歴史オントロジー構築のための史料からの人物情報抽出

石川　徹也[†]・北内　啓[††]・城塚　音也[††]

本研究の目的は，歴史資料（史料）を対象に歴史知識の構造化の基盤となる「歴史オントロジー」を構築するシステムを開発し，広く提供することによって歴史学の発展に寄与することにある．この目標を具体的に検証するために，昭和15年に時の帝国学士院において始められた明治前日本科学史の編纂成果である『明治前日本科学史』（刊本全28巻）の全文を日本学士院の許諾の下に電子化し，明治前の日本の科学技術を創成してきた科学技術者に関する属性および業績の情報を抽出することにより，前近代日本の人物情報データベースの構築を試みる．人物の属性として人名とそれに対する役職名と地名を，人物の業績として人名とそれに対する書名を，いずれもパターンマッチングなどのルールベースの手法によって抽出する．『明治前日本科学史総説・年表』を対象とした性能評価を行った結果，人名，人名とその役職名，および人名とその地名について，F値で0.8を超える結果が得られた．

キーワード：情報抽出，歴史情報，史料，歴史知識学，オントロジー

Extraction of Person Information from Historical Materials for Building Historical Ontology

Tetsuya Ishikawa[†], Akira Kitauchi[††] and Otoya Shirotsuka[††]

Our goal of this study is to contribute to the progress in historical science by developing a system for building a historical ontology from historical materials and making it available to the public. We digitize all the books of "Meiji-mae Nippon Kagaku-shi" (Pre-modern Japanese History of Science and Technology) published by Nippon Gakushiin (The Japan Academy), and extract the attribution and the works of scientists and engineers from the books to build a database of person information in pre-modern Japanese history. We extract the names of persons, positions, places, and books as the attribution and the works of persons by pattern matching. The experimental results show that the F-measures for the names of persons, positions, and places are over 0.8.

Key Words: *Information Extraction, Historical Information, Historical Materials, Knowledge-based Historical Science, Ontology*

[†] 東京大学史料編纂所前近代日本史情報国際センター, International Center for Digitization of Pre-modern Japanese Sources, Historiographical Institute, University of Tokyo
[††] 株式会社 NTT データ技術開発本部ビジネスインテリジェンス推進センタ, Business Intelligence Deployment Center, Research and Development Headquarters, NTT DATA Corporation

1 はじめに

本研究の目的は，歴史資料（史料）から歴史情報を自動抽出する方式を確立すること，および歴史知識を構造化するためにその抽出結果を歴史オントロジーとして構築し，提供することにある．歴史研究は史料内容の解読から始まる．そのために史料の収集・翻刻（楷書化）・解読の作業が伴う．ただし，史料の形態・記述は多様であり，翻刻・解読には相当の知識と経験を必要とする．国内には未解読の史料が未だ多数存在する．一方，これまでに解読された結果についても電子化されていない，あるいは機関・個人など個別に存在するために各史料を共用できないという問題があり，歴史事象の関連性の解明，すなわち歴史研究の推進そのものに支障をきたしている．この種の問題解決のために，すなわち歴史知識の構造化のために歴史オントロジーの提供が求められている．

われわれは，歴史研究のより一層の推進を目的として「歴史オントロジー構築プロジェクト」を実施している．本プロジェクトは，史料を電子化する，史料に記載されている情報を抽出，構造化して歴史オントロジーを構築する，歴史オントロジーを利用した検索・参照システムを構築するという3つの手順によって構成されている．本プロジェクトを具現化するための史料として『明治前日本科学史』（日本学士院編・刊行，全28巻）を対象に歴史オントロジーを構築する．当刊行史料は，明治前日本科学史の編纂を目的に昭和15年に帝国学士院において企図され，昭和35年に最初の巻が出版され，昭和57年に28巻目の刊行によって現在，完結している．本史料をより有効に活用するため，全巻の電子化および研究目的の利用・提供に関して日本学士院の許諾を得て，電子化に着手した．本史料は公的性が高く，歴史研究の推進という本研究の目的に適合するものである．本史料から日本の科学技術を創成してきた明治前の人物に関する情報を抽出，構造化することにより歴史オントロジーを構築する．

本研究では，プロジェクトの第一歩として，『明治前日本科学史』のうちの1巻『明治前日本科学史総説・年表』の本文を電子化したテキストから，人物の属性として人名とそれに対する役職名と地名，人物の業績として人名とそれに対する書名を抽出する．機械学習に基づく情報抽出によって十分な精度を得るには大量の正解データを作成する必要があり，多大な時間がかかることから，本研究ではルールベースの手法によって人物に関する情報を抽出する．

本稿では，2章で歴史オントロジー構築プロジェクトの全体像を示す．3章で人物に関する情報を抽出する手法について説明し，4章で実際に評価実験を行い，その結果を考察する．最後に5章で本研究の結論を示す．

