『明治前日本科学史』を対象とする歴史知識の構造化検索・参照システムの構築研究

石川 徹也＊　伊藤 直之＊＊　松本征二＊＊　新堀英二＊＊
＊東京大学 史料編纂所 前近代日本史情報国際センター
＊＊大日本印刷株式会社 情報コミュニケーション研究開発センター

本研究では、デジタル化した史料から索引語検索機能および他史料参照機能を持つ検索・参照システムを実現することを目的に、史料を対象とする索引語の自動抽出手法を構築し、既存手法との比較評価により、本手法の有効性を確認した。また、抽出した索引語を利用して、史料を対象とする検索機能および他史料を参照する機能を備えた『明治前日本科学史』検索・参照プロトタイプシステムを構築した。

Constructing Knowledge-based Historical data Retrieval and Reference System Using "*History of Science and Technology of Japan before Meiji Period (pre-modern)*"

Tetsuya Ishikawa*　Naoyuki Ito**　Seiji Matsumoto**　Eiji Shinbori**
*Historiographical Institute The University of Tokyo
**Dai Nippon Printing Co., Ltd. Media Research Center

In this paper, we propose a indexing method for historical sources (documents) to construct historical data retrieval and reference system. Result of experiment, our indexing method shows effectiveness more than other methods: TermExtract system and likelihood method. Based on the result, we developed retrieval and reference system as a prototype system. User can find historical data related science and technology in "*History of Science and Technology of Japan before Meiji Period (pre-modern)*" and also look up other historical sources.

1．はじめに

本研究には２つの目的がある。第１の目的は、昭和 15 年に時の帝国学士院において編纂が開始された『明治前日本科学史』（全 28 巻）[1]を、日本学士院の許諾の下にデジタル化し、その内容の検索および、利用者による検索結果の理解を支援する目的のため、他の多様な史資料を参照できるシステムを構築することである。第２の目的は、第１の目的の遂行を通して、歴史知識の構造化を目的とする史料の全文検索システムの機能を明確にすることである。

これらの目的のために３段階に分けて研究を推進している。初めに、確度の高い検索機能を保証するために『明治前日本科学史』のうち２刊本の全文をデジタル化し、そのデジタルデータを対象に、自動索引語抽出機能の研究を行う。次に、史料においては写真、図表の掲載が多であることから、全巻のデジタル化を目標に、それらテキスト情報以外の史料情報の最適なデジタル化形態を探る。以上が完成した段階で、『明治前日本科学史』全巻のデジタル化を図り、検索・参照システムとして提供する。

以上を踏まえ本報告では、歴史知識の構造化を目的する検索・参照システムの構想を 2．で示し、3．で史料を対象とする索引語自動抽出手法について述べ、4．で構築した『明治前日本科学史』検索・参照プロトタイプシステムを紹介し、5．で完成に向け取り組むべき研究課題を明確にする。

2．歴史知識の構造化 検索・参照システム

2.1 従来のシステム

現存する史料の情報を提供する代表的なシステムとして東京大学史料編纂所の「史料情報データベース検索システム」があげられる。本システムは、史料の記述文字を翻刻し、その内容の解読結果（例：綱文）とともに史料に関するデータを提供することで広く利用に供している[2]。

史料の記述文字の翻刻結果および、その結果に基づく史料の内容の解読結果（例：綱文）は、歴史学者の知識の下に作成されている。この編纂作業は専門家とはいえ、人間の知識に基づく方式で

あることから精度およびコストの面で限界がある。この問題を解消する方式を探るのが歴史知識学の創成であるといえる。具体的には史料内容を自動的に解読し、再生するシステムの実現にある。歴史知識学の領域は史料の記述文字を、崩し字データを下に自動的に翻刻することから始まり、史料内容の自動解読、その結果の自動生成（例：綱文の自動生成）に至る。このためには例えば「崩し字データベース」に代表される多種多様な歴史関係知識データベース(Ontology)が必要となる。しかし、これらの完成には相当な時間がかかろう。

2.2 歴史知識の構造化を目的とする検索・参照システム

本研究では歴史知識学の創成を目指し、『明治前日本科学史』を対象に図１に示すように、全文検索システムおよび、検索結果理解支援を目的とする他史資料参照システムの構築を目指す。

本研究で対象とする『明治前日本科学史』は、各技術分野の概要を総説した『明治前日本科学史総説・年表』と、天文学・数学・薬物学など全 18 の技術分野に関して、日本における各分野の発展が詳細に記述された各巻からなる。本研究では、『明治前日本科学史 総説・年表』と『明治前日本天文学史』の２刊本を XML 形式にてデジタル化し、索引語抽出手法の構築および検索・参照システムのプロトタイプシステムの構築を行った。

