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

奥山和樹 [1]　木村泰知 [1]
[1] 小樽商科大学
g2020067@edu.otaru-uc.ac.jp　kimura@res.otaru-uc.ac.jp

概要

本研究では，機械判読が困難な表について，有価証券報告書における詳細な実態を明らかにする．分析の結果，機械判読が困難であると考えられる表は対象とした有価証券報告書に合計 3,215 件，全体の約 37 %あることを確認した．困難な表は「小見出し行を含む表」「複数の Header や Attribute を持つ表」「空白セルを含む表」「非スカラ値のセルを含む表」「特殊な形の表」の 5 種類に分類できた．そのなかでも「複数の Header や Attribute をもつセルを含む表」「小見出し行を含む表」「空白セルを含む表」が多く存在していた．

1　はじめに

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 [1, 2]．

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．人間にとって判読しやすい表は，各セルに含まれる数値や値などを「自然言語文に変換しつつ理解している」と考えられる．例えば、図 1 の有価証券報告書の表に含まれる「43 銘柄」というセルは，表中の他のセルを用いて「非上場株式の銘柄数は 43 銘柄である」というような自然言語文に変換して説明することができる．

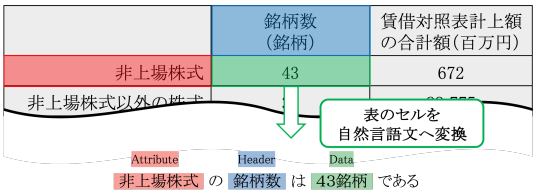

図 1　表のセルを自然言語の文に変換する例

ここで用いた表中のセルについて [Attribute][Header][Data] という分類方法に当てはめると [3]，表中の Data セルは「[Attribute] の [Header] は [Data] である」という枠組みの自然言語文になる．

このように，表中のセルから「Attribute」「Header」「Data」の三つ組を過不足なく用意し，セルを適切な自然言語の文に変換できる表は，機械的なデータ抽出に理想的である．評価型ワークショップである NTCIR-17 UFO タスクにおける表データ抽出の研究においても，この [Attribute][Header][Data] の三つ組を過不足なく用意してセルを適切な自然言語の文に変換できる表を用いている [4]．

しかしながら，有価証券報告書には図 2 のような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．

区分	前連結会計年度	
	監査証明業務に基づく報酬（百万円）	非監査業務に基づく報酬（百万円）
提出会社	29	−

図 2　表のセルを自然言語文に変換することが困難な例

NTCIR-17 UFO のサブタスクである TDE(Table Data Extraction) 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，利用データから除かれている [4]．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．また，有価証券報告書内の表の数については，正確な実態は明らかになっていない[1)] [5]．

1）2 名のアノテータが扱った表の延べ数で示されており，詳細の分析が行われていなかった．

そこで，本研究では「[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さを明らかにすることを目的とする．本稿では，表のセルを機械的に自然言語の文に変換することが困難な表はどのような特徴を持つのか，また，そうした表が有価証券報告書にどの程度含まれているかを明らかにする．

本研究の貢献は以下の 3 つである．

- 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%含まれていることを明らかにした．
- 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 つのタイプに分類した．
- 対象とした有価証券報告書には，順に複数の Header や Attribute をもつセルを含む表，小見出し行を含む表，空白セルを含む表などが多く存在することを明らかにした．

2 関連研究

政府や企業により多くの統計データが表形式で公開されている昨今，大量の統計データから知見を得るうえで，機械的な表データ抽出を行うための表の理解に関する研究が注目されている．たとえば，有価証券報告書において表形式のデータや文書から構造化情報を抽出する技術を開発する研究 [6] や，ソーシャルネットワーク上の言説の客観的根拠を統計表データから探索するシステムを構築するための機械学習による表構造認識に関する研究 [7] など，表の機械判読に関する研究が行われている．しかし，こうした機械判読を試みる研究は多いが，機械判読が難しかった事例に焦点を当てて，それがどのような表であったか，どの程度含まれていたかを分析する研究は見当たらなかった．

また，総務省は統計表における機械判読可能なデータの表記方法の統一ルールを策定するにあたって，機械判読が困難な表形式の例として，一つのセルに複数のデータが入力されているものや，セルが結合されているものなどを取り上げており[2)]，そうした書式の表が機械判読の妨げとなっていることがわかる．

2） https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000186.html

3 有価証券報告書

3.1 有価証券報告書とは

有価証券報告書とは，企業の事業概況や経理の状況などの情報を記した文書で，上場企業などには毎年度提出が義務付けられている．投資家が上場企業の中から将来の有望株を探索したり，投資先の決算状況を評価したりする上で，欠かせない情報源の一つになっている [8]．一つの文書は 7 項目ほどに節立てされていて，大きな企業のものは全体で 100 ページを超えるほど文量が多い資料である．多種多様な表が用いられており，今回分析の対象とした有価証券報告書では，一文書あたり平均 217 件，合計で 8,673 件[3)]の表が含まれていた．

