

TOPIX100 の有価証券報告書の表を対象とした  
ChatGPT による PDF から JSON への自動変換の試み  
Trial of Automatic Conversion from PDF to JSON  
using ChatGPT for the Tables of TOPIX100 Annual Security Reports

佐藤 栄作<sup>1)</sup> 木村 泰知<sup>1)</sup>  
Eisaku Sato Yasutomo Kimura

## 1 はじめに

有価証券報告書(以下、有報)は、金融商品取引所で株式を公開している会社が、事業年度ごとに外部へ公開する資料である。有報は、原則として、EDINET への電子提出が義務付けられており、PDF、および、XBRL で公開されている。有報の XBRL に付与されているタクソノミやインスタンスの情報を用いることで、必要なデータを抽出できるが、全てのデータに情報が付与されていないため、複数企業の比較分析が困難な表も存在する [1]。例えば、TOPIX100 の有報に含まれる表には、平均 222 の表 (TABLE タグ) が含まれており、企業の概況、提出会社の状況、経理の状況などの財務情報のセクションにはタグ付きの表が多いものの、財務情報以外のセクションにはタグが少ないという報告がある [2, 3]。特に、PDF に含まれる表は、機械可読な形式に変換することが困難であり、データ活用の妨げとなっている。このような状況から、複数企業の比較分析を容易にするためにも、表の構造解析を進めることが期待されている。

従来の表構造解析の研究では、表を理解するためのフレームワークとして、セルの分類、ブロックの認識、レイアウトの検出の順に分ける方法が提案されている [4]。しかしながら、このフレームワークは、最終出力までに、それぞれの段階に機械学習手法を適用する必要があるが、各段階の精度が最終結果に与える影響が大きい。この問題を解決するアプローチとしては、複数の段階を踏まずに、end-to-end で PDF から構造化されたデータへ変換する方法が考えられる。end-to-end で PDF から構造化されたデータへ変換するには、OpenAI が作成した ChatGPT を利用する方法がある。ChatGPT は、入力と出力がテキスト形式であれば、自然言語以外にも、プログラミング言語、CSV(カンマ区切りのテキスト、表形式)形式、JSON 形式なども生成可能である。

そこで、本研究では ChatGPT を用いて、PDF の表を JSON へ変換することを試みる。本研究の目的は、ChatGPT を用いて、PDF から JSON 形式への自動変換がどの程度可能であるかを明らかにすることである。本稿では、TOPIX100 に属する企業の有価証券報告書の PDF に含まれる表を対象として、JSON への自動変換した結果について述べる。

本研究の貢献は、下記の 3 つである。

- ChatGPT を用いて、PDF に含まれる「表」の構造解析に取り組んだ。具体的には、PDF の表から構造化された JSON データへ変換した。
- TOPIX100 の有報の PDF を入力として、ChatGPT により JSON を出力し、XBRL に付与されているタグを正解として、自動評価を行った。

- 自動評価は、2 つの観点 (label-texts 関係、親子関係) の評価指標を取り入れて、大規模な評価が可能であることを示した。

## 2 タスクの定義

### 2.1 タスクの概要

本タスクの目的は、PDF に含まれる表形式データを JSON 形式のデータに変換することである。すなわち、図 1 の左図で示すような PDF を入力とした際、それに含まれる表形式データの親子関係を保ったまま、図 1 の中央で示すような JSON に変換したものが出力となるシステムの構築を目指す。

PDF をテキストに変換したのち、ChatGPT を用いて JSON への変換を行う。適切に変換が行われているかの評価については、「項目名と数値がどの程度一致しているか」、「親子関係をどの程度再現できているか」という 2 つの観点から評価を行う。評価に用いるデータは、図 1 の右図で示すような XBRL から変換した JSON 形式のデータを正解データとする。

### 2.2 データセットの構築

#### 2.2.1 入力データの作成

入力データとして使用する PDF は、EDINET からダウンロード可能である。2021 年度決算かつ 2021 年 10 月 29 日時点で TOPIX100 に属する企業の有価証券報告書の PDF である。その中でも、下記の 3 つの理由から、「財務諸表等-財務諸表-貸借対照表」を対象とする。