2　歴史オントロジー構築プロジェクト

歴史研究において，史料はその手がかりとなる重要な資源であるにも関わらず，多くの史料が電子化されておらず，また電子化されていたとしても利用しやすい形で提供されていない場合が多いという問題がある．東京大学史料編纂所では史料をデータベース化し，キーワードなどの条件で検索可能なシステムを提供する取り組みを進めており，これまでにその一部が「東京大学史料編纂所データベース SHIPS」(東京大学史料編纂所 2006) として公開されている．このシステムで提供されているデータベースには，史料中の文や図が掲載されている箇所（巻やページなど），記述された出来事の日付，図に描かれた人物の名前などのメタデータが付与されているものもある．しかし，これらのメタデータのほとんどを人手で付与しているため，データベースの構築に多大な時間を費やしている．また，上記以外のメタデータ，たとえば出来事が発生した場所，人物の役職や著作といった情報はほとんど付与されていないため，「ある人物が執筆した著作の年代順一覧」などのように，検索条件や出力内容に様々な種類の情報を指定した複雑かつ柔軟な検索を行うことができない．

われわれは歴史研究のより一層の推進を目的として，史料に記載されている多様な情報をより効率的に抽出，構造化して歴史オントロジーを構築することにより，広範な歴史情報を様々な形で利用可能とするためのプロジェクト「歴史オントロジー構築プロジェクト」を進めている．本章では，本プロジェクトの全体構成，および本プロジェクトにおける歴史オントロジーの詳細について説明する．

2.1　プロジェクトの全体構成

本プロジェクトは，史料の電子化，電子化されたデータからの歴史オントロジーの構築，歴史オントロジーを利用した史料の検索・参照システムの構築という 3 つの手順によって構成されている（図 1）．

まず，史料を電子化する．対象となる史料は印刷物として刊行されている『明治前日本科学史』全 28 巻である．史料の電子化においては，その利用目的に応じた電子化方式を確立する必要がある．たとえば，史料の見た目をそのまま復元すればよいのであれば高い解像度でスキャンした画像を蓄積すればよいが，それだけではキーワードによる検索ができない．また，史料中の文を単にテキスト化するだけでは，年代や人名などの情報を利用した検索や図の参照ができない．複雑かつ柔軟な検索を実現するには，史料に記述された文や掲載されている表や図について，できるだけ論理的な構造を保持したまま XML などの構造化文書の形式で電子化する必要がある．たとえば文については，文字のテキスト化における外字の表現や，文のテキスト化における章や節，箇条書きなどの論理構造の表現といった課題がある．また，表や図については，表構造の表現，タイトルや説明文との関係付け，文中で参照している箇所との関係付け

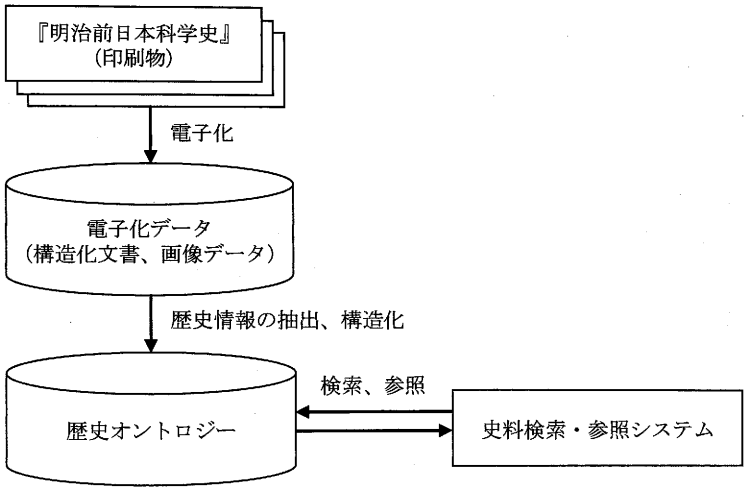

図 1 歴史オントロジー構築プロジェクトの全体構成

などが検討課題となる.

次に,電子化されたデータから歴史情報を抽出し,構造化することにより歴史オントロジーを構築する.歴史オントロジーの内容と構築手順については次節以降で述べる.

最後に,歴史オントロジーを利用した史料の検索・参照システムを構築する.歴史オントロジーがもつ情報を最大限に活用し,歴史研究に必要とされる様々な観点での検索の実現,利用者が必要とする情報を分かりやすく表示する検索結果の可視化,使いやすいユーザインタフェースなどが検討課題としてあげられる.また,検索・参照機能を提供するだけではなく,歴史オントロジーそのものを RDF などの形式で公開し,利用者が自由に利用できるようにすることも計画している.

2.2 歴史オントロジーの定義

本プロジェクトの対象となる史料『明治前日本科学史』には,主に明治よりも前の時代における日本の科学・技術の成果に関する史実(歴史的事実)が,図や表とともに各事項について時系列で記述されている.史実としては,各時代の科学・技術の推移やその内容,および科学者の業績などの人物に関する情報(以下,人物情報)が大部分を占める.特に人物情報は歴史研究におけるニーズが高く,人物情報を検索・参照可能とすることは歴史研究への貢献度が高いと考える.そこで,人物情報を可能な限り漏れなく抽出することを本プロジェクトの目標のひとつとする.ただし,史料中には一部史実でない記述,たとえば推測,疑問,感想などが記述されており,これらは抽出対象外とする.『明治前日本科学史総説・年表』における科学者の業績に関する記述部分の抜粋を図 2 に示す.

