大規模言語モデルを用いた Few-Shot プロンプティングによる J-REIT の投資物件に関する表構造認識

土井 惟成 [1,2]　田中 麻由梨 [3]
[1] 株式会社日本取引所グループ　[2] 東京大学大学院 工学系研究科　[3] 株式会社 JPX 総研
{n-doi,may-tanaka}@jpx.co.jp

概要

本研究では，J-REIT の投資物件に関する表から情報を抽出するために，HTML 形式の表を JSON 形式の構造化データに変換する手法として，大規模言語モデルを用いた Few-Shot プロンプティングに基づく手法を提案する．実験の結果，既存の Zero-Shot プロンプティングに基づく手法よりも大幅に精度が向上することを示した．特に，同一の J-REIT の表をサンプルとした場合，変換精度は 98.70%に達し，実務への応用可能性を示唆する．

1　はじめに

不動産投資信託（Real Estate Investment Trust: REIT）は，不動産を投資対象とし，賃貸収入や売却益を投資家へ分配する仕組みをもつ投資信託であり，日本では J-REIT として広く知られている [1]．J-REIT の発行体である不動産投資法人は，上場会社と同様に有価証券報告書を開示しており，その中で保有資産としての投資物件に関する詳細な情報を表形式で記載している．有価証券報告書における J-REIT 投資物件の情報としては，賃貸借契約に関する情報や担保の有無，築年数といった，不動産分析に有用な属性が多数含まれる．その一方で，有価証券報告書における J-REIT 投資物件の情報がまとめられたセクションでは，各不動産投資法人ごとに独自のレイアウトの表で記載されており，フォーマットが統一されていないという課題がある．さらに，これらの表では，財務諸表等で広く使用されている XBRL によるタグが付与されていないため，機械的な情報抽出は容易ではない．J-REIT 投資物件に関する表の例を，図 1 及び図 2 に示す．

J-REIT 投資物件の情報は，多様な用途が考えられる．学術における用途としては，Suzuki et al.[2] は，J-REIT の投資物件の情報を用いて，長期的な観点からの J-REIT における資産入替メカニズムを明らかにしている．実務における用途としては，株式会社 JPX 総研が算出している「東証 REIT 用途別指数シリーズ」において，J-REIT を「オフィス」「住宅」「商業・物流等」の 3 つのいずれかに分類するため，有価証券報告書から J-REIT 投資物件の一覧を参照することがある [3]．このように，J-REIT 投資物件の情報は，投資判断の根拠としてのみならず，リサーチや市場分析にも活用されている．

物件の名称		新宿三井ビルディング
特定資産の種類		不動産
所在地（住居表示）		東京都新宿区西新宿二丁目１番１号
土地	地積	14,448.38㎡
	用途地域	商業地域
	所有形態	所有権100%
建物	構造	鉄骨鉄筋コンクリート造陸屋根地下３階付50階建
	延床面積	179,696.87㎡
	所有形態	所有権100%
	建築時期	1974年９月30日
	用途	事務所
取得年月日		2021年１月８日
取得価格		170,000,000,000円
信託受託者		ー
建物管理会社		三井不動産株式会社
特記事項		①本投資法人は、本物件を三井不動産株式会社に賃貸し、三井不動産株式会社は転借人にこれを転貸しています。
		②本物件には、点検・記録等の管理を実施すべきアスベストを含有する吹付け材の使用が確認されていますが、安定した状態であり、健康被害を及ぼす状態ではありません。今後は状況に応じて撤去又は封じ込め等を行っていく予定です。

図 1　J-REIT 投資物件の表の例（日本ビルファンド投資法人「有価証券報告書（内国投資証券）－第 45 期 (2023/07/01 － 2023/12/31)」の一部）

