

表形式データおよびテキストを対象とした根拠抽出に関する研究

代表研究者 木 村 泰 知 小樽商科大学 商学部社会情報学科 教授

1 はじめに

近年、大規模言語モデル（LLM; Large Language Model）が注目されている。大規模言語モデルは、自然言語処理分野の専門用語であるが、2022 年 11 月に OpenAI が公開した ChatGPT の普及にともないメディアでも取り上げられるようになり、一般用語になりつつある。自然言語処理は、日本語や英語のように自然発生した言語をコンピュータが人間のように操れるようになることを目指した研究・技術の総称である。ChatGPT に代表される大規模言語モデルは、自然言語の理解と生成を目的としてテキストデータから学習されたモデルを用いて、文書生成、文書要約、文章校正、質問応答などの文書生成タスクにおいて驚異的な性能を発揮している。ChatGPT の驚くべき点は、これらの複数のタスクを高い精度で出力できることに加えて、プログラミング言語の生成、CSV のような表形式の生成、JSON 形式のような半構造化テキストの生成などが可能な点である。ChatGPT の出現による自然言語処理の劇的な進歩は目覚ましく、自然言語処理研究者にとっても「ChatGPT で自然言語処理は終わるのか？」と題するパネル討論が行われるほど大きな影響を与えている¹。

しかしながら、ChatGPT により自然言語処理の全ての課題が解決されたわけではない。特に、重要な課題として、生成過程がブラックボックスであることによる説明性の課題があることや、誤情報や偽情報などの生成による安全性・信頼性の課題がある[1]。ChatGPT は、流暢な文章を生成できるため、全てを信じてしまいそうになるが、ハルシネーション（幻覚）と呼ばれる実在しないものを本当に存在するかのように生成してしまう問題がある。このような生成による誤情報や偽情報に関する問題は、フェイクニュース、あるいは、ファクトチェックという観点からも注目されている[2]。文章の自動生成技術により、意図の有無にかかわらず、誤情報が存在する状態になっている。そのなかで、正誤を判定するには、一次情報、あるいは、信頼性の高いデータを参照しながら、根拠をみつけるアプローチがある[3]。一次情報として提供される文書の代表的な例として、政府、または、地方公共団体から公開される公共データが挙げられる。

公共データは、報道や意見の根拠を示す一次情報として利用できる信憑性の高いデータである。公共データの一部は、無償で利用でき、二次利用可能なルールが適用され、オープンデータとして公開されている。オープンデータは、機械判読性が比較的高い CSV 形式(第3段階)、すなわち、表形式データとして公開されることが多い。公共データは、表形式データ以外にも、予算の説明書、条例に関する文書、あるいは、自治体の課題(予算や条例など)に関する議論を書き起こした会議録のように、テキストとして公開されている。このような、表、あるいは、テキストは、単体で扱われるよりも、Linked Open Data の考え方にあるように、相互に紐付けて情報提供されることが期待されている。しかしながら、実際には活用があまり進んでいない。その原因は、表形式データが必ずしもリレーショナルデータベースのような形式とは限らないことに加えて、列名となる項目名の表記ゆれ、あるいは、曖昧性が存在し、さらに、金額表現の省略(千円、百万円以上のみ記述)などの問題が挙げられる。また、テキストデータは、非構造化データであるため、扱いやすい構造化データへの変換が必要となるためである。これらの問題を解決する方法としては、有価証券報告書で使われている XBRL のような XML ベースの言語で記述し、各セルの値に対して「売上高」の値であるとか「純利益」の値であるとかのように、タクソノミと呼ばれる意味情報を付与することにより構造化しやすいデータにする方法がある。有価証券報告書では、貸借対照表や損益計算書のような財務情報にこれらのタクソノミが付与されている。しかしながら、それらの意味情報が付与されている表は 30%程度である。他のアプローチとしては、深層学習を用いて、表構造の理解をする方法がある。例えば、表に含まれるセルを項目名、属性名、データ、注記などに分類することで、意味構造として捉えようとしている。しかしながら、この研究アプローチは、注目されているものの、日本語を対象としたデータセットがほとんど存在しない。

そこで、本研究では日本語の「表」および「テキスト」を対象として、表の意味構造を捉えつつ、テキストと紐付けることで根拠を抽出する。本報告書では、地方議会会議録、および、有価証券報告書を対象として、データセットを構築し、評価型ワークショップを開催し、複数の研究者に提供した結果について述べる。

¹ <https://www.anlp.jp/topics/topic230601.html>

2 地方議会会議録における根拠抽出

2-1 議会会議録と予算表を紐づける Minutes-to-Budget Linking タスク

(1) 目的

地方自治体の役割のひとつに予算編成がある。予算は、首長により予算案が作成され、議会で審議された後に成立する。しかしながら、審議結果である予算は、どのような背景に基づいて予算案が作成され、どのような議論を経て成立しているのかを把握しづらい。予算作成の議論を理解するには、参照しなければならない文書や表が数多く存在する[4][5]。例えば、対象年度の予算は、前年度の予算と比較しながら、説明される[6]。特に、正確な金額を確認する場合には予算表を参照することになる。

