国文学研究論文目録データベースの高次資源化と汎用化をめざして

相田満・栗城大地・野本忠司（国文学研究資料館）

「国文学論文目録データベース」は,8 の時代分類の下位に 128 分野が 4 階層で類別されるほか,タイトルに現れないキーワードが付与される日本文学研究論文の総合目録データベースである．分類・キーワードは,全て専門研究者達が全文閲読の上で施され,令和元年時で 60 万件超,130 年間の論文を擁する．網羅型データベースでは求め得ない詳細な検索結果を得ることが可能な専門特化型のデータベースといえる．しかし,単独組織で人力によるデータ構築には,対費用効率の面で限界がある．そこで,それまでの資源を活かしつつ,採録者の負担軽減と効率化を果たせるよう,根本的な作業の組み立てに取り組むことに至った.

具体的目標としては,①既存辞書資源の活用,②入力方法の効率化,③発展的データナビゲーションの実現である.

そして,これらの諸目標の解決のために取り組まれた方法により得られる成果は,他のデータベース構築にも可能となるよう,構築データのオープン化,手法の共有,さらに他機関との連携によりデータや類別スキームの弁別と共有を果たすことである．これにより,専門特化型のデータベースの発展に寄与することで,研究の活性化をはかる構想の一端を述べる.

Aiming for Higher-order Resources and Generalization of the Database for Rsearch Thesis in Japanese Literature

AIDA,Mitsuru / KURIKI,Daichi / NOMOTO,Tadashi(Natinal Institue of Japanese Literature)

The “Database for Rsearch Thesis in Japanese Literature” is a comprehensive catalog database of Japanese literature research papers in which 128 fields are classified into 4 levels below the 8 era classifications, and keywords that do not appear in the title are given. The classification and keywords are all read by specialist researchers after reading the entire text, and have over 600,000 papers and 130 years of papers in the REIWA the 1’st year. However, there is a limit to the cost efficiency in constructing data manually by a single organization. Therefore, we came to work on the assembly of the fundamental work so that the burden on the registrant and the efficiency improvement could be achieved while utilizing the existing resources.

Specific goals are: (1) Utilization of existing dictionary resources, (2) Efficiency of input methods, and (3) Realization of advanced data navigation.

And the results obtained by the methods that have been addressed to solve these goals can be obtained by opening the construction data, sharing methods, and collaborating with other organizations so that other databases can be constructed. And to differentiate and share classification schemes. In this way, we will describe a part of the concept of revitalizing research by contributing to the development of specialized database.

１．まえがき

多様化・肥大化の一途をたどる研究動向の中で,当該研究がどの位置にあるかを確認するために,一定の枠組みで括られる研究分野の中から情報と研究史・研究動向の概要を把握する指針となるナビゲーションシステムは不可欠である．特に人文科学分野に於いては,研究の息が長く,先行研究が 10～20 年以上前にしか存在しないということも珍しくない．そのような状況の中,数ページの中に明確な主張と新規のトピックが凝縮される扱われる学術論文の存在は,その時々の研究における新規性と,その出現頻度によって研究の趨勢が把握可能な最適の媒体といえよう.

しかし,年々研究が肥大化・細分化・学際化・国際化の一途を辿る中,適切な研究情報を入手するためのツールも不可欠となり,その中で論文情報のナビゲーションデータベースの重要性はますます重みを増しているにも関わらず,その維持・管理にかかる負荷は,情報量が増大するのと反比例するかのように,そこに費やされる人的・経済的コストは減少の一途をたどりつつあるのが実情である.

人文科学分野でよく使われる網羅的リファレンスデータベースは,そうした動向を反映して生まれたもので,現在よく使われるものでは,

Ciniii(Articles(NII)
約 2,150 万件<2019.3 時点>
J-STAGE(科学技術振興機構)
約 493 万論文<2019.10.27 時点>
Google Scholar(google)
約 1.6 億文章<2014.5 時点>
NDL ONLINE／国会図書館サーチ
図書・雑誌書誌約 1.200 万件<2018.1 時点>

などがある.

上記の内,2018 年 3 月 30 日の NII-ELS サービスの終了にともない,J-Stage へのデータ移行が進められてはいるものの,なお 2019 年時点でも途上にあるため,レファレンス情報や論文入手に多少混乱を生じているようである[1].