10001　あびこショッピングプラザ				
特定資産の概要		信託受益権の概要		
特定資産の種類	不動産信託受益権	信託受託者		三菱ＵＦＪ信託銀行株式会社
取得年月日（注4）	2003年3月4日及び2010年2月12日	信託期間満了日		2031年4月30日
取得価格	10,322百万円			
土地価格（構成割合）	5,984百万円（58.0%）			
建物価格（構成割合）	4,337百万円（42.0%）			
土地 所在地（注1）	千葉県我孫子市我孫子四丁目11番1号	建物	構造と階数（注2）（注4）	鉄骨鉄筋コンクリート造鉄骨造陸屋根地下1階付8階建ガーデンコート：鉄骨造陸屋根2階建
			建築時期（注2）	1994年10月25日ガーデンコート：2010年2月8日
面積（注2）	22,894.36㎡		延床面積（注2）	55,761.48㎡
用途地域（注3）	近隣商業地域		種類（注2）	店舗・駐車場等
所有・それ以外の別	所有権		所有・それ以外の別	所有権
賃貸借概況				
期末テナント数	48	期末総賃貸可能面積		41,453.36㎡
期末入居率	97.0%	期末総賃貸面積		40,189.10㎡
プロパティ・マネジメント会社	株式会社プライムプレイス	主要なテナント		イトーヨーカドー
担保設定の有無	株式会社イトーヨーカ堂に対する敷金返還債務を担保するため、本建物に抵当権が設定されています。			

図 2　J-REIT 投資物件の表の例（日本都市ファンド投資法人「有価証券報告書（内国投資証券）－第 44 期 (2023/09/01 － 2024/02/29)」の一部）

こうした背景から，JreitTableSet[1)][4] という表構造認識のためのデータセットが構築されている．この中では，J-REIT 投資物件の表について，PNG 形式の画像ファイル，HTML ファイル，手作業で作成した JSON ファイルの 3 形式で提供されている．この JSON ファイルは，HTML 形式の表について，項目名と値の対応やネスト構造を維持した，構造化データである．また，本データセットを用いた実験として，大規模言語モデル（Large Language Models: LLM）の一つである，OpenAI 社[2)]の ChatGPT を用いた，ゼロショット（Zero-Shot）プロンプティングによる，HTML 形式のテキストから JSON 形式の構造化データへの変換が挙げられる．しかしながら，この手法では，複雑なネスト構造に対応しきれず，変換精度は 43.97%に留まったと報告されている．

本研究では，上記の課題を踏まえ，J-REIT 投資物件に関する HTML 形式の表を，JSON 形式の構造化データに変換するための方法として，Few-Shot プロンプティング [5] に基づく手法を提案する．具体的には，従来の Zero-Shot の手法において問題となっていた，複雑なネスト構造やキーと値の対応を克服するため，J-REIT 投資物件の表をサンプルとして提示し，LLM に対し例示を参照しながら変換を行う．また，この実験においては，例示するサンプルを，同一の J-REIT とする場合と，異なる J-REIT とする場合の両方を評価することで，Few-Shot プロンプティングにおけるサンプルの選択が，変換精度へ与える影響について，示唆を得ることを目指す．

本稿の構成は以下のとおりである．まず，第 2 章では，本研究に関する先行研究を整理する．第 3 章では，提案手法となる Few-Shot プロンプティングの概要を示す．第 4 章では，実験設定と実験結果を示すと共にその考察を述べ，第 5 章において結論と今後の展望を述べる．

2 関連研究

2.1 不動産情報の抽出

不動産に関する情報抽出に関する研究として，ルールベースを用いた手法 [6] や，画像情報とテキストを組み合わせた情報抽出 [7] といった，様々な手法が提案されている．その中でも，光学文字認識技術と LLM を活用した，物件情報の PDF ファイルからの情報抽出 [8] においては，92%の情報抽出精度を達成した．さらに，同研究では，この抽出結果を活用することで，データ入力業務において 65%の業務時間削減が可能であることが報告されている．このことから，実業務への応用を見据えた不動産に関する情報抽出における目標の一つとして，92%超という水準が考えられる．

2.2 表構造認識のデータセット

表形式のデータにおけるセル間の関係性を理解し，表構造の認識と構造化されたデータの抽出を目指すタスクとして，表構造認識（Table Structure Recognition: TSR）が知られている．

表構造認識のためのデータセットとして，自然科学論文の PDF から表構造を抽出して作成された英語のデータセットである，PubTabNet[9]，年次報告書（Annual Report）の PDF から表構造を抽出して作成された英語のデータセットである，FinTabNet[10] が挙げられる．日本語のデータセットとしては，国立国会図書館が所蔵する著作権保護期間の満了したデジタル化資料の画像内の表データで構成される，NDLTableSet[11] がある．

