木簡解読支援のための文脈処理の提案と実装

西嶋 佳津* 齋藤 恵** 耒代 誠仁*** 中川 正樹*** 馬場 基**** 渡邊 晃宏****

本論文では，木簡の解読を行う専門家を支援するための文脈処理手法，およびその実装について述べる．奈良平城京跡などから出土する木簡の多くに，破損，汚損による文字情報の部分的な欠損が見られる．我々は，“和名類聚抄”に記された地名情報を木簡解読に活用するため，独自の文脈処理手法を設計し，我々が開発した木簡解読支援システム上に実装した．この文脈処理手法を用いると，9~11 文字で構成される地名情報のうち 6 文字が欠損した場合にも，約 74%の確率で上位 10 個の候補の中に正解を含めることが実験によって示された．

Proposal and Implementation of Context Analysis to Support Decoding Scripts on Mokkan

Kazu Nishijima* , Kei Saito** , Akihito Kitadai*** , Masaki Nakagawa***,
Hajime Baba**** , Akihiro Watanabe****

This paper describes a context analysis method and its implementation to assist expert readers of Mokkans. Since most of old Mokkans have been damaged, textual information on them are partially lost. We designed a method to restore the lost information by considering the address information in “Wamyo-ruijyu-syo” and implemented it as a component of our support system for Archaeologists to read Mokkans. In our experiments, the method nominates the correct text in the top 10 candidates with 74% correctness when the length of the original text is 9~11 characters and the 6 of them are missing.

*東京農工大学 工学部，**東京農工大学 大学院 工学教育部，
東京農工大学 大学院 共生科学技術研究部，*奈良文化財研究所
*Faculty of Technology Tokyo University of Agriculture and Technology.
**Graduate School of Technology Tokyo University of Agriculture and Technology.
***Institute of Symbiotic Science and Tokyo University of Agriculture and Technology.
****National Research Institute for Cultural Properties, Nara.

1. まえがき

歴史的文書の解読は，過去の出来事を知る有効な手段であり，その解読に関する研究は古くから行われてきた．近年では，広く普及している情報処理技術を用い，解読作業を支援するシステムや，古文書のデジタルアーカイビングなどの研究が国内外で進められている[1][2].

我が国では，木片に文字情報などを記した『木簡』が，奈良時代を中心に，文書の記録・伝達メディアとして広く利用されていたと考えられている．その理由としては，

- 森林資源が豊富な我が国では調達が容易であった
- 当時，紙は貴重品で高価であった
- 木簡は風雨に強く，荷札などに用いる場合にも優れていた
- 使用済みの木簡であっても表面を削れば再利用が可能であった

などが挙げられる．

これまでに，国内の遺跡などから出土した木簡は約 32 万点にのぼる．そのうち 17 万点以上は奈良平城京跡から出土したものであり，木簡が奈良時代を中心に広く利用されていたことがわかる．出土した木簡の中には荷札として利用された木簡，および中央・地方の役人などに交付された木簡などが含まれており，当時の歴史・文化などを知る貴重な資料となっている．

しかし，出土した木簡の多くに，割れや文字の掠れ，滲みなどの破損や汚損が見られ，記された文字情報を完全に留めたものは稀である．その解読は専門家にとっても容易ではない．

我々は，コンピュータによる手書き文字認識，画像処理などを用いた木簡解読支援システムを提案，実装した[3,4]．しかし，破損・汚損が著しい文字の解読には，文脈処理による支援が不可欠であり，その実現が課題であった．文脈処理とは，ある文字の字種を推定する際，その前後にある文字や文字列，および文法・単語などの言語的知識を用いることを指す．

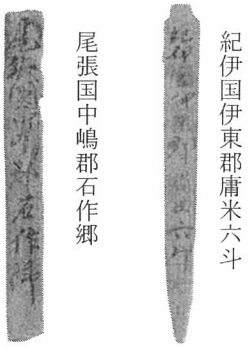


図 1. 木簡の例

本論文では，荷札などの木簡に含まれる地名情報を対象とした文脈処理の実現と木簡解読支援システムへの実装について述べる．

以下，2.では，文脈処理と木簡解読支援システムの概要と構成について述べる．3.では，今回開発を行った文脈処理モジュールの実現について述べる．4.では，文脈処理モジュールに対して行った評価について述べる．5.では今後の課題などを含めてまとめを述べる．