ここで、小樽市議会における市長の説明を例に、予算に関する金額表現処理の難しさについて説明する。図1の例にあるように、市長は「歳入」となる「市税」「個人市民税」「法人市民税」「固定資産税」「都市計画税」の税収入の説明をしつつ、見込み額を述べている。「市税」「個人市民税」などの額は、文中に記載されていないものの、別の資料に記載されており、予算表を参照することで、正確な金額を確認できる。「135億 7,350 万円」という金額表現は、前年度の市税の「2.7%」である「3億 5,280 万円」の金額が増えることを見込んで計算された対象年度の見積額である。

まず、歳入についてであります。市税につきましては、個人市民税、法人市民税などで減収が見込まれるものの、固定資産税、都市計画税などで増収が見込まれることから、2.7%、3億5,280万円増の135億7,350万円を見込みました。

2018年 予算書				2019年 予算書			
款	項	金	額	款	項	金	額
1	市 税		千円	1	市 税		千円
	1 市 民 税	5,719,200			1 市 民 税	5,600,700	減
	2 固 定 資 産 税	5,367,900			2 固 定 資 産 税	5,773,100	増
	3 軽 自 動 車 税	169,400			3 軽 自 動 車 税	180,700	増
	4 た ば こ 税	924,800			4 た ば こ 税	914,500	減
	5 特 別 地 保 有 税	4,200			5 特 別 地 保 有 税	1,000	減
	6 入 湯 税	22,800			6 入 湯 税	22,000	減
	7 都 市 計 画 税	1,012,400			7 都 市 計 画 税	1,081,500	増
		13,220,700				13,573,500	

図1 予算に関する金額表現処理の難しさの例

議会における発言は、予算編成の経緯が述べられることもあるが、予算関連の書類が配布されたことを前提にしているため、詳細な数値や正確な金額が明言されないこともある。このような場合には、予算表を参照することで、金額や数値を確認することが可能である。また、予算の増減額や割合についても、予算表の金額と紐づけて、計算することができる。さらに、予算編成に至る経緯については、提案された予算金額とともに、理由を加えて説明していることが多い[6]。

そこで、本研究では、議会における予算審議の発言と予算表を紐づけることを目的としてMBLink (Minutes-to-Budget Linking) タスクのデータセットを構築し、評価型ワークショップを開催した。MBLinkは、NTCIR-17 QA-Lab PoliInfo-4²のサブタスクのひとつである。NTCIR-17は、NIIが主催する評価型ワークショップであり、2022年7月から2023年12月まで開催された³。

(2) Minutes-to-Budget Linking タスク概要

Minutes-to-Budget Linking タスクのデータセットは、人手でアノテーションをすることで作成した。自治体の公開している予算表は、PDFで公開されていることが多い。そのため、アノテーションを行う前に、PDFからHTMLへ変換する。HTMLへの変換は、変換ツールとしてABBYY FineReader PDFを用いて自動で変換した後で、人手でデータ整形をした。

² <https://sites.google.com/view/poliinfo4/>

³ <https://research.nii.ac.jp/ntcir/ntcir-17>

下記に、MBLink の入力、出力、評価について述べる。

入力

- 議会会議録中の市長の発言文 (HTML)
- 予算説明書などに含まれる表 (HTML)

出力

- 発言文に関連する予算表を紐づけた結果 (JSON)

評価

- F 値 による評価

図 2 は MBLink において識別 ID を用いてテキストと表を紐づける例である。入力されたテキストに対して関連がある 2 つの表を紐づけている。図 3 は MBLink において HTML フォーマットで提供しているデータセットの具体例である。データセットは HTML ファイルとして与えられ、出力は JSON ファイルに SentenceID に紐づけられる表をリスト形式で含める。作成したデータセットは、参加者向けに GitHub 上で公開した⁴。

入力 (HTML ファイルにIDが付与されたもの)

data-mblink-sentence-id="otaru_r01-02-sent85"

まず、歳入についてであります。市税につ
きましては、個人市民税、法人市民税など
で減収が見込まれるものの、固定資産税、
都市計画税などで増収が見込まれること
から、2.7%、3億5,280万円増の135億7,350万円を見込みました。

data-mblink-table-id="otaru_h30-r02-01-tab136"

(歳入)			
科	目	当初予算額	
		千円	千円
1	市	税	13,573,500
2	地	方	議
3	利	子	割
		交付金	22,000

data-mblink-table-id="otaru_h30-r02-01-tab137"

(歳出)			
科	目	当初予算額	
		千円	千円
1	議	会	費
			265,567
2	総	務	費
			1,732,431

data-mblink-table-id="otaru_h30-r02-01-tab139"

歳入							
科	目	本年度			前年度		
		予算額	構成比	%	当初予算額	構成比	%
		千円			千円		
1	市	税	13,573,500	23.7	13,220,700	24.0	
2	地	方	議				
		与税	327,501	0.6	314,001	0.6	

図 2 MBLink において識別 ID を用いてテキストと表を紐づける例

入力 (HTML ファイルに ID が付与されたもの)

```
<html><body>
<p data-mblink-sentence-id="otaru_r01-02-sent85">
  まず、歳入についてであります。市税につきましては、個人市民税、法人市民税などで減収が見込まれるものの、固定資産税、都市計画税などで増収が見込まれることから、2.7%、3 億 5,280 万円増の 135 億 7,350 万円を見込みました。
</p></body></html>
```