和算家は一般に洋学に耳を傾けなかったが、本多利明のような異例もあることを前に述べた。本多は算学を関流建部派の今井兼庭に、天文を千葉歳胤に学んだ。また彼は『経世秘策』などを書いている経世家であり、『長器論』において船舶は国家の長器であることを論じ、航海術を習得して海外に飛躍すべきことを説いた。航海術はオランダがすぐれていることを知り、その航海術書を利用するため、中川淳庵・大槻玄沢などの門を叩いてオランダ語を勉強し、オランダの航海術書によって『渡海新法』（一八〇四）を書いた。蝦夷地開発を唱え、みずから同地へ舟行したときのことを『渡海日記』（一八〇一）に書いている。このように測量航海などの実利的方面からオランダの学術を取り入れた人もあるのである。

図 2　『明治前日本科学史総説・年表』における科学者の業績に関する記述部分の抜粋

抽出した情報を様々な形で利用可能とするためは，資料中の文や図などをそのまま抽出して，キーワードで検索できるようにしたり一覧表示したりするだけでは不十分であり，人名で検索したり書名を一覧表示したりといった，様々な種類の概念に基づく検索や参照が必要となる．この要件を実現するため，各種の概念を史料から抽出し，それらを構造化してオントロジーとして蓄積する．以降，史料から抽出した人物情報を格納したオントロジーを「歴史オントロジー」とよぶ．

オントロジーについては，哲学をはじめとした様々な観点からの定義が提案されている (Mizoguchi and Ikeda 1997) が，人工知能分野においては「概念間の関係を記述することによって知識を体系化したもの」と定義づけられることが多い．たとえば Standard Upper Ontology Working Group(IEEE 2003) では，オントロジーは概念 (concepts)，関係 (relations)，公理 (axioms) の 3 つの要素によって構成されており，これらの構成要素によって，ある分野に関する事物や構造を記述できるとしている．本稿では，この説明に基づいて歴史オントロジーを定義する．

歴史オントロジーで対象とする分野は日本の科学史である．「概念」は，人物情報を構成する各要素である．人物情報は，人物自体を指す情報，人物の属性，人物の業績の 3 種類に大きく分類される．人物自体を指す情報としては人名，写真，似顔絵などが，人物の属性としては役職，出身地，生没年，家族関係などが，業績としては著作，建造物，訪問先などがあげられる．したがって，歴史オントロジーにおける「概念」，すなわち人物情報の構成要素としては，人物，写真，役職，年代，書籍，建造物，場所などがあり，概念どうしを結ぶ関係としてはたとえば以下のものがあげられる．

- 人物から写真への「写真」という関係
- 人物から年代への「生年」という関係

表 1 歴史オントロジーにおける，概念どうしを結ぶ関係（一部）

概念 1	概念 2	関係	関係の種類
人物	写真	写真	人物自体
人物	似顔絵	似顔絵	人物自体
人物	役職	役職	属性
人物	場所	出身地	属性
人物	年代	生年	属性
人物	年代	没年	属性
人物	人物	父	属性
人物	人物	門弟	属性
人物	書籍	著作	業績
人物	建造物	建立	業績
人物	場所	訪問先	業績
書籍	年代	発行年	業績
建造物	年代	建築年	業績

- 人物から書籍への「著作」という関係
- 書籍から年代への「発行年」という関係

本プロジェクトで抽出対象とする歴史オントロジーの概念や関係の全体像については現在検討中であるが，これまでに候補としてあがっている概念どうしを結ぶ関係の一部を表 1 に示す．なお，表 1 において概念間の階層関係はない．

本プロジェクトでは基本的に各概念のインスタンスをその名前で表し，写真などの画像データは「人物」という概念とは別の概念として「写真」などの関係で結ぶ．したがって，人物情報のうち，人名は「人物」という概念のインスタンスとして，それ以外の情報は概念のインスタンスどうしを結ぶ関係として表現される．

「公理」は概念や関係が満たす制約条件であり，概念間の階層関係や，ある概念がもつ各関係の数（たとえば「一人の人物は生年を一つだけもつ」）などがあげられる．

2.3 歴史オントロジーの構築手順

歴史オントロジー構築の最終目標は，『明治前日本科学史』全 28 巻を電子化し，そこから人物情報を高精度に抽出することである．しかし，以下のような問題があるため，それを一度に実施するのは膨大な時間がかかる上に非効率的である．

- 人物情報には様々なものがあり，抽出すべき情報を決めるのには時間がかかる．
- 全 28 巻の電子化，特に写真や図表などの画像を電子化するのには時間がかかる．
- 高精度な抽出方式を確立するには，仮説と検証を繰り返す必要があり時間がかかる．