また，上場会社の有価証券報告書に特化したデータセットとして，NTCIR-17 の UFO タスク [12] における，Table Data Extraction のデータセットがある．さらに，関連するデータセットとして，SIG-FIN UFO-2024 タスク [13] における，表質問応答のデータセットが挙げられる．

これらをはじめとする既存のデータセットに含まれる表は，比較的単純な形式が多く，J-REIT 投資物件のような複雑なネスト構造や幅広い情報を含む表に対応するのは困難と考えられる．

2.3 構造化データへの変換

LLM の性能向上により，LLM の応用先としては，テキストからの情報抽出だけでなく，テキストから構造化データへの変換にも応用が広がっている [14, 15, 16, 17]．Bai et al.[18] は，LLM を用いた構造化データへの変換手法として，Schema-Driven Information Extraction を提案している．同研究では，SWDE（Structured Web Data Extraction）[19] をはじめとした HTML 形式の表等を対象として，事前に定義したスキーマに基づき，LLM を用いて表から情報を抽出し，JSON 形式の構造化データを生成する手法を報告している．このことから，LLM を用いた

1） https://github.com/n-doi/JreitTableSet
2） https://openai.com/

構造化データへの変換手法として，プロンプトにサンプルを入力する手法には，一定の精度の向上が期待される．

3 提案手法

本研究では，J-REIT 投資物件が記載された HTML 形式の表を JSON 形式へ変換するタスクに対し，LLM を用いた Few-Shot プロンプティングの手法を提案する．これは，LLM に対して，変換対象である HTML 形式の表に加えて，「サンプルとなる入力とその正解」を複数例与えることで，LLM に「表構造を理解し，JSON 形式の構造化データとして出力する」手順を指示するものである．

また，本研究では，Few-Shot プロンプティングで与えるサンプルとして，以下の 2 パターンの実験を行った．

異なる J-REIT のサンプル　ターゲットと異なる不動産投資法人が開示している表をサンプルに用いる．フォーマットの違いが大きいため，精度に対する効果は限定的だが，ある程度の一般化が見込まれる．

同一の J-REIT のサンプル　ターゲットと同じ不動産投資法人が開示している別物件の表をサンプルに用いる．フォーマットの構造が似ているため，高い精度が期待できる一方，他の表への汎用性が課題となる．

また，本研究では JreitTableSet における，手作業による JSON ファイルの作成方法に準拠し，以下のルールをプロンプトへ明示的に付与した．

- キーと値をどのように判断するか
- ネスト構造の表現方法
- 改行や透明罫線によるセルの結合規則
- キーが不足する場合の暫定的なキーの扱い

4 実験

4.1 データセット

本研究では，提案手法の評価に当たって，JreitTableSet を使用した．これは，2024 年 6 月末時点で上場している全 58 件の J-REIT の有価証券報告書から 10 件ずつ抽出した，合計 580 件の J-REIT 投資物件の表で構成されている．JreitTableSet には，各 J-REIT 投資物件の表に対して，以下に述べる 3 種類の形式のデータが含まれている．

PNG 形式の画像データ　PDF ファイルの有価証券報告書から手作業で表領域を切り出すことで作成した画像データ．

HTML 形式のテキストデータ　HTML ファイルの有価証券報告書から，テーブルタグ（<table>）を抽出し，投資物件ごとに整理したテキストデータ．

JSON 形式の構造化データ　人手により，複数のネストにわたってキーと値の対応付けを行った構造化データ．

本実験では，JreitTableSet における HTML 形式のテキストデータと，JSPN 形式の構造化データを使用する．そして，Few-Shot プロンプティングに基づく変換手法の優位性を評価するため，その一部をサンプルとして使用し，残りを評価対象とした．具体的には，JreitTableSet には各 J-REIT に対して 10 件の J-REIT 投資物件のデータが格納されているところ，各 J-REIT から 8 件の J-REIT 投資物件の表を評価対象とし，残り 2 件は Few-Shot プロンプティングで用いるサンプルとして使用した．従って，本実験では，合計 464 件の J-REIT 投資物件のデータを用いて変換精度の評価を行った．

また，Few-Shot プロンプティングにおいて入力するサンプル数としては，1 件または 2 件の両方で実験を行った．これにより，サンプル数が精度に与える影響についても評価した．

