Vol.2011-CH-91 No.5
2011/7/30

日本史史料における翻刻テキストの構造化支援手法

山田太造[†1]　井上聡[†2]
遠藤珠紀[†2]　久留島典子[†2]

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．翻刻を支援するため，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．翻刻データはユーザごとに管理されており，他ユーザの版を利用することができる．このとき利用元の版に変更を加えないため，他ユーザの作業に影響を与えない．またシステムは入力したテキストを自動的に構造化することが可能である．さらに史料読解を支援するため，候補文字検索機能を提供する．この機能では，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．この機能を評価するための実験を行い，その結果も示す．

A Structuring Support Method for Reprint Text of Japanese Historical Documents

Taizo Yamada,[†1] Satoshi Inoue,[†2] Tamaki Endo[†2]
and Noriko Kurushima[†2]

A decoding to reprint is one of important factors to advance studies of history and historical document, however, its work is needed very high skill and knowledge of history. Then, we developed a decoding support system due to assist decoding each historical document. Our system can manage decoded data which is interactively given by a user to an input image of a historical document. Because our system has a function of multi-versioning for decoded data, a user can use a version of other user easily without changing the version of the user. Our system can automatically structurize input text. Moreover, for assisting decode, our system has a function of character recommendation. We evaluate the effectiveness and useful of the function, and show the experimental results.

1. はじめに

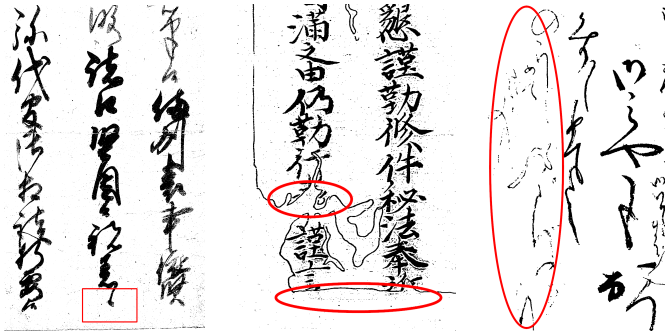
翻刻は史料の内容を正確に読解して活字にする作業およびその作業の結果としての出力である．史料に記述されている内容を確認したり，より深く史料を調査するためには翻刻は不可欠である．そのため，歴史学・史料学の研究や史料編纂などを行う上で，翻刻は重要な作業である．

国文学研究資料館における『吾妻鏡データベース』[14] や東京大学史料編纂所データベース SHIPSDB[21),23)] 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などでは翻刻したテキストを提供している．文化遺産オンライン[22)]，人間文化研究機構研究資源共有化データベース[18)]，PORTA（国立国会図書館デジタルアーカイブポータル）[15)]，SHIPSDB などの史料に関連するポータルサイトでは，史料の目録や画像などの多種多様な史料コンテンツ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ている．しかしながら，上記の史料ポータルサイトにおいて，翻刻が提供されている史料はごくわずかである．

史料には**図 1** で見られるように，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．また，史料の記述年代によって出現する言葉の用法が異なることがある．翻刻を行うためには，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．

翻刻を支援するため，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．翻刻データは入力された史料画像に自動的に関連付けて格納する．翻刻データはユーザごとに管理されており，他ユーザの版を利用することができる．このとき利用元の版に変更を加えないため，他ユーザの作業に影響を与えない．従来，版管理システムとしてよく知られているシステムとして CVS[1)] が挙げられる．文書ごとに版管理を行っており，最新版だけでなく過去の版も取得できるなど多様な文書共有を可能とする．しかしながら，CVS はユーザ個々の版の管理までサポートしていない．CMS(content management

†1 人間文化研究機構
National Institutes for the Humanities
†2 東京大学史料編纂所
Historiographical Institute, The University of Tokyo



難読文字[⋆1]
（大内義隆書状）

虫食い・欠損
（吾妻鏡紙背文書）

文字の擦れ
（大外記中原師生
母記紙背文書）

図 1 翻刻を困難にする要因

system) は wiki[3] や blog[11] に代表される web-based な文書共有システムである．テキストや画像，動画などのマルチメディアだけでなく，関連コンテンツやリンクなども管理可能である．しかしながら，ユーザが指定した DTD などのスキーマに従った XML データの出力は困難である．

