_{n-N+1}^{n-1})} \qquad (1)$$

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．

$P(c_n|c_1^{n-1})$を推定するためには学習データが必要である．本研究では，SHIPSDB における編纂史料データベースからテキストを抽出し，文字 N グラムを作成することで学習データを構築する．表 2 は抽出したテキストにおける異なり文字数と延べ文字数を抽出対象のデータベースごとに示している．これらのデータベースは西暦 700 年から 1651 年までの史料を対象としている．抽出したテキストから文字 N グラムの出現頻度を求める．例えば，N = 2のときの文字列 “爲新寺作偈化鐘” における N グラムは，“爲新”，“新寺”，“寺作”，“作偈”，“偈化”，“化鐘” である．本研究では，各データベースで N グラムt_iが出現した頻度の合計を$freq(t_i)$とする．

作成した学習データから入力文字に応じて，（1）式を計算し，そのスコアが 0 よりも大きい値であれば，その文字を候補文字とする．ランキングはスコアにもとづいて行い，上位 r 件をユーザに提示することで候補文字検索を実現する．この手順を図 7 に示す．

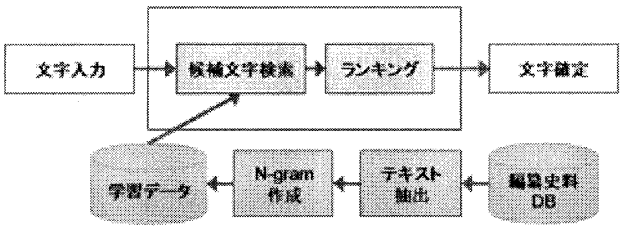

図 7　候補文字検索の手順

3.2 候補文字検索の評価実験

3.2.1 実験準備

前節で示した候補文字検索の性能を評価する．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→∞としたときのヒット率として求めた．大日本史料 6 編之 46』のテキスト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（表 3）．本実験で用いた学習データでは，テストデータで用いたテキストを除いている．

本実験で用いた候補文字検索手法は，前節で示した N グラムモデルに基づいた手法であり，これらを以下に示す．

（1）そのまま利用する手法: この手法では（1）式をそのまま利用することで候補文字のスコアを計算する．本実験では N の値を 1 から 4 まで変動させ，ヒット率への影響を確認した．

（2）スムージングを用いた手法: N グラムモデルでは，学習データに出現しない文字列に対応するためスムージングを行うことがある．N の次元が高くなるほど，クエリに対応する学習データ内の N グラムの頻度が 0 となってしまう可能性が高くなる（ゼロ頻度問題）．そこで，本実験では Modified Kneser-Ney スムージング[3][7]を用いた．これ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り，次式で N グラムの確率を計算する．

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

D(freq)はディスカウント関数であり，

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

である．また，

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

$$Y=\frac{n_1}{n_1+2n_2}$$

$$D_1=1-2Y\frac{n_2}{n_1}$$

$$D_2=2-3Y\frac{n_3}{n_2}$$

$$D_{3+}=3-4Y\frac{n_4}{n_3}$$

表 3　テストデータ

DB 名	刊本	件数	異なり文字数	延べ文字数
大日本史料総合	大日本史料 6 編 46	510	2,248	87,270

表 4　再現率

手法	N=1	N=2	N=3	N=4
再現率	1.00	0.980	0.804	0.548

手法	MKNS	SkipBigram	GappedKernel
再現率	1.00	0.986	0.986

である．ただし，$N_1\left(c_{n-N+1}^{n-1}\cdot\right)=\left|\{c_i:freq\left(c_{n-N+1}^{n-1}c_i\right)=1\}\right|$であり，$N_2\left(c_{n-N+1}^{n-1}\cdot\right)$および$N_{3+}\left(c_{n-N+1}^{n-1}\cdot\right)$も同様に定義される．また，$n_1=|\{t_i:freq(t_i)=1\}|$であり，$n_2$，$n_3$，および$n_4$も同様に定義される．本実験では N=4 としており，（2）式により 1-4 グラムまでをスムージングの対象とした．