4.2 変換手法

本実験では，以下の 3 種類の変換手法を評価する．

Zero-Shot プロンプティング（ベースライン）
先行研究 [4] を踏襲し，サンプルは一切提示せずに変換を指示する．

Few-Shot プロンプティング（異なる J-REIT）
変換対象とは異なる J-REIT の表をサンプルとして提示し，その HTML データと正解の JSON 形式の構造化データの対応関係を LLM に指示する．

Few-Shot プロンプティング（同一の J-REIT）
変換対象と同一の J-REIT が開示した別物件の表をサンプルに用いる．

LLM には，OpenAI 社の ChatGPT のうち，gpt-4o-2024-08-06[3] を使用した．当該モデルには，出力を JSON 形式とすることを保証する機能が搭載されている[4]．

3） https://platform.openai.com/docs/models/gpt-4o
4） https://platform.openai.com/docs/guides/structured-outputs/introduction

表 1 実験結果

変換手法	サンプル数	Accuracy	Precision	Recall	F1 Score
Zero-Shot	0	43.77%	48.85%	49.76%	49.25%
Few-Shot（異なる J-REIT）	1	62.28%	68.98%	68.24%	68.56%
	2	62.02%	69.47%	68.55%	68.95%
Few-Shot（同一の J-REIT）	1	98.48%	98.98%	99.10%	99.03%
	2	98.70%	99.18%	99.16%	99.15%

4.3 評価方法

本実験では，先行研究 [4] に倣い，評価指標として Accuracy（精度），Precision（適合率），Recall（再現率），F1 スコアを使用する．これらの算出においては，まず，出力結果（自動変換された JSON）と正解データ（JreitTableSet に含まれる JSON）を再帰的に比較し，その一致を真陽性（TP），出力のみに存在する部分を偽陽性（FP），正解のみに存在する部分を偽陰性（FN）として集計する．それらを用いて，各指標は以下の式で定義される．

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

以上の指標を用いることで，各項目の検出および階層構造の一致の程度を総合的に評価した．

4.4 実験結果と考察

実験結果の一覧を表 1 に示す．

まず，Zero-Shot における各評価指標は，先行研究と同様に，全体的に 50%弱に留まっていることが確認された．このことは，複雑な表に対して，サンプルの提示無しで変換タスクを実施する Zero-Shot 変換の困難さを示している．

Few-Shot（異なる J-REIT）では，サンプルを 1 件または 2 件提示することで，Accuracy は 62%前後に向上し，Precision も 69%前後に改善している．ただし，サンプルを 1 件から 2 件に増やしても，各評価指標には僅かな変化しか見られなかった．また，誤認識の要因を分析すると，出力された JSON 形式の構造化データは，サンプルで入力した JSON 形式の構造化データの構造に酷似する傾向があり，期待された構造化データとはネストの構造が異なる傾向が確認された．このことは，J-REIT 間で表のフォーマットが大きく異なる場合，サンプルによる一般化が十分には促されない可能性を示唆している．

Few-Shot（同一の J-REIT）では，いずれの評価指標も 99%前後という高い数値を示している．また，サンプルが 1 件でも既に Accuracy は 98.48%に到達しており，2 件に増やしても 98.70%と僅かな上昇に留まった．このことから，同一 J-REIT 内では表のフォーマットが高い程度で似通っており，1 件のサンプルを例として与えるだけで，ほぼ完全に変換規則を把握しうるという傾向が伺える．

以上を総合すると，Few-Shot プロンプティングは Zero-Shot プロンプティングに比べて明確な優位性があるものの，異なる J-REIT 同士では 1 件または 2 件程度のサンプル提示では限界があると考えられる．逆に，同一の J-REIT の表をサンプルに用いる場合は，ごく少数のサンプルでもほぼ完全に近い変換が可能であることが示された．

5 おわりに

本研究では，J-REIT の投資物件情報に関する HTML 形式の表を JSON 形式の構造化データに変換するタスクに対し，Zero-Shot プロンプティングの課題を克服するため，LLM を用いた Few-Shot プロンプティングを提案した．JreitTableSet を用いた実験の結果，特に同一 J-REIT の表をサンプルとして提示する場合の精度は 98.70%であり，実務への応用可能性を示唆した．

ただし，J-REIT の有価証券報告書では，一つの J-REIT 投資物件に対して，複数の表や注記が付随するケースが少なくない．そのため，複数の表における情報を統合する仕組みや，脚注及び注記を含めた総合的な情報抽出の検討が必要である．今後の課題としては，複数の表を対象とした J-REIT 投資物件の情報に関する構造化データの作成方法の検討に加えて，これらを組み合わせた J-REIT 投資物件に関する大規模なデータセットの構築が考えられる．

謝辞

本稿の作成に当たっては，株式会社日本取引所グループ，株式会社東京証券取引所及び株式会社 JPX 総研のスタッフから有益なコメントを頂いた．ここに深く感謝申し上げる．

参考文献

[1] 株式会社日本取引所グループ. REIT って何？ https://www.jpx.co.jp/equities/products/reits/outline/index.html (参照 2025-01-10).

