大規模言語モデルを用いた発掘調査報告書からの考古学情報抽出

山本湧大 [1]　武内樹治 [2]　大内啓樹 [3,4]　高田祐一 [2]
[1] 早稲田大学 [2] 奈良文化財研究所 [3] 奈良先端科学技術大学院大学 [4] 理化学研究所
asasa@suou.waseda.jp　takeuchi-m2m@nich.go.jp
hiroki.ouchi@is.naist.jp　takata-y23@nich.go.jp

概要

遺跡の発掘調査後に作成される発掘調査報告書には、出土した遺物や遺構をはじめとする重要な考古学情報が含まれている。しかし、全国の発掘調査報告書を合算すると、その数は膨大であり、人手で全てを読み解くことは困難である。そこで本研究では、発掘調査報告書から自動的に考古学情報に関する記述（出土した遺物や遺構、それらの時代や数など）を抽出する手法を検討した。まず、奈良文化財研究所が公開している発掘調査報告書の PDF を対象に、考古学情報と判断される表現に対して人手でアノテーションを施し、評価データセットを構築した。構築したデータセットに対して、ChatGPT を利用して考古学表現の抽出し、その性能を評価した。精度 (Precision) が 17 %前後、再現率 (Recall) が 30 %程度という結果となり、まだ改善の余地が大きいことが明らかになった。

1　はじめに

発掘調査報告書には、遺跡の発掘調査によって得られた成果や情報が詳細にまとめられている。例えば、調査の経緯、遺跡の位置や環境、発見された遺構や出土品、分析結果、そしてそれらに基づく考察などが含まれる。これらは、地域の図書館や博物館などで閲覧可能である。また、奈良文化財研究所の全国遺跡報告総覧[1)]では、全国の多くの報告書が PDF 化され、一般に公開されている。

こうした発掘調査報告書は膨大な量にのぼり、人力で読み切るのは困難な状態にある。推計では、その数は 10 万件弱に達するとされる [1]。また、発掘調査報告書自体も専門的な記述に終始しており、内容の把握が容易ではない。発掘調査報告書から出土した遺物や遺構、その数、時代といった考古学的に重要な情報（以降、「考古学情報」と呼ぶ）を自動抽出し、機械可読な形式に構造化できれば、計算機を用いたより高度な分析や応用への道が拓ける。

本研究では、その最初のステップとして、遺物や遺構といった考古学情報を発掘調査報告書から自動抽出することを目指す。我々はまず、奈良文化財研究所から公開されている発掘調査報

