[S3-P11]

有価証券報告書のPDFに含まれる表を対象にした構造解析の試み

佐藤栄作 (小樽商大), 木村泰知 (小樽商大)

有価証券報告書とは

- **投資判断に有用**な情報を開示するための書類
- **PDFとXBRL**の二つの形式で公開されている
- XBRLに付与された**タクソノミ**やインスタンスの情報を利用して情報抽出が可能

XBRLから**JSON**へ変換可

しかし...

タクソノミが付与されている表は全体のおよそ**18%**にすぎない

XBRLから取得できる正解データを活かして、全ての表をJSONに変換するには？

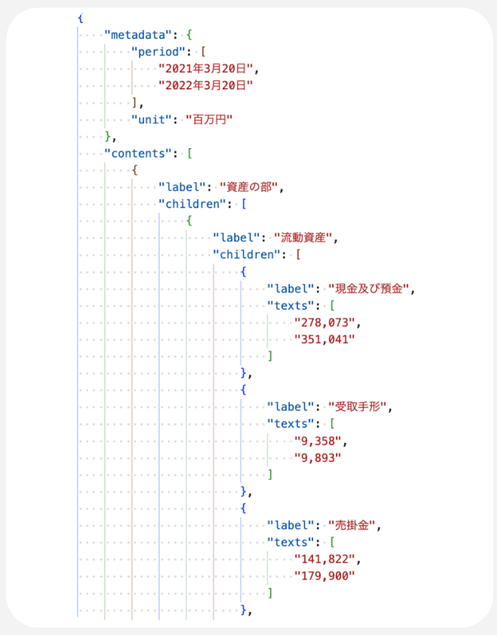
研究の目的

ChatGPTを用いて、
PDFに含まれる表をJSONに変換する

2 【財務諸表等】
(1) 【財務諸表】
①【貸借対照表】

(単位：百万円)

	前事業年度 (2021年3月20日)	当事業年度 (2022年3月20日)
資産の部		
流動資産		
現金及び預金	278,073	351,041
受取手形	9,358	9,893
売掛金	※1 141,822	※1 179,900
電子記録債権	15,816	23,280
有価証券	520,997	532,476
たな卸資産	※2 25,772	※2 49,382
その他	4,169	4,045
貸倒引当金	△16	△21
流動資産合計	995,993	1,149,999
固定資産		
有形固定資産		
建物及び構築物	4,305	4,168
工具、器具及び備品	4,134	4,264
その他	288	482
有形固定資産合計	8,728	8,915
無形固定資産		
その他	3,994	3,252
無形固定資産合計	3,994	3,252
投資その他の資産		
投資有価証券	788,061	910,216
関係会社株式	16,923	16,923
関係会社出資金	126	126
繰延税金資産	4,494	7,361
その他	1,434	1,697
貸倒引当金	△61	△61
投資その他の資産合計	810,978	936,263
固定資産合計	823,701	948,431
資産合計	1,819,694	2,098,430



ChatGPT (OpenAI API)

プロンプト等

プロンプト

```
## テーブルデータ抽出
Extract tabular data from
textualized PDF data and
output in JSON format.
Output a minimum number of tokens.
The JSON format is as follows.

 (
  "metadata": (
      "period": "point in time",
      "unit": "unit of money",
  ),
  "contents": (
      "label": "item name",
      "texts": "data",
      "children": （
            (abbreviation below)
      ）,
  )
)
```

タスクの指定

トークン数超過対策

出力のフォーマットを指定

モデル

GPT-4

- 2023年6月10日時点のモデル
- 同時期のGPT-3.5-Turboでは不可

temperature

0.0

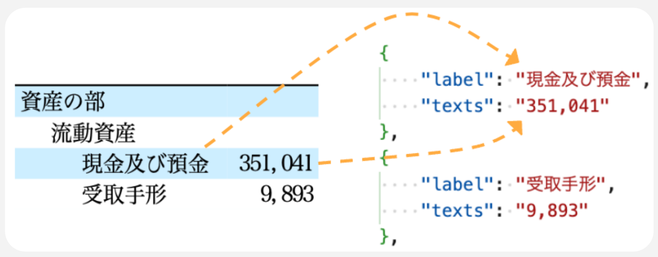
- この値が低ければ低いほど、出力のランダム性が低下

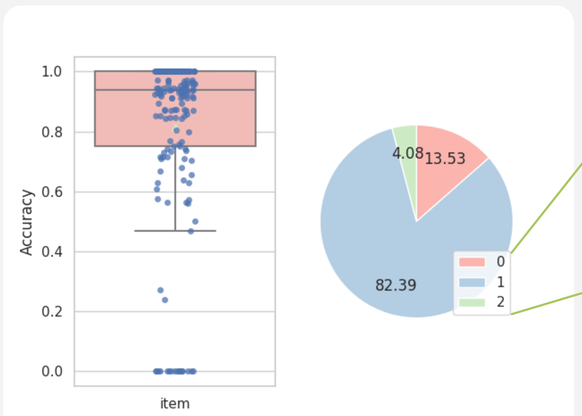
実験

- TOPIX100を構成する企業の有価証券報告書**100部**のうち、**貸借対照表**を対象として変換
- XBRLから取得したJSONを正解とし、**2種類の観点**から評価

評価1：Label-Texts関係

Label（項目名）とTexts（データ）が、どのくらい一致しているかを評価





0	LabelとTextsの両方、あるいはLabelのみが不一致
1	LabelとTextsの両方が一致
2	Textsのみ一致

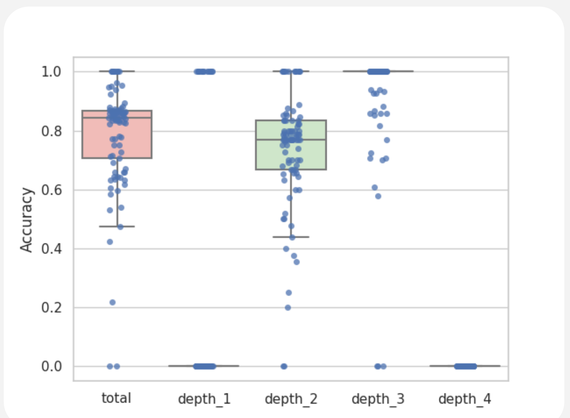
正解率のマクロ平均：0.813

評価2：親子関係

資産の部
流動資産
現金及び預金
受取手形

Parent	Child	Depth
資産	流動資産	1
流動資産	現金及び預金	2
流動資産	受取手形	2

任意の項目名とその直上の項目名が、どのくらい一致しているかを評価



Depth（階層の深さ）毎に箱ひげ図を作成

階層ごとに大きなバラつき

正解率のマクロ平均：0.784

まとめ

結果

- 項目名とデータの紐づけは、**過半数**の有価証券報告書で**90%以上**の正解率
  - ➡ 表中に「※1」「△」といった表現が存在しても、自動的に適切な変換が高精度で可能であることを示した
- 親子関係は、階層ごとに差はあるものの、**過半数**の有価証券報告書で**80%以上**の正解率
  - ➡ インデントによる視覚的な親子関係も、ある程度の再現が可能であることを示した
- XBRLを利用した、有価証券報告書内の表の大規模なデータセット構築の可能性を示した

今後の課題・取り組み

- 有価証券報告書内の他の表を対象とした**汎化性能**の検証
- LLMを用いた**表記ゆれ問題**の解決
- ChatGPTの**Functions**機能を用いた変換
- ChatGPTの**Fine-Tuning**、及びそのモデルを利用した変換
- XBRLから得られる大量の正解データを利用した、**機械学習**によるPDFからJSONへの変換