科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１６　日現在

機関番号：　１０１０１
研究種目：　若手研究(B)
研究期間：　2012～2013
課題番号：　２４７００１５６

研究課題名（和文）文書画像全文検索技術を基盤とする古文書画像翻刻支援システムの研究

研究課題名（英文）Assisting Application for Analytical Transcription of Historical Documents based on Full-text Search System of Document Images.

研究代表者
猪村　元（Imura, Hajime）

北海道大学・知識メディアラボラトリー・特任助教

研究者番号：７０６１５２１０

交付決定額（研究期間全体）：（直接経費）　　3,100,000円、（間接経費）　　930,000円

研究成果の概要（和文）：本研究では、これまでに確立した基本文字列検索技術に加えて、翻刻知識を利用し、擬似コードを用いた高度なテキスト解析を確立することにより、翻刻知識の共有による翻刻者の協働作業を可能とし、高機能、高精度な翻刻支援システムを構築することを目標とする。この目標を達成するために、平成24年度は、(1)翻刻データの蓄積によるキーワード検索の精度向上、(2)入力補完のための翻刻候補の提示技術の二つの研究課題を遂行した。平成25年度には、(3)入力補完のための翻刻候補の提示技術と翻刻エディタの開発 (4)N-gram共起頻度に基づく文書画像上でのテキスト解析技術の4つの研究課題を遂行した。

研究成果の概要（英文）：We have developed an assisting application for analytical transcription of historical documents based on full-text search technique for image-scanned documents that does not recognize individual characters. The proposed method works independently of differences in language and font because it uses a new pseudo-coding scheme based on the statistical features of character shapes. In this project, we have developed an assisting application for historical document and back-end search server as a service, relevance feedback method with HMM and font robust pseudo-coding method.
The system support transcribers of cursive manuscripts and accelerate the analysis of historical materials.

研究分野：　総合領域

科研費の分科・細目：　情報学・知覚情報処理・知能ロボティクス

キーワード：　画像情報処理　文書画像処理

１．研究開始当初の背景

近年、世界各国の大学や企業、機関による大規模な図書館や公文書館の文書の電子化プロジェクトが盛んに進められ、これにより歴史的に貴重な文献や資料の画像や、知的財産権の失効した文書の全文がウェブ上に公開されるようになってきた。その膨大な量の文書を有効活用するために、状態の良いものは OCR 技術によって機械可読なテキストへと変換される。一方で、保存状態の悪い劣化の大きな活版印刷時代の文書や手書きの古文書などの歴史的文書に対しては、原本の経年劣化や多様な字体の問題、崩し文字や続け文字によって正確な文字単位での切り出しが非常に困難なため OCR 技術の適用は限られており、膨大な量の文書に対し翻刻と呼ばれるテキストへの変換が専門家の手によって行われている。例えば、東京大学史料編纂所の保有する史料は、現在の処理ペースで翻刻を進めるとおよそ千年に及ぶ期間を要すると見積もられている。これらの文書に対して、全文検索の手法を基盤とした翻刻支援技術を提供することは重要である。

２．研究の目的

近年重要視されている歴史的に貴重な文献や資料画像に対する機械可読テキストへの人手による変換(翻刻)作業に対して、文書画像全文検索技術を基盤とした翻刻支援技術を提供することは意義が大きい。申請者らは、入力補完のための翻刻候補の提示技術、文書画像上でのテキスト解析を利用した古文書分析補助の 2 種類の翻刻支援技術を確立することを目標とする。そのために、翻刻データの蓄積によるキーワード検索の精度向上、入力補完のための翻刻候補の提示技術、N-gram 共起頻度に基づく文書画像上でのテキスト解析技術の 3 つの課題を解決する。これによって翻刻者の協働作業を可能とし、より高機能、高精度な翻刻支援システムを実現する。本研究計画における成果は、史料編纂所の保有するような手書き日本語古文書のみならず、国会図書館近代デジタルライブラリー等に所蔵される劣化の大きな活字文書、英語文書に対しても適用可能であり、種々の文書画像の翻刻を加速し文書活用の促進を図る。

