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　６ 年　 ６ 月 １９ 日現在

機関番号：３４４０６
研究種目：研究活動スタート支援
研究期間：2021～2023
課題番号：２１Ｋ２１３２３

研究課題名（和文）自然言語処理による日本列島全体における考古遺跡の時空間動態の把握

研究課題名（英文）Understanding the spatio-temporal dynamics of archaeological sites across the Japanese archipelago using natural language processing

研究代表者
坂平　文博（Fumihiro, Sakahira）

大阪工業大学・情報科学部・准教授

研究者番号：７０５７８１２９

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：本研究はマクロ的な歴史現象に関する考古学研究のために、発掘調査報告書の文章に対して自然言語処理技術を適用し、発掘調査報告書間の類似度を算出することで、類似度の高い、つまり文化的に関連する遺跡同士を抽出する手法を開発した。具体的には既存の考古学研究論文において類似している遺跡として分類されている遺跡の発掘調査報告書に対して自然言語処理を行い類似度を算出して、既存の考古学研究論文における分類との整合性を確認できた。この結果により本アプローチは、専門用語の表記揺れが多くコーパスも十分に整備されていない発掘調査報告書に対して、有効なものであることがわかった。

研究成果の学術的意義や社会的意義
本手法を開発することで考古学研究者は、遺物や遺構など複雑な要素で構成される文化の伝播を研究する際に、従来のように発掘調査報告書を1冊ずつ虱潰しに読んだうえで必要な情報かどうかを判断するという多大な労力から解放されることで、容易に長期間の広域における大局的な展開の把握が可能となる。さらに、従来では見落としていた遺跡の情報が抽出される場合もあり、再発見や再解釈につながることが期待できる。さらに、この成果は個別の研究のみならず、考古学の研究方法の変革や学術文書検索システムのアルゴリズム改良にも貢献することが期待できる。

研究成果の概要（英文）：This study developed a method to extract culturally related archaeological sites with high similarity by applying natural language processing (NLP) techniques to excavation reports and calculating the similarity between these reports. Specifically, NLP was applied to the excavation reports of sites classified as similar in existing archaeological research papers, and their similarity was calculated to verify the consistency with the classifications in the existing research papers. As a result, it was found that this approach is effective for excavation reports, which often contain many variations in the notation of technical terms and have not yet been sufficiently developed as a corpus.

研究分野：計算考古学

キーワード：自然言語処理　考古遺跡　発掘調査報告書

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

日本考古学において建築土木工事に伴い毎年 9 千件程度の発掘調査が行われており、大量の発掘調査報告書が刊行されている。このことは考古学研究における資料の豊富さを示す反面、研究者はそれぞれ形式が異なる大量の発掘調査報告書の内容を網羅することが困難になっている。その結果、考古学の研究テーマは長期間の日本列島全体を対象としたマクロ的な研究よりも、時期や地域を限定したミクロ的な研究にならざる得ない現状がある。

しかしながら、例えば、縄文から弥生への文化変容などの歴史的事象は長期間にわたる日本列島全体での漸進的な現象であるため、遺跡群総体を対象に統計解析等を用いた定量的な研究アプローチが有効であると考えらえる。ただし、そのためには、発掘調査報告書を 1 冊ずつ虱潰しに読んだうえで必要な情報かどうかを判断したうえで、データを抽出するという多大な労力が必要とされる作業が発生する。このために、このようなマクロ的な研究は特定の代表的な資料の解釈による定性的なアプローチが主であり、定量的なものについては一部を除きほとんど存在しないのが現状である。

２．研究の目的

本研究はマクロ的な歴史現象に関する考古学研究のために、発掘調査報告書の文章に対して自然言語処理技術を適用し、発掘調査報告書間の類似度を算出することで、類似度の高い、つまり文化的に関連する遺跡同士を抽出する手法を開発する。具体的には、遺跡毎の発掘調査報告書の文章そのものに対して自然言語処理を用いてベクトル化にすることによって、遺跡間の類似性を算出する方法を試みる。このように発掘調査報告書から遺跡の特徴をベクトル化することで、様々な統計解析等の手法が適用可能となるため、マクロ現象に対する定量的な研究アプローチが可能になる。

本手法を開発することで考古学研究者は、遺物や遺構など複雑な要素で構成される文化の伝播を研究する際に、従来のように発掘調査報告書を 1 冊ずつ虱潰しに読んだうえで必要な情報かどうかを判断するという多大な労力から解放されることで、容易に長期間の広域における大局的な展開の把握が可能となる。さらに、従来では見落としていた遺跡の情報が抽出される場合もあり、再発見や再解釈につながることが期待できる。さらに、この成果は個別の研究のみならず、考古学の研究方法の変革や学術文書検索システムのアルゴリズム改良にも貢献することが期待できる。