- 正解データを作成するためのタクソノミが付与されている
- 複数ページに跨った表が存在する
- それぞれの項目が親子関係を有している

また、後述する ChatGPT の使用トークン数を抑えるため、対象範囲を「貸借対照表-資産の部」に限定した。

有価証券報告書の PDF は、一つのファイルあたり 140~200 ページ前後である。そこから対象範囲のページを抽出した結果、入力となる PDF は 1 ページあるいは 2 ページとなった。したがって、有価証券報告書の PDF 内における、貸借対照表-資産の部の部分を抽出した 100 個の PDF を入力データとして用いる。

#### 2.2.2 正解データの作成

正解データとして使用するデータは、入力データと同一箇所 (貸借対照表-資産の部) の XBRL から変換した JSON 形式のデータである。使用する XBRL は PDF と同様に、EDINET からダウンロード可能なものである。また、XBRL から JSON への変換には、xbrl-leader<sup>1)</sup>を用いた。

1) 国立大学法人 北海道国立大学機構 小樽商科大学

1) <https://github.com/teatime77/xbrl-reader>

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)
<b>資産の部</b>		
<b>流動資産</b>		
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	965,933	1,149,999
<b>固定資産</b>		
<b>有形固定資産</b>		
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
<b>無形固定資産</b>		
その他	3,994	3,252
無形固定資産合計	3,994	3,252
<b>投資その他の資産</b>		
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

入力: PDF

```
{
  "metadata": {
    "period": [
      "2021年3月20日",
      "2022年3月20日"
    ],
    "unit": "百万円"
  },
  "contents": [
    {
      "label": "資産の部",
      "children": [
        {
          "label": "流動資産",
          "children": [
            {
              "label": "現金及び預金",
              "texts": [
                "278,073",
                "351,041"
              ]
            },
            {
              "label": "受取手形",
              "texts": [
                "9,358",
                "9,893"
              ]
            },
            {
              "label": "売掛金",
              "texts": [
                "141,822",
                "179,900"
              ]
            },
            {
              "label": "電子記録債権",
              "texts": [
                "15,816",
                "23,280"
              ]
            },
            {
              "label": "有価証券",
              "texts": [
                "520,997",
                "532,476"
              ]
            },
            {
              "label": "たな卸資産",
              "texts": [
                "25,772",
                "49,382"
              ]
            },
            {
              "label": "その他",
              "texts": [
                "4,169",
                "4,045"
              ]
            }
          ]
        },
        {
          "label": "固定資産",
          "children": [
            {
              "label": "有形固定資産",
              "children": [
                {
                  "label": "建物及び構築物",
                  "texts": [
                    "4,305",
                    "4,168"
                  ]
                },
                {
                  "label": "工具、器具及び備品",
                  "texts": [
                    "4,134",
                    "4,264"
                  ]
                },
                {
                  "label": "その他",
                  "texts": [
                    "288",
                    "482"
                  ]
                }
              ]
            },
            {
              "label": "無形固定資産",
              "children": [
                {
                  "label": "その他",
                  "texts": [
                    "3,994",
                    "3,252"
                  ]
                }
              ]
            }
          ]
        },
        {
          "label": "投資その他の資産",
          "children": [
            {
              "label": "投資有価証券",
              "texts": [
                "788,061",
                "910,216"
              ]
            },
            {
              "label": "関係会社株式",
              "texts": [
                "16,923",
                "16,923"
              ]
            },
            {
              "label": "関係会社出資金",
              "texts": [
                "126",
                "126"
              ]
            },
            {
              "label": "繰延税金資産",
              "texts": [
                "4,494",
                "7,361"
              ]
            },
            {
              "label": "その他",
              "texts": [
                "1,434",
                "1,697"
              ]
            },
            {
              "label": "貸倒引当金",
              "texts": [
                "-61",
                "-61"
              ]
            }
          ]
        }
      ]
    }
  ]
}
```

