S2-P24

有価証券報告書を対象とした質問応答タスクのデータセット構築とLLMを用いた手法の評価

佐藤栄作 (小樽商科大学)　木村泰知 (小樽商科大学)

背景・目的

有価証券報告書に代表される金融文書には、表が多く含まれる
➡ 金融文書を対象とした表関連タスクが提案されている

FinQA（Chenら 2021）
米国企業の財務諸表を用いた、金融ドメイン特有の**複雑な数値的推論**を目的としたQAデータセット。

TAT-QA（Zhuら 2021）
米国企業の財務諸表を用いた、**テキストと表のハイブリッドデータ**に対するQAデータセット。

課題①
- そもそも表や数値を高精度に抽出することは難しい
- 金融ドメインにおける日本語のタスクやデータセットが不十分

有価証券報告書には、タクソノミとインスタンスと呼ばれる表からのデータ抽出に利用可能な情報が付与されている

課題②
- 一部（全体の18%）の表にしかタクソノミが付与されていない

有報に付与されたタクソノミを活かした表関連タスクの提案と、データセットの構築を目指す

Table Retrieval タスク
文書 から 質問文 を用いて 表 を検索・抽出
- 入力：文書（有報）と質問文
- 出力：質問の答えが含まれる表
- 評価：Accuracy

Table QA タスク
表 から 質問文 を用いて セル を検索・抽出
- 入力：有報に含まれる表と質問文
- 出力：質問の答えとなるセルの内容
- 評価：Accuracy

本研究の目的
1. 提案タスクで用いるデータセットの自動構築方法を提案する
2. 2つのタスクの難易度を明らかにして、改善点を示す

データセットの構築

使用する有報の条件
- 2021年度のTOPIX100の企業
- 2021年度提出分の**100文書**

データの分割
- **東証33業種分類の大分類**を考慮し、文書単位で分割
- **train : valid : test = 7 : 1 : 2**

作成するデータ

```
"question-train791": {
    "question": "株式会社バンダイナムコホールディングスの2020年度時点における「純資産額、経営指標等」は？",
    "doc_id": "S100ISF1",
    "table_id": "S100ISF1-0101010-tab2",
    "cell_id": "S100ISF1-0101010-tab2-r8c7",
    "value": "454684000000"
},
```

質問文以外の作成方法
- table_idとcell_idは、データセット作成者により、文書内の全ての表、セルに対して機械的に付与される。

doc_id	EDINETの書類番号を参照
table_id	HTMLの <table> タグに付与
cell_id	HTMLの <th> タグおよび <td> タグに付与
value	有報CSVの「値」列を参照

質問文の作成

タクソノミとインスタンスが記述されたCSV

要素ID	項目名	コンテキストID	相対年度	連結・個別	期間・時点	値
jpcrp_cor:NetSales...	売上高、経営指標等	Prior2YearDuration	前々期	その他	期間	379...0
jpcrp_cor:NetSales...	売上高、経営指標等	Prior1YearDuration	前期	その他	期間	414...0
jpcrp_cor:NetSales...	売上高、経営指標等	CurrentYearDuration	当期	その他	期間	438...0

大和ハウス工業の2020年度の「売上高、経営指標等」は？

- CSVから**項目名**、**コンテキスト情報**を取得して、ルールベースで連結し一文にする
- テンプレート：**{企業名}の{コンテキスト情報}における「{項目名}」は？**

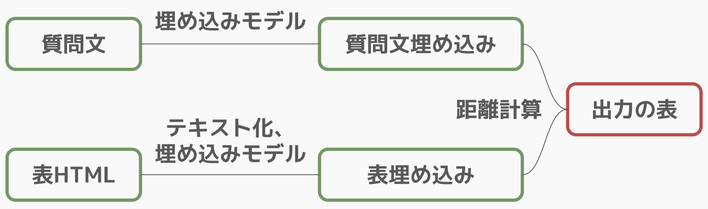
データの総数
- データの総数は**65,176件**
  タスクごとでは、約32,600件
- **validデータ**を用いて、ベースライン手法を評価

	train	valid	test	total
Table Retrieval	22,982	3,131	6,474	32,587
Table QA	22,982	3,132	6,475	32,589
total	45,964	6,263	12,949	**65,176**