2. 木簡解読支援と文脈処理

2.1 文脈処理を用いた文字情報の推定

木簡の解読では，木簡の現物もしくは木簡を撮影した画像（木簡画像）に記されている文字情報の解読を行う．しかし先述のように，文字情報を完全に留めている木簡は稀である．

木簡の文字情報が欠損している可能性がある場合，専門家は欠損した文字に対していくつかの仮定を行い，検証を行うことで解読を試みる．この作業は専門家の非常に高度な知識や豊富な経験を基に行われる作業であり，現在のコンピュータの情報処理技術を用いて完全に置き換えることは非現実的である．

しかし，仮定や検証の際に専門家の行う作業の一部に対して，計算機が有する情報記憶・管理の能力を生かした支援を行うことは有効であると考えられる．たとえば，字形からでは解読が困難な文字に対して，文脈処理手法を用いることで，文字情報の欠損を伴う木簡の解読に有効であると考えられる．

本研究で提案・実現する文脈処理手法は，古代の地名に用いられた用語，および国，郡，郷／里などの階層構造に関する知識を用いて，木簡の解読を支援するものである．木簡は，荷札などとして広く用いられたため，その多くに地名に関する記述が見られる．このような地名に関する情報を，コンピュータによる文字認識，画像処理と組み合わせて提供することができれば，木簡解読の有効な支援になるものと考えられる．

さらに，当時においても 4000 通り以上と考えられている地名の組合せに対して，個々の地名に縁の深い情報（特産物，人名，方言など）を関連付けて，それらを適宜参照できるようにすることで，文脈処理を用いた支援の効果を高めることができると考えられる．

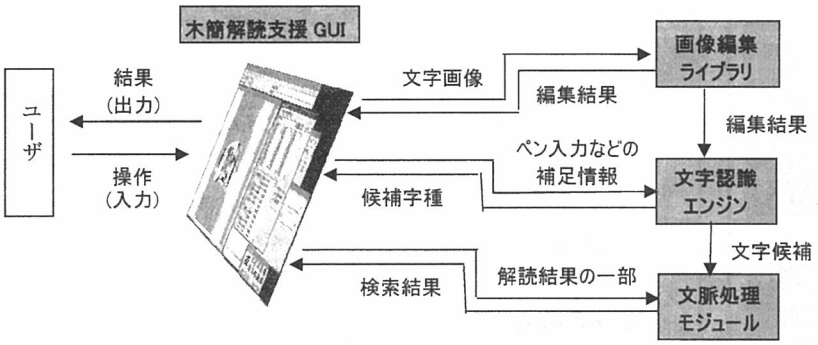


図 2. 木簡解読支援システムの構成

2.2 木簡解読支援システム

我々が開発を進めてきた木簡解読支援システムは，文字認識，画像処理などの，解読作業の支援が可能と考えられる情報処理技術を，ユーザである専門家に対話的に提供する．本システムの目的は，解読作業全体におけるユーザの負担を軽減し，ユーザの発想を助け，木簡の解読を支援することにある．

本論文で述べる文脈処理についても，木簡解読支援システムの一部として実装することを想定して開発を行った．

文脈処理をモジュールとして追加した場合のシステム全体の構成を図 2 に示す．

木簡解読支援システムは，画像処理ライブラリ，文字認識エンジン，木簡解読支援グラフィカルユーザインタフェース（木簡解読支援 GUI），および今回追加を行った文脈処理モジュールから構成される．

画像処理ライブラリは，木簡画像から文字を表す墨の部分（墨部）の抽出を行う作業を支援することで，木簡画像の視認性を高めると共に，後述する文字認識処理の精度向上を実現する．木簡の破損や汚損の状況は様々であるため，4 種類の墨部抽出機能を提供することで対応する．