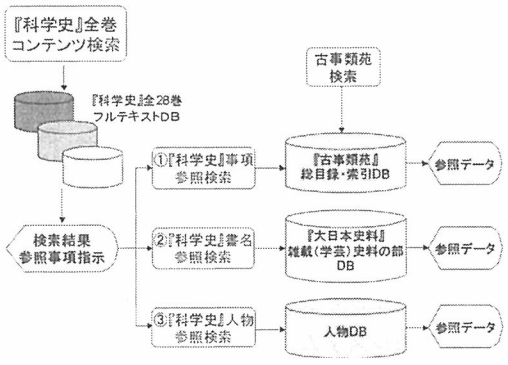

図１『明治前日本科学史』検索・参照システム 構想

2.2.1 検索対象および検索結果の出力内容

本研究で構想している史料検索・参照システムの概要を示す。

① 検索対象

検索の対象は「科学技術事項」、「科学技術者（人名）」、「著作物（書名）」の 3 事項であり、これらを検索要求として利用者は史料を検索することが可能である。

② 検索結果の出力内容

検索要求の期待、すなわち検索結果として求められる内容は、検索対象に係わる“解題”(歴史的事象説明)である。この解題は対象史料の特性から、分野別(各巻)において中心的に解題されている。しかし、多くの事項は他の解題において引用・参照されている。通常の全文検索システムでは、検索対象が索引語として認識されている場合は、全てが検索結果として表示されるが、本システムおいては検索対象史料の特性から、『明治前日本科学史 総説・年表』編(巻)の検索結果に出現する索引語のみを他巻の索引語にハイパーリンクし、他巻における出現箇所を出力する。この結果、『明治前日本科学史 総説・年表』編(巻)で概要を把握し、他巻において詳細を把握できる利点がある。ただし、全巻を対象とする各巻独立の検索は、別途用意している。

2.2.2 検索結果理解支援「参照」の対象

検索結果の“解題”内容を直感的に理解できるとは限らない。そこで、検索結果内に出現する歴史的事項、史料、人物に関して、その対象を指示することで他の史資料における該当事項を参照する機能を設定した。歴史的事項に関しては“明治前を対象とする史料準拠の百科事典”である『古事類苑』[3]を、史料内容の参照のために『大日本史料』[4]の「雑載(学芸)史料の部」を、人物に関しては本所構築の『個人史 DB』を検索参照できるようにした。ただし、『古事類苑』については『総目録・索引』編のみを検索参照としていることから本文確認は冊子に当たることになる。

以上のことを踏まえ、本研究では、歴史知識の構造化を目的とする検索・参照システムを目指し、プロトタイプシステムを構築した。このシステムについては４．において詳細を述べる。

3. 自動索引システムの設計

叢書とはいえ、当該検索対象史料は我が国の科学史の編纂物として唯一の刊本であることから、索引語の自動設定のための学習データとなるテキストデータは存在しない。そのため、機械学習で得た学習データを利用して索引語を自動抽出することは困難である。

そのため、本研究では機械学習等のように抽出対象のテキストと類似するテキスト（学習デー

タ）を用いずに、史料データからできるだけ多くの索引語を自動的に抽出する手法として、品詞接続規則による複合語生成手法を検討し構築した。構築した品詞接続規則により複合語を生成する方式の索引語抽出性能の評価のために、比較対象として既存の用語抽出手法である TermExtract 方式[5]と、likelihood 方式[6][7]を選定し比較評価実験を行った。評価実験では、上記の３方式により、『明治前日本科学史 総説・年表』編を対象に索引語の自動抽出を行い、抽出された索引語について、歴史学専攻のポストドクターにより、索引語としての妥当性の判定を実施することで３方式の精度比較を行った。

3.1 品詞接続規則による複合語生成

本研究では、学習データもしくは抽出パターンをあらかじめ作成せず、自動的に索引語を抽出する手法として、品詞接続規則により複合語を生成する方式を構築し、『明治前日本科学史 総説・年表』編から索引語の自動抽出を行った。

本手法では、対象のテキストデータに対して形態素解析システム『茶筌』[8]により形態素解析処理を行い、分割された各形態素に付与される品詞情報を用いて複合語を生成する。形態素解析し、名詞と判定された形態素が連続で出現した箇所について、茶筌の出力する品詞下位分類(IPA 品詞体系[9])に基づき、経験則により設定した連結ルール（表１）を用いて語の連結・非連結を決定し、複合語を生成する（図２）。図２の例では、対象テキスト中の「関孝和著」という文字列が、形態素解析処理によって、「関」「孝和」「著」の３つの形態素に分割され、それぞれの形態素に品詞情報が付与されることを示している。この結果に対して、表１の名詞連結ルールを基に、「関孝和」が複合語として生成され、索引語として用いられる。ただし、形態素解析によって未知語と判定された語については、無条件に連結するものとした。本手法では、上記の処理の結果、生成された複合語をすべて索引語とした。

