様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ８ 日現在

機関番号：３４３１５
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ００４５７

研究課題名（和文）Research on Knowledge Extraction from Ancient Mongolian Historical Documents using Deep Learning

研究課題名（英文）Research on Knowledge Extraction from Ancient Mongolian Historical Documents using Deep Learning

研究代表者

バトジャルガル　ビルゲサイハン（BATJARGAL, BILIGSAIKHAN）

立命館大学・衣笠総合研究機構・研究員

研究者番号：３０７２５３９６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、デジタル化された古代モンゴル文字文書からの情報抽出方法を提案した。提案方法では、深層学習技術を利用して歴史的文書から深層特徴を抽出した。抽出した特徴を用いて、モンゴルの歴史的文書での古代モンゴル語単語の解釈および同じ形の異なる文字を認識するための特徴をエンコードできる閲覧検索システムを作成した。注釈付きの伝統的モンゴル文字の原文テキストおよび原文のスキャン画像を含む古代文字歴史的文書のデジタル版は、学術ツールとして使用できると考えられる。さらに提案方法は、モンゴルの歴史的文書だけでなく、甲骨文字や浮世絵にも適用された。

研究成果の学術的意義や社会的意義
本研究は歴史的文書の分析にかかる時間と手間を軽減することができると考えられる。現代モンゴル語の文書には含まれない隠れた知識を伝統的モンゴル文字の古文書から発見できると考えられる。提案方法を用いた歴史的文書の分析結果を分かりやすく表示するシステムはどんなユーザーにも利益をもたらすことが期待されている。さらに、歴史書類のデジタル化を研究対象にしている学者に大きく貢献すると考えられる。

研究成果の概要（英文）：In this research, an information extraction and analysis method for digitized ancient Mongolian historical documents is proposed. The proposed method extracts features from historical manuscripts by utilizing deep learning techniques. Extracted deep features are utilized for building retrieval systems that encode the interpretations of ancient Mongolian words, as well as features for recognizing different letters with the same shape. Digital representations of ancient historical manuscripts with annotated ancient Mongolian texts along with the scanned images of manuscripts could be used as scholarly tools. The proposed methods were applied not only to Mongolian historical manuscripts but also to Oracle bone script and Ukiyo-e, a Japanese traditional art.

研究分野：Digital Humanities

キーワード：deep learning　historical documents　traditional Mongolian　machine learning

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

人文系では、歴史書類を分析し、知識を得られるのは重要なことである。人文系研究者らより膨大なテキストを短時間で早急にテキスト分析できることが高く要求されている。この場合は、コンピュータは早急に操作できるので適している。既存の歴史書類をコンピュータで分析でき、その書類を完全に表示するシステムは人文系研究者らのもう一つの要求である。時代に伴い、歴史書類は再度写され、修正される中、間違いが発生し、かなりの変更が生じ、それに研究者らは各自の説明をつける。このような全ての状況を含むコンピュータアプリケーションは困難である。最近、モンゴル文字で書かれた少数の歴史書類はデジタル化され、公開された。しかし、これらの古代モンゴル文字書類に対応できる自然言語処理のツールが存在しないため上記書類の分析が未だにできてない。そのため、コンピュータを用いた分析手法がもっとも重要である。歴史書類から情報抽出可能な深層学習技術の提案が必要とされる。

２．研究の目的

本研究では、デジタル化された伝統的モンゴル文字の歴史書類から深層学習技術を使って情報抽出方法を提案する。歴史書類の分析にかかる時間と手間を軽減することを目的とし、古代モンゴル語の特徴を深層学習技術を使用して抽出する。深層学習モデルは伝統的モンゴル文字書類を分析し、特徴を抽出して特徴ベクトルとして保存する。そして抽出された特徴を用い、デジタル・ヒューマニティーズ研究のための Web ベースのプロトタイプシステムを開発する。このシステムの①モンゴルの歴史書類の古代モンゴル語単語の解釈できる機能②同じ形の異なる文字を認識するための特徴をエンコード化したテキストおよび③原文のスキャン画像は、歴史書類のデジタル化を研究対象にしている学者に大きく貢献すると考えられる。

