大規模言語モデルを用いた有価証券報告書の表質問応答

司龍 張引 王小天 宇津呂武仁
筑波大学大学院 システム情報工学研究群
{s2420833, s2465023, s2320811}_@_u.tsukuba.ac.jp
utsuro_@_iit.tsukuba.ac.jp

概要

本論文の目的は，有価証券報告書の表から情報を抽出するタスク（NTCIR-18 U4 Task）に参加するためのシステムの構築である．NTCIR-18 U4 タスクは，表検索サブタスクと表質問応答サブタスクに分けられる．本研究は，表質問応答サブタスクに注目した．表質問応答サブタスクでは，多くの自然言語処理タスクで優れた結果を示している最先端の大規模言語モデル（LLM）を利用し，既存の事前学習済みのモデルを上回り，最高の性能を達成することが期待される．

1 はじめに

有価証券報告書は，上場企業が金融商品取引法に基づいて作成・提出する法定書類であり，企業の財務状況や事業内容、経営方針などを投資家や株主に明らかにするためのもので，企業の透明性を高め，投資判断に役立つ情報を提供することが目的である．企業別の有価証券報告書を分析することで，企業間の比較が可能である．

有価証券報告書は通常，年度ごとに作成され，提出期限は決算終了後 3 ヶ月以内とされている．企業が提出した有価証券報告書は，金融庁の EDINET（電子開示システム）を通じて一般に公開されており，誰でも閲覧することが可能である．EDINET で公開されている有価証券報告書は，主に XBRL の形式であり，EDINET のサイトもしくは API を通じて，有価証券報告書を直接にダウンロードすることができる．有価証券報告書の XBRL ファイルにはタクソノミやインスタンスの情報が含まれており，このタクソノミを解析することで，構造的な正確さを保つデータを抽出可能である．

しかしながら，企業間の有価証券報告書を比較する際に，二つの課題があった [1]．一つ目は，有価証券報告書中の表は，全てがタクソノミが付与されているわけではないという点である．木村らの研究によると，TOPIX100 の企業の有価証券報告書は，約 18%しかタクソノミが付与されていない [2]．二つ目は，企業が独自に定義することができるタクソノミや，一部の企業のみ使用しているタクソノミが存在している点である．この二つの課題によって，タクソノミを機械的に読み，解析することでデータを抽出することが極めて困難である．また、奥山らは EDINET 上で公開された有価証券報告書中の表 8673 件を分析した結果，約 37%の表が機械判読困難で複雑な構造を持つため，これらの表からデータ抽出には新たな手法を探る必要がある [3]．

以上を踏まえて，木村らは NTCIR-18 U4 タスク[1)] を提案した．U4 タスクは，与えられた自然言語の質問を用いて，有価証券報告書の文書中から表を見つける表検索サブタスクと，表からデータを見つける表質問応答サブタスクで構成される．最終的には文書中から所望のデータを高精度で得られるようになり，企業間比較が容易になることが期待できる．本研究では，U4 タスクに参加し，機械的に抽出困難なデータを大規模言語モデル（Large Language Model 以下：LLM) を用いる手法で抽出することを提案する．

2 関連研究

有価証券報告書に対する表構造解析の研究は従来から行われている．木村らは，有価証券報告書に含まれる表や文章から構造化の情報を抽出することを目的とする UFO タスクを提案した [4]．UFO タスクは，TDE サブタスクと TTRE サブタスクから構成される．TDE サブタスクは，表からデータ抽出の前段階として，表のタイプを識別するために，セルを分類するサブタスクである．TTRE サブタスクは，有価証券報告書中の表の説明のテキストと表のセルを結びつけるサブタスクである．しかし，これらのサ