上記の索引語抽出手法を『明治前日本科学史総説・年表』全文に適用した結果、10,804 語の索引語が抽出された。

3.2 索引語抽出手法の評価

品詞接続規則による複合語生成に基づく索引語抽出手法の性能評価のため、比較対象として専門用語抽出ツール TermExtract、likelihood 方式を選定し比較評価を行った。評価では、『明治前日

表 1 名詞連結ルール（抜粋）（連結:○ 非連結:×）

	名詞区分	例
○	名詞-一般	
○	名詞-固有名詞-一般	
○	名詞-固有名詞-人名-一般	「お市の方」
○	名詞-固有名詞-人名-姓	「伊藤」「山田」
○	名詞-固有名詞-人名-名	「太郎」「花子」
○	名詞-固有名詞-組織	「経済産業省」「NHK」
○	名詞-固有名詞-地域-一般	「アジア」「トリノ」「京都」
○	名詞-固有名詞-地域-国	「日本」「オーストラリア」
×	名詞-代名詞-一般	「それ」「ここ」「あいつ」
×	名詞-代名詞-縮約	「ありゃ」「こりゃ」「こりゃあ」
×	名詞-副詞可能	「金曜」「一月」「午後」
○	名詞-サ変接続	「インプット」「愛着」「悪化」
○	名詞-形容動詞語幹	「健康」「安易」「駄目」
×	名詞-ナイ形容詞語幹	「申し訳」「仕方」「とんでも」
×	名詞-数	「0」「1」「2」「何」
×	名詞-非自立-一般	「あかつき」「暁」「かい」
×	名詞-非自立-副詞可能	「あいだ」「間」「あげく」
×	名詞-非自立-助動詞語幹	「よう」「やう」「様(よう)」
×	名詞-非自立-形容動詞語幹	「みたい」「ふう」
×	名詞-特殊-助動詞語幹	「そう」
○	名詞-接尾-一般	「おき」「かた」「方」
×	名詞-接尾-人名	「君」「様」「著」
○	名詞-接尾-地域	「町」「市」「県」
○	名詞-接尾-サ変接続	「化」「視」「分け」
×	名詞-接尾-助動詞語幹	「そう」
○	名詞-接尾-形容動詞語幹	「的」「げ」「がち」
×	名詞-接尾-副詞可能	「後(ご)」「以後」
○	名詞-接尾-助数詞	「個」「つ」「本」
○	名詞-接尾-特殊	「(楽し)さ」「(考え)方」
×	名詞-接続詞的	「(日本)対(アメリカ)」
×	名詞-動詞非自立的	「ごらん」「ご覧」

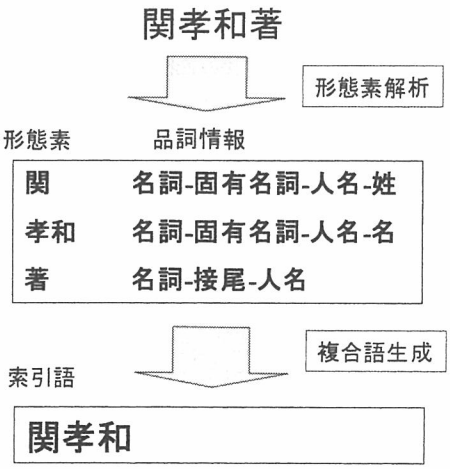

図 2 品詞情報による複合語生成

本科学史 総説・年表』編を対象に索引語の自動抽出を行い、抽出された索引語について、歴史学専攻のポストドクター２名による索引語としての妥当性の判定を実施し 3 方式の精度比較を行った。

TermExtract は、形態素解析による単語分割、複合語の生成、統計量による各複合語の重要度の計算、という３つのステップを踏み、テキスト中から専門用語を抽出する。likelihood 方式は、形態素解析による単語分割の後、名詞 2-gram（隣り合って出現する 2 つの名詞の組）について尤度比検定を行うことで、各名詞 2-gram の相関スコア（correlation）を算出し、算出されたスコアの大小により名詞の連結・非連結を決定する。比較対象とした上記の２手法は、ともにテキスト中での単語の出現頻度に基づく統計値による抽出方式である。