３．研究の方法

本研究では、１３世紀から１６世紀までのデジタル化された古代モンゴル文字の歴史書類から情報抽出する方法を提案するため以下のことが必要とされる。

a) 伝統的モンゴル文字の訓練データ、テストデータセットおよび言語資源を作成する。
b) a の資源を使用して深層学習モデルを構築し、古代モンゴル語テキストに適用する。
c) 古代モンゴル語単語の翻字や解釈を表示できる Web ベースシステムを開発する。

まずは、辞書や注釈訓練データのような言語資源が作成される。これらは情報抽出や情報認識等の次の処理に必要な資源として使用できる。現在、使用可能な少数の電子辞書の一つはツェベル氏の辞書である。約 30,000 語を含む古代・現代モンゴル語辞書は現代モンゴル文字および伝統的モンゴル文字で書かれている。本研究では古代コーパスの共起回数や単語回数のような統計情報を比較する辞書を構築する。古代モンゴル文字の歴史書類のスキャン画像からすべての文字を画像として抽出してタグ付けする。そして、古代モンゴル語コーパスを形態素解析し、全トークンを注釈して訓練データが作られ、学習のための深層学習モデルに入力する。古代モンゴル文字歴史書類の自然言語処理ツールや品詞データがないため、「Qad-un ündüsün quriyang γui altan tobči -Textological Study」(Choimaa, 2002)から手動でコンパイルされた単語のインデックスを用い、「小」アルタン・トプチの全単語を先に注釈した。本研究では、古代モンゴル語の文法に基づく深層学習モデルを用いた情報抽出手法を提案する。本手法は、トークンあるいは古代モンゴル語の特徴を深層学習に入力することによって、デジタル化された伝統的モンゴル文字の歴史書類を分析する。トークンはほとんど空白で区切られた単語になるが、古代モンゴル語にはトークンを分析するためのいくつかのユニークな特徴がある。本研究では、以下の特徴が考慮される：

- 現在のトークンの先行情報：前のトークンが世代または王朝の情報、貴族の継承されたまたは生涯のタイトルおよび伝統的記述的表現な場合、現在のトークンは人物名である可能性が高くなる。
- 接尾辞:伝統的モンゴル文字では人物名および生き物には特別な接尾辞や複数接尾辞を使う(Chinggaltai, 1963)。接尾辞は屈折語の一部になるが、ほとんどの接尾辞は単語の語幹または他の接尾辞から狭い隙間によって分けて書かれている。
- 文章の始まり：通常、名詞または人物名は文章の先頭にある。
- トークンの終わり：最後の母音文字の「a」または「e」は語幹の不可欠な部分であるが、「a」または「e」が先行する子音から狭い隙間によって分けて書かれている。

伝統的モンゴル文字文書からの情報認識・情報抽出する深層学習モデルとして文字認識に幅広く応用されている Convolutional Neural Networks（畳み込みニューラルネットワーク）を使った。訓練データとして、「小」アルタン・トプチのデジタル化テキストおよび原文のスキャン画像を用いた。

４．研究成果

本研究で得られた研究結果に基づき、抽出された特徴およびその他の情報を用い、Web ベースのプロトタイプシステムを開発した。本提案システムは、現代モンゴル語の文書には含まれない隠れた知識を伝統的モンゴル文字の古文書から発見できる社会的意義があると考えられる。または、注釈付きの伝統的モンゴル文字の原文テキストおよび原文のスキャン画像を含む古代モンゴル文字の歴史書類のデジタル版は、学術ツールとして使用できると考えられる。図１に示すように、モンゴル年代記である①「Qad-un ündüsün-ü quriyangγui altan tobči neretü sudur」（The Altan Tobchi or the Golden Summary: Short history of the Origins of the Khans）（1604 年頃に書かれた）あるいは「小」アルタン・トプチ②「Asaraγči neretü-yin teüke」「The Asragch nertiin tüükh or The Story of Asragch）（1677 年頃に書かれた）あるいは「アサラクチ・ネレティン・トゥーフ」のデジタル化テキストおよび原文のスキャン画像は閲覧・検索可能になった。「小」アルタン・トプチは約 16,200 語を含む 164 ページがある。または、「アサラクチ・ネレティン・トゥーフ」は約 17,250 語を含む 130 ページがある。

