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

平成　２９　年　　６　月　１５　日現在

機関番号：　３４３１５
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３０１６６

研究課題名（和文）Research on visualization and information extraction from ancient Mongolian historical documents

研究課題名（英文）Research on visualization and information extraction from ancient Mongolian historical documents

研究代表者

バトジャルガル　ビルゲ（Batjargal, Biligsaikhan）

立命館大学・総合科学技術研究機構・研究員

研究者番号：３０７２５３９６

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：本研究では、デジタル化した古代モンゴル文字文書の固有表現の抽出方法を提案した。歴史的文書の分析に必要な時間と手間の軽減を目的とし、個人名や地名の固有表現を、サポートベクタマシンを用い抽出した。抽出した固有名詞やその他情報を用い、伝統的モンゴル文字古文書のデジタル版を作った。抽出された固有名詞、解説や字訳は、TEIガイドラインに基づきエンコードを行った。これを基に、デジタル・ヒューマニティーズ研究用Webベースプロトタイプシステムを開発した。本システムでは、伝統的モンゴル文字の原文テキストやそのラテン文字訳を表示・検索可能であり、更にハイライトされた固有表現や原文のスキャン画像を表示可能である。

研究成果の概要（英文）：In this research, we proposed a named entity extraction method for digitized ancient Mongolian documents. Named entities such as personal names and place names were extracted by employing Support Vector Machine that aims to reduce the labor-intensive analysis on historical text. Using the extracted results, we built a digital edition of a Mongolian historical manuscript written in traditional Mongolian script. The Text Encoding Initiative guidelines was adopted to encode the named entities, commentaries and transliterations. A web-based prototype was developed for digital humanities scholarship. The proposed prototype can display and search traditional Mongolian text and its transliteration in Latin letters along with the highlighted named entities and the scanned images of the source manuscript. We believe the proposed system will have a social significance for digging the hidden knowledge from ancient Mongolian historical documents that is not available in modern Mongolian documents.

研究分野：　Digital Humanities

キーワード：　historical documents　traditional Mongolian　name entity extraction　digital library　machine learning

１．研究開始当初の背景

人文系では、いくつかの歴史書類を分析し、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、もちろん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デジタル版の存在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、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デジタル版に含むのは困難である。

近日中、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必要である。歴史書類から情報抽出可能なテキストマイニング方法の提案が必要とされる。

２．研究の目的

本研究では、デジタル化された歴史的モンゴル文字書類から古代・現代語辞書を使って情報抽出方法を提案する。歴史書類の分析に必要な時間と手間を軽減することを目的とし、人物名や地名等の固有表現を、テキストマイニング手法を使用して抽出する。抽出された固有名詞、解説および字訳は、TEI ガイドラインに基づいてエンコードを行う。

これを基に、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。このシステムの固有表現を可視化できる機能、TEI エンコード化テキストおよび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の歴史書類の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) 訓練データ、テストデータセットおよび言語資源、伝統的モンゴル文字およびキリル文字辞書を作成する。
b) a の資源を使用し、固有表現抽出を行う、TEI ガイドラインに基づきタッグする。
c) 歴史的人物、古代地名、学者らのコメントや翻字を表示できる Web ベースシステムの開発。

図１では、提案する方法のメインステップおよび構造を示す。

まずは、i)古代・現代（伝統的モンゴル文字およびキリル文字）辞書および対訳コーパスや ii)注釈訓練データのような言語資源が作成される。古代および現代二言語辞書は情報抽出や固有表現抽出等の次のステップに必要な有効的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。さらに、古代歴史書類の対応コーパスから我々は古代語辞書を作る。古代歴史書類の対応コーパスは古代および現代両言語で書かれたテキストを含む。「元朝秘史」のような古代モンゴル歴史書類はインターネットで公開された現代モンゴル語の翻字の電子テキストを含む。我々は現代および古代コーパス両方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。

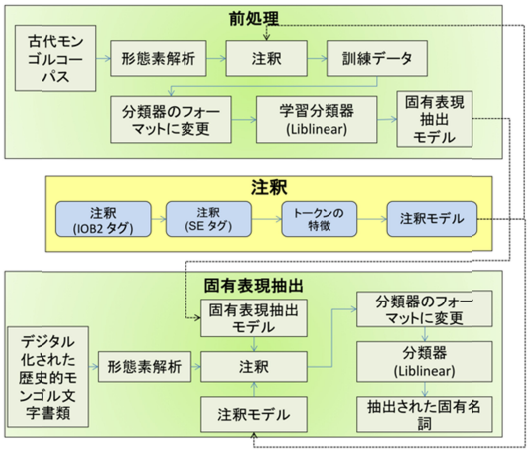

図１.提案する方法のメインステップおよび構造

これは前処理タスクからはじまる古代モンゴル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、学習のためのサポートベクタマシン（SVM）にベクトルを入力する。なぜなら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人物名のインデックスを用い、我々は「小」アルタン・トプチの全人物名を先に注釈した。