（3）skip-bigram を用いた手法: N グラムモデルでは連続する文字列を用いる．つまり文字間の距離が 1 となる．skip-bigram[3]では，ある文字列において，その文字列での出現順で任意の文字の組，つまり文字間の距離が 1 以上である文字の組を扱う．たとえば，文字列“足利義満”であれば，“足利”，“足義”，“足満”，“利義”，“利満”，“義満”の 6skip-bigram である．skip-bigram は文字列 s_1，s_2 の類似度を次式の $F(s_1,s_2)$で類似度を計算する．

$$F(s_1,s_2)=\frac{(1+\beta^2)R(s_1,s_2)P(s_1,s_2)}{R(s_1,s_2)+\beta^2 P(s_1,s_2)} \quad (3)$$

$$R(s_1,s_2)=\frac{SKIP(s_1,s_2)}{C(m,2)}$$

$$P(s_1,s_2)=\frac{SKIP(s_1,s_2)}{C(n,2)}$$

ここで m，n はそれぞれ s_1，s_2 の文字列長を示す．$\beta=P(s_1,s_2)/R(s_1,s_2)$である．$SKIP(s_1,s_2)$は s_1，s_2 の skip-bigram がマッチした回数を示す．たとえば，s_1 を“足利義満”，s_2 を“足利義詮”であるとき，$SKIP(s_1,s_2)$は 3 である．また $R(s_1,s_2)=3/6$，$P(s_1,s_2)=3/6$ なので，$F(s_1,s_2)=0.5$となる．本実験では，ユーザが入力した文字列を s_1 とし，学習データから（3）式の値が 0 よりも大きな文字列を候補文字 s_2 とする．候補文字のランキングは（3）式の値にもとづいて並び替えることにした．

また本方法を用いるときは，前節の N グラムと同様に特徴ベクトルを作成していくが，N グラムの代わりに skip-bigram をカウントしている．

（4）文字列カーネルを用いた方法: 本手法は SVM（Support Vector Machine）[1]でデータの分類を行うために用いられる文字列カーネルを用いる．文字列カーネルは，2 つの文字列 s_1，s_2 の類似度を計算することができ，これを用いて方法（3）と同様の方法で候補文字を検索する．本実験では文字列カーネルとして Gapped-weighted subsequences kernel[9]を用いた．この文字列カーネルは

次式で定義される．

$$K_N(s_1,s_2)=\sum_{u\in\Sigma^N}\sum_{i:u=s_1[i]}\sum_{j:u=s_2[j]}\lambda^{l(i)+l(j)} \quad (4)$$

ここでΣ^Nはすべての文字Σで長さ N である文字の部分列，l(i)は文字列 s に対するインデックス i での部分列長を示す．この方法では，λによって文字間の距離に応じた重み付けを行うことができ，λ<1であれば文字間の距離に応じて重みを低減し，λ=1であれば方法（3）と同様に文字間の距離を無視する．たとえば，s_1 を“足利義満”，s_2 を“足利義詮”，N=2，λ=1/2であるとき，Φ("足利義満")およびΦ("足利義詮") を計算し，のち（4）式より $K_2(s_1,s_2)=\lambda^5+2\lambda^4=0.15625$となる．

3.2.2 実験結果

前節で示した候補文字検索の各手法でのヒット率を図に，再現率を表 4 に示す．図 8 では，x 軸はランクを，y 軸はヒット率を示している．

方法（1）では N の値を 1 から 4 まで変動させた．r≤50 のとき N=3 がもっともヒット率が高く，r=1 では 0.31，r=5 のとき 0.51，r=20 のとき 0.65 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 0.294 であり，N=3 に近いヒット率であったが，r=5 で 0.42，r=20 で 0.496 であり，r≥5 以上では N=3 でのヒット率には及ばなかった．表 4 はから 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．特に N=4 では極端にその値が低下している．本実験での再現率はヒット率の最大値を示している．そのため 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．

方法（3）では学習データに現れるテキスト特徴として離れた文字間の特徴も扱う．そのため，再現率は N=2 よりも少しではあるが改善している．しかしながら r≤100 のときのヒット率は N=1 よりは高いが，N=4 よりも低いヒット率だった．これは，方法（3）は文字列全体としての類似度を計算するために提案された手法であり，skip-bigram は文字の出現順は取り入れているものの，文字間の距離は考慮していない．そのため，非常に雑音が多くなってしまい，提示する候補文字が誤りやすくなったためだと考えられる．

方法（4）では方法（3）と対比した場合，文字間の距離に応じて重みを修正している．ヒット率の結果としては，N=3 に比べた場合，r=1 のときは雑音が多いため N=3 に比べかなり低いヒット率だったが，r の値が 5-20 のときは少し低い，さらに r=50 のときは少し高くなった．N=2 と比べた場合では，r=1 のときは同等の結果であったが，r≥5 以上であれば常にヒット率が高かった．この結果より，方法（3）のように単に離れた文字の情報をそのまま利用するよりも，文字間の距離に応じた重み付けを行う方がヒット率は高くなることがわかった．