出力: JSON

```
{
  "label": "資産",
  "children": [
    {
      "label": "流動資産",
      "children": [
        {
          "label": "現金及び預金",
          "texts": "351,041"
        },
        {
          "label": "受取手形",
          "texts": "9,893"
        },
        {
          "label": "売掛金",
          "texts": "179,900"
        },
        {
          "label": "電子記録債権",
          "texts": "23,280"
        },
        {
          "label": "有価証券",
          "texts": "532,476"
        },
        {
          "label": "棚卸資産",
          "texts": "49,382"
        },
        {
          "label": "その他",
          "texts": "4,045"
        }
      ]
    },
    {
      "label": "固定資産",
      "children": [
        {
          "label": "有形固定資産",
          "children": [
            {
              "label": "建物及び構築物",
              "texts": "4,168"
            },
            {
              "label": "工具、器具及び備品",
              "texts": "4,264"
            },
            {
              "label": "その他",
              "texts": "482"
            }
          ]
        },
        {
          "label": "無形固定資産",
          "children": [
            {
              "label": "その他",
              "texts": "3,252"
            }
          ]
        }
      ]
    },
    {
      "label": "投資その他の資産",
      "children": [
        {
          "label": "投資有価証券",
          "texts": "910,216"
        },
        {
          "label": "関係会社株式",
          "texts": "16,923"
        },
        {
          "label": "関係会社出資金",
          "texts": "126"
        },
        {
          "label": "繰延税金資産",
          "texts": "7,361"
        },
        {
          "label": "その他",
          "texts": "1,697"
        },
        {
          "label": "貸倒引当金",
          "texts": "-61"
        }
      ]
    }
  ]
}
```

正解: JSON

図 1 入力, 出力, 正解データの例

2.3 評価方法

本タスクでは, PDF から JSON への変換が適切に行われているかを評価するために, 「label-texts 関係の正解率」と, 「親子関係の正解率」という 2 つの観点で評価する。

2.3.1 評価 1: label-texts 関係の評価

label-texts 関係とは, 図 2 に示されるように, 項目名 (label) と数値データ (texts) の対応関係を指す。label-texts 関係の正解率を評価するにあたり, 正解データの label-texts 関係を, 出力データの label-texts 関係はどのくらい再現できているのかを算出する。評価は以下の 4 ステップからなる。

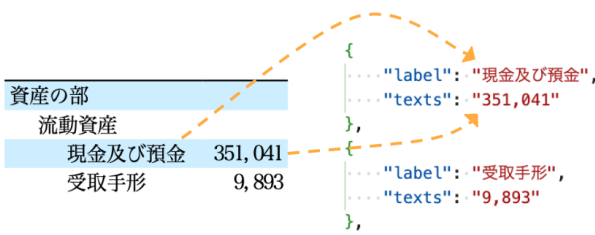


図 2 label-texts 関係の例

Step1 label-texts の対応表を作成する

正解データ, 出力データそれぞれに含まれるすべてのデータに対し, 項目名と数値データを対応させた表を作成する。

Step2 一致する texts を検索する

正解データの texts に対し, 出力データの texts に一致しているものがあるかどうかを記録する。ここでの検索対象を label ではなく texts とする理由は, 数値データである texts の方が表記揺れ, 変換ミスが少なく, 正確に比較ができると考えたためである。しかし, texts が重複する場合も考えられるため, その場合は評価の対象外とする。

Step3 タグ付けを行い, 表記揺れリストを作成する

出力データ中に一致する texts が存在しないものを 0, texts と label の両方が一致するものを 1, 一致する texts は存在するも, label が異なるものを 2 としてタグを付与する。2 のタグが付与されたペアは, label の表記揺れが懸念されるため, 図 3 に示すように, 表記揺れを考慮したリストを作成する。表記揺れの機械的判断は困難であるため, 人手で表記揺れだと認められたもの以外はリストから排除する。具体的には, 図 3 の左図下方の「流動資産合計 ※1」に含まれる米印は, 本来 label に付与されるものではないため, 右図の人手で調整したリストからは除外されている。

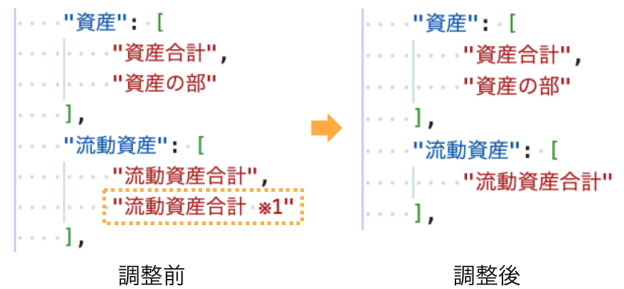


図 3 表記揺れリストの例

Step4 再びタグ付けを行い, 正解率を算出する

作成した表記揺れリストを用いて, 再び Step3 と同様のタグ付けを行う。正解率の算出には, 以下の評価式を用いる。

$$\text{Macro-Average Accuracy} = \frac{1}{N} \sum_{i=1}^N \text{Accuracy}_i$$

N は対象となる PDF のファイル数, すなわち表の総数を表し, Accuracy は表 i の正解率である。ただし, JSON への変換ができなかったクラスは除いている。