`<table border="1" data-mblink-table-id="otaru_h30-r02-01-tab136">
<tr><td>.... </td>
</tr>
</table>`

`<table border="1" data-mblink-table-id="otaru_h30-r02-01-tab139">
<tr><td>.... </td>
</tr>
</table>`

出力 (提出ファイル: ID を紐付けた JSON ファイル)

```
{
  "sentenceID": "otaru_r01-02-sent85",
  "tableIDs": [ "otaru_h30-r02-01-tab136", "otaru_h30-r02-01-tab139", ... ]
}
```

図 3 MBLink において HTML フォーマットで提供しているデータセットの具体例

⁴ <https://github.com/poliinfo4/PoliInfo4-FormalRun-Minutes-to-Budget-Linking>

(2) 評価型ワークショップの結果

NTCIR-17 QA Lab-PoliInfo-4 Task のサブタスクとして Minutes-to-Budget Linking を 2022 年 7 月から 2023 年 12 月まで開催した。参加チームは、3 チームであり、自動評価を行うリーダーボードへ投稿した。図 4 にリーダーボードの例を示す。

リーダーボード							
ログイン・参加申し込み STANCE CLASSIFICATION-2 MINUTES-TO-BUDGET LINKING QUESTION ANSWERING-2	概要・提出		FORMAL RUN	DRY RUN	LATE SUBMISSION		
	ID	Rank	Team Name	Download	Submitted at	F-measure ↓	More...
	188	1	fuys		2023/07/14 22:30	0.3371	▼
	112	-	fuys		2023/07/09 15:15	0.2860	▼
	137	2	AKBL		2023/07/12 13:52	0.2577	▼
	148	-	AKBL		2023/07/13 05:03	0.2555	▼

図 4 MBLink の Leaderboard の例

3 有価証券報告書における根拠抽出

3-1 有価証券報告書とは

有価証券報告書(有報)には、決算書に含まれる情報をはじめ、企業の概況、事業の状況などが含まれる。有報は、第一部「企業情報」と第二部「提出会社の保証会社等の情報」に分けられる。有報はおよそ 200 ページから 300 ページ程度であり、ほぼ第一部の内容で占められる。第一部は、第 1「企業の概況」第 2「事業の状況」第 3「設備の状況」第 4「提出会社の状況」第 5「経理の状況」第 6「提出会社の株式事務の概要」第 7「提出会社の参考情報」の 7 つのセクションから構成されている。

有報は XML ベースの言語である XBRL⁵で記述され、損益計算書、貸借対照表、キャッシュフロー計算書などに含まれる財務情報にタグが付与されている。XHTML で表現された内容に<ix:nonNumeric>のような「ix:」で始まる独自タグが埋め込まれており、独自タグを除くと通常の HTML として取り扱い可能であり、ブラウザでの表示も可能である。第一部「企業情報」のうち財務情報のタグが付与されている表は、18%程度であり、最も多い第 1「企業の概況」でも 31%程度である[7]。

(1) 表の理解に関する研究

近年、表の理解に関する研究が注目されている。国際会議 KDD21 では、表の理解に関するチュートリアルとして”From Tables to Knowledge: Recent Advances in Table Understanding”が開催され、表の構造、表のセマンティックモデル、ニューラル表現学習について解説された⁶。表の理解には、表のタイプ分類[8]、表のエリア抽出[9]、表の意味構造の解析[10]、セルの表記統一などの課題がある。中野ら[11]は、文書中の数値の真偽を検証するために、文章中の数値から統計データ中のセルを特定する問題を定式化した。西田ら[8]は、ウェブ上に存在する表を対象に、Header, Attribute, Data の観点から、Vertical Relational (VR), Horizontal Relational (HR), Vertical Entity (VE), Horizontal Entity (HE), Matrix (M), Other (O) の 6 つに分類している。最近では、深層学習を用いて解決するアプローチも増えている⁷。

⁵ 金融庁 企画市場局 企業開示課：報告書インスタンス作成ガイドライン, 2024 年版 EDINET タクソノミ, XBRL 作成ガイド, <https://www.fsa.go.jp/search/20231211.html>

⁶ <https://usc-isi-i2.github.io/KDD21Tutorial/>

⁷ <https://usc-isi-i2.github.io/KDD21Tutorial/3-KDD2021-FT2K-Representation-Learning.pdf>

表の理解に関する Shared task は、ICDAR 2013 Robust Reading Competition[12]⁸などがある。

(2) テキストと表を紐づける研究

テキストと表を紐づける研究には、主張や意見などの分類タスクにおいて、入力として与えられたテキストを分類し、その根拠となる数値や金額などを表から抽出するものがある。FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information)[3]は、Wikipedia に記述された Claim(主張)の真偽を検証するために、Evidence(根拠)を Wikipedia の表から見つけるタスクである。FEVEROUS では、例えば Claim に「Roberto Fico は、総投票数の 57.6% で 57,119 票を獲得しました。」と入力文が与えられた場合、表から投票数が「61,819」であることを見つけ、Verdict(評決)としては Refuted(反論)に分類することになる。日本語のテキストと表を対応付ける Shared task としては、NTCIR-16 QA Lab-PoliInfo-3[13] のサブタスクである Budget Argument Mining がある。