文字認識エンジンは，古代の文字を対象とした文字認識の機能を提供する．ここでは，木簡の破損・汚損に伴う文字の欠損を考慮した認識手法を採用すると共に，簡単な情報をユーザが補うことで認識精度を高める手法が実現されている．

木簡解読支援 GUI は，木簡解読支援システムの各モジュール／エンジンの機能をユーザに提供すると共に，解読作業の途中経過及び結果を管理，保存する．

なお、本研究で実現する文脈処理モジュールは，解読が困難な文字に対して，その周辺にある文字の解読結果，及び地名に関する情報を用いた文脈処理を提供し，木簡の解読を支援する．詳細は次章で述べる．

3．文脈処理の実現

3.1 内部処理

文脈処理モジュールは，木簡の解読を支援するための文脈処理の実現を目的として，設計，開発を行ったものである．現時点では，奈良時代の地名に関する記述を対象とした文脈処理を実装している．

図 3 に，地名部分に解読が困難な箇所を含む木簡に対して，解読が可能な文字，”国父郡”をキーとする検索を行う例を示した．ユーザは，後述する文脈処理 GUI を介して，文脈処理モジュールが提示する地名の中から可能性が高いものを選択する．次に，データ構造とデータベース検索の詳細を示す．

(1) 地名データベース

奈良時代の地名は，国名，郡名，及び郷／里名から構成されており，その組合せは 4,000 通り以上あったと考えられている．我々は，辞典“和名類聚抄”に記されている地名情報が，奈良時代の地名を考える上で最も適していると考え，これらをデータベース化した．

また，当時の国，郡，郷／里の地名は，現在の国，都道府県，市町村と同様に階層構造となっている．我々は地名データを，[5,6]などを参考に，階層化されたデータベースの構築に適したトライ構造を用いて表現した．さらにそれを拡張することで，地名データベースに対する高速な検索を実現した．

(2) 検索

文脈処理モジュールでは，解読結果の一部を利用してデータベースを検索し，データベースの各地名文字列に対して，可能性の高さを評価値として計算する．この評価値の高いものから文脈処理 GUI に優先的に表示し，解読の支援を行う．

評価値は，解読結果の一部の文字をキーとして算出するだけでなく，その文字の順序にも着目した算出方法をとっている．

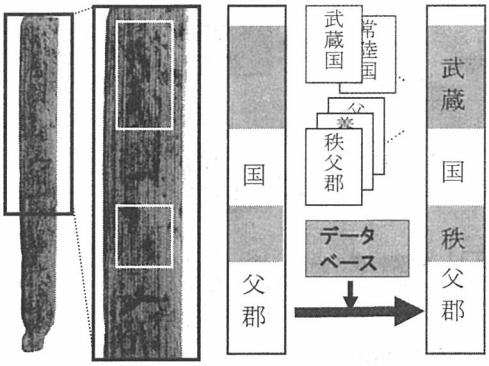


図 3. 文脈的情報が効果的な場合の例

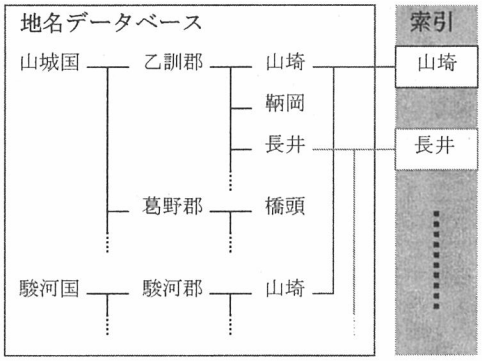
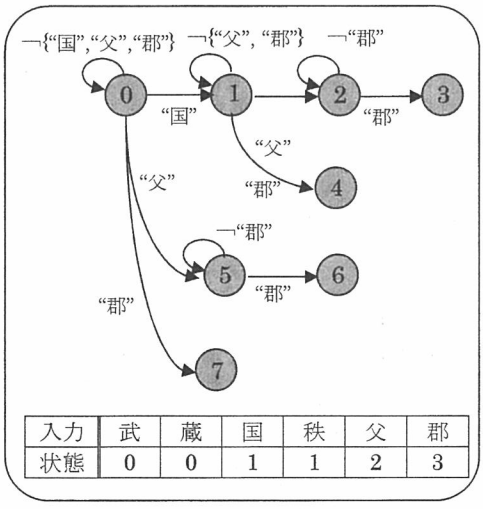