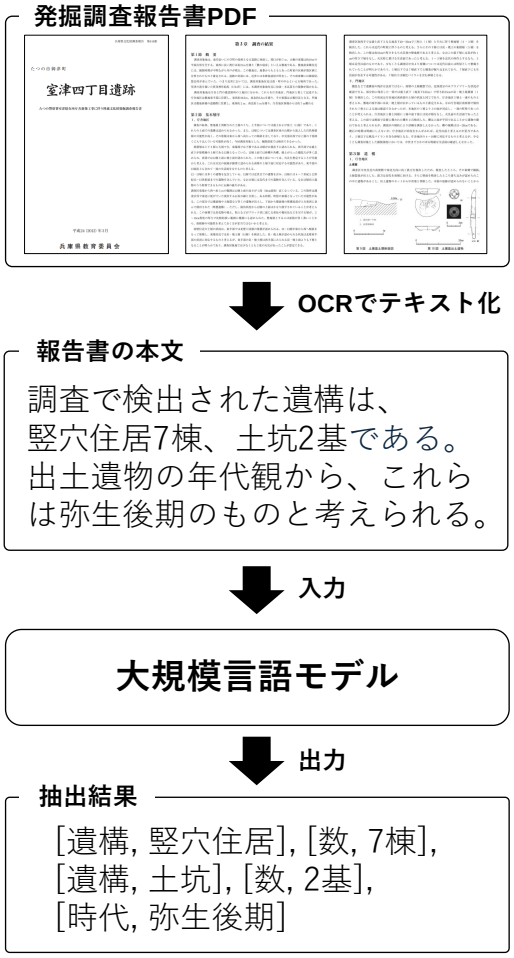

図 1: 考古学情報の抽出過程の概要

1)　https://sitereports.nabunken.go.jp/ja

告書を対象とし、各調査報告書に記載された考古学情報に関する表現に人手でアノテーションを施して評価データセットを構築した。構築したデータセットに対して、ChatGPT（GPT-4o および GPT-4o-mini）[2] の抽出精度を評価した。その結果、いずれのモデルにおいても精度 (Precision) が 17 %前後、再現率 (Recall) が 30 %程度という結果が得られ、まだ改善の余地が大きいことが明らかになった。

2 関連研究

本研究では、考古学情報の自動抽出のために、固有表現抽出（Named Entity Recognition：NER）技術を活用する。考古学分野では、発掘調査報告書をはじめとする膨大な文書データの自動処理が求められている。しかし、自然言語処理技術の導入はそれほど進んでいない状況にある。数少ない自然言語処理技術の応用例として、文書埋め込み技術（Doc2Vec [2]）を用いた研究 [3] がある。また Richards ら [4] は、大量の未刊行調査報告書を対象にテキストマイニング手法を適用し、「時代」では 0.98 の F1 値を達成し、遺物や遺構といった「物理的物体」と「場所」はそれぞれ 0.81 と 0.85、「材質」についてはあいまいさの影響により 0.63 にとどまるものの、高い処理効率と精度を得たと報告されている。Yuan ら [5] は、中国考古学に関する大規模テキストを対象として BiLSTM-CRF モデル [6] を用い、遺跡名・位置・文化類型・時代の 4 種類のエンティティを抽出した。その結果、F1 スコア 0.8787 という高い精度を達成している。さらに、BiGRU-Dual Attention モデル [7][8] を用いて、遺跡–位置・遺跡–時代・遺跡–文化といったエンティティ間の関係を抽出したところ、F1 スコア 0.8805 を得たと報告されている。

本研究と最も近いのは Brandsen ら [9] の研究である。BERT [10] をベースとした固有表現抽出モデルを構築し、考古学表現の自動抽出に取り組み、F1 値で 0.735 を報告している。Brandsen らは教師データを作成して学習したモデルを使用している一方、本研究では教師データを使用せずに最新の大規模言語モデル (LLM) でどの程度の結果が達成できるかを検証する。

2） https://openai.com/ja-JP/chatgpt/overview/

表 1: それぞれのスパンラベルの基準

ラベル	定義
遺物	発掘調査などで実際に出土した「モノ」（例：土器、石器など）。
遺構	発掘調査などで実際に見つかった「形跡・構造物」（例：溝、柱穴など）。
遺跡	歴史的な遺物・遺構が存在する場所や範囲。(例: 吉野ケ里遺跡など)
時代	遺物や遺構が属する年代や時代区分を示す（縄文時代、弥生時代、中世など）。
数	出土した遺物や遺構の「数量」を示す。
調査区	遺跡内をいくつかに分割して調査を行う場合、その区画のこと。
過去の調査の遺物	過去に行われた調査で出土した「モノ」。
過去の調査の遺構	過去に行われた調査で見つかった「形跡・構造物」。

3 評価データセット構築

本研究では、全国遺跡報告総覧の発掘調査報告書の PDF データを利用した。報告書は全部で 13 冊[3] で、PDF データは pdfminer[4] を用いてテキストデータ化した。その後、遺物 (例：土器、黒曜石製石刃) や遺構 (例：住居跡、井戸) といった考古学表現に対して人手でラベルを付与してアノテーションデータを作成した。アノテーション作業は Label Studio [11] を用いた。

3.1 解析対象箇所の抜粋

発掘調査報告書は一般的に以下の構成から成る: (1) タイトル、(2) 調査の経緯、(3) 遺跡の位置および環境、(4) 発掘された遺構および出土品の詳細、(5) 分析結果、(6) 考察、(7) 抄録（遺跡の情報が記載）。本研究のアノテーション作業では、(1) タイトル、(4) 発掘された遺構および出土品の詳細、(7) 抄録の部分を抜粋して行った。これらのセクションは遺跡の基本情報や具体的な発掘成果が集約されている。