網羅的なリファレンスデータベースに対して,機関が独自に採録を行う「専門特化型データベース」がある.

日本文学研究論文の総合目録データベースである「国文学論文目録データベース」も専門特化型の 1 つに加える事ができる.

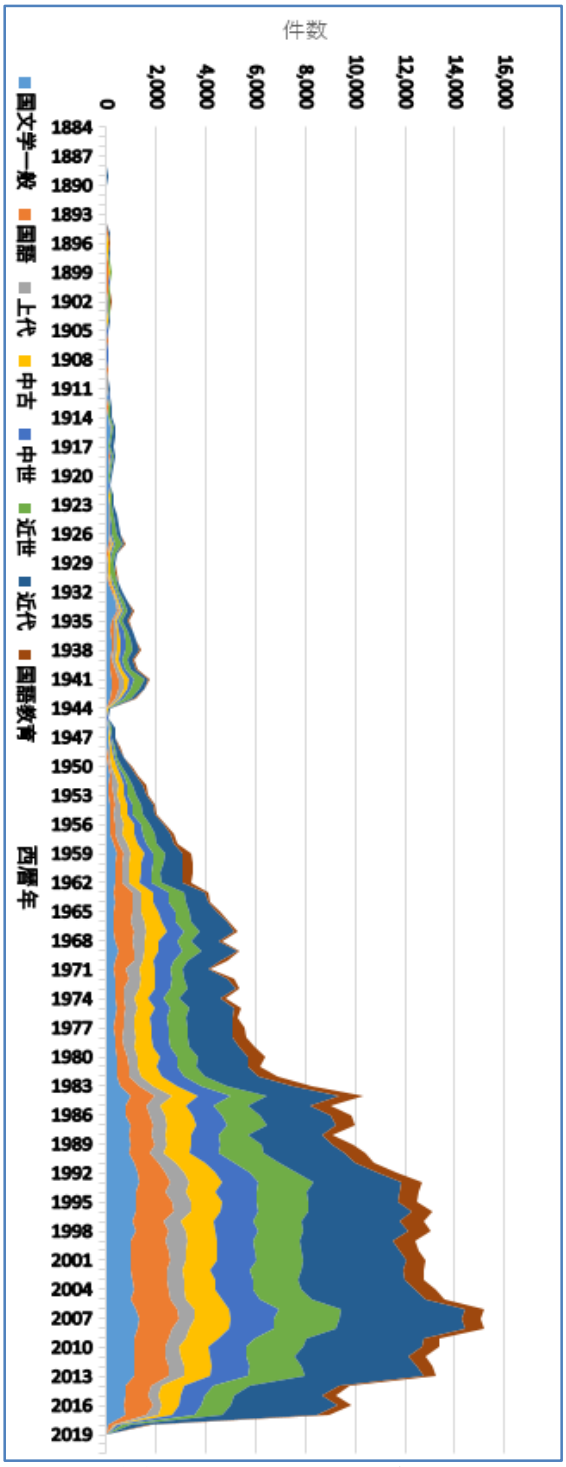


図 1　国文学論文 DB 時代・分野の推移（件数）

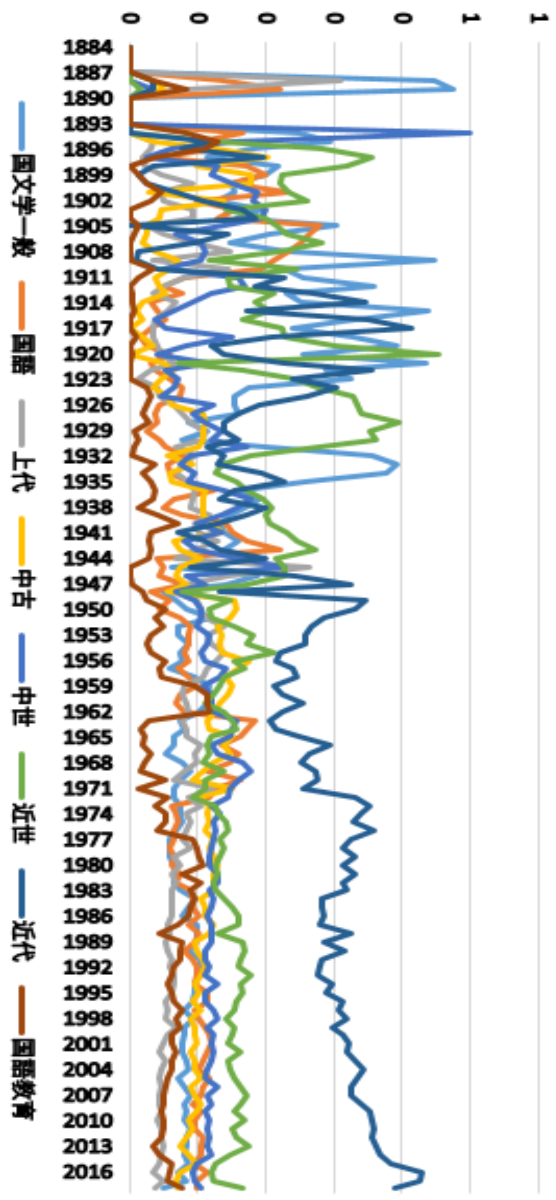


図 2　国文学論文 DB 時代・分野の推移（比率）

２．国文学論文目録データベースの特徴

本論で扱う日本文学研究論文の総合目録データベース「国文学論文目録データベース」は,2019元年５月に登録データ 60 万件を超え,9 月 13 日の更新で 604,776 件に達している．採録されるデータの収載期間は 130 年に及び,,斯界では最も息の長い部類に入る主題書誌型の文献目録データベースである．

収録データの上限 1888 年（明治 21）は,東京大学創設（1876 年）とともに生まれた和漢文学科から分離独立（1885 年）した和文学科が,国文学科と改称された年にあたる．論文目録データベースが覆う 130 年間は,まさに国学・和学から発展して近代国文学が標榜され始めた明治から大正・昭和・平成時代の全てをあたり,まさに国文学研究の研究史を,研究論文の輩出状況面から追証可能なデータベースといえる．