図 4. 地名のデータ構造（一部）



入力	武	蔵	国	秩	父	郡
状態	0	0	1	1	2	3

図 5 ． 生成するオートマトンと状態遷移の例

破損・汚損の著しい木簡では，ある文字と，他の文字の間に挟まれている文字が何文字かの判別が難しい場合がある．入力として得る解読結果の一部を最大限に利用するため，検索と評価のアルゴリズムは，複数キーによる全文検索を行う際に有効な Aho-Corasick 法（AC 法）を参考にして設計した．AC 法では，検索するキー文字列からマシン AC と呼ばれる有限オートマトンを生成する．この有限オートマトンに検索対象を入力すると，1 つのデータを 1 回走査するだけで，複数のキーワードを同時に検出することができる．今回設計したアルゴリズムはこの考え方を利用し，検索と評価値の計算を並列に行う．

入力として得られる解読結果の一部の文字列は，部分的な解読結果の寄せ集めであり，ある文字と次の文字の間に，別の文字が存在する可能性をはらんでいる．図 3 の例では，”国父郡”は部分的な情報で，この部分の正しい解読結果は“武蔵国秩父郡”である．そのため，部分的なの解読結果が例えば『ABC』であった場合，『＊A＊B＊C＊』（ただし，＊は 0 個以上の文字）のように扱う必要がある．

我々のアルゴリズムでは，部分的な解読結果から，図 5 のようなオートマトンを生成する．初期状態を 0 とし，各地名データをこのオートマトンに順次入力し，図 5 の例のように状態遷移させることで，検索および評価値の算出を行う．状態 3,4,6,7 のような終端ノードに達する，または何らかの理由で状態のリセットが行われた場合，そのノードの保持する一定の値を評価値に加算する．

ノードの保持する値は，オートマトンの生成時に計算する．この値は，経験的に求めた次の式により算出される．

$$value = 2^{p-1}$$

ただし，*value* は特定のノードの保持する値を，*p* は状態 0 のノードから特定のノードに到達するまでに経由する自己遷移でない状態遷移パスの数を，それぞれ示す．

図 5 の例では，この算出方法により，状態 1 のノードは値に 1，状態 2 のノードは値に 2，状態 3 のノードは値に 4 を持つ．

データの検索が終了したら，データを評価値の高い順にソートし，文脈処理 GUI に出力する．

3.2 文脈処理 GUI

文脈処理 GUI では，既に解読済みの一部の結果が文字入力エリアに入力され，それを含む国名，群名，郷／里名が候補としてガイドエリアおよび候補出力エリアに表示される（図 6）．さらに，ガイドエリアや候補表示エリアの地名候補を選択することで，その地名を中心におい