本システムでは，史料読解の支援するために，難読文字など読めない文字に対して候補文字を提示する候補文字検索機能を提供する．本機能は，入力文字列の次に出現する文字を学習データに基づいて検索することで候補文字を提示する．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[19]，SMART-GS[12]，Mokkanshop[16] 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．電子くずし字字典データベース[20] や文字管理システム[24] 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[7),17] による古文書翻刻支援システムがあり，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，出現文字を提示する機能を提供する．例えば，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-gram 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-gram の出現頻度に応じてランキングし，候補文字を推奨する．しかし，難読文字が出現したとき，対話的にこの機能を利用することができない．本研究ではユーザの要求に応じて即座に候補文字検索を行うことができる．

本論文は以降，次のように構成している．2 節で本研究での翻刻データの構造を示す．3 章で翻刻支援システムの概要（史料検索機能，翻刻編集機能および候補文字検索機能）を示す．また候補文字検索機能の評価実験の結果も示す．

⋆1 矩形で示した部分は “候” と読む．

2. テキスト構造

ここでは本論文における翻刻データの構造を示す．**図 2** にこの構造の概要を示す．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データを翻刻データと呼ぶ．要素 *doc* は翻刻データのルートであり，次式で定義する．

$$doc = (doc_{id}, \{image_1, \ldots, image_l\}) \quad (1)$$

ここで，doc_{id} は翻刻データ doc の識別子，$\{image_1, \ldots, image_l\}$ は doc に割り当てられた史料画像群を示す．要素 $image$ は史料画像を示しており，その識別子と翻刻テキスト $\{text_1, \ldots, text_m\}$ で構成され，要素 $image$ は次式で定義される．

$$image = (image_{id}, \{text_1, \ldots, text_m\}) \quad (2)$$

要素 $text_u$ は，翻刻作成者 u がある史料画像 $image$ に対して作成し付与した翻刻テキストを示す．(2) 式の定義により，史料画像 $image$ に対する翻刻テキストはただ 1 つではなく，翻刻テキストを作成者ごとに作成・管理される構造としてある．ある史料に対する翻刻は，翻刻作成者ごとの研究・利用目的の差異により，翻字，付与されるアノテーション，メモなど，記述される内容が翻刻作成者ごとに異なることが多い．また，(2) 式では 1 史料画像に対して翻刻作成者に 1 つだけしか翻刻テキストを保持できない．しかしながら，次式のように翻刻テキスト群を 2 次元で表現することにより翻刻作成者は 1 史料画像に対して複数の翻刻テキストを付与することが可能となる．

$$image = (image_{id}, \{text_{1,1}, \ldots, text_{1,o} \ldots, text_{m,1}, \ldots, text_{m,o}\}) \quad (3)$$

要素 $text$ を次式で定義する．

Vol.2011-CH-91 No.5
2011/7/30

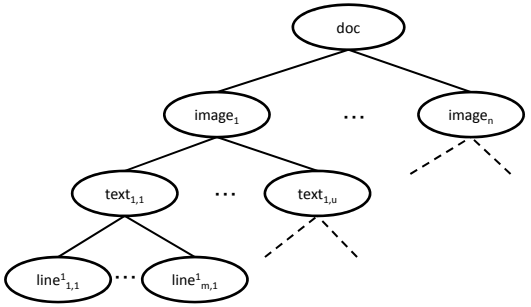


図 2 テキスト構造

表 1 行に対するアノテーション

種類	説明
指定なし	史料の本文における行
ウハ書	タイトル，主題，宛所など
裏書	差出など
行間補書	行間にある，行に対する補足記述
追筆	追記など．本文と追筆の筆者が同じであるとは限らない
頭書	頭書，標注など

$$text = (\{line_1, \ldots, line_n\}, textnote) \quad (4)$$

(4) 式は翻刻テキスト $text$ は n 個の行 $line$ と翻刻テキストに対するメモ書き $textnote$ で構成されることを示す．翻刻テキスト内の行を示す要素 $line$ は管理上での最小単位としている．要素 $line$ は，その行番号 $line_{num}$，その行での記述内容 $line_{str}$ および行に対するアノテーション $line_{annotation}$ で構成され，この定義を次式に示す．

