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　２ 年　 ６ 月 １９ 日現在

機関番号：３４３１５
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ００４５２

研究課題名（和文）多言語ディジタルアーカイブにおける言語横断レコード同定手法の研究

研究課題名（英文）Research on Cross-Language Record Linkage for Multilingual Digital Archives

研究代表者
前田　亮（Maeda, Akira）

立命館大学・情報理工学部・教授

研究者番号：２０３５１３２２

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，近年急速にディジタル化が進んでいる人文系資料のディジタルアーカイブの有効活用を目的として，世界中に散在するディジタルアーカイブ間の同一実体レコードを自動的に発見し，それらをリンクさせる技術を確立することを目標として研究を行った．その成果として，多言語かつ異種のディジタルアーカイブから同一実体を表す関連レコードを発見して動的にリンクさせることで，これまで個別にアクセスするしか方法がなかった複数のディジタルアーカイブに対して統一的なアクセス手段を提供するという当初の目標を達成することができた．

研究成果の学術的意義や社会的意義
本研究の成果により，これまでデータベース間のリンクがほとんど存在せずばらばらに存在していた，世界中に散在する人文系ディジタルアーカイブが自動的に連携されることになる．また，同一の実体を表す関連レコードが集約されることにより，ある著作に対するメタデータの補完や誤りの発見にも繋がることが期待できる．これらより，これまでの人文系研究のように特定の分野や言語に閉じることなく，これらの壁を越えた，従来の人文学研究の方法論にとらわれない新たな研究手法への発展が期待でき，人文系研究の進展に貢献できると考えている．

研究成果の概要（英文）：In this research, we established a technique for automatically finding and linking records of identical entities among digital archives in the world, with the objective of effectively utilizing digital archives of humanities materials which are rapidly digitized in recent years. As a result, we achieved our research goal of providing unified access means for multiple digital archives which could only be accessible individually, by finding and dynamically linking related records of identical entities among multilingual and diverse digital archives.

研究分野：図書館情報学

キーワード：メタデータ　浮世絵　レコード同定　エンティティリンキング　情報図書館学　多言語処理　ディジタルアーカイブ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，国内外の図書館・博物館・美術館・文書館などにおいて，資料のディジタル化および公開が進んでいる．これらは，通常は各機関が個別にディジタルアーカイブとして公開を行っており，データベースによってユーザインタフェース，提供言語，メタデータスキーマなどが異なるのが通常であり，そのままでは統合利用は困難なのが現状である．一部では標準的なメタデータスキーマ（Dublin Core，CIDOC CRM など）の適用も見られるが，現状では有効に活用されているとは言い難い．

複数ディジタルアーカイブの統合利用を可能としたシステムとして，人間文化研究機構の研究資源共有化システムや，国立国会図書館サーチ，Europeana などが存在するが，これらは，基本的に，事前に統合システム側でメタデータを収集する「ハーベスティング」もしくは横断検索のための登録作業が必要であり，多大な人的コストを要するのが現状である．

一つあるいは複数のデータベース中から同一の実体を表すレコードを自動的に見つけ出す問題は 1950 年代後半から研究されている歴史の長い研究分野である．大規模なデータベースの管理や統計的な分析に不可欠の技術として，統計学，データ工学，人工知能などの分野で現在までに多くの研究が行われている．しかしながら，従来のレコード同定技術は，基本的に同言語で記述されたレコードが対象であり，別言語で記述されたレコード間の同定を扱った研究はほとんど見られない．

２．研究の目的

本研究では，近年急速にディジタル化が進んでいる人文系資料のディジタルアーカイブの有効活用を目的として，世界中に散在するディジタルアーカイブ間の同一実体レコードを自動的に発見し，それらをリンクさせる技術の開発を目指して研究を行った．ここで言う同一実体レコードとは，ある著作に対する，異なる表現形を含む別データベースのレコード（たとえば複数枚刷られた版画作品や，ある古典籍の原本・写本・版本・現代語訳・外国語訳などを含む）のことを指す．本研究では特に，ディジタルアーカイブによってメタデータの記述に用いられる言語が異なっている場合でも，同一実体を表すレコードを自動的に発見する技術，すなわち言語横断レコード同定技術の確立を目指した．