前頁の図 1 は,採録論文数を各時代・分野別件数で積み上げたグラフ,図 2 はそれぞれが全体件数の中で,どれ程の比率を占めているかを示したものである．図 1 で最新年度のデータ数が少ないのは,まだ遡及入力が続いているためで,1990 年代以降は,概ね 1 万～1 万 2,3 千件のデータ量状況で推移しており,いずれはその数値に落ち着くものと予想される．

図 2 は各時代・分野がその年の総件数の何%を占めてきたかを示したものである．大体に於いて近代文学の研究が最も多いが,1900 年代初頭は,近世文学の論文が最も多いが,当時は井原西鶴等の浮世草子や黄表紙などの草双紙が研究対象とは認知されず,採り上げられたトピックで目立つ者では,平田篤胤や本居宣長,芭蕉などの俳諧などが多く,近年は逆に見直されつつある分野も少なくない．

「国文学研究論文目録データベース」は,『国文学研究文献目録』（東京大学国語国文学研究室）と『国文学年鑑』（国文学研究資料館）の冊子データの内,雑誌と論集に収載される論文の主題書誌をデータベースにしたものである．

ここでいう主題書誌（subject bibliography）とは,たとえば,図書として刊行されるもの（「単行書誌」）,図書の一部（「参考文献一覧」など）,専門雑誌などに記事として掲載されるもの,ウェブサイトに掲載されるものなどがある.特定のテーマやトピックに関する文献リストで,それぞれの分野の専門家や研究者が,非売品や私家版なども含め関連文献を網羅的に採録し,各文献に解説（解題）を付していることがあるが,国文学論文目録データベースは,キーワードとして,ジャンル分けされた主題と,作品・作者を主とするキーワードが,採録者の閲読を経て付される点に特徴がある．