$$line = (line_{num}, line_{str}, line_{annotation}) \quad (5)$$

$line_{annotation}$ は $line$ に対するアノテーションである．例えば，ウハ書，裏書，頭書などがある．行に対するアノテーションとその説明を**表 1** に示す．要素 $line_{str}$ は史料画像上にある文字列とそれに対するアノテーションで構成される．文字列に対するアノテーションとしては，人名注，地名注などの用語・表記に関するアノテーションと，抹消，細字双行などの体裁に関するアノテーションがある．これを**表 2** に示す．

⋆2 多くの場合，本文に “〇” と記述し，その右もしくは左に内容を記述してある．

表 2 文字列に対するアノテーション

分類	種類	説明
用語・表記	人名注	人名・組織名など
	地名注	地名，場所に関連する事項など
	説明注	文字列に対する説明
	校訂注	文字列に対する校訂・修正
体裁	抹消	抹消，取消
	細字双行	2 カラムで記述された，直前の文字列に対する注
	傍書	行の右もしくは左にある記述
	補入	本文に記述すべきだった内容⋆2
	補書	補足記述

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
官宇紀太</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 3 翻刻データの定義と出力例

3. 翻刻支援システム

本節では，我々が開発している翻刻支援システムについて述べる．この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．本システムは，翻刻フローに従って**図 3** に示した XML データを作成・管理することで前節で示した翻刻データを具現化している．そのため，翻刻データを作成者や翻刻の作成目的を単位として管理することができ，さらに他作成者の翻刻の閲

（a）史料検索機能　　（b）翻刻編集機能

図 4　翻刻支援システム

覧や利用が可能である．本システムでは，この XML データの作成を支援するため，1) 史料検索機能，2) 翻刻編集機能および 3) 候補文字検索機能を提供している．

3.1　史料検索機能

本システムでは，1) 史料目録システムでの検索手段 (finding aids)・史料識別子，2) 翻刻テキストに対する検索および 3) 翻刻作成者を用いて史料を検索することができる．このユーザインタフェースを**図 4**（a）に示す．この検索では，検索結果は (1) 式の doc の集合となる．2) および 3) は翻刻してある史料がなければ検索することができないため，未翻刻の史料を検索するためには 1) の方法しか利用できない．本システムでは，既存の史料目録機能を持つシステムである HI-CAT を利用することで，史料目録検索を行っている．これにより，本システムから非常に重要であるが大変複雑な史料目録の管理・検索を切り離すことができる．

翻刻テキストの検索は，(5) 式における $line_{str}$，$line_{annotation}$ および (4) 式における $textnote$ に対する全文検索を行なうことで実現している．

翻刻作成者の検索は，作成者を特定した翻刻テキストに対する検索である．翻刻作成者が u である翻刻を検索する場合，(2) 式において $i = u$ である $text$ が検索対象となる．翻刻作成者を指定しない検索では，史料画像 $image$ における $text_{i,\ldots,m}$ のうち最新版 $text_{latest}$ のみが検索対象となる．

検索結果は doc を単位としており，検索結果表示ではヒットした doc_d とともにその史料画像 $image_{1,\ldots,n}$ を表示する．1 史料に関連する史料画像が 100 を越えることもしばしばある．そのため，翻刻テキストや作成者で絞り込んだ史料画像群のみを表示する機能も有している．

3.2　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・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 4（b））．翻刻作成者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．

任意に選択した史料画像上の行自体および文字列に対して，表 1 および表 2 に示したアノテーションを付与することができる．この付与はアノテーションの種類とそのアノテーションの内容を対話的に記述することで簡易に行なうことができる．