そこで，以下の手順で段階的に歴史オントロジーを構築する．

(1) 数巻程度の本文をテキスト形式で電子化する．
(2) 電子化されたテキストを対象として，人物情報のうち特に歴史研究に必要とされる情報を高精度に抽出する方式を確立する．
(3) 全巻について，テキストと画像の両方を電子化する．
(4) 全巻の電子化データを対象として，(2) と同様に人物情報のうち特に歴史研究に必要とされる情報を抽出する．
(5) 抽出する人物情報を拡大し，全巻を対象として様々な人物情報を高精度に抽出する方式を確立する．

われわれは現在，上記の手順のうち (1), (2) について取り組みを進めている．また (3) のうち，画像の電子化方法について検討中である．

3 人物情報の抽出

本研究では，『明治前日本科学史』全 28 巻のうちの 1 巻『明治前日本科学史総説・年表』を対象として，人名，人物の属性として人名とそれに対する役職名と地名，および人物の業績として人名とそれに対する書名を抽出する．これらは基本的かつ重要な情報であり，また史料中の記述量も多いため評価実験で多くの知見が得られる見込みが高い．役職名には，「将軍」のような官職（役人の職業）の名前と「医師」のような一般の職業名の 2 種類がある．地名には，その人物の出身地，所属する組織の地域，国籍などがあり，歴史研究への活用のためにはそれらが区別されている方が好ましいが，本研究では区別せず，すべて地名として抽出する．ある人物の役職名や地名は時間の経過にしたがって変化する場合があるため，一人に対し複数の役職名や地名が抽出されることがある．したがって，人物の役職名と地名の抽出結果は，〈人名，役職名〉，〈人名，地名〉のいずれかの組の列として表現される．書名についても，一人が複数の書籍を書く場合があるため，人物が書いた書籍の抽出結果は，〈人名，書名〉の組の列として表現される．

人名のような固有表現を抽出する方法については，大きく分けてルールベースの手法 (Rau 1991) と機械学習に基づく手法 (Asahara and Matsumoto 2003; McCallum and Li 2003) の 2 種類が提案されている．機械学習に基づく手法は，学習のための正解データが必要となる．史料には人名の索引が掲載されているが，科学者のみが掲載の対象となっており政治家の人名は含まれない．また，史料の文中には姓や名のみが出現する場合がある．索引の人名は姓名（フルネーム）のみであるため，姓や名を人名として抽出するのは困難である．したがって，史料の正解データを作成するのには膨大な時間を必要とする．IREX(Sekine and Isahara 2000) の公開データなどの人名タグ付きコーパスを正解データとして利用する方法もある．しかし，上記のコーパスは 1994 年から 1995 年の新聞記事を対象としている．一方，史料中の人名は主に明治

よりも前の時代のものであり，人名を構成する文字や形態素が大きく異なるため，高い抽出精度の実現を期待できない．そこで，本研究では，ルールベースの手法により人名を抽出する．

また，人物に対する役職名のような関係を抽出する情報抽出についても，ルールベースの手法と機械学習に基づく手法 (Sudo, Sekine, and Grishman 2003; Greenwood, Stevenson, Guo, Harkema, and Roberts 2005) の 2 種類があるが，上記の固有表現抽出と同様の理由で，ルールベースの手法により人物の属性や業績を抽出する．

人名の抽出手順としては，人手で作成した形態素列の抽出パターンを利用したパターンマッチングによって人名を抽出したあと，大域的情報を利用してさらなる人名の抽出と名寄せを行う．また，人物の属性と業績もパターンマッチングによって抽出する．これらの手法について以下に説明する．

3.1 形態素列のパターンマッチングによる情報抽出

図 3 に示す手順で形態素列のパターンマッチングによる情報抽出を実行する．まず，史料中の文に対して形態素解析を実施する．その結果に対して，各形態素の出現形，基本形，品詞，字種などの情報を用いて，正規表現に似た形式で形態素列を表現したパターンにマッチする形態素列を抽出する．このとき，形態素列のパターンがパターンマッチング処理に埋め込まれていると，パターンの修正にともなうパターンマッチング処理の修正に時間がかかる．そこで，形態素列のパターンとパターンマッチング処理を分離し，パターンのみを修正すれば，それに応じたパターンマッチング処理が実行できるようにした．たとえば，図 3 の抽出パターンの 1 行目は，「名詞—固有名詞—地域」の品詞をもつ形態素と基本形が「都」「府」「県」のいずれかの形態素からなる形態素列（たとえば「岩手県」），あるいは基本形が「北海道」である形態素の