3.2 アノテーション基準

表 1 の 8 つのラベルを定義し、該当する言語表現の文章中における開始・終了位置とラベル

3） 付録の表 4 に各報告書の詳細を記述している。
4） https://pypi.org/project/pdfminer/

表 2: 本データセットの記述統計

総文字数	224,688
総単語数	194,319
総エンティティ数	2,400
ラベルごとの数	
遺物	847
遺構	756
遺跡	80
時代	242
数	396
調査区	35
過去の調査の遺物	2
過去の調査の遺構	42
総セグメント数	481
総エンティティ数 / 総セグメント数	4.990
エンティティを一つ以上含むセグメント	290

を付与した。また、当該発掘調査にて出土した遺物や遺構の情報を知りたいので、遺物や遺構の名前が書いてあっても当該発掘調査にて出土していないものにはラベルをつけない。これは、過去の発掘調査の情報が書かれることもあるが、それらと当該発掘調査での情報とを区別する目的がある。作成したアノテーションデータの記述統計を表 2 に示す。

4 実験

構築したデータセットに対して、大規模言語モデルが遺物や遺構といった考古学的な専門情報をどの程度正確に抽出できるかを調査する。

4.1 対象システムと実験設定

システムとして、対話型大規模言語モデル ChatGPT（OpenAI Chat completions API[5]）を利用し、GPT-4o[6] および GPT-4o-mini[7] のモデルを使用し、0-shot で推論を実行した。また、各実験設定のプロンプトは Han ら [12] と片山ら [13] が場所参照表現抽出に利用したプロンプトを参考にした。プロンプトは図 2 に示す。ChatGPT のハイパーパラメータについては、すべてデフォルト値を用いた。

5) https://platform.openai.com/docs/guides/text-generation
6) https://platform.openai.com/docs/models#gpt-4o
7) https://platform.openai.com/docs/models#gpt-4o-mini

Read the given passage of an archaeological site report and find out all words/phrases that indicate the following eight types of entities:
遺物: Excavated artifacts,
遺構: Excavated remains,
遺跡: Names of archaeological sites,
時代: Eras of excavated artifacts and remains,
数: The count of excavated artifacts and remains,
調査区: Areas of archaeological investigation,
過去の調査の遺物: Artifacts excavated in previous investigations,
過去の調査の遺構: Remains excavated in previous investigations.
Answer in the format of '[“entity type 1”, “entity name 1”], [“entity type 2”, “entity name 2”], ...' in the order of appearance without any explanation. If no entity exists, then just answer ”[]”. Note that artifacts and remains that have not been excavated are also written in the passage. In that case, extract no words/phrases.
Passage:
Answer:
Sentence: (INPUT_SENTENCE)
Answer:

図 2: 実験で使用した ChatGPT へのプロンプト

4.2 実験結果

表 3 の結果（精度、再現率、F1 値）を見ると、二つのモデル（GPT-4o および GPT-4o-mini）間で大きな性能差は見られなかった。全般的に、いずれのモデルにおいても精度より再現率が高い傾向が示され、一部のラベルでは再現率が 6 割を超える結果が得られた。一方、精度が 3 割を超えたのは GPT-4o の遺物ラベルのみであり、その他のラベルに関しては精度が低かった。

なお、ラベルごとの詳細を検討すると、「時代」ラベルにおける抽出結果が最も良好であった一方、「過去の調査の遺物」および「過去の調査の遺構」ラベルに対しては正解が一つも得られなかった。これら 2 つのラベルは、過去の調査で判明した遺物や遺構といった複数の要素や文脈情報を含み得るため、その意味合いが複雑化していると考えられる。