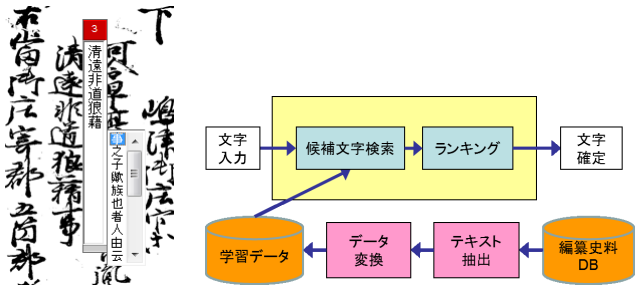
各行の画像上での配置情報を保持するため，この配置位置およびテキストフィールドのサイズを記述している．本研究で扱う画像はラスタイメージ（raster image）であるため，ラスタイメージ上での座標 (x,y) を記述位置として扱う．各行は行番号を持っており，これにより行の並び順を決定する．行の順番を修正することもできる．

図 4（b）のようにテキストを入力していくと，本システムは自動的に XML データを構成している．本システムは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．ここでの出力対象は現在の翻刻対象の史料画像に対する自分の翻刻テキストのみである．1 史料全体の翻刻データを出力する場合，史料検索機能を用いる．検索結果の状態に応じて 1 史料の翻刻データを出力することができる．『島津家文書源頼朝下文文治二年八月三日条』に対する XML での出力例を図 3 に示す．

履歴表示では，対象史料画像の翻刻テキスト対して検索・表示・利用できる機能である．翻刻テキストを保存すると，新たな版に生成される．自分で作成した過去の版や他作成者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版として格納される．そのため，ある作成者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．

3.3　候補文字検索機能

翻刻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



（a）検索例　　（b）手順

図 5　候補文字検索の例と手順

装している．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（**図 5**（b））．本ユーザインターフェースでは，図 5（a）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．

4. 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文字列が $c_1^{n-1} = c_1, \cdots, c_{n-1}$ 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，1) 連続文字列を用いた方法および 2) 不連続文字列を用いた方法のを検討した．1) は文字 n-gram モデルを利用した方法であり，文字の連続性に着目した方法である．2) は Skip-bigram モデル[9] に見られる不連続文字列（部分列）を用いた方法であり，文字列の近似性に着目した方法である．本節では，c_n を推定するための学習データとその抽出方法，および両者の候補文字検索の手法を示す．

4.1 学習用テキストデータの抽出

本研究では，SHIPSDB 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．**表 3** は本研究で対象としたデータベースのデータ件数，異なり文字数および延べ文字数を示しており,『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．

表 3　学習用編纂史料データベース

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

4.2 文字の連続性に基づいた候補文字検索

文字 n-gram モデルは，文字列 c_1^n から確率 $P(c_n \mid c_1^{n-1})$ を求めるとき，文字 n-gram 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[4] ため，次式のように示すことができる．

$$P(c_n \mid c_1^{n-1}) \approx P(c_n \mid c_{n-N+1}^{n-1}) \tag{6}$$

また，$P(c_n \mid c_{n-N+1}^{n-1})$ は，学習データ中に出現する n-gram から最尤推定を行うと，

$$P(c_n \mid c_{n-N+1}^{n-1}) = \frac{freq(c_{n-N+1}^{n})}{freq(c_{n-N+1}^{n-1})} \tag{7}$$

となる．ここで $freq(c)$ は文字 c の出現頻度を示す．本研究では（7）式の $P(c_n \mid c_{n-N+1}^{n-1})$ を推定の尺度として用い，候補文字検索を行う際，この尺度に応じてランキングし，上位 r 件の結果を候補文字としてユーザに提示する．

n-gram 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-gram 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，本研究では Modified Kneser-Ney スムージング[5),8)] 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-gram の確率を計算する．

$$P_{KN}\left(c_n \mid c_{n-N+1}^{n-1}\right) = \frac{freq\left(c_{n-N+1}^{n}\right) - D\left(freq\left(c_{n-N+1}^{n}\right)\right)}{\sum_{c_n} freq\left(c_n \mid c_{n-N+2}^{n-1}\right)} + \gamma\left(c_{n-N+1}^{n-1}\right) P_{KM}\left(c_n \mid c_{n-N+2}^{n-1}\right) \tag{8}$$