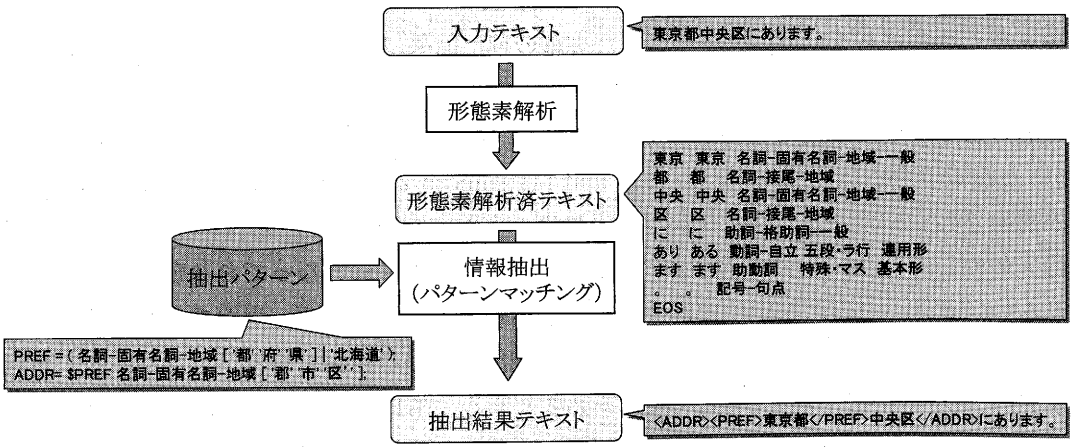

図 3 形態素列のパターンマッチングによる情報抽出手順

いずれかにマッチし，PREF というタグを付与することを意味する．

人物の属性や業績のように，固有表現どうしの関係を抽出する際は，固有表現（人名，役職名，地名，書名）を抽出するパターンとそれらをまとめあげるパターンを作成し，固有表現を抽出した後でそれらをまとめあげるという二段階の処理を行う．

パターンマッチングによる情報抽出については，情報抽出に関する評価型会議である MUC (Grishman and Sundheim 1996) をはじめとして様々な手法が提案されている．たとえば，(西野，落谷 1998) では本稿と同様に人名とその職業名を抽出しており，その際，職業名のリストや人名の直後に出現する「さん」といった語句を手がかり句を利用している．われわれの抽出パターンでは，このような手がかり句に加えて，人名などの固有表現を構成する形態素の特徴も利用している．固有表現を抽出するパターンに利用した主な特徴を表 2 に示す．なお，「パターンの例」は実際のパターンを分かりやすいように一部書き換えてある．たとえば，表 2 の上から 2 番目のパターンの例では，構成する形態素の字種を利用することにより，「ウィリアム・アダムズ」のようなカタカナを含む人名を抽出できる．また，『明治前日本科学史』には古い時代の人名が数多く出現し，形態素解析誤りが頻繁に発生する．たとえば，「桂川甫周」を形態素解析すると，「桂川」は品詞が「人名—姓」の形態素となるが，「甫」と「周」はそれぞれ別の形態素に分割されてしまう．このような場合，表 2 の上から 3 番目のパターンの例のように，形態素の文字数を指定することにより，「桂川甫周」を人名として正しく抽出できる．

固有表現をまとめあげるパターンについては，固有表現と固有表現の間やその前後に出現する形態素の特徴をもとにパターンを作成した．人名と地名をまとめあげるパターンとそれにマッチする形態素列の例を表 3 に示す．パターンの例において，$PERSON, $PLACE はそれぞれすでに抽出された人名，地名の形態素列を表す．

固有表現を抽出するパターンとそれらをまとめあげるパターンをあわせ，全部で約 90 個の抽出パターンを作成した．

表 2　形態素列のパターンに利用した主な特徴

抽出対象	特徴	パターンの例
人名	構成する形態素の品詞	<品詞：人名-姓> <品詞：人名-名>
人名	構成する形態素の字種	<字種：カタカナ> "・" <字種：カタカナ>
人名	構成する形態素の文字数	<品詞：人名-姓> <文字数：1> <文字数：1>
人名	直前に出現する語句	直前 = "和算家"
人名	直後に出現する語句	直後 = "神父"
役職	構成する語句そのもの	"和算家" OR "蘭医" OR "藩主"
地名	構成する形態素の品詞	<品詞：地域>
地名	直後に出現する語句	直後 = "藩"
書名	前後に出現する語句	直前 = "『" AND 直後 = "』"

3.2　大域的な情報を利用した情報抽出と名寄せ

人手で作成した形態素列のパターンを利用したパターンマッチングによって人名を抽出する場合，すべての人名を抽出するためのパターンを網羅的に記述することは困難であり，多くの抽出漏れが発生してしまう．そこで，パターンマッチングによる抽出結果に対し，大域的な情報を利用して抽出漏れを削減する．

史料中に出現する人名は，初出時は姓名（フルネーム）で出現し，その後方で同一人物を表す人名が姓名，姓，名のいずれかの形で出現することが多い．そこで，形態素列のパターンマッチングによる人名の抽出結果に対し，さらに追加で人名の候補となる形態素列をパターンマッチングにより抽出し，その中で人名となる形態素を判定する．人名の候補となる形態素列として，漢字のみで構成される形態素が連続する形態素列，およびカタカナのみで構成される形態素列と「・」（ナカグロ）が交互に出現する形態素列を抽出する．抽出された人名の候補のうち，以下のいずれかの条件を満たすものを人名と判定する．