Budget Argument Mining では、地方議会における予算に関する議論マイニングとして、議会会議録のテキストに含まれる金額表現を予算表と結びつけるデータセットを作成し、GitHub で公開している⁹。

3-2 テキストに関連するセルと紐づけるタスク (NTCIR-17 UFO TTRE)

(1) 目的

フリーフォーマットの文章に含まれるテキストと表から特定の情報を抽出する研究は、従来から研究が行われている。Zhong らは、表形式およびテキストに関する質問応答において、表とテキストとの関連性を考慮するために、表とテキストといった異種情報間の構造的なつながりのモデル化に取り組んでいる[14]。Chen らは、自然言語理解と数値計算の両方を必要とする機械読解タスクを対象として、テキストに対する数値推論(加算、減算、並べ替え、カウントなど)に取り組んでいる[15]。Dae らは、テキストと表のリンクによる文書の読みやすさの向上に関する研究を行っている[16]。

しかしながら、これらの研究は英語の文書を対象としており、有価証券報告書特有の構造的な問題や、日本語の表にしばしば含まれるインデントに意味を持たせるような記述、セル結合による小見出し表といった独特な表現に対応することができない。また、上記の従来研究は、発展途上であり、精度が低いものが多いことから、日本語にそのまま適用することは容易でない。

そこで、本研究では、有価証券報告書に含まれる内容の根拠となる数値を見つけるために、本文中のテキストに関連する表のセルを紐づけることを目的としたデータセットを構築する。具体的には、図 5 に示すように、「平均年間給与」を説明しているテキストに対して、関連する表に含まれる「平均年間給与(円)」および「8,000,000」のセルを紐づけることを目指す。TTRE (Text-to-Table Relationship Extraction)¹⁰は、有報の表に含まれる値と関連するテキストを紐づける関係抽出タスクである。このタスクの目的は、有報に含まれる内容の根拠となる数値を見つけ、本文中のテキストに関連する表のセルを紐づけ、セルが項目名であるのか数値であるのかを分類することである。このようなテキストと表を紐づける研究は、主張・意見・説明などを述べる文の根拠となる数値や金額を抽出する際にも役立つ。

⁸ <https://iapr.org/archives/icdar2013/>

⁹ <https://github.com/poliinfo3/>

¹⁰ <https://sites.google.com/view/ntcir17-ufo/subtasks/text-to-table-relationship-extraction>

平均年間給与 は、賞与及び基準外賃金を含んでいます。

従業員数（人）	平均年齢（歳）	平均年間給与（円）
1,000	38	8,000,000

図 5 テキストに関連する表に含まれるセルを紐づける例

(2) タスク概要

TTRE は、有価証券報告書に含まれる内容の根拠となる数値を探すためのタスクであり、その目的は本文中のテキストに関連する表のセルを紐づけ、セルが項目名なのか数値なのかを分類することである。テキストと表を紐づける研究は、有価証券報告書だけでなく、主張・意見・説明などの文の根拠となる数値や金額の抽出にも利用される。テキストと対応する表のセルの紐づけ方について説明する。図 6 に TTRE の最終イメージを示す。下記の HTML のように、オレンジ色のテキストをクリックすると、関連する表のセルが強調表示される。例として、「平成 28 年 3 月期の 1 株当たり配当額 80.00 円」をクリックすると、関連する表の「第 77 期」「1 株当たり配当額(円)」という項目名はピンク色で、対応する値の「80.00」は緑色で表示される。

Text

平成28年3月期の1株当たり配当額80.00円 は、創立60周年記念配当10.00円を含んでいます。
The dividend per share of 80.00 yen for the fiscal year ended March 31, 2016 includes a commemorative dividend of 10.00 yen for the 60th anniversary of the Company's founding.

Table

回次 Period	第76期 The 76th (2015)	第77期 The 77th (2016)	第78期 The 78th (2017)
...	...	...	...
1 株当たり配当額 (円) Dividend per share (yen)	60.00	80.00	92.00

Name

Value

図 6 TTRE のイメージ

TTRE タスクにおける入力、出力、評価指標について述べる。図 7 に TTRE タスクのながれを示す。入力是有報のテキストと表であり、対象となるテキストに予め ID が付与されている。したがって、TTRE タスクでは表セルのリンク元となるテキストを抽出する必要はない。出力はテキストに対応する表のセルに付与されている ID を選択するとともに、項目名(Name)、あるいは、数値(Value)のいずれかのラベルを付与する。評価指標として、リンクセルの推定結果の正確性、網羅性を評価するために、以下の式で定義される Precision, Recall, F1 値を各テキストの結果に対して算出する。

入力

- 有報に含まれるテキスト (HTML 形式)
- 有報に含まれる表 (HTML 形式)

出力

- テキストに関連する表のセル ID
- 選択したセル ID の分類 (Name あるいは Value)