３．研究の方法

（１）検証対象

この手法の有効性の担保のためには、そもそも発掘調査報告書間の自然言語処理上の類似度が、文化的に関連する遺跡の類似度を表すことが可能かどうかを検証する必要がある。そのため、既存の考古学的研究（藤尾 2013）で分類されている板付遺跡、有田遺跡、四箇遺跡、田村遺跡の４つの遺跡を対象とした。日本における現在の考古学的研究において、遺物や遺構を包括した遺跡全体を対象としたマクロな研究は少ないが、藤尾 2013 では様々な遺物や遺構を含む多面的な観点から遺跡の特徴を包括的に概観していたうえで、４つの遺跡を分類している。そのため、本研究における検証に適している考えられるため、藤尾 2013 の分類を採用した。その分類において、四箇遺跡と田村遺跡は在来の狩猟採集民が稲作農耕化した集落である一方で、板付遺跡と有田遺跡は弥生稲作民の移住によって稲作農耕が専業化したした集落であるとしている。

以上を踏まえたうえで、具体的な検証内容は次の通りである（Sakahira et al. 2023）。

- 四箇遺跡と田村遺跡の発掘調査報告書が、自然言語処理上において他の遺跡よりも互いに類似しているかどうか
- 板付遺跡と有田遺跡の発掘調査報告書が、自然言語処理上において他の遺跡よりも互いに類似しているかどうか

（２）文書データ

全国遺跡報告総覧から板付遺跡、有田遺跡、四箇遺跡、田村遺跡の発掘調査報告書、最大 128 冊の PDF ファイルを入手し、OCR（Optical Character Recognition）処理を行ったうえでテキストデータに変換した。

（３）自然言語処理

文書間の類似度を得るために各文書の分散表現を取得した。分散表現とは単語の周辺文脈を含めて単語の意味をベクトルとして表現する方法である。本研究において、分散表現を用いる利点として、表記揺れや OCR 処理の誤認識に対応可能であるからである。例えば、表記揺れに関しては発掘報告書間で刊行された時期や作成者によって「縄文土器」「縄文式土器」「縄紋土器」などのような表記揺れが存在し、OCR の誤認識に関しては「夜臼式」を「夜目式」「夜白式」と変換してしまう場合がある。

Doc2Vec により 300 次元の分散表現を取得したうえで、類似度の指標としてコサイン類似度を用いて、四箇遺跡と田村遺跡、板付遺跡と有田遺跡のコサイン類似度が他よりも高いかどうかを検証する他よりも高いかどうかを検証した。

（４）検証方法

検証については、発掘調査報告書の特徴を考慮し、次の４つの異なるアプローチで実施した。

検証１：同一遺跡の発掘調査報告書の分冊間の関係性

発掘報告書は 1 つの遺跡に対して調査年度毎に複数の分冊で刊行されている場合がある。そこで、同一遺跡の発掘報告書の分冊のそれぞれの文書に対する自然言語処理上の類似度を算出して、分冊間の関係性を把握する。

検証２：同一遺跡毎に発掘報告書を結合した文書における遺跡間の関係性

同一遺跡における発掘報告書の分冊を結合した文書を対する自然言語処理上の類似度を算出して、遺跡間の関係性を把握する。

検証３：対象時代の記述を抽出した文書における遺跡間の関係性

今回の検証の対象となる期間である縄文時代と弥生時代の遺物や遺構に関する文章を手作業で抽出したうえで結合した文書を対する自然言語処理上の類似度を算出して、遺跡間の関係性を把握する。