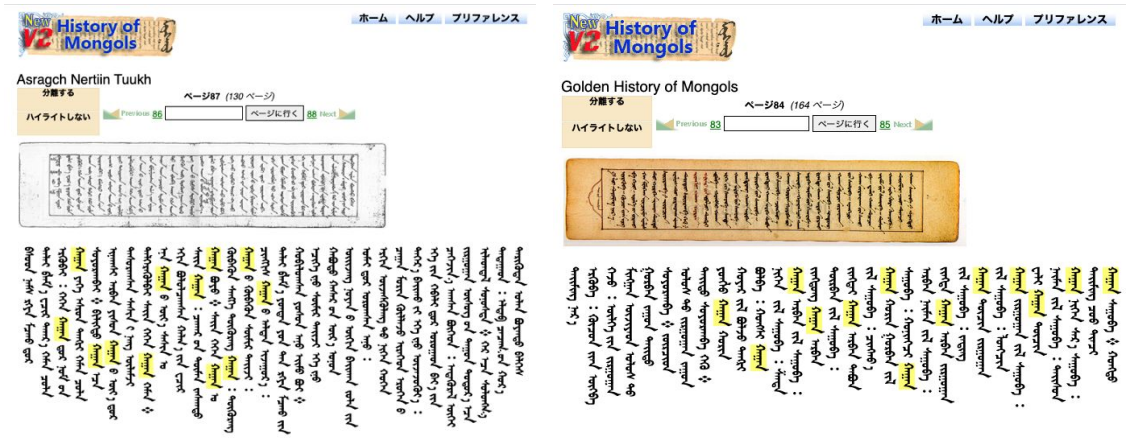

図１.伝統的モンゴル文字歴史書類の閲覧・検索できるシステム

図２と図３では、「小」アルタン・トプチから抽出された固有名詞等の情報がハイライトされている。図３に示すように、伝統的モンゴル文字を左に、それに対応するラテン文字訳を右にそれぞれ表示し、比較できるように設定した。

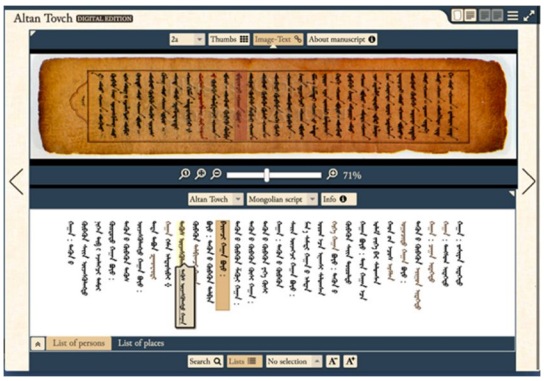

図２.ハイライトされた固有表現および画像とテキストリンクのプロトタイプ

図３.ハイライトされた固有表現および字訳テキスト版

提案方法はモンゴルの歴史的書類だけでなく、甲骨文字や浮世絵にも適用された。甲骨文字の認識、篆書体による蔵書印の文字認識および印文情報検索に深層学習を適用する研究を共同で行ってきた。さらに、深層学習を用いて浮世絵を初心者ユーザに推薦する研究を共同で行ってきた。上記の研究成果について、第一著者として分担執筆の本の章１件、国際会議発表３件、国内会議発表６件および共著者として雑誌論文４件、研究報告１件、国際会議発表１２件、国内会議発表１５件で紹介してきた。

＜引用文献＞

① Chinggaltai. (1963). A Grammar of the Mongol Language. New York: Frederick Ungar Publishing Co.

② Choimaa, Sharav. (2002). Qad-un ündüsün quriyangγui altan tobči (Textological Study). vol. 1. Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia, Urlakh Erdem. (in Mongolian).

③ Jamba erke daicing (1677). Asaraγči neretü-yin teüke. (in Mongolian).