評価

- Name の Precision, Recall, F1 のマクロ平均

- Value の Precision, Recall, F1 のマクロ平均

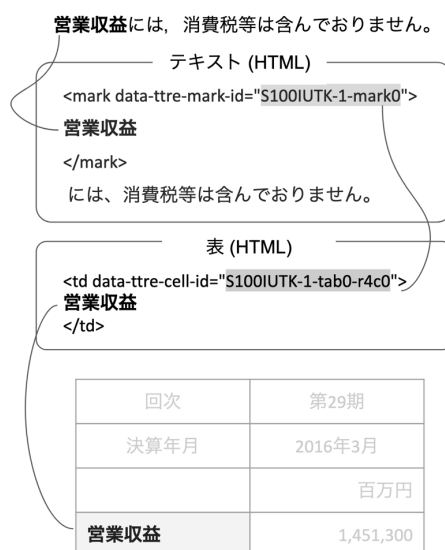


図 7 TTRE タスクの流れ

表 1 に TTRE データセットの統計を示す。データセットは Training データと Test データに分けて公開している。Training データおよび Test データのテキスト、表、セルについての数を記述している。

表 1 TTRE データセットの統計

	テキスト	表	セル
Training	3,402	1,726	80,644
Test	1,875	1,125	47,517
合計	5,277	2,851	128,161

(3) 評価型ワークショップの結果

NTCIR-17 UF0 Task のサブタスクとして TTRE を 2022 年 7 月から 2023 年 12 月まで開催した。本タスクの参加者は 4 チームであった。図 8 にリーダーボードの例を示す。

リーダーボード									
ログイン・参加申し込み TABLE DATA EXTRACTION TEXT-TO-TABLE RE	概要・提出 FORMAL RUN DRY RUN V20230613 DRY RUN V20230501								
	ID	Rank	Team Name	Download	Submitted at	Name	Value	Total ↓	More...
	95	1	fuys		2023/07/28 22:27	0.2707	0.1943	0.2325	▼
	97	-	fuys		2023/07/28 23:56	0.2707	0.1920	0.2314	▼

図 8 TTRE のリーダーボードの例

3-3 各セルの意味を理解するための分類タスク (NTCIR-17 UFO TDE)

(1) 目的

表構造解析は、自然言語処理において、テキストに記述されている根拠を表データから抽出する技術として注目されている。表データは、データベースのような整然とした構造化データだけではなく、不規則または半構造化された形式で存在する。表のデータを構造化されたデータへ変換する方法としては、HTML や XML などのマークアップ言語を解析する手法などがある。また、RDF (Resource Description Framework) や OWL (Web Ontology Language) などの技術を使用してデータに構造と意味を与え、セマンティック・ウェブの技術を利用する方法などがある。しかしながら、データベース以外の表データは、人間には判読しやすいものの、機械判読可能な形式へ変換する処理が課題となっている。表構造の理解は、異なる企業間で比較可能な表を見つけ、項目と値の正しい組み合わせを抽出するだけでなく、表を用いた質問応答や含意関係認識、ファクトチェックなどの応用にも活用が見込まれる。

そこで本研究では、非財務情報に関する記述を各社間で比較しやすい形式にまとめるために、比較可能な表を見つけ正しい項目と値を抽出することを目的とする。具体的には、有価証券報告書に含まれる表の各セルを Metadata, Header, Attribute, Data の 4 種類のクラスに分類する。例として、図 9 に示される表では、主キーとしての会社名が各行に記載されており、名称や資本金などの属性が各列に記載されている。このような認識を基にクラス分類を行う。このクラス分類により「[関係会社の状況] について、[ABC (株)] の [資本金] は [1 億円] である」といった形式の自然言語文が生成できる。

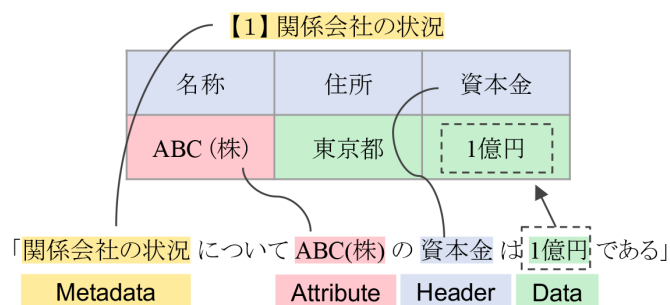


図 9 表のクラス分類を行うことで自然言語文として理解できる例

我々は Understanding of non-Financial Objects in Financial Reports (以下、UFO) タスクにおける表データ抽出 TDE(Table Data Extraction)¹¹サブタスクとして評価型ワークショップの形で取り組むこととした。TDE は、有価証券報告書に含まれる表を取り扱うためのタスクであり、その目的は表の構造を理解することを目的としている。

(2) タスク概要

TDE タスクでは、有価証券報告書に含まれる表を扱うために、有価証券報告書に含まれるテーブルセルを 4 つのクラス(Metadata, Attribute, Header, Data)に分類する。TDE タスクにおける入力、出力、評価指標を下記に示す。

入力

- 有報に含まれる表(HTML 形式)

¹¹ <https://sites.google.com/view/ntcir17-ufo/subtasks/table-data-extraction/>

出力

- 各セルの分類 (Metadata, Header, Attribute, Data のいずれか)