### 2.3.2 評価 2: 親子関係の評価

評価の複雑化を避けるため、ここではあるデータとその直下、あるいは直上のデータとの関係、すなわちネストの深さの差が 1 であるデータを親子関係と定義する。よって、親子関係の正解率を評価するにあたり、正解データの親子関係を、出力データの親子関係がどのくらい再現できているのかを算出する。評価は以下の 2 ステップからなる。

#### Step1 親子の組の対応表を作成する

正解データ、出力データの両方に対し、図 4 の右図に示すような、親 (parent) と子 (child) の対応表を作成する。具体的には、正解データの対応表には、親の階層の深さも併せて記録する。

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

図 4 親子関係の例

#### Step2 対応表を元に比較し、正解率を算出する

正解データの parent,child のペアに対し、出力データの parent,child のペアが一致していれば True、どちらか片方でも一致していなければ False を記録する。parent,child はどちらも項目名であるため、表記揺れにより正確な判定がなされない可能性がある。そこで、評価 1 の Step3 で作成した表記揺れのリストを活用することにより、表記揺れを考慮した判定を可能とする。

正解データの親子データのなかで、出力データの親子データと一致するものの割合を算出する。加えて、階層ごとの正解率も算出する。

$$\text{Macro-Average Accuracy} = \frac{1}{N} \sum_{i=1}^N \text{Accuracy}_i$$

N は対象となる PDF のファイル数、すなわち表の総数を表し、Accuracy は表 i の正解率である。ただし、JSON への変換ができなかったクラスは除いている。

## 3 ChatGPT による PDF から JSON への変換

本研究では、OpenAI 社が提供する ChatGPT の API を用いて、以下の流れで PDF から JSON への変換を行う。まずは PDF をテキストに変換し、次にそのテキストを ChatGPT により JSON に変換する。

### 3.1 PDF からテキストへの変換

ChatGPT への入力、現時点でテキストデータに限定されているため、PDF をテキストデータに変換する必要がある。本研究では、PDF 内の構造化データを識別し、その箇所をテキスト化することが可能である、Unstructured<sup>2)</sup>を用いて、PDF を図 5 に示すようなテキストに変換した。

対象とする 100 の PDF のうち、テキストデータへの変換が成功したのは 93 の PDF であった。その他の 7

2) <https://www.unstructured.io/>

(単位: 百万円) 前事業年度 当事業年度 (2021年3月20日) (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 Δ16 Δ21 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

図 5 PDF から変換したテキストデータの例

つの PDF は、データを得ることができなかった。原因としては、行の色分けがされていなかったことにより、Unstructured が行、ひいては構造化データを認識することができなかったためと考えられる。

### 3.2 テキストデータから JSON への変換

API の使用期間は、2023 年 6 月 10 日 03 時 50 分～12 時 10 分である。model は GPT-4 を用い、temperature は 0 に設定した。用いたプロンプトは以下の通りである。

```
## 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)
    ),
  )
)
```

変換精度の向上をねらい、one-shot となるようプロンプトを記述した。one-shot として与える例は、可能な限り抽象的にすることでプロンプトの汎化性能の向上を図った。加えて、トークン過多に起因する API のタイムアウトによりレスポンスが返らないことが多発したため、「Output a minimum number of tokens.」の一文を記述した。

また、本来であれば JSON には波括弧が用いられるが、波括弧はプロンプトの特殊文字であるため丸括弧で代用した。

### 3.3 出力結果

入力とした 100 の PDF に対し、JSON として扱うことができる形での出力に成功したのは、89 の PDF であった。出力に失敗した 11 の PDF ファイルのうち、7 つの PDF において PDF からテキストの変換に失敗していた。残りの 4 つの PDF は、テキストへの変換は成功したものの、JSON への変換に失敗していた。

## 4 実験

### 4.1 結果

表 1 にそれぞれの評価方法における実験結果を示す。評価 1 と設定した label-texts 関係の評価について、正解率のマクロ平均は 0.813 であった。

評価項目	評価 1	評価 2
Accuracy	0.813	0.784

図 6 に示すとおり、付与したラベルの割合は、0 (不一致) が 13.53%、1 (完全一致) が 82.39%、2 (texts のみ一致) が 4.08% となった。