３．研究の方法

本研究では，世界中に散在するディジタルアーカイブ間の同一実体レコードを自動的に発見し，それらをリンクさせる技術の実現に向けて，主に以下の研究を行った．

- 複数ディジタルアーカイブにおける言語横断レコード同定技術の開発
- 複数ディジタルアーカイブに対する言語横断エンティティリンキング手法の開発

（1）複数ディジタルアーカイブにおける言語横断レコード同定技術の開発

本研究では，言語が異なる複数ディジタルアーカイブから同一作品を自動的に同定する手法について研究を行った．

浮世絵は木版画であるため，同一作品が複数のデータベースに所蔵されていることが多くあるが，メタデータスキーマや記述言語の違いから，同一作品を見つけ出すことは容易ではない．そこで本研究では，メタデータの特定の項目（作品の題名など）を用い，題名の音訳や英訳など，表記や言語が異なる場合であっても同一作品を自動的に見つけ出すための手法を開発した．

具体的には，比較対象のデータベースの全レコードから，まず浮世絵の作者（絵師）のメタデータを用いて同一作者に絞り込みを行う．次に，検索対象の作品名と絞り込み後の対象レコードの作品名の類似度を求め，一定以上の類似度を持つレコードを同一作品と推定する．作品名の類似度の計算手法として，以下の手法を提案した．

①　作品名の文字 n-gram と辞書・シソーラスとのマッチングに基づく手法

まず，対象の浮世絵の日本語作品名を文字 n-gram に分割する．その後，分割した全 n-gram を対訳辞書などを用いて目標言語に翻訳する．さらに，翻訳した単語に対してシソーラスを用いて類義語を取得する．最後に，取得したすべての訳語と類義語について，同定対象の作品名と単語単位でマッチングを行うことで類似度を計算する．

②　固有名詞の逆翻字（back transliteration）に基づく手法

まず，目標言語のメタデータにおいて翻字された単語の原言語への逆翻字を行うことにより，二言語の翻字単語組のリストを構築する．次に，この翻字単語組リストを用いて，原言語メタデータ中の固有名詞の抽出および翻字を行う．

③　単語分散表現（word embedding）を用いた意味的マッチングに基づく手法

まず，原言語から翻訳された各メタデータに対して，目的言語メタデータにおいてマッチングの可能性のある候補を取得する．次に，単語分散表現を用いてメタデータ間の意味的マッチングを行う．

④　言語横断型の単語分散表現を用いた意味的マッチングに基づく手法

まず，単語分散表現を用いて各言語によるメタデータをベクトルで表現する．次に，各言語のベクトル空間の間のマッピングを学習することにより，異なる言語のメタデータ類似度の計算を行う．その際，作品名に含まれる各単語の分散表現を加算したものを作品名のベクトルとして用いる手法および，起点言語と目標言語の作品名に含まれる全ての単語の組み合わせについてマッチングを行う手法の 2 種類の手法を提案した．

（2）複数ディジタルアーカイブに対する言語横断エンティティリンキング手法の開発

本研究では，ディジタルアーカイブ内のメタデータなどのテキスト中で言及されているエンティティ（実体）から，それを説明する別言語のディジタルアーカイブのレコードに自動的にリンクする言語横断エンティティリンキングの研究を行った．
提案手法は，次の手順からなる．まず，原言語の文書中からキーフレーズ候補を抽出する．次にキーフレーズの難易度を判定し，言語レベルのタグを付与する．次に，抽出した原言語のキーフレーズを目標言語に翻訳する．次に，翻訳したキーフレーズに対応する目標言語の知識ベース記事候補を抽出する．さらに，対象の原言語文書を目標言語に翻訳し，その名詞ベクトルを作成する．この名詞ベクトルと目標言語の知識ベース記事候補のコサイン類似度を計算し，類似度が最大となる記事を，原言語文書中のキーフレーズに対応する目標言語の知識ベース記事としてリンクする．

４．研究成果

（1）複数ディジタルアーカイブにおける言語横断レコード同定技術の開発

本研究で提案した言語横断レコード同定の各技術について，評価実験を行った．実験結果の概要を以下に示す．

①　作品名の文字 n-gram と辞書・シソーラスとのマッチングに基づく手法

江戸東京博物館が公開している浮世絵の日本語作品名 76 件とメトロポリタン美術館が公開している浮世絵の英語作品名 450 件を用いた実験の結果，ランキングの 1 位のみを正解とした場合の正解率で 67.05%，正解を 10 位以内とした場合の正解率で 77.63%という結果となった．