① 索引語抽出精度の比較評価

各手法により抽出された語の、索引語としての妥当性判定を実施した結果を表２に示す。表２における正解語数とは、各抽出手法による抽出語のうち、索引語として妥当であると判定された語の数であり、正解率は抽出語数における正解語数の割合を示している。

実験により、本研究の手法は likelihood 方式に比べ、正解語数、正解率ともに上回ることが分かった。また、TermExtract 手法と比較すると、正解率では劣るが、抽出された正解語数では上回っていることが分かった。

表 2 各抽出手法の正解率比較

	本研究手法	TermExtract	likelihood方式
抽出語数	10,804	8,867	9,715
正解語数	6,757	6,036	5,571
正解率	62.54%	68.07%	57.34%

② 索引語種の比較評価

次に、正解率の高かった本研究手法（品詞接続規則による複合語生成）および TermExtract について、より詳細な評価を行った。抽出語のうち索引語として妥当であると判定された語（正解語）に対して、索引語種別を人手により割り当てた。設定した索引語種別を表３に示す。

本研究手法と TermExtract の２手法により抽出された正解語中の索引語種別の内訳を調べた結果を表４に示す。

表 4 正解語の索引語種別内訳

	本研究手法	TermExtract
正解語数	6,757	6,036
① 科学用語	2,962	2,908
② 歴史用語	529	526
③ 人名	1,004	997
④ 書名	459	431
⑤ 地名	429	441
⑥ 年代	652	78
⑦ 正語	722	655

本研究手法は TermExtract と比較して、「⑤地名」以外の索引語種別において、より多くの正解索引語を抽出できたことが確認できた。

以上の結果から、本研究で想定している検索・参照システムによる索引語検索・参照という用途においては、索引語の網羅性が最重要であるとの考えから、本研究の検索・参照システムに用いる索引語の抽出には、本研究にて構築した「品詞接続規則により複合語を生成する方式」を採用することとした。

表 3 索引語種別

索引語種別	説明	抽出例
① 科学用語	科学技術の専門用語	ニュートン物理学　宣明暦　鎌倉彫
② 歴史用語	歴史用語	遣唐使　大宝律令　和銅改元
③ 人名	個人名、天皇	伊能忠敬　大久保石見守　醍醐天皇
④ 書名	著書、	阿蘭陀本草和解　魯西亜志略　垂仁記
⑤ 地名	国名、地方名、建物名	伊予国　多田銀山　アラブ圏
⑥ 年代	年号、西暦年	明治三年　慶長年間　一八五三年
⑦ 正語	①～⑥以外の正語	外人教師　ロシア語　輸出品

４．『明治前日本科学史』検索・参照システム

３．で示した索引語抽出手法により抽出された索引語を用いて、『明治前日本科学史』を対象とする検索・参照システムをウェブアプリケーションにてプロトタイプシステムとして実装した。

プロトタイプシステムには２．で示した索引語検索機能および他史資料参照機能を実装した。現時点では、『明治前日本科学史 総説・年表』と『明治前日本天文学史』のフルテキストを検索でき、他史料として『古事類苑』の『総目録・索引』編および『大日本史料』の「雑載(学芸)史料の部」が参照可能である。

索引語検索機能では、「科学技術事項」、「科学技術者（人名）」、「著作物（書名）」の３つの索引語種別を検索することができる。利用者は検索したい索引語種別をシステムに指定できる。検索が実行されると、検索要求にマッチした索引語のリストとともに、各索引語に対して史料内の出現位置が提示され、同時に提示されるハイパーリンクを選択することで本文を参照することができる（図３）。本文閲覧画面では、テキストデータとしてデジタル化された史料を閲覧することができる。また、本文閲覧画面では、テキスト中の各索引語箇所にハイパーリンクが挿入されており、利用者はハイパーリンクを選択することで、索引語の『明治前日本科学史』他巻における出現箇所および他史料（『古事類苑』、『大日本史料』）における出現箇所を参照できる（図４）。