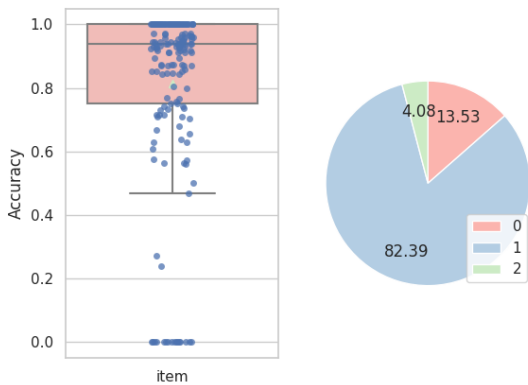


図 6 左：label-texts 関係を評価した箱ひげ図  
右：付与されたタグの割合

評価 2 と設定した親子関係の評価について、表 1 のとおり正解率のマクロ平均は 0.784 であった。図 7 では、最も左が階層の深さを考慮しなかった場合の箱ひげ図、その他は深さごとに正解率を評価した箱ひげ図が示されている。

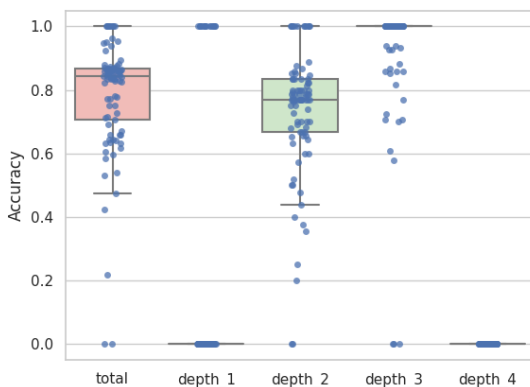


図 7 親子関係を評価した箱ひげ図

### 4.2 考察

評価 1 の結果からは、比較的高い精度で項目名と数値データの対応付けはできていると思われる。外れ値に正解率が 0 であるデータが複数見られるが、その原因はテキストから JSON に変換する際の ChatGPT の出力である。本来は不要である「※」を label、あるいは texts に含めてしまうことや、texts を文字列型ではなく数値型で変換してしまうことが原因と考えられる。そのため、トークン数との折り合いを付けながらも、「※を含めないで」や「texts は全て文字列型で」というプロンプトを加えることができれば、より安定した出力が期待できる。この部分か解決できれば、付与されたタグの割合を鑑みて、さらに 4 ポイント以上は正解率の上昇が見込める。

評価 2 の結果からは、こちらも比較的高い精度で親子関係の対応付けはできていると見られるも、第一層では正解率が二極化している。このように、上位の階層で親子関係が一致しなければ、それよりも下層のデータに辿り着くことができないため、大きな改善点だと言える。また、第四層に関しては一つのデータも一致していないため、複雑にネストした表を扱う際には課題が残る。

ただ、ChatGPT モデルのアップデートにより、推論能力や許容トークン数の向上が見られれば、単純に正解率の上昇も期待できる。加えて、今回の実験で自動化せざるを得なかった「label の表記揺れの判断」の自動化も望める。また、新たに発表された ChatGPT API のアップデートでは、テキストから構造化データを抽出する関数ができるようになるなど、ChatGPT が構造化データの抽出に活用される気運が高まりつつある。

## 5 おわりに

本稿では、TOPIX100 の有価証券報告書の表を対象として、ChatGPT による PDF から JSON への自動変換の性能評価について述べた。実験では、TOPIX100 に属する企業の有価証券報告書の PDF に含まれる表を対象として、ChatGPT により、JSON への自動変換した結果について 2 つの観点から性能評価を行った。1 つ目の性能評価は、項目名とデータの一致に対する評価であり、正解率が 0.813 であった。2 つ目の性能評価は、項目名の親子関係の抽出についての評価であり、正解率が 0.784 であった。この結果から、ChatGPT による PDF から JSON 形式への変換を利用できる可能性を示した。

### 謝辞

本研究は JSPS 科研費 21H03769、および、電気通信普及財団の助成を受けたものである。

### 参考文献

- [1] 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.
- [2] 佐藤栄作, 梶縁, 木村泰知. Topix100 の有価証券報告書に含まれる表形式データの分析. 情報処理学会 第 21 回情報科学技術フォーラム (FIT2022), 9 2022.
- [3] 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.
- [4] Jay Pujara, Arunkumar Rajendran, Majid Ghasemi-Gol, and Pedro A Szekely. A common framework for developing table understanding models. In *ISWC (Satellites)*, pp. 133–136, 2019.