$D(freq)$ はディスカウント関数であり，

$$D(freq) = \begin{cases} 0, & if freq = 0 \\ D_1, & if\ freq = 1 \\ D_2, & if\ freq = 2 \\ D_{3+}, & if\ freq \geq 3 \end{cases} \tag{9}$$

である．また，

$$\gamma\left(c_{n-N+1}^{n-1}\right) = \frac{D_1\left(N_1\left(c_{n-N+1}^{n-1}\cdot\right)\right) + D_2\left(N_2\left(c_{n-N+1}^{n-1}\cdot\right)\right) + D_{3+}\left(N_{3+}\left(c_{n-N+1}^{n-1}\cdot\right)\right)}{\sum_{c_n} c_{i-n+1}^{n}} \tag{10}$$

$$\begin{aligned} Y &= \frac{n_1}{n_1 + 2n_2} \\ D_1 &= 1 - 2Y\frac{n_2}{n_1} \\ D_2 &= 2 - 3Y\frac{n_3}{n_2} \\ D_{3+} &= 3 - 4Y\frac{n_4}{n_3} \end{aligned} \tag{11}$$

である．ただし，$N_1(c_{n-N+1}^{n-1}\cdot) = \mid \{c_i : freq(c_{n-N+1}^{n-1})\} \mid$ であり，$N_2(c_{n-N+1}^{n-1}\cdot)$ および $N_{3+}(c_{n-N+1}^{n-1}\cdot)$ 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．

4.3 部分列を用いた候補文字検索

n-gram モデルでは連続する文字列を用いる．この場合文字間の距離が 1 となる．skip-bigram モデル[9] 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 モデルでは文字列 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．

$$F(s_1, s_2) = \frac{(1+\beta^2)R(s_1, s_2)P(s_1, s_2)}{R(s_1, s_2) + \beta^2 P(s_1, s_2)} \tag{12}$$

$$R(s_1, s_2) = \frac{SKIP(s_1, s_2)}{C(m, 2)} \tag{13}$$

$$P(s_1, s_2) = \frac{SKIP(s_1, s_2)}{C(n, 2)} \tag{14}$$

ここで m，n はそれぞれ s_1，s_2 の文字列長を示す．$\beta = \frac{P(s_1,s_2)}{R(s_1,s_2)}$ 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” とするとき，$SKIP(s_1, s_2)$ 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．このとき $s_1 = c_1, \ldots, c_{n-1}, c_n$ とする．

本研究では，ユーザが入力した文字列を c_1^{n-1} とし，その直後に出現する文字 c_n とした．候補文字検索ではこの c_n の推定が問題となる．s_2 は学習データ内の任意の文字列である．すべての文字間の関係を扱うには計算量が膨大になりすぎるため，本研究では一定距離内（ウィンドウサイズ）の文字間のみを学習データの対象としている．また，c_n の候補は $a \in \Sigma$（ここで Σ をアルファベットとする）となってしまうため，すべての文字 a を計算してしまう．そこで，(12) 式の値が 0 よりも大きな文字列を候補文字 s_2 とした．次式のように，すべての s_2 について (12) 式を計算し，c_n の候補ごとに総計することで候補文字のスコア $score_{SB}(c_n)$ を算出する．

$$score_{SB}(c_n) = \sum_{s_2} F(c_n \mid c_1^{n-1}, s_2) \tag{15}$$

このスコアに応じて降順にソートすることで候補文字のランキングを行った．また本方法を用いるとき，学習データを作成するとき，前節の n-gram モデルと同様に特徴ベクトルを作成していくが，n-gram の代わりに skip-bigram をカウントしている．

skip-bigram モデルでは，skip-bigram の出現頻度だけではなく照合する 2 つの文字列の文字長も考慮しているが，文字間の距離は考慮していない．gap-weighted string kernel[10] は文字列カーネルの 1 つとして知られており，ギャップサイズに応じて部分列の重みを低減させることができる．文字列カーネルは SVM（Support Vector Machine）[2),6)] でデータの分類を行うために用いられている．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る．この文字列カーネルは次式で定義される．

$$K_N(s_1, s_2) = \sum_{u \in \Sigma^N} \sum_{i:u=s_1[i]} \sum_{j:u=s_2[j]} \lambda^{l(i)+l(j)} \tag{16}$$