分類器のフォーマットに変更後、そのデータを学習分類器に入力し確率行列を返す（例えば、モデル）。分類器は既知のクラス（例えば、人物名）を有する訓練データで学習させる。これらの重みは確率行列に保存され、次のステップで未確認な固有表現を分類するために使用される。

次のステップでは、古代モンゴル語の文法に基づく技法と統計モデルを用いた固有表現抽出手法を提案する。提案手法では訓練データおよび伝統的モンゴル文字およびキリル文字の対訳辞書が利用される。本手法は SVM によってトークンの分類とグループ化を行う。固有表現の特徴を SVM に入力することによって、SVM の分類器は、トー

クンが固有表現になる確率を計算する。トークンの機能は、このトークンが固有表現であるかどうかに手がかりとなる可能性がある。固有表現を分類するためにいくつかの特徴が必要です。

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、伝統的な記述的表現な場合、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：*** 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).
- ***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視覚的に分離される。

最終的な作業は、古代モンゴルの歴史書類から人物名や地名等の固有表現を抽出してタグ付けすることである。抽出された固有名詞、解説および字訳は、TEI ガイドラインに基づいてエンコードを行う。固有表現のタグは<placeName>と<persName>になる。

前のタスクで得られた研究結果に基づいて、Web ベースのプロトタイプシステムを開発し、インターネット上、無料で公開する。

４．研究成果

伝統的モンゴル文字歴史書類から固有表現を抽出する提案方法を評価するため、５分割交差検証で評価実験を行い、適合率、再現率および F 値を測った。テストデータとして、モンゴル年代記である 「小」アルタン・トプチのデジタル化テキストを用いた。「小」アルタン・トプチは約 16,200 語を含む 164 ページがある。

SVM のソフトウェアとして LIBLINEAR 機械学習ライブラリーを用い学習した。(Fan, Chang, Hsieh, Wang, & Lin, 2008)。

実験の結果、SVM を用いた提案手法は適合率が 0.6993、再現率が 0.5679、F 値が 0.6268 だった。

抽出した固有名詞およびその他の情報を用い、「小」アルタン・トプチのデジタル版を作成した。抽出された固有名詞、解説および字訳は、TEI ガイドラインに基づいてエンコードを行った。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がハイライトされている。オープンソースソフトウェア Edition Visualization Technology (EVT) (Del Turco, et.al, 2014)を用い、モンゴル文字で書かれたモンゴルの歴史書類のデジタル版を作成した。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

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。

参考文献

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 Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research, 9, 1871-1874.

④ Del Turco, R. R., Buomprisco, G., Pietro, C. D., Kenny, J., Masotti, R., and Pugliese, J. (2014). Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital

Editions, Journal of the Text Encoding Initiative, Issue 8, Dec 2014, DOI: 10.4000/jtei.1077.

５．主な発表論文等

〔雑誌論文〕（計１件）

① Biligsaikhan Batjargal. Design and Prototype Implementation of a Federated Search System for Multiple Japanese Humanities Databases. *International Journal of Advances in Computer Science and Its Applications*、査読無、Vol. 6、No. 1、2016、pp. 43-47.

〔学会発表〕（計１８件）

① Yuting Song、Taisuke Kimura、Batjargal Biligsaikhan and Akira Maeda. Cross-Language Record Linkage by Exploiting Semantic Matching of Textual Metadata. 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（DEIM2017）、岐阜県高山市、2017年３月７日.

② 木村 泰典、Yuting Song、Biligsaikhan Batjargal、木村 文則、前田 亮. 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. 人文科学とコンピュータシンポジウム論文集、pp. 233-238、国文学研究資料館 ・ 国立国語研究所、東京都　立川市、2016年12月10日.

③ Zhang Chi、Batjargal Biligsaikhan、前田 亮. 動画コメントシステムにおける適切なドット絵の推薦手法の提案. 第24回インタラクティブシステムとソフトウェアに関するワークショップ（WISS2016）、長浜ロイヤルホテル、滋賀県　長浜市、2016年12月7日.

④ Yuting Song、Taisuke Kimura、Biligsaikhan Batjargal and Akira Maeda. Proper Noun Recognition in Cross-Language Record Linkage by Exploiting Transliterated Words. *In Proceedings of the 20th International Conference on Asian Language Processing (IALP 2016)*、pp. 83-86、Tainan、Taiwan、2016年11月22日.

⑤ Yuting Song, Taisuke Kimura, Biligsaikhan Batjargal and Akira Maeda. Cross-Language Record Linkage using Word Embedding driven Metadata Similarity Measurement. *In Proceedings of the 15th International Semantic Web Conference (ISWC2016) Posters & Demonstrations Track*、Kobe、Japan、2016年10月19日.