評価

- Accuracy による評価

表 2 に TDE データセットの統計を示す。データセットは、TOPIX100 に上場している 20 社を選び、3 人のアノテーターが HTML 形式の有価証券報告書の全ての表にアノテーションを行った。このデータセットでは、アノテーションに成功したすべてのセルを使用した。つまり、「[メタデータ]に関して、[属性]の[ヘッダ]は[データ]である」という形で記述できないような非常に複雑な表は除外した。その結果、有価証券報告書に含まれる全テーブルの 44.5%がタスクに使用された。20 社のうち 1 社の異なる年の有価証券報告書が Train と Test のデータセットに使用されたため、表 2 の企業数の合計が 20 社ではなく 21(=13+8)になっているのは Training と Test において同じ企業が 1 社含まれているためである。ただし、有価証券報告書の重複はなく、章、表、セルも重複していない。

表 2 TDE データセットの統計

項目	Training	Test
企業数	13	8
有価証券報告書数	25	15
章数	252	190
表 (有価証券報告書内)	5,683	3,716
セル (有価証券報告書内)	247,302	173,732
表 (タスクで使用)	2,530	1,660
セル (タスクで使用)	66,369	45,499
ヘッダー	13,930	8,949
属性	11,236	7,611
データ	41,060	28,895
メタデータ	143	44

(3) 評価型ワークショップの結果

NTCIR-17 UFO Task のサブタスクとして TDE を 2022 年 7 月から 2023 年 12 月まで開催した。本タスクの参加者は 4 チームであった。図 10 にリーダーボードの例を示す。

リーダーボード								
ログイン・参加申し込み TABLE DATA EXTRACTION TEXT-TO-TABLE RE	概要・提出							
	FORMAL RUN		DRY RUN		LATE SUBMISSION			
	ID	Rank	Team Name	Download	Submitted at	Accuracy ↓	F-measure	More...
	81	1	KSU		2023/07/27 19:14	0.9537	0.9537	▼
	100	2	FA		2023/07/29 08:09	0.9343	0.9343	▼

図 10 TDE のリーダーボードの例

(4) 今後の課題

有価証券報告書にはセルが結合されている表や属性が多数設定されている表も多く存在しており、機械的なデータ抽出が困難な事例がある。NTCIR-17 UFO のサブタスクである TDE では、このような [Attribute][Header][Data] の三つ組を過不足なく用意できない表は、利用データから除いていた。TDE サブタスクのデータセットでは、セルの属性分類が不可能だった表を対象としていない。

そこで、本節では「[Attribute]の[Header]は[Data]である」という形式で、セルを自然言語の文に変換する際の困難な表を明らかにする。有価証券報告書の困難なパターンは大きくわけて5つのタイプに分類できることを確認した。

タイプ1 小見出し行

小見出し行とは表中にふくまれる見出しである。小見出しの行には Data (値) が含まれない。小見出し業は、値を持たず下の行を修飾している場合が多い。例えば、図 11 の「資産の部」や「流動資産」といった行について、その下の「現金・預金」や「預託金」などの行を修飾する役割を持っていて、その行には何もデータが記されていないものがある。

その行にはDataセルを持たないが、 下のHeaderなどを修飾するための行がある			面 値 金 種
2019			40
当期発生額			
税効果調整前	8,324		20
税効果額	△2,338		△
税効果調整後	5,986		2

図 11 小見出しの例

タイプ2-1 5つ組

5つ組とは、表に含まれる Data(値)を4つ組 (Metadata, Header, Attribute, Data) だけでは表現できず、5つ目の情報が必要となる表である。例えば、図 12 のように、上下2段の組みができており、Header と Data が上下に対応するような形では、5つ組のセットが必要で機械的なデータ抽出が難しくなる。

銘柄	当事業年度	前事業年度		当社の
	株式数(株)	株式数(株)		
ナブテスコ株式会 社	貸借対照表計上額 (百万円)	貸借対照表計上額 (百万円)	(保有 確保す (定量	
	3,760,000	3,760,000		
富士電機株式会 社	19,026	9,381	(保有 確保す (定量	有
	2,684,200	2,684,200		
津上精密機床 (中 国) 有限公司	12,374	6,574	(保有 確保す (定量	無
	6,200,000	6,200,000		

図 12 5つ組の例

タイプ2-2 結合されたセル

結合されたセルとは、セルが結合され、複数の行や列に跨っているセルである。例えば、図 13 の「純営業利益」や「経常利益又は経常損 (△)」といったセルは、それぞれ4つの列にまたがっているほか、「ホールセール部門」というセルは「グローバル・マーケット」と「グローバル・インベストメント・バンキング」のセルを包含することを結合によって表現している。こうしたセルの結合によるデータの表現は判読が難しい。

	純営業収益				経常利益又は経常損失(△)			
	2020年 3月期	2021年 3月期	対前年同期 増減率	構成比率	2020年 3月期	2021年 3月期	対前年同期 増減率	構成比率
リテール部門	166,430	169,505	1.8%	36.3%	6,405	20,070	213.3%	15.6%
ホールセール部門	172,289	215,860	26.2%	46.2%	29,034	74,737	96.5%	58.1%
グローバル・マーケッ ツ	121,301	161,730				62,777	122.7%	48.8%
グローバル・インベス トメント・バンキング	50,988	54,129				11,021	18.1%	8.6%
アセット・マネジメント 部門	48,091	51,145	6.3%	11.0%	26,580	32,775	23.3%	25.5%