方法（2）では，N=4 をベースとした Modified Kneser-Ney スムージングを用いている．ヒット率の結果から，いずれの r の値においても，他の方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．これらの結果より，方法（3）および方法（4）で利用した不連続文字列をベースとした方法よりも連続した文字列をベースとした方が候補文字検索としては有効であることがわかった．Modified Kneser-Ney スムージングでは，出現しない N グラムを単に線形に補間するのではなく，他の N グラムの出現頻度に応じて N グラムの確率値をディスカウントしている．さらに低次元での N グラムの出現頻度も考慮している．そのため，N グラム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．

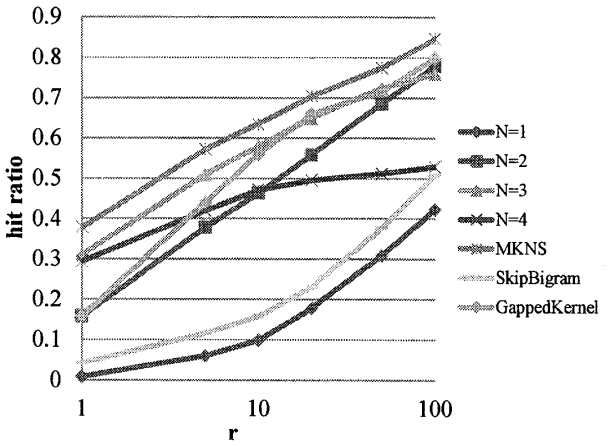

図 8　各手法での候補文字のヒット率

本システムでの候補文字検索機能では，これらの実験より方法（2）による N グラム手法にもとづく検索手法を用いた．また，r=20 と r=50 のヒット率を比べた場合，ヒット率の差は約 0.07 である．提示した候補文字のうち，上位 20 件で 0.65 の確率で正解文字を見つけることができるが，次の 30 件では約 0.07 の確率でしか見つけ出すことができない．そこで，候補文字の提示は 20 件とした．

4. 関連研究

翻刻支援に関するシステムや研究としては（1）史料ポータルサイトなどの翻刻対象や関連する史料を提供する検索システム，（2）文字画像を対象とした検索システムや文字画像処理を行うシステム，（3）テキスト特徴に基づく支援システムがある．

（1）は史料名，年代，形状などの史料メタデータや史料画像などの史料情報を提供しており，例えば，先にあげた文化遺産オンライン，人間文化研究機構研究資源共有化データベース，PORTA，SHIPSDB などがある．研究資源共有化データベースでは，人間文化研究機構の各研究機関で管理している史料の横断検索ができる．国文学研究資料館における『吾妻鏡データベース』では，検索結果はヒットした行のテキストとその史料画像が示される．翻刻を行う場合，これらのシステムを用いて同年代や同じ所蔵先などの史料と対比しながら翻刻を進めることができる．

（2）は文字画像の読解を支援する機能を有しており，さらに 2 つのタイプに分類することができる．1 つは，文字列画像を検索することで翻刻を支援するシステムであり，例えば，トランスメディア[17]，SMART-GS[9]などがあり，画像処理を行うことで，文字画像をクエリとして同型の文字画像を検索することができる．また，OCR 処理を行うこ

とで画像からテキストに変換する方式もあるが，日本史史料では草書体での記述，漢文調での記述，平仮名だけでの記述などの要因により顕著な成果はない．もう 1 つは，テキストから文字画像を検索するシステムである．例えば，SHIPSDB 電子くずし字字典データベースや文字管理システム[21]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．図 1 のように史料に欠損・破損がある場合には用いることができないが，参考データとして利用しながら翻刻を進めていくことができる．

本研究における候補文字検索機能は（3）に分類される．（3）は既存の翻刻データからテキスト特徴を抽出し，それを学習データ用いることで，出現文字を推奨する機能を提供する．例えば，HCR（Historical Character Recognition）プロジェクト[5] [15]による古文書翻刻支援システムがある．このシステムでは，近世の借金証文類の史料を対象とした難読文字などの読解支援を行う．翻刻を行う際，難読文字など読めない文字が出現した箇所をマークアップする．文字 N グラム を学習データとして用い，マークアップした箇所の前後の文字列に対する前方・後方一致検索を行う．この結果を文字 N グラムの出現頻度に応じてランキングし，候補文字を提示する．しかし，難読文字が出現したとき，即座にこの候補文字検索機能を用いることができない．また N グラムの確率値をスムージングするなど，ゼロ頻度問題に対応していないため，出現しない N グラムには対応できない．

本システムでは 1 つの史料に対して，翻刻データは記述者ごとに管理されており，史料名や年代などのようなメタデータと関連が保持される．このデータ構造は Owlery[1]におけるメタデータ構造に近い．記述の定義としては TEI[11]がある．これは人文科学研究で扱うテキストの電子文書化のガイドライン，特に情報交換や共有のための共通フォーマットなどを定めている．SGML・XML による記述方法や記述支援のためのアプリケーションも存在する．本研究での翻刻データは TEI よりも下位に相当する位置づけであり，また TEI では考慮されていない各種アノテーションも取り扱っている．