3.2 有価証券報告書に含まれる表

有価証券報告書には損益計算書，貸借対照表，キャッシュフロー計算書など企業の財務情報をまとめた表が含まれるほか，SDGs への取り組みやコーポレートガバナンスの状況，事業等のリスクといった，非財務情報に関する図表を各社が用意している．これらの非財務情報に関する表は各社が図表を交えたフリーフォーマットの文章で説明しており，特に表についてはそれぞれが表す内容の範囲や列の構成も統一されてはいない． したがって，こうした多種多様で複雑な表を多く含む文書から情報を機械的に抽出することは特に困難になっていて，複雑な表の機械判読を目指す研究に適した資料の一つといえる．

4 表の分析と集計

4.1 目的

研究では，初めに対象とした有価証券報告書内に含まれる全ての表に対して，人力による表の分析“アノテーション”を行い，セルを [Header][Attribute][Data] に判別することを試みた．そのうえで，機械判読が困難な表がどの程度含まれたかを集計し，機械判読が困難だと判断された表に対して，判読が困難だと考えられた理由について表のタイプ分けを行い，それぞれのタイプがどのような割合で含まれているかを分析した．本研究では，

3） 入れ子になった table 要素や，本文中のレイアウトのためだけに使われた table 要素などは，門脇ら [5] の計上方法に準じて対象としていない．

表の分析“アノテーション”の結果をもとに，機械判読が難しいと判断された表を集計し，その内容を分類することで，有価証券報告書に含まれる機械判読が困難な表の実態を明らかにすることを目的とする．

4.2 表の分析“アノテーション”の方法

初めに行った表の分析“アノテーション”は，有価証券報告書 1 文書につきそれぞれ 3 名，合計で 8 名のアノテーターが行った．アノテーションは，表を構成する主なブロックとして [Header]，[Attribute]，[Data] の三つが挙げられることにもとづいて [3]，表の各セルをそれらいずれかのクラスに分類する作業である．このアノテーションによって分類された表中の各セルから [Header]，[Attribute]，[Data] の 3 種類の組を取り出し，「[Attribute] の [Header] は [Data] である」という型の自然言語文で表現することによって，各 Data セルをそれ以外のセルで説明できるようになる．作業には門脇らが作成した図 5 のような分析ツール [5] を用い，人手によって対象とした 8,673 件全ての表のセルを色分けすることで分類した．

しかし，対象とした文書の中には前述のように表構造が複雑で，[Header]，[Attribute]，[Data] の 3 つ組を過不足なく取り出して自然言語文に変換することができない表も存在する．こうした表の存在は表の機械判読を試みる研究において課題となる．アノテーションを行った表のうち，そのような自然言語文への変換が難しいセルを含む表や，セルが 3 種類の属性のどれに該当するか判断が難しいセルを含む表に対しては，別途機械判読が困難な表として記録した．

分析の対象としたのは，TOPIX100 に含まれる企業のうち，京セラ，村田製作所など 20 社が，2020 年度から 2021 年度に提出した有価証券報告書 40 文書であり，なるべく幅広い業界などから企業を選出することで，文書内容に偏りが出ないよう配慮した．

4.3 集計の結果

アノテーションを行い，自然言語文に変換して説明することができないセルを含む表がどの程度あるかを集計した結果，過半数のアノテーターが自然言語の文に変換できないセルを含むと判断した表は合計 3,215 件あった．これは対象とした 20 社の 40 文書の有価証券報告書に含まれた合計 8,673 件の表のうち，全体の約 37 %に相当する量であった．

4.4 分類の結果

アノテーションで自然言語文に変換できないセルを含む表であると判断されたものについて，なぜセルを自然言語文に変換できないのか，その詳細を調査した[4]．

タイプ 1 小見出し行

1 つ目は小見出し行を含む表である．その行には Data を持たず下の行を修飾している場合があり，三つ組を過不足なく用意できない．例えば，図 6 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，その行自身には何もデータが記されていないものがある．

タイプ 2-1 5 つ組

2 つ目は 5 つ組があれば表現できる表である．図 3 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．

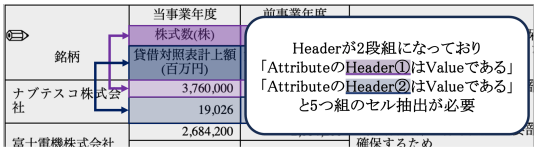

図 3 複数の Header や Attribute をもつセルを含む表の例

タイプ 2-2 結合されたセル

3 つ目は結合されたセルを含む表である．セルが複数の行や列に跨っていると，それを機械的には認識できずデータ抽出が難しい．例えば図 4 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・マーケッツ」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．

タイプ 3 空白セル

4 つ目として図 7 のような表の体裁を整えるための空白セルを含む表もあった．このような場合は，セルにデータを含まないためデータ抽出を行う際の障害になる．

4） 当初は 6 種類のタイプで分類していたが，今回の調査により，5 種類のタイプにまとめる方が適切であることが明らかになった．そのため，2-1，2-2 と表記している．