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　4件／うち国際共著　3件／うちオープンアクセス　4件）

１．著者名 Kangying Li、Biligsaikhan Batjargal、Akira Maeda	４．巻 Vol. HistoInformatics
２．論文標題 Character Segmentation in Asian Collector's Seal Imprints: An Attempt to Retrieval Based on Ancient Character Typeface	５．発行年 2021年
３．雑誌名 Journal of Data Mining and Digital Humanities	６．最初と最後の頁 1-19
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Song Yuting、Batjargal Biligsaikhan、Maeda Akira	４．巻 Vol. 30, No. 03
２．論文標題 Learning Japanese-English Bilingual Word Embeddings by Using Language Specificity	５．発行年 2020年
３．雑誌名 International Journal of Asian Language Processing	６．最初と最後の頁 1-14
掲載論文のDOI（デジタルオブジェクト識別子） 10.1142/S2717554520500149	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Yuting Song、Biligsaikhan Batjargal、and Akira Maeda	４．巻 Vol. 1: 52
２．論文標題 An Application of Cross-Language Record Linkage Techniques to Digital Cultural Collections	５．発行年 2020年
３．雑誌名 Asia-Japan Research Academic Bulletin	６．最初と最後の頁 1-7
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yuting Song、Biligsaikhan Batjargal、and Akira Maeda	４．巻 17巻1号
２．論文標題 Cross-Language Record Linkage based on Semantic Matching of Metadata	５．発行年 2019年
３．雑誌名 日本データベース学会英文論文誌（DBSJ Journal）	６．最初と最後の頁 1-18
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yuting Song、Biligsaikhan Batjargal、and Akira Maeda	4．巻 27
2．論文標題 Recognition and Transliteration of Proper Nouns in Cross-Language Record Linkage by Constructing Transliterated Word Pairs	5．発行年 2017年
3．雑誌名 International Journal of Asian Language Processing	6．最初と最後の頁 111-125
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計36件（うち招待講演　0件／うち国際学会　15件）

1．発表者名 王 嘉韻、Batjargal Biligsaikhan、前田 亮、川越 恭二、赤間 亮
2．発表標題 深層学習モデルに基づく浮世絵画像検索システムの開発
3．学会等名 第10回知識・芸術・文化情報学研究会
4．発表年 2021年

1．発表者名 Li Kangying、Batjargal Biligsaikhan、前田 亮、赤間 亮
2．発表標題 浮世絵レコードのクロスモーダル多言語横断検索に向けて：Multilingual-BERTによる作品情報の特徴埋め込み抽出の試み
3．学会等名 第10回知識・芸術・文化情報学研究会
4．発表年 2021年

1．発表者名 Kangying Li、Biligsaikhan Batjargal、Akira Maeda、and Ryo Akama
2．発表標題 Artwork Information Embedding Framework for Multi-source Ukiyo-e Record Retrieval
3．学会等名 The 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)（国際学会）
4．発表年 2020年

１．発表者名
Yuting Song、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
A Preliminary Attempt to Evaluate Machine Translations of Ukiyo-e Metadata Records

３．学会等名
The 22nd International Conference on Asia-Pacific Digital Libraries (ICADL 2020)（国際学会）

４．発表年
2020年

１．発表者名
Kangying Li、Biligsaikhan Batjargal、Akira Maeda、and Ryo Akama

２．発表標題
Toward Exploring Artist Information from Seal Images in Ukiyo-e Collections

３．学会等名
The Digital Humanities 2020 Conference（国際学会）

４．発表年
2020年

１．発表者名
Jiayun Wang、Biligsaikhan Batjargal、Akira Maeda、Kyoji Kawagoe、and Ryo Akama

２．発表標題
Making Ukiyo-e Easier to Discover: A Recommender System for Digital Archives

３．学会等名
The Digital Humanities 2020 Conference（国際学会）

４．発表年
2020年

１．発表者名
Yuting Song、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Improving Japanese-English Bilingual Mapping of Word Embeddings based on Language Specificity