検証４：ベクトル演算による対象時代の分散表現を用いた遺跡間の関係性

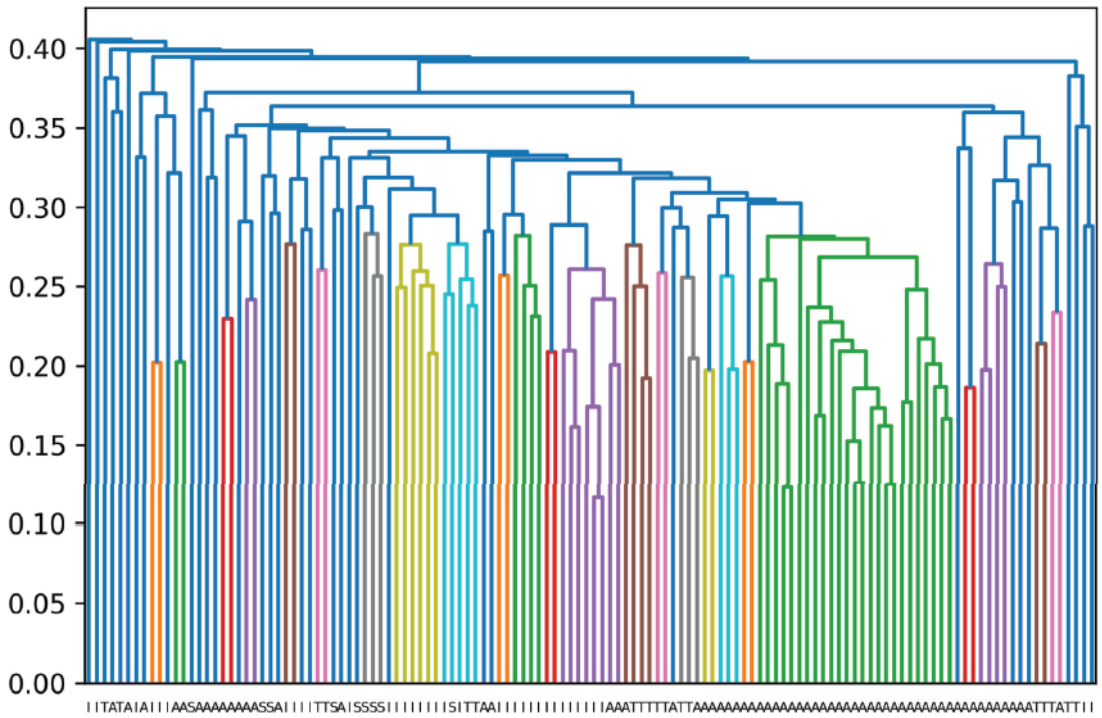
ベクトルの加法構成性を利用して、発掘報告書の全文章のベクトルから、縄文時代と弥生時代以外の遺物や遺構に関する文章のベクトルを差し引いた後の分散表現を用いて、遺跡間の関係性を把握する。

４．研究成果

（１）検証結果

検証１：同一遺跡の発掘調査報告書の分冊間の関係性

発掘調査報告書の 128 冊の分散表現に基づくテンドログラムは以下の図の通りである。同一遺跡の発掘調査報告書の分冊がクラスタ内に混在していることがわかる。つまり、同一遺跡であっても発掘調査報告書の各分冊は、自然言語処理上は類似関係にないことがわかった。



I：板付遺跡、T：田村遺跡、S：四箇遺跡、A：有田遺跡

検証２：同一遺跡毎に発掘報告書を結合した文書における遺跡間の関係性

同一遺跡毎に発掘報告書を結合した文書に対する分散表現を用いた遺跡間のコサイン類似度は以下の表の通りである。四箇遺跡は田村遺跡との類似度が最も高く、板付遺跡は有田遺跡の類似度が最も高くなっている。

	板付遺跡	田村遺跡	四箇遺跡	有田遺跡
板付遺跡				
田村遺跡	0.329			
四箇遺跡	0.344	0.337		
有田遺跡	0.367	0.286	0.318	

検証３：対象時代の記述を抽出した文書における遺跡間の関係性

対象時代の記述を抽出した文書における分散表現を用いた遺跡間のコサイン類似度は以下の表の通りである。類似関係は検証２と同様であったが、類似度の値は全体的に検証２よりも高くなっている。

	板付遺跡	田村遺跡	四箇遺跡	有田遺跡
板付遺跡				
田村遺跡	0.396			
四箇遺跡	0.373	0.481		
有田遺跡	0.416	0.452	0.358	

検証４：ベクトル演算による対象時代の分散表現を用いた遺跡間の関係性

ベクトル演算による対象時代の分散表現を用いた遺跡間のコサイン類似度は以下の表の通りである。類似関係は検証２や検証３と同様であったが、類似度の値は全体的にそれらよりもさらに高くなっている。

	板付遺跡	田村遺跡	四箇遺跡	有田遺跡
板付遺跡				
田村遺跡	0.524			
四箇遺跡	0.506	0.602		
有田遺跡	0.575	0.501	0.485	

以上の結果から、研究成果として次の知見が得られた。

- 発掘調査報告書の自然言語処理による遺跡間の類似度評価には、発掘調査報告書を分冊毎に別々に自然言語処理を行うよりも、同一遺跡で分冊を結合した後に自然言語処理を行った方が適している
- 特定の時代を対象とした遺跡間の類似度評価において、その時代に該当する情報の文章のみを対象に自然言語処理を行うことは、その時代を対象とする遺物や遺構に基づく遺跡の特徴と高い整合性を示す
- ベクトルの加法構成性を用いて、特定の時代を対象とした遺跡間の類似度を評価することは有効である

以上の結果から、表記揺れや OCR 誤認識を含む発掘調査報告書においても、自然言語処理に基づく類似度の近さは、研究者が考える遺跡の類似度の近さを反映できる可能性が示唆された。

本手法は個別の研究のみならず、考古学研究の他の研究テーマにおける文献調査や他の人文社会科学分野での研究に応用可能である。

（２）今後の課題

今後は、他の考古学研究者が考える他の遺跡間の類似性についても検証する必要がある。なお、当初計画していた縄文から弥生への文化変容についての約 1 万年間の日本列島全体の遺跡を対象とした類似度評価は、上記の検証結果を踏まえると、当該時代の情報のみの文章もしくは分散表現を得る必要が出てきため、今後の課題とすることになった。

また、パラメータを変えると得られる分散表現は異なるために類似関係も変わるが、その最適なパラメータの問題については、今後の課題としたい。さらに、最近の大規模言語モデルの進展を考慮すると、これらを用いた検証も今後行う必要があると考える。

＜＜引用文献＞＞

藤尾慎一郎, 2014, 『弥生文化像の新構築』, 吉川弘文館.