次に、図 3 に実際の解析例を示す。図 3a では、遺物や遺構の記述を並列的に示した文において

表 3: 評価結果比較。 表中の数値のうち、0.3 を超えるものを太字で示した。

評価モデル	GPT-4o			GPT-4o mini		
	精度	再現率	F1 値	精度	再現率	F1 値
全体	0.172	**0.343**	0.229	0.176	0.286	0.218
遺物	**0.304**	0.248	0.274	0.267	0.197	0.227
遺構	0.200	**0.362**	0.258	0.215	0.299	0.250
遺跡	0.106	**0.635**	0.181	0.102	**0.676**	0.177
時代	0.249	**0.544**	**0.342**	0.232	**0.382**	0.289
数	0.209	**0.377**	0.269	0.228	**0.359**	0.279
調査区	0.039	**0.333**	0.069	0.024	0.273	0.044
過去の調査の遺物	0.000	0.000	0.000	0.000	0.000	0.000
過去の調査の遺構	0.000	0.000	0.000	0.000	0.000	0.000

今回の調査で検出された遺構は、竪穴住居7棟、古墳8基、土坑14基、陥し穴状遺構2基、溝跡11条、井戸1基、ピット65個、性格不明遺構1基である。個々の遺構については第3節以降で詳述することとし、ここでは調査区内の地形について簡単に触れておきたい。

(a) 正解スパンを見逃している例

以降の建物跡である。本宮熊堂B第20次調査で、9世紀中期の竪穴住居と重複して検出されたRB012掘立柱建物跡(9世紀中期以降と報告)を除き、住居跡と重複しているものはない。竪穴住居跡の密集しない場所に建てられ、細谷地遺跡第16。17次調査や今回の調査では、掘立柱建物跡の周辺では畝間状遺構も確認されている。畝間状遺構は、灰白色火山灰の混入から検出されることが多く、撹乱や削平などの影響から調査時に確認できなかっただけで本来もっと存在していた可能性がある。このことから、これまで指摘されてきたようにこれらの掘立柱建物跡は倉的な施設であることが、より濃く示されたと思われる。ただし、奥州市の林前Ⅱ遺跡の「集落とは異なる空間に設けられた掘立柱建物跡群」のようなものではなく、規模以外には規則性があまり見られず、集落(周辺住居)に密着した存在のもの(倉)だったと思われる。第4・5次調査では、掘立柱建物跡と周辺の竪穴住居跡の直接的な関係はないとしていたが、当初、本調査区においては軸を同じとする点、遺構ごとにおける間隔などから、2棟の掘立柱建物跡とRA073竪穴住居跡およびRA074竪穴住居跡とは同時期に

(b) 誤ってスパンを抽出した例

図 3: 本研究で作成したアノテーション（黄色）と、GPT-4o による抽出結果（紫）

の抽出漏れを示している。この例の他にも、たとえば「土器と黒曜石が出土した」という文において、土器のみを認識し、黒曜石を見落とすといった事例が確認された。図 3b では、出土していない遺物や遺構、あるいはそれらに付随する時代や数量を誤って抽出してしまう事例が多く見受けられた。本タスクが一般的な固有表現抽出とは異なり、当該調査において出土したもののみを対象としていることに起因すると考えられる。

5 おわりに

本研究では、発掘調査報告書を対象とし、当該調査において出土した遺物や遺構、さらにそれらの数量や時代に関する考古学表現の抽出を試みた。大規模言語モデル ChatGPT（GPT-4o および GPT-4o-mini）を用いて性能評価実験を行い、その性能には改善の余地が大きいことが明らかとなった。特に精度（precision）の向上が今後の課題として浮き彫りになった。原因としては、各ラベル間の文脈的な違いをモデルが適切に把握できないことや、報告書本文にしばしば含まれる「実際には出土していない」遺物や遺構を誤って抽出してしまうことが挙げられる。

今後は、few-shot 学習のプロンプトを導入し、モデルにアノテーション例を少数提示することで文脈解釈の手がかりを与え、誤抽出の削減や適合率の向上を狙いたい。具体的には、実際に出土した遺物・遺構の典型的な記述例や、誤抽出されがちな「仮定的な遺物・遺構」の例をプロンプト中に示し、どのような文脈であれば「実際に出土した」と判断できるのかをモデルに学習させる方策が考えられる。

謝辞

本研究は JSPS 科研費 JP23K24904 の助成を受けたものです.