３．学会等名
The Digital Humanities 2020 Conference（国際学会）

４．発表年
2020年

１．発表者名
佐藤 英男、Yuting Song、Biligsaikhan Batjargal、前田 亮

２．発表標題
異言語の映画データベース間における同一作品の言語横断レコード同定手法

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム（DEIM2020）

４．発表年
2020年

１．発表者名
Biligsaikhan Batjargal

２．発表標題
アート・リサーチセンター所蔵資料データベースのオープンデータ化への取り組み

３．学会等名
第68回 国際ARCセミナー（Web配信）

４．発表年
2020年

１．発表者名
Biligsaikhan Batjargal

２．発表標題
白川フォントおよび漢字検索システム開発共同プロジェクト

３．学会等名
第61回　ARCセミナー，立命館大学アート・リサーチセンター

４．発表年
2019年

１．発表者名
Kangying Li、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Character Segmentation in Collector’s Seal Images: An Attempt on Retrieval Based on Ancient Character Typeface

３．学会等名
The 5th International Workshop on Computational History (HistoInformatics 2019)（国際学会）

４．発表年
2019年

１．発表者名
Jiayun Wang、Biligsaikhan Batjargal、Akira Maeda、Kyoji Kawagoe、and Ryo Akama

２．発表標題
A Graph-based Recommender System for Ukiyo-e Prints

３．学会等名
The 13th International Conference on Metadata and Semantics Research (MTSR 2019)（国際学会）

４．発表年
2019年

１．発表者名
Yuting Song、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Title Matching for Finding the Identical Metadata Records in Different Languages

３．学会等名
The 13th International Conference on Metadata and Semantics Research (MTSR 2019)（国際学会）

４．発表年
2019年

１．発表者名
Yuting Song、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Improving Japanese-English Bilingual Mapping of Word Embeddings based on Language Specificity

３．学会等名
The 2019 International Conference on Asian Language Processing (IALP 2019)（国際学会）

４．発表年
2019年

１．発表者名
王 嘉韻、Biligsaikhan Batjargal、前田 亮, 川越 恭二

２．発表標題
デジタルアーカイブのためのグラフベースの深層学習による推薦システム

３．学会等名
人文科学とコンピュータシンポジウム（じんもんこん2019）

４．発表年
2019年

１．発表者名
李 康穎、Biligsaikhan Batjargal、前田 亮、赤間 亮

２．発表標題
落款印および関連情報の検索システムの構築：人物情報と人物関係ネットワークの自動抽出に向けて

３．学会等名
人文科学とコンピュータシンポジウム（じんもんこん2019）

４．発表年
2019年

１．発表者名
Yuting Song、Biligsaikhan Batjargal and、and Akira Maeda

２．発表標題
Metadata Similarity Calculation in Cross-Language Record Linkage based on Cross-lingual Embedding Models

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム （第17回日本データベース学会年次大会）

４．発表年
2019年

１．発表者名
李 康穎、Biligsaikhan Batjargal、前田 亮

２．発表標題
古代文字検索のためのフォントからの字形特徴量の抽出および活用可能性の検討

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム （第17回日本データベース学会年次大会）

４．発表年
2019年

１．発表者名
王 嘉韻、Biligsaikhan Batjargal、前田 亮、川越 恭二

２．発表標題
浮世絵ディジタルアーカイブのための分散表現による作品の関連性に基づいた推薦システム

３．学会等名
第11回データ工学と情報マネジメントに関するフォーラム （第17回日本データベース学会年次大会）

４．発表年
2019年

１．発表者名
Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu、and Akira Maeda

２．発表標題
Creating a Digital Edition of Ancient Mongolian Historical Documents

３．学会等名
Digital Humanities 2018（国際学会）

４．発表年
2018年

１．発表者名
Biligsaikhan Batjargal

２．発表標題
伝統的モンゴル文字で書かれた歴史書類のデジタル版の作成

３．学会等名
第52回　ARCセミナー，立命館大学アート・リサーチセンター

４．発表年
2018年

１．発表者名
Biligsaikhan Batjargal

２．発表標題
専門性の深化を目的とした人文系大規模データベースの構築 -ポータルデータベースと横断検索システムによる世界規模の所蔵品検索・閲覧システム-

３．学会等名
国際シンポジウム「デジタル時代における人文学の学術基盤をめぐって」

４．発表年
2018年

１．発表者名
前田 亮、バトジャルガル ビルゲサイハン、李 康穎

２．発表標題
古代文字のディジタル化とその活用の可能性

３．学会等名
2018年度日本古文書学会大会「古文書学への招待―ひらかれる研究の窓」

４．発表年
2018年

１．発表者名
Kangying Li、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Ownership Stamp Character Recognition System Based on Ancient Character Typeface