ここで Σ^N はすべての文字 Σ で長さ N である部分列，$l(i)$ 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，$\lambda < 1$ であれば文字間の距離に応じて重みを低減し，$\lambda = 1$ であれば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”，$N = 2$，$\lambda = 1/2$ であるとき，Φ(“足利義満”) および Φ(“足利義詮”) を計算し，のち (16) 式より $K_2(s_1, s_2) = \lambda^5 + 2\lambda^4 = 0.15625$ となる．本研究ではこの特徴を利用し，skip-bigram と同様の方法で候補文字を検索する．すなわち，次式で示すスコア $score_{GK}(c_n)$ をもとに検索結果をランキングすることで候補文字検索を行なう．

表 4 テストデータ

データベース名	刊本	時代区分	異なり文字数	延べ文字数
大日本史料総合 DB	大日本史料 6 編 46	6	2202	83001

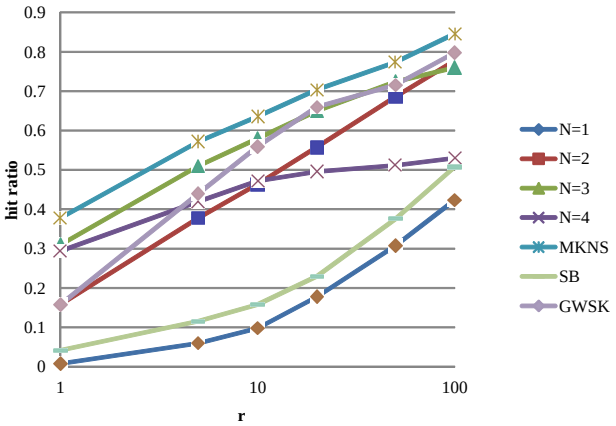


図 6 各種法での候補文字のヒット率

$$score_{GK}(c_n) = \sum_{s_2} K_N(c_n \mid c_1^{n-1}, s_2) \tag{17}$$

4.4 評価実験

4.4.1 実験準備

ここでは各種の候補文字検索機能の性能を評価する．評価する指標としては，検索結果にヒット率と再現率とした．ヒット率は検索結果の上位 r 以内に正解データが含まれる確率 (正解が含まれていた件数)/(テストデータ件数) として求めた．再現率は $r \to \infty$ としたときのヒット率として求めた．**表 4** に示したテキストを用い，表 4 のテキストから任意の位置にある文字を 500 箇所を選択し，これをテストデータとした．4.1 節で提示した n-gram モデルの学習には，表 3 から表 4 を除いたテキストを用いた．4.2 節で示した 2 つの手法では $n = 1, \ldots, 4$ とした．4.3 節で示した 2 つの手法ではウィンドウサイズを 7 とした．また，(16) 式の計算では $\lambda = 1/2$，$N = 2$ とした．

4.4.2 実験結果

候補文字検索の各方法，すなわち，$N = 1, \ldots, 4$ での文字 n-gram モデルを用いた方法

表 5 再現率

手法	N=1	N=2	N=3	N=4	MKNS	SB	GWSK
再現率	1.00	0.980	0.804	0.548	1.00	0.986	0.986

($N = 1, \ldots N = 4$)，文字 n-gram モデルに対して Modified Kneser-Ney スムージングを適用した方法（MKNS），skip-bigram モデルを用いた方法（SB）および gap-weighed string kernel を用いた方法（GWSK）のヒット率を**図 6** に，再現率を**表 5** に示す．図 6 では，x 軸はランクを，y 軸はヒット率を示している．

$N = 1, \ldots, 4$ のうち $r \leq 50$ のとき $N = 3$ がもっともヒット率が高かった．$rge100$ 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，$r \geq 5$ 以上では $N = 3$ でのヒット率には及ばなかった．表 5 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．再現率は各方法のヒット率の上限を示しているため，$N = 4$ では 0.548 を越えるヒット率を実現することができない．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．