5. おわりに

本研究では，日本史史料の研究で不可欠な作業である翻刻を支援するため，翻刻データの構造とそれを作成支援・管理を行うシステムについて述べた．まず，本システムで扱う翻刻データの定義と XML による記述方法を示し，翻刻データをオーサリングするための史料画像・翻刻テキストの検索機能と翻刻編集機能を示した．さらに，翻刻の読解を支援する候補文字検索機能について示し，その有効性を評価する実験を行った．その結果，Modified Kneser-Ney スムージングを用いた文字 N グラムモデルに基づく検索手法であれば，上位 5 件以内でおよそ 58%，上位 20 件以内でおよそ 70%のヒット率であることを確認した．

史料編纂の質・スピードを向上させる上で，翻刻の支援は不可欠だと考えている．翻刻を支援する方法は本論文で述べた方法以外にも，各種史料目録データベースとの連携，関連史料の提示，用語辞書の利用，など解決すべき課題が多い．また候補文字検索では，文字 N グラムモデルだけではなく，単語 N グラムの利用，時代ごともしくは史料の種別ごとの N グラム特徴の利用によるテキスト特徴を用いる方法が考えられる．他方，トランスメディアや SMART-GS のような文字画像の特徴に基づいた方法もある．これらの方法を取り込むことで，候補文字検索機能はさらに向上すると考えている．

謝辞

研究の一部は，日本学術振興会科学研究費基盤研究（A）（21240022），および若手研究（B）（21700274）の助成を受けたものである．

参考文献

[1] Aihara, K., Yamada, T., Kando, N., Fujisawa, S., Uehara, Y., Baba, T., Nagata, S., Tojo, T., Awaji, T., Adachi, J.: Owlery: A Flexible Content Management System for “Growing Metadata” of Cultural Heritage Objects and Its Educational Use in the CEAX Project, Proceedings of ICADL 2006, pp.22-31 (2006).

[2] Boser, B., Guyon, I., Vapnik, V.: A Training Algorithm for Optimal Margin Classifiers. Proceedings of 5th COLT, pp.144-152 (1992).

[3] Chen, S., Goodman, J.: An empirical study of smoothing techniques for language modeling. Proceedings of the ACL96, pp. 310–318, 1996.

[4] Chin-Yew, L., Franz, O.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. Proceedings of ACL '04, 605(2004).

[5] 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.

[6] Jelinek, F.: Self-organized language modeling for speech recognition, Readings in Speech Recognition, Morgan Kaufmann, pp. 450-506 (1990).

[7] James, F.: Modified Kneser-Ney Smoothing of n-gram Models, RIACS Technical Report 00.07, http://www.riacs.edu/navroot/Research/TR_pdf/TR_00.07.pdf (2000).

[8] Khy, S., Ishikawa, Y. and Kitagawa, H.: Novelty-based Incremental Document Clustering for On-line Documents. Proceedings of 22nd International Conference on Data Engineering Workshops (ICDEW06), p.40 (2006).

[9] Lodhi, H., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. The Journal of Machine Learning Research, 2, pp. 419-444, 2000.

[10] SMART-GS: a tool for humanistics. http://www.shayashi.jp/SMART-GS/mainjp.html.

[11] TEI: Text Encoding Initiative, http://www.tei-c.org/index.xml.

[12] 石川佳治，北川博之：忘却の概念に基づくインクリメンタルな文書クラスタリング手法，電子情報通信学会技術研究報告. DE, データ工学，101, 192, pp. 145-152 (2001).

[13] 国文学研究資料館：吾妻鏡本文検索. http://ocelot.nijl.ac.jp/dlib/azuma/.

[14] 国立国会図書館：デジタルアーカイブポータル. http://porta.ndl.go.jp/portal/dt.

[15] 山田奨治，柴山 守：N グラム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).

[16] 人間文化研究機構：研究資源共有化データベース. http://www.nihu.jp/kyoyuka/tougou/index.html.

[17] 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).

[18] 東京大学史料編纂所：東京大学史料編纂所データベース. http://www.hi.u-tokyo.ac.jp/ships/.

[19] 文化庁：文化遺産オンライン. http://bunka.nii.ac.jp/.

[20] 加藤友康：研究成果報告書「WWW サーバによる日本史データベースのマルチメディア化と公開に関する研究」，http://www.hi.u-tokyo.ac.jp/personal/kato/index.htm(1999).

[21] 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).