条件 a　任意の箇所に出現する人名と同じ文字列の形態素列

条件 b　ある箇所に出現する人名よりも後ろに出現し，かつその人名の先頭または末尾の部分文字列となっている形態素列

条件 b に合致する形態素については，その前に出現する（フルネームの）人名と同一人物であるという判定（名寄せ）も同時に行う．

表 3　人名と地名をまとめあげるパターンの例

パターンの例	パターンにマッチする形態素列の例
$PLACE "人" $PERSON	ポルトガル人マノエル・ゴンサロ
$PERSON "は" $PLACE "人で"	シドッティはイタリア人で
$PLACE "の" $PERSON	薩摩の島津重豪
$PERSON "は" $PLACE "の人"	橋本宗吉は大坂の人
$PLACE "侯" $PERSON	越前侯松平慶永

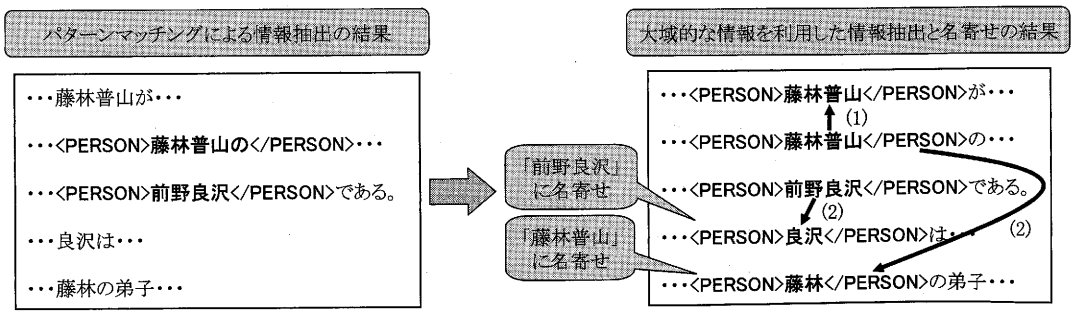

図 4　大域的な情報を利用した情報抽出と名寄せの例

大域的な情報を利用した情報抽出と名寄せの例を図4に示す．条件aにより，一番上の「藤林普山」が人名と判定される．また条件bにより，「良沢」と「藤林」が人名と判定され，さらに「良沢」は「前野良沢」と，「藤林」は「藤林普山」とそれぞれ同一人物であると判定される．

4 評価実験

史料からの人名，人物の属性，人物の業績の抽出精度を評価する実験を行った．

4.1 実験条件

評価用データとして『明治前日本科学史総説・年表』の本文（5096文，約21万8千字）を使用し，人名，人物の属性（人名とそれに対する役職名と地名），人物の業績（人名とそれに対する書名）の抽出精度を評価した．評価のための正解作成にあたっては，システムが抽出した誤りや漏れを人手で修正することにより，すべての情報を人手で抽出するのと比べて短時間で正解を作成することができた．

人名については，パターンマッチングのみを行った場合と，パターンマッチングのあと大域的な情報を使って抽出漏れを削減した場合の精度を評価した．また，現代の文書を対象に機械学習を行う手法として，CaboCha[1]の固有表現抽出機能において，CaboChaに付属する，毎日新聞記事のタグ付きコーパスで学習したモデルを使って人名を抽出した場合との精度を比較した[2]．パターンマッチング，CaboChaとも形態素解析にはChaSen[3]を利用した．

人物の属性と人物の業績については，パターンマッチングのみを行った場合の精度を評価した．人物の属性については，人名とそれに対する役職名として〈人名，役職名〉の組を，人名とそれに対する地名として〈人名，地名〉の組を抽出し，その抽出精度を評価した．また人物の業績については，人名とそれに対する書名として〈人名，書名〉の組の抽出精度を評価した．

評価尺度として，以下の式で算出される再現率，適合率，F値を測定した．

$$\text{再現率 } R = \frac{\text{一致件数}}{\text{正解件数}}, \quad \text{適合率 } P = \frac{\text{一致件数}}{\text{出力件数}}, \quad F\text{値} = \frac{2PR}{P+R}$$

上記の件数の算出にあたっては，人名については出現箇所を区別し，人物の属性と人物の業績については出現箇所を区別せず算出した．たとえば人名の正解については，同じ人名が複数の箇所に出現する場合はそれぞれ別のものとして正解件数をカウントした．また，出力された人名を正解と比較し，同じ人名が同じ箇所に出現する場合にのみ一致するとして，一致件数を算出した．逆に人物の業績については，同じ〈人名，書名〉の組が複数の箇所から抽出された場

[1] http://chasen.org/~taku/software/cabocha/
[2] CaboChaの固有表現抽出機能は地名も抽出可能だが，今回は比較対象としなかった．今後の課題としたい．
[3] http://chasen.naist.jp/hiki/ChaSen/