1) https://sites.google.com/view/ntcir18-u4/home

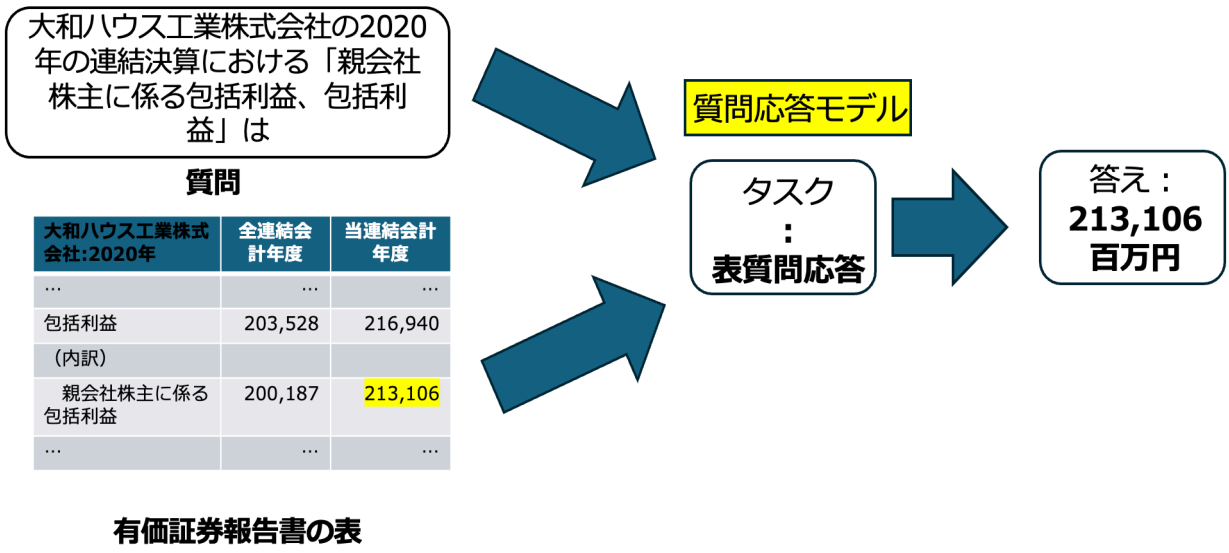

図 1　有価証券報告書中の表からの情報抽出タスク (NTCIR-18 U4 タスク) の枠組み

ブタスクのいずれも，質問に対する必要なデータを有価証券報告書から直接抽出することには言及していない．

有価証券報告書以外の表構造に関する研究も広く行われている．Pan らは大規模の表コーパスから関連する表を検索し、表のセルから質問を答えるシステムを提案し，金融文書に対する実験も行った [5]．しかし，論文で使用された表コーパスは前処理済みで英語のものであり，日本語でかつ複雑な構造を持つ有価証券報告書の表に対して同様の効果が得られるかどうかは疑問が残る．

それ以外、複数の表から所望の表を検索する研究も広く取り組まれた．Herzig らは，表の構造的特徴を解析し，それに基づく表検索手法を提案した [6]．しかし、Wang らの研究では，表構造に特化した検索手法と通常の文書検索手法を比較した結果，有意な差が認められず，特別な検索モデルを構築する必要はないとしている [7]．

また，一つの表を対象に与えられた自然言語の質問に対して，答えを導く表質問応答タスクも従来から研究されてきた．Heizig らは Encoder ベースの BERT モデルを改良し，表構造も入力として扱うことができる事前学習済みモデル TAPAS を提案した [8]．さらに，Gu らは TAPAS より一層進み，表構造だけではなく，表の操作を理解できる Encoder ベースのモデル Pasta を提案した [9]．

さらに，LLM の登場以降，さまざまな自然言語タスクにおいて従来最先端の事前学習済みモデルの精度を大きく上回り，そのため表質問応答タスクへの応用も進んでいる．Jiang らは，LLM の推論能力を活用し，知識グラフや表、データベースなどの構造化データに対する質問応答システムを開発し，事前学習済みモデルと同等の精度を達成した [10]．また，Wang らは最先端の LLM の推論手法 CoT(Chain of Thought) を表推論に拡張し，表データへの理解を特化した手法として Chain-of-Table を提案した [11]．

既存研究の多くは，事前学習済みモデルの活用に留まっている．大規模言語モデル（LLM）を用いた表質問応答の研究も散見されるが，日本語の金融文書，特に有価証券報告書を対象とした場合，同等の精度を達成できるかは未解明である．そこで本研究では，最新の LLM を用いて，有価証券報告書に含まれる表を対象とした表質問応答システムを提案し，その有効性を検証する．

3　データセット

本節では，NTCIR-18 U4 タスクで使用されるデータセットの収集方法，およびデータセットの構築の流れについて述べる．

データセットに用いる有価証券報告書は，2021 年度時点での TOPIX100 算出対象企業の 2020 年度の有価証券報告書であり，総数は 100 文書である [12]．EDINET から EDINET API v2 を用いてダウンロードした HTML と，HTML 形式を CSV 形式に変換した表を用いてデータセットを構築した．次に，CSV 形式の表を用いて質問文テンプレートを作成し

て，ChatGPT を使って自然な日本語に変換し，さらに表検索サブタスクおよび表質問応答サブタスクに適した複数の言い換えバリエーションを生成した．これにより，単一の質問文テンプレートから様々な表現の質問文が作成され，最終的に表検索タスクは質問文 14,637 個，表質問応答タスクは質問文 14,639 個が作られた．