Sakahira F., Yamaguchi Y., Terano T., 2023, Understanding Cultural Similarities of Archaeological Sites from Excavation Reports Using Natural Language Processing Technique, Journal of Advanced Computational Intelligence and Intelligent Informatics, 27(3), 394-403.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4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 Sakahira Fumihiro、Yamaguchi Yuji、Terano Takao	４．巻 27
２．論文標題 Understanding Cultural Similarities of Archaeological Sites from Excavation Reports Using Natural Language Processing Technique	５．発行年 2023年
３．雑誌名 Journal of Advanced Computational Intelligence and Intelligent Informatics	６．最初と最後の頁 394～403
掲載論文のDOI（デジタルオブジェクト識別子） 10.20965/jaciii.2023.p0394	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 Sakahira Fumihiro、Tsumura Hiroomi	４．巻 10
２．論文標題 Tipping points of ancient Japanese Jomon trade networks from social network analyses of obsidian artifacts	５．発行年 2023年
３．雑誌名 Frontiers in Physics	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fphy.2022.1015870	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 Fumihiro Sakahira, Yuji Yamaguchi, Takao Terano	４．巻 -
２．論文標題 Evaluating Cultural Similarities Extracted from Excavation Reports of Archaeological Sites with Natural Language Processing	５．発行年 2022年
３．雑誌名 Proceedings of the Fourteen Japan-China International Workshop on Information Technology and Control	６．最初と最後の頁 -
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 Sakahira Fumihiro、Hiroi U	４．巻 66
２．論文標題 Designing cascading disaster networks by means of natural language processing	５．発行年 2021年
３．雑誌名 International Journal of Disaster Risk Reduction	６．最初と最後の頁 102623～102623
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ijdrr.2021.102623	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 －

〔学会発表〕　計8件（うち招待講演　2件／うち国際学会　2件）

1．発表者名
Fumihiro Sakahira, Yuji Yamaguchi, Takao Terano

2．発表標題
Plenary Lecture: Evaluating Cultural Similarities Extracted from Excavation Reports of Archaeological Sites with Natural Language Processing

3．学会等名
The Fourteen Japan-China International Workshop on Information Technology and Control Applications（招待講演）（国際学会）

4．発表年
2022年

1．発表者名
廣井悠，坂平文博

2．発表標題
リアルタイム型災害連鎖未来予測システムの検討

3．学会等名
第17回防災計画研究発表会ー防災分野におけるオープンソースインテリジェンス（OSINT）創出と利活用推進

4．発表年
2022年

1．発表者名
坂平文博，山口雄治

2．発表標題
発掘調査報告書の自然言語処理による文化的に類似した遺跡の抽出の試み

3．学会等名
第76回日本人類学会大会・第38回日本霊長類学会大会連合大会

4．発表年
2022年

1．発表者名
岸本幹史，大澤僚也，山口雄治，坂平文博，津村宏臣

2．発表標題
文化的相転移の定量評価手法の提案

3．学会等名
日本文化財科学会第39回大会

4．発表年
2022年

１．発表者名
大澤僚也，岸本幹史，山口雄治，坂平文博，津村宏臣

２．発表標題
遺伝的アルゴリズムを応用した進化シミュレーション

３．学会等名
日本文化財科学会第39回大会

４．発表年
2022年

１．発表者名
坂平文博，廣井悠

２．発表標題
新聞記事に基づく災害因果ネットワークにおけるカスケード効果の評価

３．学会等名
日本セキュリティ・マネジメント学会第35回全国大会（招待講演）

４．発表年
2022年

１．発表者名
Fumihiro Sakahira, Hiro'omi Tsumura

２．発表標題
Social Network Analysis of Ancient Japanese Obsidian Artifacts with Reduced Sampling Bias

３．学会等名
International Conference on Computer Applications & Quantitative Methods in Archaeology（国際学会）

４．発表年
2022年

１．発表者名
津村宏臣，坂平文博，原尚幸

２．発表標題
ネオーパレオデモグラフィ創成のプロトコル

３．学会等名
考古学研究会第68回総会・研究集会

４．発表年
2022年

〔図書〕　計0件

〔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