SB 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．しかしながらヒット率は他方法に比べかなり低かった．SB に対し方法 GWSK は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 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．この結果より，SB のように文字間の距離をそのまま利用するよりも，そ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．

MKNS では，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，再現率は $N = 1$ と同等である．これらの結果より，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．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，他次元の 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．さらに低次元での n-gram の出現頻度も考慮している．

本システムでの候補文字検索機能では，これらの実験結果より MKNS を採用した．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．そこで，候補文字の提示は 20 件とした．

5. おわりに

翻刻は歴史学や史料学の研究において重要な作業である．本研究では，史料画像と翻刻を関連づけて格納する翻刻データの格納方式，その翻刻データを作成するためのユーザインターフェース，および史料読解のヒントとなる情報を提示する候補文字検索機能を示し，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．また，候補文字検索機能の有効性を評価するための実験を行った結果，Modified Kneser-Ney スムージングを用いた n-gram の方法であれば，検索結果の上位 5 件で 0.57，上位 20 件で 0.72 のヒット率であることがわかった．

翻刻データを XML で出力しているが，可読性が高いとは言えない．そのため，PDF のような人間が見てわかりやすい出力を考慮したい．また，他システムとの共有を考える上では TEI[13] のような人文科学を中心とした文化での情報交換や共有のための共通フォーマットでの出力も必要だと考えている．

候補文字検索では，本研究ではテキストの特徴を用いたが，他の方法の例として，SMART-GS や Transmedia のような史料画像内に出現する文字列画像の検索を可能とするシステムがある．このような方式と融合させることで，翻刻支援機能をさらに向上させることが可能だと考えている．また，n-gram モデルなど最尤推定を行なうモデルでは過学習の問題があるため，それに対応する予定である．

謝辞 研究の一部は，日本学術振興会科学研究費基盤研究（S）（20222001），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) Berliner, B.: CVS II: parallelizing software development, *Proceedings of the Winter 1990 USENIX Conference*, pp.341–352 (1990).
2) Boser, B.E., Guyon, I.M. and Vapnik, V.N.: A training algorithm for optimal margin classifiers, *Proceedings of the fifth annual workshop on Computational learning theory(COLT '92)*, ACM Press, pp.144–152 (1992).
3) Canningham, W.: WikiWiki (2005). http:/c2.com/cgi/wiki?WikiWikiWeb.
4) Chelba, C. and Jelinek, F.: Self-organized language modeling for speech recognition, *Readings in Speech Recognition*, Morgan Kaufmann, pp.450–506 (1990).
5) 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).
6) Cristianini, N. and Shawe-Taylor, J.: *An Introduction to Support Vector Machine*, Cambridge University Press, Cambridge, U.K. (2000).
7) 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.
8) James, F.: Modified Kneser-Ney Smoothing of n-gram Models, Technical report, RIACS Technical Report 00.07 (2000). http://www.riacs.edu/navroot/Research/TR_ pdf/ TR_ 00.07.pdf.
9) Lin, C.Y.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, *Proc. 42nd Meeting of the Association for Computational Linguistics (ACL'04)* (2004).
10) Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N. and Watkins, C.: Text classification using string kernels, *J. Mach. Learn. Res.*, Vol.2, pp.419–444 (2002).
11) Rosenbloom, A.: Introduction, *Commun. ACM*, Vol.47, pp.30–33 (2004).
12) SMART-GS: SMART-GS: a tool for humanistics. http://www.shayashi.jp/SMART-GS/mainjp.html.
13) TextEncodingInitiative: TEI: P5 Guidelines. http://www.tei-c.org/Guidelines/P5/.
14) 国文学研究資料館：吾妻鏡本文検索. http://ocelot.nijl.ac.jp/dlib/azuma/.
15) 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.
16) 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).
17) 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).
18) 人間文化研究機構：研究資源共有化データベース. http://int.nihu.jp/.
19) 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).
20) 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.u-tokyo.ac.jp/ships_help/OSIDE/W34/.
21) 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/.
22) 文化庁：文化遺産オンライン. http://bunka.nii.ac.jp/.
23) 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」(1999). http://www.hi.u-tokyo.ac.jp/personal/kato/index.htm.
24) 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).