３．研究の方法

本研究では、文書画像全文検索技術を基盤として、翻刻者に対する翻刻候補テキストの提示による入力補完、および、文書画像上でのテキスト解析による古文書分析補助を可能とする古文書画像翻刻支援システムの構築を目標とする。具体的には、以下の 3 項目の研究課題を実現する。

(1)　翻刻データの蓄積によるキーワード検索の精度向上

(2)　入力補完のための翻刻候補の提示技術

(3)　N-gram 共起頻度に基づく文書画像上でのテキスト解析技術

上記(1)では、翻刻データ（文字列画像と翻刻テキストの組）を蓄積、共有することで、既存の検索技術の検索精度を向上させる技術を確立する。ここで、1 文字に対する 10 位までの平均順位正答率を、現在の 0.43 から、現在の 3 文字に対する正答率である 0.85 程度まで向上させる。上記(2)では課題(1)にて開発した技術を利用して、高精度な翻刻候補の提示技術を実現する。これは、翻刻者が次に翻刻する文字列をドラッグして選択した際に、画像上での全文検索により類似した形状の文字列を検索し、検索結果の文字列画像上に過去に付加した翻刻テキストがあればこれを順位付けして提示し、翻刻者は選択することで入力を可能とするものである。上記(3)では、
文書画像上での全文検索技術を応用した重要語の抽出とその共起頻度に基づく文書画像上でのテキスト解析技術を開発する。さらに、課題(1)、(2)によって達成される翻刻データの蓄積と、検索精度の向上を利用して解析精度の向上を図る。

４．研究成果

本研究では、これまでに確立した基本文字列検索技術に加えて、翻刻知識を利用し、擬似コードを用いた高度なテキスト解析を確立することにより、翻刻知識の共有による翻刻者の協働作業を可能とし、高機能、高精度な翻刻支援システムを構築することを目標とする。この目標を達成するために、平成 24 年度は、(1)翻刻データの蓄積によるキーワード検索の精度向上、(2)入力補完のための翻刻候補の提示技術の二つの研究課題を遂行した。平成 25 年度には、(3)入力補完のための翻刻候補の提示技術と翻刻エディタの開発 (4)N-gram 共起頻度に基づく文書画像上でのテキスト解析技術の 2 つの研究課題を遂行した。

(1)翻刻データの蓄積によるキーワード検索の精度向上: 劣化印刷文書画像や手書き草書体古文書画像へ適用可能な全文検索技術をもとに、これを汎用プロトコルに準拠した Web サービスとして検索サーバの構築を行った。さらに翻刻データを蓄積・共有することで、既存の検索技術の検索精度を向上させる技術の研究開発を行った。具体的には、ユーザが画像上で翻刻テキストを入力した際に、文書画像上での文字列の位置情報とユーザが入力した翻刻テキストをサーバに送信し、蓄積する。これらの蓄積された翻刻データを用いて、HMM 学習による適合性フィードバックを利用した検索精度の向上技術を確立した。特に、日本語文書の場合には 1 文字、2 文字のキーワードに対処する必要があるため、1 文字に対する 10 位までの平均順位正答率を、現在の 0.43 から、2 文字，3 文字に対

する正答率に近い 0.8 程度まで向上させた（図 1）。

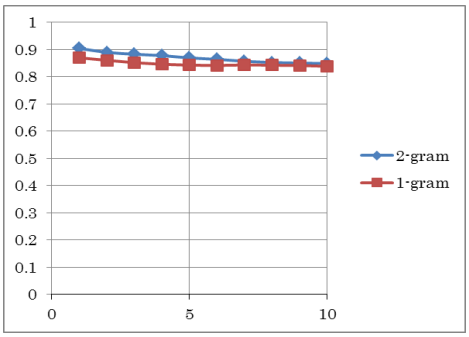

図 1 10 位までの順位正答率

(2) 入力補完のための翻刻候補の提示技術：課題(1)にて開発した技術を利用して、翻刻候補の提示技術を開発した。具体的には、GUI 上で対話的に検索、翻刻テキストを入力して翻刻作業が可能なユーザ・インターフェイスの開発を行った（図 2）。これは、ユーザの文書上での単語選択、入力作業に応じてバックエンドにて検索と候補の提示順位の計算を行い表示するもので、新規性の高いものであるといえる。