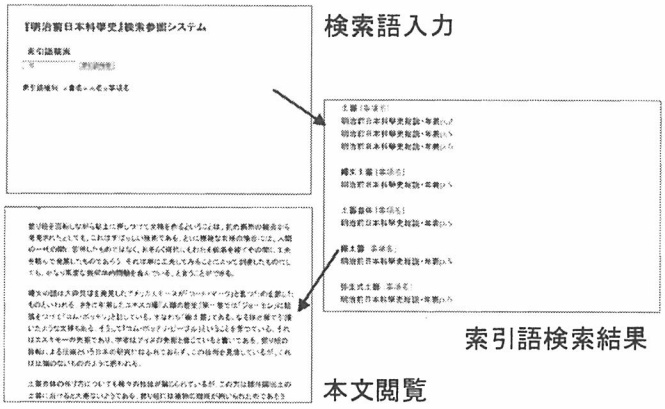

図３ 『明治前日本科学史』検索・参照システム：検索機能

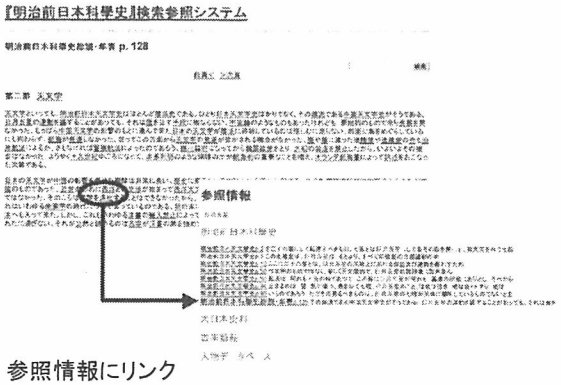

図４ 『明治前日本科学史』検索・参照システム：参照機能

5．あとがき

1）結論

“わが国のモノ作り”の原点を知るには『明治前日本科学史』を紐解くのが有効である。当該史料はこれまで刊本のみであり、今は絶版となっている。この度、日本学士院の理解の下に全巻デジタル化の許諾をいただき、全文を対象にデジタル化研究ができるようになった。そこで、歴史知識の構造化を目的とする本文検索と検索結果理解支援を目的に他史資料を参照する検索・参照システムを構築する研究を開始し、現時点で、『明治前日本科学史 総説・年表』編および『明治前日本天文学史』編のデジタル化を行い、基本となるプロトタイプシステムを構築し、システムの有効性等の検証研究を進めている。本論文においては研究の構想を含めた全体像を提案し、同時に、本研究で検討・構築した自動索引機能について手法の内容と評価結果を示した。

2）今後の研究

システムの完成および精度向上を目指しての今後の研究として、以下の点がある。

① 「科学技術事項」、「科学技術者」、「著作物」の 3 事項の索引抽出に関して、歴史・科学技術対象事項に限る判定研究。

② 上記 3 事項の解題以外の引用・参照出現における重要度判定の研究。このことは『総説・年表』を対象とする検索結果から各巻へのリンクのための分野判定にも関わる。

③ テキスト以外の写真・絵図のデジタル化方式の検討、それらのデータの検索・参照システムでの効果的な出力方法の研究。

3）検索機能の拡張

① 歴史用語からの検索に限らず、現代の科学技術用語から検索できることは重要な機能であると考えられる。そこで学術用語検索システムである『オンライン学術用語集』[10]から本研究で構築したシステムの検索を計画している。

②現状では、『古事類苑』については『総目録・索引』編のみを検索参照としている。システム上にて 『古事類苑』の対象全文出力が可能になると利便性は飛躍的に向上すると考えられる。

謝辞

当研究には多くの組織のご理解・ご支援をいただいている。ここに記し感謝を申し上げる。『明治前日本科学史』全 28 巻のデジタル化に許諾をくださった「日本学士院」、この研究の機会をくださった「史料編纂所」（東京大学）の前所長保立道久教授および現所長の横山伊徳教授、そして産学連携研究の推進を支援くださっている「情報コミュニケーション研究開発センター」（大日本印刷株式会社）前センター長斎藤雅氏および現センター長中川清貴氏に感謝する。

参考文献

[1] 日本科学史刊行会編、『明治前日本科学史刊行図書』、日本学術振興会発行

[2] http://www.hi.u-tokyo.ac.jp/ships/

[3] 『古事類苑』、吉川広文館発行

[4] 『大日本史料』、東京大学史料編纂所編纂・発行

[5] http://gensen.dl.itc.u-tokyo.ac.jp/

[6] 5.3.4 likelihood ratios, pp.172-175 *In*: Manning C. D. and Schutze H.: Foundations of Statistical Natural Language Processing, MIT, ISBN 0-262-13360-1, 1999.

[7] Dunning, T. : Accurate Methods for the Statistic of Surprise and Coincidence, Computational Linguistics, 19, pp.61-74, 1993.

[8] http://chasen.naist.jp/hiki/ChaSen/

[9] 5. 品詞体系 pp. 16-22 ipadic ユーザーズマニュアル

[10] http://sciterm.nii.ac.jp/

[11] 相田 満：日本文化のオントロジー「古事類苑」のデータベース化のために、平成 17 年度研究成果報告 日本文学国際共同研究(ICJS)研究集会 pp.34-58(25)、2006,3.