図 3 が各論文データの採録画面である．各論文に付された時代・主題・分野・キーワードなどは,冊子編集の時代には,それが各レコードの配列・

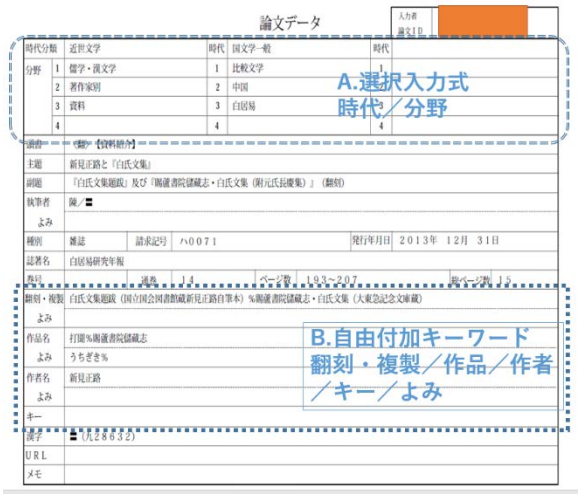


図 3　データ採録例

に反映されており,毎年 I 万数千件を収載する冊子の編集は,専門研究者の監修の元に 2007 年（平成 17）まで続けられてきた歴史を持つ．.

こうして積み上げられてきたデータと,分類・配列のために付与されてきたキーワード群と配列のために施されてきた分類語群は,すべてエキスパートが採録した優良な学習データと位置づけることができる．

これを図示すれば,図 4 の通りとなる．

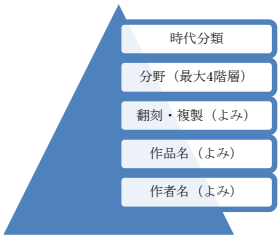


図 4　論文に付加されるエキスパート情報

こうして採録されたデータは,現在,次のように分類されている．現在公開されるデータベースメニューでは,各時代分類と分野別に採録されたデータ総数が示されており,数字の所をクリックすると,そのデータが絞り込まれて出力される．内訳は,以下の通り．

時代分類／分野
国文学一般　上代文学　中古文学　中世文学　近世文学　近代文学　国語　国語教育
国文学一般(54031)
一般(449)演劇・芸能(3079)沖縄文学(148)歌謡(867)芸能(461)古典文学(10477)詩歌(676)詩歌・歌謡(558)書評・紹介(3664)説話・昔話(4387)南島文学(2334)俳諧(1051)比較文学(2193)文学論(584)文学論・国文学論(8373)民俗学(7387)目録(345)目録・その他(4217)和歌(2781)
中古文学(62648)
一般(6808)歌謡(991)漢文学(2160)軍記(262)国語(2457)書評・紹介(2656)説話(3451)日記・随筆(7165)物語(24746)歴史物語(1859)和歌(10093)
近世文学(92849)

一般(23147)演劇・芸能(8371)演劇・芸能・芸能(3318)歌舞伎(136)漢文学・儒学(638)狂歌・狂文(1080)国学・和歌(8989)国語(2570)儒学・漢文学(5622)書評・紹介(3466)小説(15495)浄瑠璃(414)川柳・狂歌(347)川柳・雑俳(2178)俳諧(4363)連歌・俳諧(11967)和歌(130)和歌・和文(618)
国語(60199)
一般(6025)一般及び雑(300)音声・音韻(288)音声・音韻・アクセント(2450)敬語(955)言語生活(4049)語彙・意味(6710)辞書・資料(646)辞書・資料・訓点語(1552)書評・紹介(2326)対照研究(3685)日本語(371)日本語教育(6879)文字・表記(3290)文体・文章(2004)文法(12130)方言(6539)
上代文学(37856)
一般(6484)歌謡(1171)漢文学(419)古事記・日本書紀(4625)国語(1659)祝詞・宣命(325)書評・紹介(1831)神話(2309)風土記(1136)万葉集(17897)
中世文学(72302)
キリシタン文学・語学(1036)一般(8822)演劇・芸能(8654)演劇・芸能・芸能(1498)歌謡(905)漢文学(949)軍記物語(7758)国語(2490)書評・紹介(3066)唱導・縁起(310)小説(522)随筆(402)説話(1130)説話・唱導・縁起(4407)日記・紀行・随筆(3860)能(797)能・狂言(415)仏教文学(1107)仏教文学・神道(4861)物語・小説(3671)歴史物語・史論(534)連歌(2733)和歌(12375)
近代文学(182851)
一般(27705)演劇・芸能(4163)近代詩(5193)国語(588)作家別(23840)詩(2518)児童文学(4261)時評・展望(260)書評・紹介(10897)小説(24857)大衆文学(436)短歌(9398)著作家別(61230)俳句(4764)評論(2741)
国語教育(36564)
ことば(846)一般(16569)言語事項(1194)古典（古文・漢文）(319)国語教育（古典）(479)作文(926)作文・書写(298)書くこと(1407)書写・書道(365)書評・紹介(1622)読むこと(4798)読解・読書(2259)表現(1219)理解(3150)話すこと・聞くこと(1113)

また,採録方法にも違いがある.たとえば,網羅型 DB は全分野に及ぶ採録だが,NDL は書誌情報,目次,奥付.内容の読み込みはなく,Cinii は OPAC,連携機関の情報,書誌情報から取り込まれたデータが元になっており,NII が独自に作ることはなく,J-STAGE も同様である.