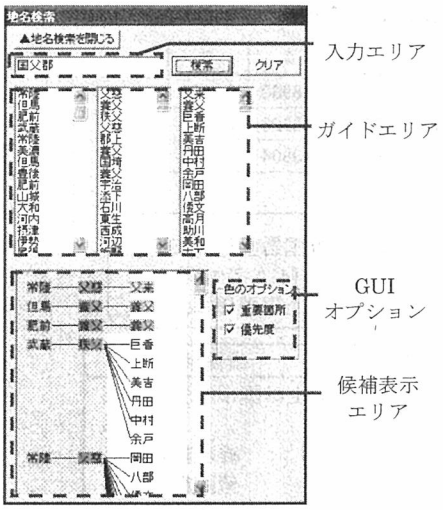


図 6 ． 文脈処理 GUI

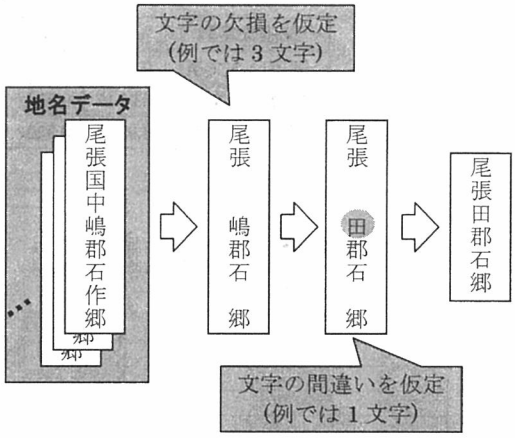


図 7 ． 実験データの作成

た地名の組合せ情報が出力され，文脈的情報の追跡を行うことができる．また，GUI オプションを操作することにより，システムの算出した評価値の大小などの表示を，ユーザの好みや必要に応じて切り替えることができる．

文脈処理 GUI は，木簡解読支援 GUI の一部として実装される．通常は小さなバーの状態に畳み込まれており，必要に応じて展開ボタンをクリックすることで，文脈処理 GUI が利用可能になる．

文脈処理 GUI を追加することにより，木簡解読支援システムが既に有する機能と文脈処理を連携させた解読作業が可能となる．

4．評価実験

文脈処理モジュールは，部分的な解読結果の情報から，より広い範囲の解読結果を推定する．この推定の精度について定量的な評価を行った．

実験に利用するデータには，実在する木簡の解読結果を利用することが望ましい．しかし，出土した木簡に記された地名だけを用いた場合，統計的な偏りが発生する可能性がある．

そこで，本研究では"和名類聚抄"に記された地名（4,000 種類以上）を実験に利用した．“和名類聚抄”は奈良時代に近い平安時代の辞書と考えられており，木簡が作成された当時の地名を多数収録していると考えられる．

次に，実験データの作成方法を次に述べる．ここでは，上記の地名を表す文字列（地名データ）の，9~11 文字からなるものに対して，

① 1~6 文字を欠損させたもの
② 1~6 文字を欠損させ，残りのうち 1 文字を誤字に変更したもの
③ 1~6 文字を欠損させ，残りのうち 2 文字を誤字に変更したもの

の 3 種類を作成し，文脈処理モジュールの精度を評価するための実験データとした（図 7）．なお，欠損は木簡に記された文字が読めない場合に，また誤字は木簡に記された文字が間違って読まれた場合に，それぞれ対応する．

文脈処理モジュールは，4,000 種類以上の地名のすべてを推定の対象とするように構成した．これを用いて文字情報の欠損，および誤字を含む実験データを処理し，欠損した文字の補完，および誤字の訂正が正確に行えるかを調べた．

なお，文脈処理モジュールは，欠損が発生する前の文字列の文字数，および誤字や欠損が発生している位置の情報は一切利用しない．

評価方法は次のように設定した．文脈処理モジュールは，解読結果の補完，および訂正の結果として，実験データごとに評価値の高い順に候補となる地名データを出力する．そのため，出力された評価値の高い地名データのうち，上位 10 個中に正しい地名が含まれた場合，正確な推定が行われたものと判断した．

実験の結果を表 1 と図 8 に示す．9~11 文字の地名情報のうち，5 文字，6 文字が欠損していると仮定した場合，それぞれ約 90%，約 74%の確率で正解が上位 10 個に含有された．この結果から，文脈処理モジュールを木簡解読支援システムに追加することで，地名を表す文字列の半分以上が欠損している場合でも，解読の支援が可能になると考えられる．

また，正答が上位 10 個に含有されなかった実験データを解析したところ，その大半に共通した特徴が見られることがわかった．それは，地

表 1. 各実験における正答の上位 10 位含有率

欠損文字数	1	2	3	4	5	6
間違い 0 文字	0.9976	0.9946	0.9863	0.9620	0.8980	0.7369
間違い 1 文字	0.9949	0.9842	0.9450	0.8321	0.6328	0.3099
間違い 2 文字	0.9810	0.9362	0.8255	0.6410	0.3804	0.1678