３．学会等名
The 20th International Conference on Asia-Pacific Digital Libraries (ICADL2018)（国際学会）

４．発表年
2018年

１．発表者名
Jiayun Wang、Biligsaikhan Batjargal、Akira Maeda、and Kyoji Kawagoe

２．発表標題
A Recommender System in Ukiyo-e Digital Archive for Japanese Art Novices

３．学会等名
The 20th International Conference on Asia-Pacific Digital Libraries (ICADL2018)（国際学会）

４．発表年
2018年

１．発表者名
李 康穎、Biligsaikhan Batjargal、前田 亮

２．発表標題
古代文字フォント字形の特徴抽出に基づく蔵書印の検索支援

３．学会等名
人文科学とコンピュータシンポジウム（じんもんこん2018）

４．発表年
2018年

１．発表者名
李 康穎、Batjargal Biligsaikhan、前田 亮

２．発表標題
生成モデルによる篆書体の文字認識手法の提案

３．学会等名
第10回データ工学と情報マネジメントに関するフォーラム（第16回日本データベース学会年次大会）

４．発表年
2018年

１．発表者名
李 康穎、Biligsaikhan Batjargal、前田 亮

２．発表標題
篆書体による蔵書印の文字認識の試み

３．学会等名
第７回 「知識・芸術・文化情報学研究会」

４．発表年
2018年

１．発表者名
Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu、and Akira Maeda

２．発表標題
Creating a Digital Edition of Mongolian Historical Documents

３．学会等名
International Conference on Culture and Computing (Culture and Computing 2017)（国際学会）

４．発表年
2017年

１．発表者名
Yuting Song、Taisuke Kimura、Biligsaikhan Batjargal、and Akira Maeda

２．発表標題
Linking the Same Ukiyo-e Prints in Different Languages by Exploiting Word Semantic Relationships across Languages

３．学会等名
Digital Humanities 2017（国際学会）

４．発表年
2017年

１．発表者名
李 康穎、Batjargal Biligsaikhan、前田 亮

２．発表標題
古代文字フォントの画像データに基づく手書き篆文文字の検索支援

３．学会等名
人文科学とコンピュータシンポジウム

４．発表年
2017年

１．発表者名
Biligsaikhan Batjargal

２．発表標題
Providing Bilingual Access to Early Japanese Book Databases - Utilization of Linked Open Data

３．学会等名
Open Cultural Heritage Scholarship Workshop（デジタル文化財ビジネスとオープンデータ：ローマから日本へ）（国際学会）

４．発表年
2017年

１．発表者名
李 康穎、バトジャルガル ビルゲサイハン、前田 亮

２．発表標題
白川フォントの画像データに基づく手書き篆書文字検索支援

３．学会等名
第８回横幹連合コンファレンス

４．発表年
2017年

１．発表者名
Song Yuting、Biligsaikhan Batjargal、前田 亮

２．発表標題
複数言語からなるデジタルコレクションからの同一浮世絵作品の同定手法

３．学会等名
第８回横幹連合コンファレンス

４．発表年
2017年

１．発表者名
Biligsaikhan Batjarga、前田 亮

２．発表標題
日本の人文系データベースへのバイリンガル並列アクセスの実現 -横断検索システムの開発-

３．学会等名
第８回横幹連合コンファレンス

４．発表年
2017年

１．発表者名 Biligsaikhan Batjargal
２．発表標題 国文学研究資料館の「新古典籍総合目録データベース」のバイリンガル化対応の試み
３．学会等名 第44回　ARCセミナー，立命館大学アート・リサーチセンター
４．発表年 2017年

〔図書〕　計1件

１．著者名 Biligsaikhan Batjargal	４．発行年 2018年
２．出版社 IntechOpen	５．総ページ数 174
３．書名 Cross-Lingual and Cross-Chronological Information Access to Multilingual Historical Documents. In Sammy Beban Chumbow, Editor, Multilingualism and Bilingualism	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関