対して,国文学・国語・国語教育と関連分野に特化する国文学論文目録 DB は,所蔵資料からのマニュアルに基づくデータ採録がなされて, 1 論文につき,5 (必須) -30 目以上のデータが採録されている.その結果,国文学に特化する具体的な内容の検索に於いて,たとえば検索語「太宰治」では,

・NDL 3263 件　論文内容の情報無.不要情報多し.
・J-STAGE　976 件　ダウンロード可能論文
・Cinii 3032 件　同上〃
・国文学論文 DB 3845 件　論文内容情報有

また,検索語「歌枕」については,

・NDL 1135 件　論文内容の情報無.不要情報多し
・J-STAGE　597 件.　ダウンロード可能論文
・Cinii 545 件　同上〃
・国文学論文 DB 817 件　論文内容情報有

と,国文学論文目録データベースがデータ数 60 万件に対して,比較対象の分母は桁違いであるにも拘わらず,健闘あるいは凌駕するものも少なくない.

ただし,国文学論文目録データベースにおいても『源氏物語』と「宮沢賢治」のデータがあまりにも多いために検索過多の悩みを持つものもある.

その問題については以前ふれたことがあるが[3],過検索と,採録負担の軽減のために進めているのが,検索自動化の試行である.以下にその概要を示す.

3．キーワード・カテゴリ付与の自動化

1992 年度からのオンライン公開開始当初は,汎用機によっていたため,完全・前後方一致検索しかできなかった.そのため,タイトルの分かち書き処理を施した上で,キーワードの抽出を行っていた.

汎用機時代の当時は HAPPINESS（(株) 平和情報) データベースによる自動キーワード付加処理がなされていたが,当時は対応する文字種自体も JIS208-1978 にとどまっており,限界もあった.しかし現在は,多様なツール群も現在は整い,このツールを運用する為に整えてきた語彙辞書も,ようやく本来の機能を発揮できる状況になりつつあるといってよかろう.

.

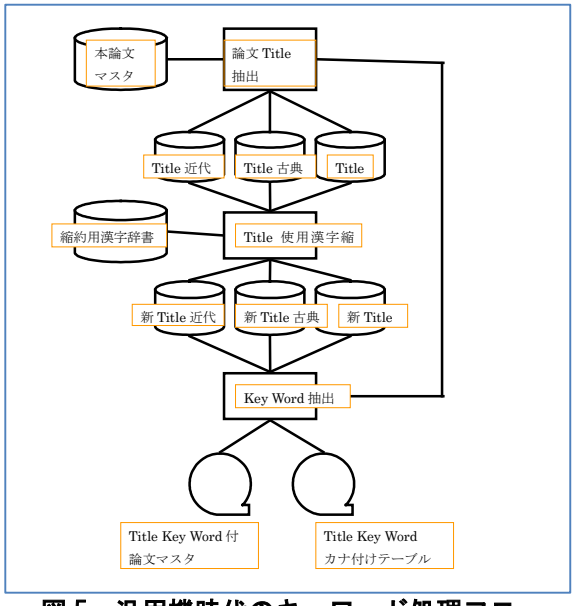


図 5　汎用機時代のキーワード処理フロー

図 5 は汎用機時代の論文タイトルからキーワードを抜き出すための処理フローである.

たとえば,1 本の論文から適切な語彙を切り取り,そこから抜き出されたキーワードを抽出

するためには,抽出を容易にするために,論文タイトルに分かち書き処理をしなくてはならなかったのが当時の問題であった.

たとえば,

「『源氏物語』作者紫式部の生涯」

というタイトルを認識させるには,『』内はデフォルトで 1 語彙として認識させるとして,他は以下のように分かち書きをしなくてはならないことになる.