図 13 結合されたセルの例

タイプ 3 空白セル

空白セルとは、Header, Attribute, Data でもなく、値が含まれないセルである。例えば、図 14 のように、表の体裁を整えるための空白セルが使われている。

	前事業年度 (2020年 3月31日)	当事業年度 (2021年 3月31日)
法定実効税率	30.5%	30.5%
(調整)		
交際費等永久に損金に算入されない項目	0.0	0.0
受取配当金等永久に益金に算入されない項目	△39.7	△14.2
住民税等均等割等	0.1	0.1
評価性引当額	7.6	△5.5
税率変化による繰	△0.1	0.0
外国税額控除	0.0	△1.2
その他	4.0	△0.8
税効果会計適用後の法人税等の負担率	2.4	8.9

図 14 空白セル

タイプ 4 非スカラ値

非スカラ値は単一の値をもたないセルである。例えば、図 15 のように Data セルに複数の情報を含む文章が記されている。

	重要なリスク	主な想定シナリオ
1	国内外の経済危機、金融・資本市場の 混乱	① リーマンショック級の世界金融危機が発生し、当社グループ保有資産の価値が大幅に下落する。 ② 地政学リスクの顕在化等により金融・資本市場の混乱が生じ、当社グループ保有資産の価値が大幅に下落する。
	Dataセルに複数の情報を含む 文章が記されている。	① 政府の信用力低下により日本国債が暴落し、当社グループ保有資産の価値が大幅に下落する。
2	巨大地震	① 首都直下地震の発生により、多額の保険金支払が発生する。また、当社グループの事業継続に重大な影響が生じるほか、当社グループ保有資産の価値が大幅に下落する。

図 15 非スカラ値の例

タイプ 5 特殊な形

特殊な形とは、これまで紹介した例に当てはまらない形である。例えば、図 16 はある企業の各分野における事業内容と主要関連会社について表したものであるが、複雑な入れ子構造や結合が多用されている。

事業区分及び主要製品		主要会社
ゲーム&ネットワークサービス		
	家庭用ゲーム機 ソフトウェア ネットワークサービス事業	(株)ソニー・インタラクティブエンタテインメント Sony Interactive Entertainment LLC Sony Interactive Entertainment Europe Ltd.
音楽		
音楽制作	パッケージ及びデジタルの音楽制作物の販売 アーティストのライブパフォーマンスからの収入	(株)ソニー・ミュージックエンタテインメント Sony Music Entertainment
音楽出版	楽曲の詞、曲の管理及びライセンス	Sony Music Publishing LLC
映像メディア・プラットフォーム	アニメーション作品及びその派生ゲームアプリケーションの制作・販売 音楽・映像関連商品のサービス提供	
映画		
映画製作	映画作品の製作・買付・配給・販売	
テレビ番組制作	テレビ番組の制作・買付・販売	CRF Holdings Inc.

多数の結合や入れ子構造など
著しく複雑な表構造を独自に用いている

図 16 特殊な形

機械判読が困難な表の分析結果

これら 6 つに分類された自然言語文に変換できないセルを含む表について、対象とした有価証券報告書にそれぞれのタイプの表がどの程度含まれたかを調査することによって、更なる詳細な実態の解明を試みた。

表がどのタイプに該当するかは、その文書を担当した 3 名のアノテーターが判断したが、表によっては判断が分かれることもあったため、過半数である 2 名以上のアノテーターが該当すると判断したタイプのみを対象とした。また、1 つの表が複数のタイプに該当する場合も多く存在した。なお、2 つ目の「5 つ組があれば表現できる表」と 3 つ目の「結合されたセルを含む表」について、それぞれのタイプに該当すると過半数のアノテーターが判断した表の内容を著者が確かめたところ、類似した書式の表が 2 つのタイプの間に混在していることがわかった。そのため、いずれのタイプも複数の Header や Attribute を持つという共通した性質を持つことから、2 つのタイプを統合し「複数の Header や Attribute を持つ表」として扱うのが適切だと考えられる。

分析の結果、以下の 2 つが明らかとなった。

- ・ 機械判読が困難な表を、小見出しや空白セル、非スカラ値のセルを含む表、複数の Header や Attribute をもつセルを含む表、特殊な形の表の 5 つのタイプに分類できた。
- ・ 機械判読が困難な体裁の表が、対象とした有価証券報告書には約 37%含まれていた。

4 おわりに

本研究では、日本語の「表」および「テキスト」を対象として、表の意味構造を捉えつつ、テキストと紐付けることで根拠を抽出した。本報告書では、地方議会会議録、および、有価証券報告書を対象として、データセットを構築し、評価型ワークショップを開催し、複数の研究者に提供した結果について述べた。

しかし、本研究にはいくつかの課題も残されている。XBRL のような XML ベースの言語によるデータ構造化や、深層学習を用いた表構造の理解といったアプローチは有望であるが、実際に利用する際にはタクソノミの追加・削除や企業独自などにより比較が困難な状況がある。また、LLM を用いて解析する場合にもトークン制限により一つの表を解釈できないなどの問題があり、研究の余地がある。