[2] Masatomo Suzuki, Seow Eng Ong, Yasushi Asami, and Chihiro Shimizu. Long-run renewal of reit property portfolio through strategic divestment. **The Journal of Real Estate Finance and Economics**, Vol. 66, No. 1, pp. 1–40, 2023.

[3] 株式会社 JPX 総研. 東証指数算出要領（東証 REIT 指数・東証 REIT 用途別指数 東証インフラファンド指数編）. https://www.jpx.co.jp/markets/indices/line-up/files/cal2_26_reit.pdf (参照 2025-01-10).

[4] 土井惟成. Jreittableset: J-reit の投資物件に関する表構造認識のためのデータセットの構築. じんもんこん 2024 論文集, 第 2024 巻, pp. 233–240, nov 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[6] Lien Vi Pham and Son Bao Pham. Information extraction for vietnamese real estate advertisements. In **2012 Fourth International Conference on Knowledge and Systems Engineering**, pp. 181–186, 2012.

[7] Emilia Apostolova and Noriko Tomuro. Combining visual and textual features for information extraction from online flyers. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1924–1929, Doha, Qatar, October 2014. Association for Computational Linguistics.

[8] 本郷槙一, 叶内晨, 齊藤佑太郎, 岩成達哉. 大規模言語モデルを用いたマイソク pdf からの情報抽出. 言語処理学会第 30 回年次大会発表論文集, pp. 2886–2890, 2024.

[9] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: Data, model, and evaluation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, **Computer Vision – ECCV 2020**, pp. 564–580, Cham, 2020. Springer International Publishing.

[10] Xinyi Zheng, et al. Global Table Extractor (gte): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 697–706, 2021.

[11] 青池亨. Ndltableset ：デジタル化資料中の表領域の構造化を目的としたデータセットの構築及び機械学習手法の検討. じんもんこん 2023 論文集, 第 2023 巻, pp. 231–236, dec 2023.

[12] 門脇一真, 木村泰知, 加藤誠, 近藤隆史, 乙武北斗. 有価証券報告書を対象とした表構造解析のためのデータセットの構築に向けて. 人工知能学会第二種研究会資料, Vol. 2023, No. FIN-030, pp. 100–105, 2023.

[13] 木村泰知, 佐藤栄作, 門脇一真, 乙武北斗. 有価証券報告書の表を対象としたコンペティションの提案. 人工知能学会第二種研究会資料, Vol. 2024, No. FIN-033, pp. 135–141, 2024.

[14] Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, 2022.

[15] Hanbum Ko, Hongjun Yang, Sehui Han, Sungwoong Kim, Sungbin Lim, and Rodrigo Hormazabal. Filling in the gaps: LLM-based structured data generation from semi-structured scientific data. In **ICML 2024 AI for Science Workshop**, 2024.

[16] Aishwarya Vijayan. A prompt engineering approach for structured data extraction from unstructured text using conversational llms. In **Proceedings of the 2023 6th International Conference on Algorithms, Computing and Artificial Intelligence**, ACAI '23, p. 183–189, New York, NY, USA, 2024. Association for Computing Machinery.

[17] Mara Schilling-Wilhelmi, Martiño Ríos-García, Sherjeel Shabih, María Victoria Gil, Santiago Miret, Christoph T. Koch, José A. Márquez, and Kevin Maik Jablonka. From text to insight: Large language models for materials science data extraction, 2024.

[18] Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, and Alan Ritter. Schema-driven information extraction from heterogeneous tables, 2024.

[19] Qiang Hao, Rui Cai, Yanwei Pang, and Lei Zhang. From one tree to a forest: a unified solution for structured web data extraction. In **Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '11, p. 775–784, New York, NY, USA, 2011. Association for Computing Machinery.