切り分け後

『源氏物語』／作者／紫式部／の／生涯

国立国語研究所で蓄積される形態素辞書では,

「源氏／物語」「紫／式部」

が切り分けられてしまい,こうした概念分析のためには別に辞書を整えなくてはならない.

現在,その分析のための辞書の構築も進めてはいるが,それと併行して,デフォルトのシステムによる評価実験を進めている.

評価を進めているのは MeCab と Python を利用しての分類実験と,Apple で開発している Create ML の Natural Language Processing を利用した機械学習モデルの実験,またアルゴリズム：Conditional random field を使用した分類付与実験と 3 通りで実現可能性の検証実験である.汎用辞書でも一部には時代分類で 8 割を超える正当を得ているものもあるが,辞書の鍛え方次第では,今後更に良好な結果を得ることが期待される.

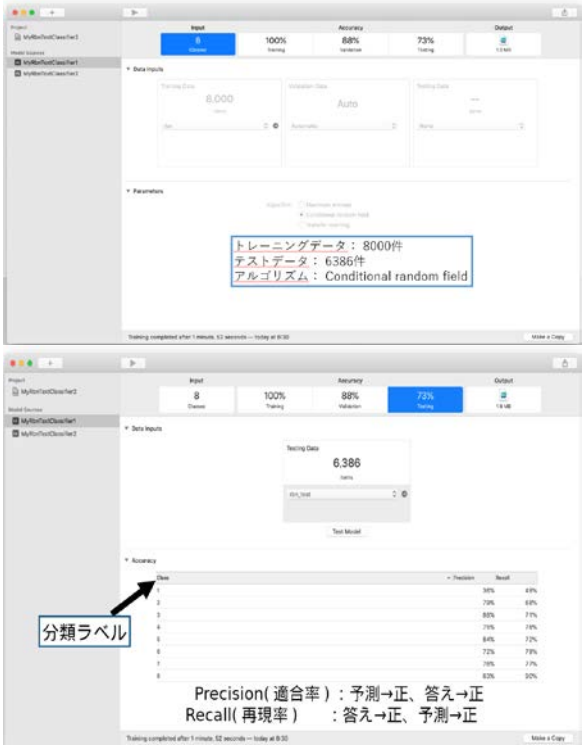
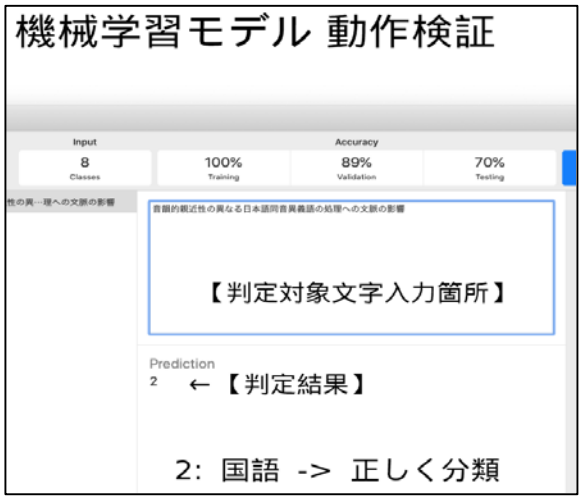


図 6　論文カテゴリの自動付与実験①



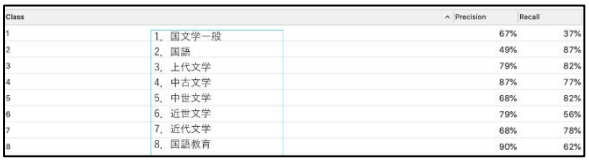


図 7　図 6 の結果

図 8　類似ベクトル表示を web api 化したもの

図 8 は『源氏物語』の注釈書「河海抄」で時代分類・関連語の表示を柳宗利氏が試みたものである.

島崎健は論文執筆者,孟津抄は河海抄と同じく源氏物語の古注釈書だが,「はいからさんが通る」はマンガ版源氏物語『あさきゆめみし』の作者,大和和紀の連想から→源氏物語→河海抄から連想が働いたと想像され,近代文学ジャンルに扱われ

る論文ではこのような関連語がヒットすることもあり得たと思われる.

次に示したのは,MeCab と Python を利用しての分類結果で冒頭の数字が分野を示す.