⑥ Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu and Akira Maeda. Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script. *In Conference Abstracts of Digital Humanities 2016*、pp. 734-735、Krakow、Poland、2016年7月13日.

⑦ Taisuke Kimura、Yuting Song、Biligsaikhan Batjargal、Fuminori Kimura and Akira Maeda. Identifying the Same Ukiyo-e Prints from Databases in Dutch and Japanese. *In Conference Abstracts of Digital Humanities 2016*、pp. 822-824、Krakow、Poland、2016年7月13日.

⑧ Yuting Song、Taisuke Kimura、Biligsaikhan Batjargal and Akira Maeda. An Approach to Build a Proper Noun Dictionary for Record Linkage across Humanities Databases in Different Languages. 第8回データ工学と情報マネジメントに関するフォーラム（DEIM2016）、ヒルトン福岡シーホーク、福岡県　福岡市 2016年3月1日.

⑨ Biligsaikhan Batjargal. Design and Prototype Implementation of a Federated Search System for Multiple Japanese Humanities Databases. *In Proceedings of the Third International Conference on Advances in Computing, Control and Networking - ACCN 2015*、pp. 80-84、Bangkok, Thailand、2015年12月28日.

⑩ 木村 泰典、Biligsaikhan Batjargal、木村 文則、前田 亮. 多言語の浮世絵データベース間における同一作品の同定手法の提案. 人文科学とコンピュータシンポジウム論文集、pp. 117-124、同志社大学京田辺校地夢告館、京都府京田辺市、2015年12月20日.

⑪ Biligsaikhan Batjargal、Akira Maeda and Ryo Akama. Providing Bilingual Access to Multiple Japanese Humanities Databases: Text Retrieval Using English and Japanese Queries. *In Proceedings of the 6th International Conference of Digital Archives and Digital Humanities (DADH2015)*、pp. 431-442、Taipei, Taiwan、2015年12月1日.

⑫ Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu and Akira Maeda. Personal Name Extraction from Mongolian Historical Documents Using Machine Learning. *In Proceedings of the 6th International Conference of Digital Archives and*

Digital Humanities (DADH2015)、pp. 419-430、Taipei, Taiwan、2015年12月1日.

⑬ Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu and Akira Maeda. An Approach to Named Entity Extraction from Mongolian Historical Documents. *In Proceedings of the International Conference on Culture and Computing (Culture and Computing 2015)*、pp. 205-206、Kyoto、Japan、2015年10月20日.

⑭ Taisuke Kimura、Biligsaikhan Batjargal、Fuminori Kimura and Akira Maeda. Finding the Same Artworks from Multiple Databases in Different Languages. *In Conference Abstracts of Digital Humanities 2015*、Sydney、Australia、2015年7月1日.

⑮ 木村 泰典、Biligsaikhan Batjargal、木村 文則、前田 亮. 言語が異なる浮世絵データベース間における同一作品の同定手法の提案. 第77回情報処理学会全国大会講演論文集、pp. 639-640、京都大学、京都府　京都市、 2015年3月18日.

⑯ 発表者：山路正憲 and Biligsaikhan Batjargal、人文系でデータベースの共同研究を管理するプラットフォーム構築について. 第4回　知識・芸術・文化情報学研究会、立命館大学梅田キャンパス（大阪府）、2015年2月7日

⑰ Biligsaikhan Batjargal、Takeo Kuyama、Fuminori Kimura and Akira Maeda. Identifying the Same Records across multiple Ukiyo-e Image Databases Using Textual Data in Different Languages. *In Proceedings of Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries (TPDL 2014)*、pp. 193-196、London, U.K.、2014年9月10日.

⑱ Biligsaikhan Batjargal、Garmaabazar Khaltarkhuu、Fuminori Kimura、and Akira Maeda. An Approach to Named Entity Extraction from Historical Documents in Traditional Mongolian Script. *In Proceedings of Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries (TPDL 2014)*、pp. 489-490、London, U.K.、2014年9月9日.

〔図書〕（計１件）

① Biligsaikhan Batjargal、Akira Maeda and Ryo Akama. 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*、464(351-367)、National Taiwan University Press、2016.

〔その他〕
ホームページ等 HP
http://www.dl.is.ritsumei.ac.jp/tmsdl/
http://www.dl.is.ritsumei.ac.jp/fessu/#lang=ja
http://www.dl.is.ritsumei.ac.jp/Shirakawa/search/

６．研究組織
(1)研究代表者
バトジャルガル ビルゲ（BATJARGAL, Biligsaikhan）
立命館大学・総合科学技術研究機構・研究員
研究者番号：３０７２５３９６