合，それらをまとめて1とカウントした．

4.2 実験結果

人名の抽出結果について，パターンマッチングのみを行った場合，パターンマッチングのあと大域的な情報を利用した場合，およびCaboChaを使った場合の結果を表4に示す．もっとも精度の高いものを太文字の数字で表している．また，役職名，地名，書名の抽出結果について，パターンマッチングのみを行った場合の結果を表5に示す．

4.3 考察

人名の抽出精度については，パターンマッチングのあと大域的な情報を使った場合がもっとも高い精度であった．パターンマッチングのみの場合の精度と比較すると，適合率が若干下がったものの，再現率が大幅に向上しており，大域的な情報を利用することによって抽出漏れを削減できたことが分かる．

CaboChaを使った場合のF値は0.702であった．現代の日本語の文書を対象として機械学習を行った場合の人名の抽出精度は0.87前後と報告されている (Asahara and Matsumoto 2003)．本実験の抽出対象文書である史料に出現する人名は，現代の人名と比べて人名を構成する文字や形態素が大きく異なることが，CaboChaを使った場合の抽出精度が低かった原因と考える．

役職名，地名，書名の抽出精度はいずれも，適合率に比べて再現率が低かった．役職名，地名，書名それぞれの抽出漏れの例を表6, 7, 8に示す．「抽出漏れの箇所」欄において，太字部分が抽出された人名を表し，[]（カギ括弧）で囲まれた語句が抽出できなかった（つまり抽出漏れの）役職名，地名，書名を表している．「解決方法」欄には抽出するための方法の案を示した．役職名，地名，書名とも，形態素列の抽出パターンの追加で抽出可能となる抽出漏れだけでなく，係り受け解析や照応解析など，形態素解析以外の自然言語処理が必要とされる抽出漏れも

表4 人名の抽出結果の比較

手法	再現率	漏れ件数	正解件数	適合率	誤り件数	出力件数	F値
パターンマッチング	0.616	806	2101	**0.864**	204	1499	0.719
パターンマッチング+大域的情報	**0.790**	433	2101	0.843	310	1969	**0.815**
CaboCha	0.668	697	2101	0.739	496	1900	0.702

表5 役職名，地名，書名の抽出結果（パターンマッチングのみ）

抽出対象	再現率	漏れ件数	正解件数	適合率	誤り件数	出力件数	F値
役職名	0.703	44	148	0.981	2	106	0.819
地名	0.772	28	123	0.960	4	99	0.856
書名	0.492	99	195	0.980	2	98	0.655

表 6　役職名の抽出漏れの例

抽出漏れの箇所	解決方法
アルメイダはポルトガルの［商人］で	形態素列の抽出パターンの追加
ドイツ人**アンドリース・クライエル**はオランダの［商館長］として	形態素列の抽出パターンの追加
京都の［製薬店主］**遠藤元理**の	形態素列の抽出パターンの追加
有馬頼は山路主住について学んだ［数学者］であるが、久留米の［藩主］であったからであろうか、	数学者：係り受け解析の利用 藩主：照応解析の利用（主語の省略）
［写真師］という職業が現われるのは一八六〇年代の初めで、横浜で**下岡蓮杖**、長崎で**上野彦馬**がこれを始めた。	照応解析の利用（「これ」の先行詞）

表 7　地名の抽出漏れの例

抽出漏れの箇所	解決方法
吉田光由は［京都］の角倉家の出で、	形態素列の抽出パターンの追加
［佐賀藩］では藩士**杉谷雍介**が	形態素列の抽出パターンの追加
［伊豆の韮山］の医師**矢田部卿雲**の ※「韮山」の品詞の解析誤り	形態素辞書の追加
［アメリカ］人**ウィリアム・ブレーキ**および**パンペリー**を ※パンペリーがアメリカ人であることの抽出漏れ	係り受け解析の利用（並列情報）
小林義信は字を**謙貞**といい［長崎］の人である。	係り受け解析の利用

表 8　書名の抽出漏れの例

抽出漏れの箇所	解決方法
筑前の人**河野禎造**は［『舎密便覧』］（一八五六）を書いた	形態素列の抽出パターンの追加
吉雄常三に［『西説観象経』］（一八二二）および［『遠西観象図説』］（一八二三）の著がある。	形態素列の抽出パターンの追加
松永良弼は関流の円理を整理して［『円理乾坤之巻』］を書いた	係り受け解析の利用
深根輔仁が醍醐天皇の勅命によって撰した［『本草和名』］二十巻を	係り受け解析の利用
ネットーはフライベルクの鉱山学校に学んだ人で、…（中略）…。彼は一八七九年［『日本鉱山編』］を著し、	照応解析（「彼」の先行詞）