②　固有名詞の逆翻字（back transliteration）に基づく手法

江戸東京博物館が公開している浮世絵の日本語作品名 2,555 件とメトロポリタン美術館が公開している浮世絵の英語作品名 3,408 件を用いた実験の結果，ベースラインとなる日本語形態素解析器を用いた手法と比較して，適合率で 10.11 ポイント，再現率で 17.12 ポイントの向上が見られた．

③　単語分散表現（word embedding）を用いた意味的マッチングに基づく手法

江戸東京博物館が公開している浮世絵の日本語作品名 203 件とメトロポリタン美術館が公開している浮世絵の英語作品名 3,398 件を用いた実験の結果，ベースラインとなる Soft-TFIDF を用いた手法と比較して，適合率で 14.29 ポイント，再現率で 13.95 ポイントの向上が見られた．

④　言語横断型の単語分散表現を用いた意味的マッチングに基づく手法

江戸東京博物館が公開している浮世絵の日本語作品名とメトロポリタン美術館が公開している浮世絵の英語作品名のうち 173 組を用いた実験の結果，ベースラインとなる浮世絵作品名を学習に使用しない場合と比較して，ランキングの 1 位のみを正解とした場合の適合率で 26.8 ポイントの向上が見られた．

（2）複数ディジタルアーカイブに対する言語横断エンティティリンキング手法の開発

キーフレーズの難易度に基づく日本語・中国語間の言語横断エンティティリンキングの実験を行った．10 件の日本語の新聞記事を対象とし，中国人の日本語学習者 18 名が選んだキーフレーズのうち，日本語能力試験 N1 レベル，N2 レベルでそれぞれ半数以上が選択したキーフレーズを正解とした．キーフレーズ抽出の評価実験の結果，既存手法と比較して F 値が N1 レベルで 35 ポイント，N2 レベルで 33 ポイントの向上が見られた．正解記事取得の評価実験では，27 件のキーフレーズ中 22 件について正しい中国語記事を取得でき，正解率は約 81%であった．最終的なキーフレーズへのリンク付与の実験では，ランキング 1 位のみを正解とした場合で約 72%，10 位以内を正解とした場合で約 95%であった．

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Yuting Song, Biligsaikhan Batjargal, and Akira Maeda	４．巻 17
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	６．最初と最後の頁 1-8
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 －

１．著者名 Xiang Song, Jialiang Zhou, Fuminori Kimura, and Akira Maeda	４．巻 2
２．論文標題 A Japanese-Chinese Cross-Language Entity Linking Method with Entity Disambiguation Based on Document Similarity	５．発行年 2016年
３．雑誌名 International Journal of Knowledge Engineering	６．最初と最後の頁 122-127
掲載論文のDOI（デジタルオブジェクト識別子） 10.18178/ijke.2016.2.3.065	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 －

〔学会発表〕　計34件（うち招待講演　2件／うち国際学会　19件）

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
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
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
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, Batjargal Biligsaikhan, 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

４．発表年
2018年

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

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（DEIM2019）

４．発表年
2019年

１．発表者名
Li Kangying, Batjargal Biligsaikhan, 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（DEIM2019）

４．発表年
2019年

１．発表者名
王 嘉韻, Batjargal Biligsaikhan, 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（DEIM2019）

４．発表年
2019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
第五十一回 日本古文書学会大会

４．発表年
2018年

１．発表者名
前田 亮

２．発表標題
日本文化資源デジタルアーカイブへの多言語情報アクセス技術

３．学会等名
「アジア芸術学」の創成　国際ワークショップ／ 東アジア文化研究のフロンティア

４．発表年
2019年

１．発表者名
Akira Maeda

２．発表標題
Management of Digital Database of Cultural Heritage

３．学会等名
TOR Seminar and Workshop "Teknologi Digital Dalam Pengelolaan Warisan Budaya"（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Tomoaki Urata and Akira Maeda