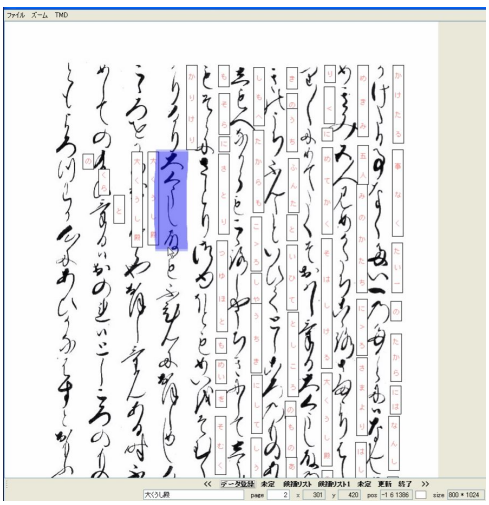

図 2 翻刻支援 GUI アプリケーション

(3)入力補完のための翻刻候補の提示技術と翻刻エディタの開発：H24 年度に開発した文書画像全文検索のための検索 Web サービスに対し並列処理による高速化を図り、5 文字の検索文字列の場合に平均 100msec での応答が可能な検索サーバ構築を行った。翻刻支援のための候補提示を対話的、かつリアルタイムで行うためには、大量の検索要求をバックグラウンドで処理する必要がある。また，これらの研究では研究者が文書画像自体の公開に関する権利を有して無い場合も多い。このような様々な手書き古文書解読・翻刻支援アプリケーションへのトランスメディア全文

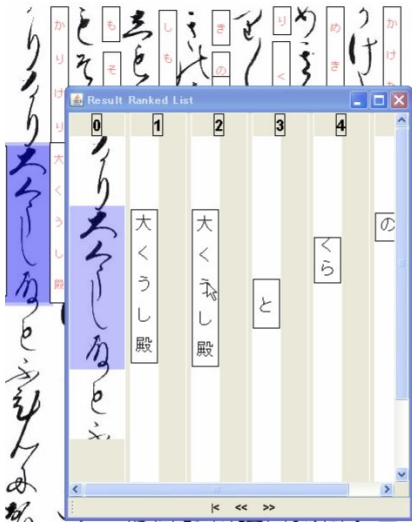

図 3 翻刻候補提示詳細ビュー

検索機能の組み込みを考えたとき、検索機能を共通基盤としてサービス化することは重要である。ICDAR2009 にてワード・スポッティング・プロトコル(以下, WSP) [Terasawa 09] として定義したプロトコルに準拠したサーバアプリケーションサービスとした。これにより、翻刻範囲の指定と同時に動的にバックエンドでの検索を行い、ほぼタイムラグ無く翻刻候補を提示することが可能となり、動的に順位付けされて表示される翻刻候補を見ながら翻刻範囲を決定できる翻刻エディタを実現した（図 3）。

(4)N-gram 共起頻度に基づく文書画像上でのテキスト解析技術：国立国会図書館にて公開されている近代デジタルライブラリーの文書に対して頻出語抽出のための技術を開発した。具体的には、異なるフォント間の差異を吸収して検索を実現するための文字画像特徴量の特徴空間における近傍構造に基づいた擬似コード生成手法を開発し、2-gram 検索精度評価にて平均適合率 80%以上を達成し、N-gram 頻出語の抽出を可能とした。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　　0 件）

〔学会発表〕（計　　1 件）
[1]Hajime Imura, Assisting Application for Analytical Transcription of Historical Documents, International Workshop on Information Search, Integration and Personalization, Oct. 2012, Sapporo

〔図書〕（計　　0 件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.meme.hokudai.ac.jp/transmedi
a/

６．研究組織
(1)研究代表者
猪村　元（IMURA, Hajime）
北海道大学・知識メディアラボラトリー・
特任助教
研究者番号：70615210

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：