今後の展望として、本研究で得られた知見を基に、生成過程の説明性や誤情報の生成に関する課題にも取り組み、信頼性の高い自然言語処理技術の確立に貢献していきたい。本研究が自然言語処理の発展に寄与し、実社会における応用が進むことを期待する。

【参考文献】

- [1] 乾健太郎, ChatGPT の出現は自然言語処理の専門家に何を問いかけているか, 自然言語処理, 30 巻 2 号, pp.273-274, 2023.
- [2] Kotonya, N. and Toni, F., Explainable Automated Fact-Checking: A Survey, Proceedings of the 28th International Conference on Computational Linguistics, pp.5430-5443, 2020.
- [3] Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., and Mittal, A.: The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task, in Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), pp. 1–13, Dominican Republic 2021.
- [4] Yasutomo Kimura, Hokuto Ootake, and Minoru Sasaki, Budget argument mining dataset using Japanese minutes from the National Diet and local assemblies. LREC 2022,2022.
- [5] 木村泰知, 永渕景祐, 乙武北斗, 佐々木稔, 予算項目に関連する議論を対応づける Budget Argument Mining のデータセット構築. 情報処理学会 第 249 回自然言語処理研究会, 2021.
- [6] 梶縁, 木村泰知. 地方自治体の予算を対象にした金額表現の分析. 第 38 回ファジィシステムシンポジウム(FSS2022), 9 2022.
- [7] 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.
- [8] Nishida, K., Sadamitsu, K., Higashinaka, R., and Matsuo, Y.: Understanding the Semantic Structures of Tables with a Hybrid Deep Neural Network Architecture, in Singh, S. and Markovitch, S. eds., Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, pp. 168–174, AAAI Press 2017.
- [9] Sun, K., Rayudu, H., and Pujara, J.: A Hybrid Probabilistic Approach for Table Understanding, in Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), pp. 4366–4374 2021.
- [10] Vu, B., Knoblock, C., Szekely, P., Pham, M., and Pujara, J.:A Graph-Based Approach for Inferring Semantic Descriptions of Wikipedia Tables, in The Semantic Web – ISWC 2021, pp. 304–320, 2021.
- [11] 中野 優, 加藤 誠:被引用統計データのセル特定データセットの構築, 第 14 回データ工学と情報マネジメントに関するフォーラム (DEIM 2022) 2022.
- [12] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L. G. i., Mestre, S. R., Mas, J., Mota, D. F., Almaz'an, J. A., and Heras, de las L. P.: ICDAR 2013 Robust Reading Competition, in 2013 12th International Conference on Document Analysis and Recognition, pp. 1484–1493 2013.
- [13] Kimura, Y., Shibuki, H., Ootake, H., Uchida, Y., Takamaru, K., Ishioroshi, M., Yoshioka, M., Akiba, T., Ogawa, Y.,Sasaki, M., Yokote, K., Kadowaki, K., Mori, T., Araki, K., Mitamura, T., and Sekine, S.: Overview of the NTCIR-16 QA Lab-PoliInfo-3 Task, Proceedings of The 16th NTCIR Conference, pp.156–174 2022.
- [14] Zhong, W., Huang, J., Liu, Q., Zhou, M., Wang, J.,Yin, J., and Duan, N.: in Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), pp. 4531–4537 2022.
- [15] Chen, K., Xu, W., Cheng, X., Xiaochuan, Z., Zhang, Y.,Song, L., Wang, T., Qi, Y., and Chu, W.: Question Directed Graph Attention Network for Numerical Reasoning over Text, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6759–6768, Online 2020.
- [16] Kim, D. H., Hoque, E., Kim, J., and Agrawala, M.: Facilitating Document Reading by Linking Text and Tables, in Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, p. 423–434, New York, NY, USA 2018.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
非構造化データを活用する試み ―議会会議録と有価証券報告書を例として―	計画行政	2023 年 12 月
BERT ベース分類器と SHAP を用いた地方議会議員の議会発言における特徴分析	知能と情報	2023 年 8 月
Deeds, not words? Speech and re-election of Japan's local legislators	Constitutional Political Economy	2023 年 4 月
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 (NLP2024)	2024 年 3 月
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 (NLP2024)	2024 年 3 月
OUC at NTCIR-17 UFO TDE and TTRE	Proceedings of The 17th NTCIR Conference}	2023 年 12 月
Overview of the NTCIR-17 QA Lab-PoliInfo-4 Task	Proceedings of The 17th NTCIR Conference	2023 年 12 月
Overview of the NTCIR-17 UFO Task	Proceedings of The 17th NTCIR Conference	2023 年 12 月
国会および地方議会会議録をコーパスとした事前学習済み言語モデルの構築と検証	第 20 回テキストアナリティクス・シンポジウム	2023 年 9 月
researchmap の情報を用いた異なる研究分野間の研究者マッチングの試み	第 22 回情報科学技術フォーラム (FIT2023)	2023 年 9 月
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	第 22 回情報科学技術フォーラム (FIT2023)	2023 年 9 月
NTCIR-16 QA Lab-PoliInfo-3 Fact Verification の報告およびデータセット公開	人工知能学会全国大会(第 37 回)	2023 年 6 月
階層的な分類構造を活用した SHAP による地方議会会議録における特徴表現抽出の試み	人工知能学会全国大会論文集	2023 年 6 月