２．発表標題
An Entity Disambiguation Approach Based on Wikipedia for Entity Linking in Microblogs

３．学会等名
6th International Congress on Advanced Applied Informatics (IIAI AAI 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
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
Tomoaki Urata and Akira Maeda

２．発表標題
An Entity Disambiguation Approach Based on Wikipedia and Word Embeddings for Entity Linking in Microblogs

３．学会等名
International MultiConference of Engineers and Computer Scientists 2018 (IMECS2018)（国際学会）

４．発表年
2018年

１．発表者名
Song Yuting, Batjargal Biligsaikhan, 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
バトジャルガル ビルゲサイハン, 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
浦田 智昭, 前田 亮

２．発表標題
マイクロブログを対象にしたエンティティリンキングにおける語義曖昧性解消

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（DEIM2018）

４．発表年
2018年

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

２．発表標題
Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script

３．学会等名
Digital Humanities 2016（国際学会）

４．発表年
2016年

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

２．発表標題
Identifying the Same Ukiyo-e Prints from Databases in Dutch and Japanese

３．学会等名
Digital Humanities 2016（国際学会）

４．発表年
2016年

１．発表者名
Tomoaki Urata and Akira Maeda

２．発表標題
Entity Linking of Artists Names in Japanese Music Articles

３．学会等名
5th International Congress on Advanced Applied Informatics (IIAI AAI 2016)（国際学会）

４．発表年
2016年

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

２．発表標題
Cross-Language Record Linkage using Word Embedding driven Metadata Similarity Measurement

３．学会等名
15th International Semantic Web Conference (ISWC 2016) Posters and Demonstrations Track（国際学会）

４．発表年
2016年

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

２．発表標題
Proper Noun Recognition in Cross-Language Record Linkage by Exploiting Transliterated Words

３．学会等名
20th International Conference on Asian Language Processing (IALP 2016)（国際学会）

４．発表年
2016年

１．発表者名
Akira Maeda

２．発表標題
Towards Integrated Multilingual Access to Diverse Digital Libraries and Archives

３．学会等名
Fifth International Conference on Digital Libraries (ICDL2016)（招待講演）（国際学会）

４．発表年
2016年

１．発表者名
木村 泰典, Yuting Song, Biligsaikhan Batjargal, 木村 文則, 前田 亮

２．発表標題
異言語の浮世絵データベースにおける描写的作品名に対応した同一作品の同定手法の提案

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
2016年

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

２．発表標題
Cross-Language Record Linkage by Exploiting Semantic Matching of Textual Metadata

３．学会等名
第9回データ工学と情報マネジメントに関するフォーラム

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
王 嘉韻, Biligsaikhan Batjargal, 前田 亮, 川越 恭二, 赤間 亮

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

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

４．発表年
2019年

１．発表者名
Jialiang Zhou, Fuminori Kimura, and Akira Maeda

２．発表標題
Cross-language Entity Linking Adapting to User's Language Ability

３．学会等名
International MultiConference of Engineers and Computer Scientists 2017 (IMECS2017)（国際学会）

４．発表年
2017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
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
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 Identical Metadata Records in Different Languages

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

４．発表年
2019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

〔図書〕　計2件

１．著者名 Fuminori Kimura, Jialiang Zhou, and Akira Maeda	４．発行年 2018年
２．出版社 Springer Singapore	５．総ページ数 397
３．書名 Japanese-Chinese Cross-Language Entity Linking Adapting to User's Language Ability (In Sio-Iong Ao, Haeng Kon Kim, Oscar Castillo, Alan Hoi-Shou Chan, and Hideki Katagiri, editors, Transactions on Engineering Technologies, chapter 28, pp.383-397)	

１．著者名 Biligsaikhan Batjargal, Akira Maeda, and Ryo Akama	４．発行年 2016年
２．出版社 National Taiwan University Press	５．総ページ数 464
３．書名 Providing Bilingual Access to Multiple Japanese Humanities Databases: Text Retrieval Using English and Japanese Queries (In Jieh Hsiang, editor, Digital Humanities: Between Past, Present, and Future, pp. 351-367)	

〔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
研究分担者	バトジャルガル　ビルゲサイハン (Batjargal Biligsaikhan) (30725396)	立命館大学・衣笠総合研究機構・研究員 (34315)	