構築されたデータセットを 7:1:2 の比率で訓練、検証、テストデータに分割した．具体的な構成 [13] は表 1 に示す．

本研究は、NTCIR-18 U4 タスクの表質問応答サブタスクに着目し，LLM を用いた表質問応答手法を提案した．

4 表質問応答タスク

4.1 入力と出力

表質問応答タスクは，質問文および所望データが含まれる表の Table ID を入力として受け取り，所望データの値もしくは Cell ID を出力とするタスクと定義されている．ここでは LLM の使用を考慮し，出力は Cell ID だけではなく回答となる値も認められた．

入力	1. 表形式データ (Table ID)
	2. 質問文
出力	データ (Cell ID or Value)
評価	Accuracy

入出力の一例は，以下に示すとおりである．

入力	1. S100ITAZ-0105020-tab135
	2. 大和ハウス工業株式会社の 2019 年度の連結決算における「その他、特別利益」を示すセルは?
出力	S100ITAZ-0105020-tab135(Cell ID) or 45468000000(値)

4.2 評価指標

与えられた質問文に対して，どの程度正確に所望データを出力できるかを調べるため，評価指標は下式の Accuracy を採用した．

$$\text{Accuracy} = \frac{\text{一致した質問の数 (}\textit{Cell ID}\text{ もしくは値)}}{\text{総質問数}}$$

4.3 LLM による表質問応答

関連研究でも述べられているように，LLM はさまざまな自然言語処理タスクにおいて非常に高い精度を達成している．したがって，与えられた質問に対する表からのデータ抽出タスクにおいても，同様に高精度な結果が期待される．そのため，今回の表質問応答タスクには LLM を使用した．

4.4 プロンプトの与え方

LLM へのプロンプトは，システムプロンプトとユーザープロンプトに分けられる．システムプロンプトは，モデルの振る舞いや応答スタイルを設定するための内部的な指示を意味する．一方で，ユーザープロンプトは，ユーザーがモデルに対して行う具体的な質問やリクエストを指し，それに基づいてモデルが応答を生成する．今回のタスクにおける LLM のシステムプロンプトとユーザープロンプトは図 2 と図 3 に示されている．LLM の出力には確率的な変動が伴うため，システムプロンプトにおいて出力形式を制限するルールを導入し，さらに，モデルの応答におけるランダム性を抑制するために，Temperature パラメータを 0 に設定した．これにより，出力の再現性と安定性を高める．

4.5 大規模言語モデルによる実験

本節では，表形式データに対する質問応答タスクにおける大規模言語モデル（LLM）の性能を評価した．具体的には，OpenAI 社の GPT シリーズ、Anthropic 社の Claude シリーズ、Google 社の Gemini シリーズ、および xAI 社の Grok シリーズの各 LLM を対象とし，1441 件の検証用質問データを用いて，その回答精度を検証した．実験では，図 2 および図 3 に示すプロンプトを各 LLM に入力し，生成された回答結果から，セル ID の正答率（Accuracy）およびセル値の正答率（Accuracy）をそれぞれ計測した．セル値の正答率については，最も高い性能を示した 8-shot 設定の結果を採用した．実験結果を表 2 に示す．

cell id と value ともに最も高い精度を達したのは，Anthropic 社の最新モデル「claude3.5-Sonnet」だった．Accuracy は両方とも 93%を達成した．

さらに，Value 予測における各モデルの性能を，0-shot、1-shot、5-shot、8-shot という異なる設定で評価した結果を，表 3 にまとめた．

表 1 U4 タスクのデータセットの訓練・検証・テストデータの文書・表・質問文の数

	文書数	表数	質問文数(表検索)	質問文数(表質問応答)
訓練データ	70	2,478	10,299	10,300
検証データ	10	320	1,441	1,441
テストデータ	20	683	2,897	2,898
合計	100	3,281	14,637	14,639

表 2 LLM の実験結果

モデル	バージョン名	Cell_id Accuracy	Value Accuracy
GPT-4o	gpt-4o-2024-11-20	0.873	0.840
Claude3.5-Sonnet	claude-3-5-sonnet-20241022	**0.939**	**0.937**
Claude3.5-Haiku	claude-3-5-haiku-20241022	0.669	0.699
Gemini-2.0	gemini-2.0-flash-exp	0.837	0.910
Gemini-1.5	gemini-1.5-flash	0.616	0.795
Grok2	grok-2-1212	0.851	0.921
Grok	grok-beta	0.865	0.908