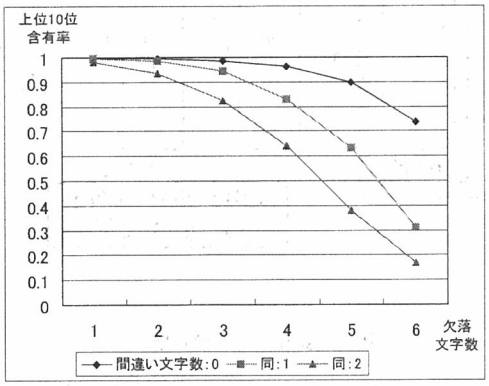


図 8. 各実験における正答の上位 10 位含有率

名データから実験データを生成する際，地名に頻出な漢字が多く実験データに残ってしまい，同程度の評価値で大量の候補が出力されるケースであった．代表例を図 9 に示す．

このケースによる精度低下を軽減するには，検索の段階で，出現確率の高い地名が上位に出やすくなるよう重みをつけ，多く評価値を割り振る処理を加える方法が考えられる．

5．あとがき

本論分では，木簡の解読を行う専門家を支援するための文脈処理手法，およびその実装について述べた．

木簡に記された情報は，地名だけに留まらない．地名情報に地域情報をリンクさせることで，地域的な特産品や人名などの情報を統括的に操作できる機能を追加すれば，より文脈処理としての解読支援はより有効になると考えられる．

また，今後の課題として，扱う情報量が増えても，簡単な操作で必要なものだけを出力できるユーザインタフェースの開発が挙げられる．

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 S: No.15102001 の補助による．

『上野国吾妻郡長田郷』
↓(4 文字欠損)
『上国郡田郷』

順位	候補地名
1	上総国市原郡江田郷
2	山田郷
3	畔蒜郡新田郷
4	望陀郡磐田郷
⋮	⋮
9	上野国片岡郡若田郷
10	多胡郡八田郷
⋮	⋮
14	吾妻郡長田郷
⋮	⋮

図 9. 10 位以内に含有されない代表例

参考文献

[1]Min Soo Kim, Kyu Tae Cho, Hee Kue Kwag and Jin Hyung Kim, "Segmentation of Handwritten Characters for Digitalizing Korean Historical Documents," Proc. 6th International Workshop on Document Analysis Systems, pp.114-124, Sep.2004.

[2]Basilios Gatos, Kostas Ntzios, Ioannis Pratikakis, Sergios Petridis, T. Konidaris and Stavros J. Perantonis, "A Segmentation-Free Recognition Technique to Assist Old Greek Handwritten Manuscript OCR," Proc. 6th International Workshop on Document Analysis Systems, pp.63-74, Sep.2004.

[3]未代誠仁, 齋藤恵, 蜂谷大翼, 中川正樹, 馬場基, 渡辺晃宏,"木簡解読支援システムの基本設計と試作", 人文科学とコンピュータシンポジウム論文集, 5-A-1, pp.215-220, Sep. 2004.

[4]Akihito Kitadai, Kei Saito, Daisuke Hachiya, Masaki Nakagawa, Hajime Baba and Akihiro Watanabe, "Support System for Archeologists to Read Scripts on Mokkans,"Proc.8th ICDAR, vol.2, pp.1030-1034, Aug. 2005.

[5]丸山勝美, 古賀昌史, 嶋好博, 藤澤浩道, "手書き漢字住所認識のためのエラー修正アルゴリズム", 情報処理学会論文誌, Vol.35, No.6, pp-1101-1110, Jun.1994.

[6]古賀昌史, 嶺竜治, 酒匂裕, 藤澤浩道, "トライ辞書を用いた語彙情報駆動型の印刷地名単語列認識方法", 電子情報通信学会論文誌, Vol.J86-D-II, No.9, pp.1297-1307, Sep.2003.