（総レコード数 596,403　[2019 年]. 月時点）

1,文学反響 Ｌｉｔｅｒａｒｙ Ｅｃｈｏｉｎｇ 垣内松三
4,源氏物語のモデル 源氏物語 紫式部 手塚昇
5,室町時代の小歌と閑吟集 閑吟集 志田義秀
4,大鏡に関する考察 大鏡 藤村作
2,ノリといふ語 芳賀矢一
1,浦島伝説 久松潜一
8,古典講読用書様式の一提案 藤村作
8,聴方教授について 田中末広
1,国語研究室焼失主要書目録 目録 橋本進吉
1,生の象徴としての短詩 岩城準太郎
2,日本語教育　政策について 保科孝一 国語政策について
4,源氏物語のモデル 承前　源氏物語 紫式部 手塚昇
1,古典の本文整定 山岸徳平
6,契沖の文学批判 契沖 久松潜一
4,大鏡に関する考察 承前　大鏡 藤村作
5,室町時代の小歌と閑吟集 承前　閑吟集 志田義秀

他にキーワードの自動付与に向け，以下の二つの条件の下で予備実験を行った．(A) 論文タイトル，著者名のみを用いて，8 カテゴリ（一般，国語，上代，中古，中世，近代，教育）への分類．(B) ウィキペディア (jawiki) から論文タイトルに最もよく合致したページを論文の代理 (サロゲート) と見なし，上記 8 カテゴリーへの分類．データポイントの総数は 52,1929，このうち 30,000 点をテスト用，残りの 49,1929 点を学習用に用いた．モデルはオープンソースの fasttext (https://fasttext.cc/) を利用した．以下結果である．評価指標は P@1（値の範囲は 0-1，高いほど良い），epoch 数は 20，その他のパラメータはデフォルト値を用いた．

```
----------
条件(n=2)  one-vs-all  softmax
A                0.897       0.870
B                0.671       0.675
----------
条件(n=3)  one-vs-all  softmax
A                0.911       0.874
B                0.673       0.653
```

ここで，one-vs-all，softmax は分類器のタイプ，前者はカテゴリ毎にバイナリ分類器を構成する，後者は 8 カテゴリを一挙に分類する．n=2/3 は入力データの表現形式，n=2 は 2 グラム，n=3 は 3 グラムを用いた．

上の予備実験の結果，A/n=3/one-vs-all の下で，最良の結果 0.911 (P@1) を得た．またウィキペディアによる拡張の効果がないことが判明した．ここで，P@1 とは，モデルが出力した 1 位候補のうち，正解であったものの割合を指す．例えば，10 回の試行のうち，正解が 1 つ含まれれば 0.1 となる．

今後は，精度向上のため使用する辞書において，専門家の作ったエキスパート辞書を活用と育成が期待されるが，先述の通り，現在，語彙辞書の整備を進めており，さらに実験用の論文 PDF データの作成も小規模ながら進めてきている．

今後は蓄積されたデータを整備した上で公開することで，論文目録データベースデータとは別の面での基盤データが公開される事も期待され，そのことがさらに別のデータ整備に繋がる好循環が生まれることを，本実験の取り組みを通して痛感する次第である．.

参考文献

[1] “ITmedia NEWS「CiNii から論文が消えた」研究者に困惑広がる”. https://www.itmedia.co.jp/news/articles/1704/05/news086.html (参照 2019-10-24).
[2] 中野真樹,渡辺由貴,国立国語研究所「日本語研究・日本語教育文献データベース」の有用性, 国立国語研究所論集 5 2013.5 pp67-76(5), info:doi/10.15084/00000504.
[3] 相田満，目録データベースの高次化によるデータマイニングを可能とするために―複数種のオントロジ辞書の利用・接合により検索効率の向上を試みる，情処学会論文集じんもんこん 2004,pp151-158 https://www.ipsj.or.jp/kenkyukai/font.pdf, (参照 2018-09-24).
[4] “論文の著作権の取り扱い”. https://www.ipsj.or.jp/copyright/ronbun/, (参照 2018-09-24).
[5] 人文太郎，情報花子．人文学の情報学的考察．情報処理学会論文誌, 2012, Vol. 0, No. 0, p.9-10.
[6] 人文次郎．人文科学とコンピュータ．人文太郎(編)，人文科学とコンピュータの歴史，じんもんこん出版, 2010, p.3-26.