あることが分かる．たとえば係り受け解析については，パターン中に形態素間の係り受け関係も記述できるようにし，形態素列のパターンにマッチする形態素列を求めたあと，それらが係り受け関係にあるものを求めるという方法がある．表 8 の上から 3 番目の例の場合，「<人名>"は"」と「<書名> "を書いた"」という形態素列のパターンにマッチする形態素列として「松永良弼は」と「『円理乾坤之巻』を書いた」をそれぞれ抽出したあと，前者の形態素列「松永良弼は」が後者の形態素列の先頭の文節「『円理乾坤之巻』を」に係ることから，松永良弼の著書と

して『円理乾坤之巻』を抽出できる．

役職名，地名，書名それぞれの抽出精度を比較すると，書名の再現率が役職名，地名に比べて低かった．それぞれの抽出漏れを分析したところ，役職名，地名と比較して，書名は人名と離れた位置に出現する場合が多いことが分かった．このような抽出漏れは，形態素列のパターンでは正しく抽出することができず，係り受け解析や照応解析が必要となるものが多いことが，書名の再現率が低かった原因だと考える．

5 おわりに

歴史資料を対象として歴史オントロジーを構築するシステムを開発するための第一歩として，『明治前日本科学史』の一部を電子化し，史料中の科学技術者に関する属性および業績の情報を抽出することにより，前近代日本の人物情報データベースの構築を試みた．人名とそれに対する役職名，地名，書名をルールベースの手法によって抽出する方法を提案し，『明治前日本科学史総説・年表』を対象とした精度評価を行った結果，人名，人名とそれに対する役職名，人名とそれに対する地名についてはF値で0.8を超える結果が得られた．

今後の課題としては，抽出精度を向上させるために，機械学習によって情報抽出を行うこと，係り受け解析や照応解析の結果を形態素列の抽出パターンや機械学習に利用することを考えている．また，人物の属性や業績として抽出する情報を拡大するとともに，抽出対象データについても『明治前日本科学史総説・年表』以外の巻を対象とした抽出と評価を行う予定である．

参考文献

Asahara, M. and Matsumoto, Y. (2003). "Japanese Named Entity Extraction with Redundant Morphological Analysis." In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, pp. 8–15.

Greenwood, M. A., Stevenson, M., Guo, Y., Harkema, H., and Roberts, A. (2005). "Automatically Acquiring a Linguistically Motivated Genic Interaction Extraction System." In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, pp. 46–52.

Grishman, R. and Sundheim, B. (1996). "Message Understanding Conference-6: A Brief History." In *Proceedings of the 16th Conference on Computational Linguistics (COLING-96)*, pp. 466–471.

IEEE. "IEEE P1600.1 Standard Upper Ontology Working Group (SUO WG).", `http://suo.ieee.org/`.

McCallum, A. and Li, W. (2003). "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons." In *Proceedings of The Seventh Conference on Natural Language Learning (CoNLL-2003)*, **4**, pp. 188–191.

Mizoguchi, R. and Ikeda, M. (1997). "Towards Ontology Engineering." In *Proceedings of Joint Pacific Asian Conference on Expert Systems/Singapore International Conference on Intelligent Systems (PACES/SPICIS '97)*, pp. 259–266.

西野文人，落谷亮 (1998). "新聞記事からの人物・企業情報の抽出." 情報処理学会自然言語処理研究会 (NL-127), pp. 125–132.

Rau, L. F. (1991). "Extracting Company Names from Text." In *Proceedings of Seventh IEEE Conference on Artificial Intelligence Applications*, pp. 29–32.

Sekine, S. and Isahara, H. (2000). "IREX: IR and IE Evaluation Project in Japanese." In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, pp. 1475–1480.

Sudo, K., Sekine, S., and Grishman, R. (2003). "An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition." In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 224–231.

東京大学史料編纂所. "東京大学史料編纂所データベース SHIPS.", `http://www.hi.u-tokyo.ac.jp/ships/`.

略歴

石川　徹也：1971 年慶応義塾大学大学院修士課程修了．富士フイルム（株）足柄研究所，図書館短期大学，図書館情報大学，文部省在外研究員 (UCLA, IU)，筑波大学等を経て，現在，東京大学史料編纂所前近代日本史情報国際センター特任教授（研究開発主査）．歴史知識学の創成研究に従事．工学博士（早稲田大学），筑波大学名誉教授，筑波大学大学院図書館情報メディア研究科共同研究員．情報文化学会学会賞（2005 年 8 月），言語処理学会優秀発表賞（2006 年 3 月），Eügen Wuster Special Prize（UNESCO, 2006 年 7 月）．

北内　　啓：1998 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年，NTT データ通信（株）入社．現在，（株）NTT データ技術開発本部において自然言語処理の研究開発に従事．言語処理学会，情報処理学会各会員．

城塚　音也：1988 年東京大学文学部言語学科卒業．同年日本電信電話株式会社入社，音声言語，知識処理技術の研究開発に従事．現在，（株）NTT データ技術開発本部において自然言語処理の研究開発を担当．

(2007 年 10 月 10 日 受付)
(2008 年 1 月 25 日 再受付)
(2008 年 3 月 18 日 採録)