参考文献

[1] 独立行政法人国立文化財機構奈良文化財研究所. 2024『都道府県別の発掘調査報告書総目録 全都道府県分の整理完了および公開について』, 2024.

[2] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, **Proceedings of the 31st International Conference on Machine Learning**, Vol. 32 of **Proceedings of Machine Learning Research**, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[3] F. Sakahira, Y. Yamaguchi, and T. Terano. Understanding cultural similarities of archaeological sites from excavation reports using natural language processing technique. **Journal of Advanced Computational Intelligence and Intelligent Informatics**, Vol. 27, No. 3, pp. 394–403, 2023.

[4] Julian D. Richards, Douglas Tudhope, and Andreas Vlachidis. **Text Mining in Archaeology: Extracting Information from Archaeological Reports**, pp. 240–254. June 2015.

[5] Wenjing Yuan, Lin Yang, Qing Yang, Yehua Sheng, and Ziyang Wang. Extracting spatio-temporal information from chinese archaeological site text. **ISPRS International Journal of Geo-Information**, Vol. 11, No. 3, p. 175, 2022. This article belongs to the Special Issue Machine Learning and Deep Learning in Cultural Heritage.

[6] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. **arXiv preprint arXiv:1508.01991**, 2015.

[7] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics**, Vol. 54, No. 1, pp. 207–212, August 2016.

[8] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Long Papers)**, Vol. 54, No. 1, pp. 2124–2133, August 2016.

[9] Alex Brandsen, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. Can bert dig it? named entity recognition for information retrieval in the archaeology domain. **Journal on Computing and Cultural Heritage (JOCCH)**, Vol. 15, No. 3, pp. 1–18, 2022.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Label Studio. Label studio: Open source data labeling tool. Accessed: 2024-12-27.

[12] Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. An empirical study on information extraction using large language models, 2024.

[13] 片山歩希, 東山翔平, 大内啓樹, 渡辺太郎. 歴史的日本語試料を対象とした場所参照表現抽出-「おくのほそ道」を例として-. 情報処理学会研究報告 (Web) (IPSJ Technical Report (Web)), Vol. 2023-NL-258, No. 9, p. 3, 2023.

A　発掘調査報告書の詳細情報

表 4: 発掘調査報告書の詳細情報

遺跡名	発行年	地域	対象時代	遺跡の種別	文字数	単語数
細谷地遺跡第 19・20 次	2010	岩手県	平安、近世	集落跡	16,136	13,998
駒板 3 遺跡	2009	岩手県	縄文	集落跡	16,751	14,627
飯岡才川遺跡第 7・13 次・細谷地遺跡第 12 次・矢盛遺跡第 9 次	2008	岩手県	縄文、古代、近世、近世以降	集落跡、墓域、狩猟場、墓城	35,301	30,954
合羽山遺跡	2009	埼玉県	縄文、中世	集落跡、墓跡	27,890	22,068
宮沢原下遺跡	2007	岩手県	縄文、平安	狩場跡	9,320	8,305
吉田館遺跡	2007	岩手県	縄文、古代、中世、近世、近代	集落跡、散布地、城館跡	26,896	24,454
すくも山遺跡	1998	岡山県	古墳、中世	城郭址、墓、古墳	15,710	13,994
上東遺跡	2001	岡山県	弥生〜中世	集落跡	13,393	11,965
新宮東山古墳群	1996	兵庫県	古墳前期	方墳	23,531	21,008
竹原遺跡	1999	兵庫県	奈良、平安、鎌倉、室町	集落跡	9,558	8,334
養久山墳墓群	1991	兵庫県	奈良、平安、鎌倉、室町	古墳、墓	6,650	5,431
清水遺跡	1999	兵庫県	弥生中期	集落跡	10,366	9,231
室津四丁目遺跡	2012	兵庫県	中世	集落	13,186	9,950