表 3 の結果から，GPT4o を除くすべてのモデルで，shot 数が増えるにつれて精度が向上する傾向が確認された．このことから，Few-shot 学習は LLM の精度向上に有効な手法であると言える．

4.6 実験の誤り分析

表 3 は，LLM の出力例によくある誤りのパターンをまとめたものである。このうち，単位の誤りは回答の正規化を修正することである程度解決可能であると考えられる。符号誤りは，答えが損失の要素を示す場合，付与すべきマイナス記号が付与されていない誤りを指す。しかし，HTML 上ではマイナス記号が付いていないものの，解答にはマイナス記号が付与されているという状況であり，これをどのように評価するかは検討すべきである。その他の誤りは，手法やモデルの性能に起因するものであり，性能面での改善の余地がある。

5 おわりに

本研究では，有価証券報告書の表データに対して，LLM を用いた質問応答手法を提案した．予備実験においては，その精度を検証し，有用性を確認した．今後は，誤り分析を行った結果を踏まえて，LLM へのプロンプト改善や few-shot 手法による例示を活用することで，さらなる精度向上を目指す．また，検索タスクにおいても，さまざまな検索モデルを試す予定である．

参考文献

[1] 佐藤栄作, 木村泰知. 有価証券報告書に含まれるデータの企業間比較における課題について. 言語処理学会第 30 回年次大会発表論文集, pp. 885–889, 2024.

[2] 木村泰知, 近藤隆史, 門脇一真, 加藤誠. UFO: 有価証券報告書の表を対象とした情報抽出タスクの提案. 人工知能学会第二種研究会資料, Vol. 2022, No. FIN-029, pp. 32–38, 2022.

[3] 奥山和樹, 木村泰知. 有価証券報告書を対象とした機械判読が困難な表構造の分析. 言語処理学会第 30 回年次大会発表論文集, pp. 874–879, 2024.

[4] Y. Kimura, H. Ototake, K. Kadowaki, T. Kondo, and M. P. Kato. Overview of NTCIR-17 UFO task. In **Proc. 17th NTCIR**, pp. 12–15, 2023.

[5] F. Pan, M. Canim, M. Glass, A. Gliozzo, and P. Fox. CLTR: An end-to-end, transformer-based system for cell-level table retrieval and table question answering. In **Proc. 59th ACL and 11th IJCNLP**, pp. 202–209, 2021.

[6] J. Herzig, T. Müller, S. Krichene, and J. Eisenschlos. Open domain question answering over tables via dense retrieval. In **Proc. NAACL**, pp. 512–519, 2021.

[7] Z. Wang, Z. Jiang, E. Nyberg, and G. Neubig. Table retrieval may not necessitate table-specific model design. In **Proc. SUKI**, pp. 36–46, 2022.

[8] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In **Proc. 58th ACL**, pp. 4320–4333, 2020.

[9] Z. Gu, J. Fan, N. Tang, P. Nakov, X. Zhao, and X. Du. PASTA: Table-operations aware fact verification via sentence-table cloze pre-training. In **Proc. EMNLP**, pp. 4971–4983, 2022.

[10] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J. Wen. StructGPT: A general framework for large language model to reason over structured data. In **Proc. EMNLP**, pp. 9237–9251, 2023.

・システムプロンプト

指示
あなたは、有価証券報告書に詳しいプロフェッショナルです。
これから、質問文と表形式データの組が付与されます。
その表を読み取り、質問に対する回答を表の中から抽出してください。
- 回答が数値の場合は、単位を考慮した位取りを行なってください。
- 必ず回答のみを出力してください。
- 回答前のセルIDの部分も必ず出力してください。
- 複数の回答は禁止される。

以下は質問文と回答の例です
- 質問文
-- 株式会社資生堂の2019年時点における「純資産額、経営指標等」は？
- 回答の例
-- S100L0PD-0101010-tab3-r8c6:427,838百万円

・ユーザープロンプト

質問文
{質問文（データセットのquestion）}
表データ
{テキスト化した表（HTML Text）}

図 2　cell id 予測用のプロンプト文

・システムプロンプト

指示
あなたは、有価証券報告書に詳しいプロフェッショナルです。
これから、質問文と表形式データの組が付与されます。
その表を読み取り、質問に対する回答を表の中から抽出してください。
- 回答が数値の場合は、単位を考慮した位取りを行なってください。
- 必ず回答のみを出力してください。
- 回答前のセルIDの部分も必ず出力してください。
- 複数の回答は禁止される。

・ユーザープロンプト

質問文
{質問文（データセットのquestion）}
表データ
{テキスト化した表（HTML Text）}

図 3　value 予測用のプロンプト文