ベースライン手法の評価・分析

Table Retrieval タスク

ベースライン手法
- 埋め込みモデルを用いて、**質問文 & テキスト化した表の埋め込み**を作成する。
- 質問文に最も近い距離の表を出力とする。



埋め込みモデル
- **text-embedding-3-small**（1536次元）
- **text-embedding-3-large**（3072次元）

表HTMLのテキスト化
- **Cell Text**：セルのテキストを**タブと改行で連結**
- **HTML Text**：**<table>, <tr>, <th>, <td>タグ**以外を削除し、**colspan, rowspan属性**以外も削除
- **Markdown Text**：**Markdown形式**に変換

評価

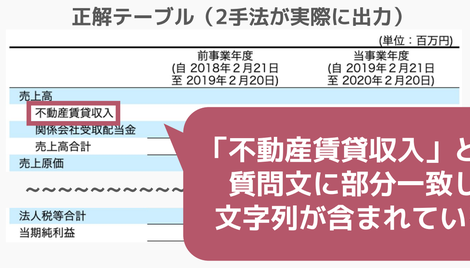
手法	Accuracy
text-embedding-3-small + Cell Text	**0.2111**
text-embedding-3-large + Cell Text	0.1833
text-embedding-3-small + HTML Text	0.1843
text-embedding-3-large + HTML Text	0.1418
text-embedding-3-small + Markdown Text	0.1233
text-embedding-3-large + Markdown Text	0.1383

- **small > large** より、埋め込み表現の次元数が重要か
- **Cell > HTML > Markdown** より表構造は性能向上に寄与しない？

出力の正例・負例

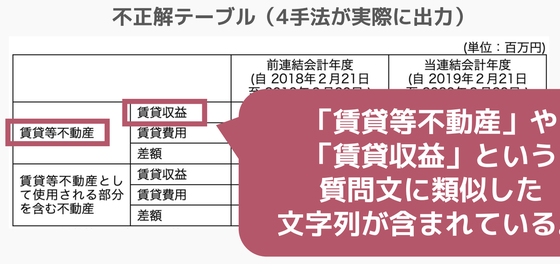
question：**株式会社ニトリホールディングスの2019年の個別決算のNonConsolidatedMemberにおける「不動産賃貸収入、営業活動による収益」は？**

正解テーブル（2手法が実際に出力）



「不動産賃貸収入」という質問文に部分一致した文字列が含まれている。

不正解テーブル（4手法が実際に出力）



「賃貸等不動産」や「賃貸収益」という質問文に類似した文字列が含まれている。

Table QA タスク

ベースライン手法
- **表テキスト化した表を、質問文とともにLLMに入力し**、Valueを出力させる。
- 表HTMLのテキストとして、表の構造を加味できるよう、**HTML Textを使用する**。

使用モデル

モデル	詳細モデル名
GPT-4o	gpt-4o-2024-05-13
GPT-3.5-turbo	gpt-3.5-turbo-0125
Gemini 1.5 Pro	gemini-1.5-pro-001
Gemini 1.5 Flash	gemini-1.5-flash-001
Claude 3 Opus	claude-3-opus-20240229
Claude 3 Haiku	claude-3-haiku-20240307
Claude 3.5 Sonnet	claude-3-5-sonnet-20240620

プロンプト
- **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熟考してください。
- 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。
```

- **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）}
```

評価

モデル	Accuracy
GPT-4o	0.6475
GPT-3.5-turbo	0.3493
Gemini 1.5 Pro	0.5744
Gemini 1.5 Flash	0.4898
Claude 3 Opus	**0.7471**
Claude 3 Haiku	0.3209
Claude 3.5 Sonnet	0.7216

モデルの性能差が如実に表れる

誤りパターン

パターン	説明
行・列誤り	誤った行や列を参照してしまう。
単位誤り	値部分はあっているものの、単位や位取りが誤っている。例えば、パーセント、倍、人、小数点以下を含む金額表現など。
符号誤り	値が損失の要素を示す場合のマイナス記号が付与されない。
ハイフン誤り	ハイフンを0と回答してしまう。
不要な要約	答えが長い文章である場合、回答を勝手に要約してしまう。
不要な情報	回答のセルを特定する過程で必要な情報を回答に含めてしまう。

本研究のTOPIX100の有報を対象とした**NTCIR-18 U4タスク**、TOPIX500の有報を対象とした**SIG-FIN UFO-2024タスク**をShared Taskとして開催中（参加募集中）です。
データセットは右のQRコードより、GitHubにて公開しています。