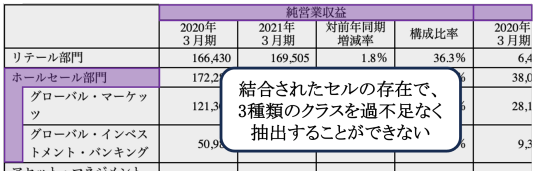

図 4　結合されたセルを含む表の例

タイプ 4　非スカラ値

5 つ目に単一の値をもたない，非スカラ値のセルを含む表があった．例えば図 8 のように Data セルに複数の情報を含む文章が記されていて，3 つ組による自然言語文への変換ができなくなっている．

タイプ 5　特殊な形

6 つ目に，これまで紹介した例のほか企業に応じて複雑で特殊な形の表が含まれる場合があった．図 9 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て，アノテータ全員が自然言語文への変換が難しいと判断した．

4.5　機械判読が困難な表の実態

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 名以上のあのテーターが該当すると判断したタイプのみを計上した．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．

ここで，6 つのタイプについて 3 人の評価者による判定がある程度一致していることを確かめるため，フライスの Kappa 値を表 1 に示す．ここでは，一つの表が複数のタイプに該当するケースが多くあることを考慮し，それぞれのタイプにおいて，ある表がそのタイプに該当するかどうかの 2 択の判断が 3 人の評価者の間でどの程度一致したかを算出している．表の通り，「5 つ組があれば表現できる表」と「結合されたセルを含む表」は，「複数の Header や Attribute を持つ表」として統合した方が高い一致度を示すことから，機械判読が困難な表を 5 種類に分離することが適切だったといえる．

表のタイプ	Kappa
小見出し行を含む表	0.26
複数の Header や Attribute を持つ表	0.60
5 つ組があれば表現できる表	0.43
結合されたセルを含む表	0.55
空白セルを含む表	0.28
非スカラ値のセルを含む表	0.63
特殊な形の表	0.21
その他	0.14

表 1　それぞれのタイプにおける評価者 3 人の判定一致度

これらの結果，対象とした 8,673 件の表のうち，小見出し行を含む表は 1,666 件，複数の Header や Attribute を持つセルを含む表は 2,020 件，表の体裁を整えるための空白セルを含む表は 1,503 件，非スカラ値のセルを含む表は 186 件，特殊な形の表が 227 件あった (表 2)．なお，同一の表が複数のパターンに同時に該当する場合は，重複してカウントしているため，割合の合計は 100%にならない．

表のタイプ	表の数	割合
機械判読が容易	5458	62.9%
小見出し行	1666	19.2%
複数の Header や Attribute	2020	23.3%
空白セル	1503	17.3%
非スカラ値	186	2.1%
特殊な形	227	2.6%
その他	461	5.3%

表 2　5 種類の表の割合

5　おわりに

本研究では，有価証券報告書の一部において，「[Attribute] の [Header] は [Data] である」とセルを自然言語の文に変換することが難しい複雑な表が約 37 %含まれていることを明らかにした．また，そうした表を 5 種類に分類すると，複数の Header や Attribute を持つセルを含む表や，小見出し行を含む表などが多く存在することを確認した．

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] 西田光甫, 吉永直樹, 西田京介. 非構造知識検索を用いた自己適応型固有表現認識. 情報処理学会 自然言語処理研究会, 札幌, 2023. 情報処理学会.

[2] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: fact extraction and verification over unstructured and structured information. **CoRR**, Vol. abs/2106.05707, , 2021.

[3] Elvis Koci, Maik Thiele, Oscar Romero, and Wolfgang Lehner. Cell classification for layout recognition in spreadsheets. In Ana Fred, Jan Dietz, David Aveiro, Kecheng Liu, Jorge Bernardino, and Joaquim Filipe, editors, **Knowledge Discovery, Knowledge Engineering and Knowledge Management**, pp. 78–100, Cham, 2019. Springer International Publishing.

[4] Yasutomo Kimura, Hokuto Ototake, Kazuma Kadowaki, Takahito Kondo, and Makoto P. Kato. Overview of ntcir-17 ufo task. **Proceedings of The 17thNTCIRConference(12 2023)**, 2023.

[5] 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.

[6] 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.

[7] 松井侑祐, 宮森恒. 統計表データを用いた動向情報の根拠探索システムの検討. 2014 年度 情報処理学会関西支部 支部大会 講演論文集, 第 2014 巻, sep 2014.

[8] 首藤昭信. リスク情報開示と企業価値. 専修ビジネス・レビュー, Vol. 3, No. 1, pp. 61–67, 03 2008.

付録

分析の対象の 20 社

株式会社村田製作所，東海旅客鉄道株式会社，ソニーグループ株式会社，株式会社日本取引所グループ，ソフトバンク株式会社，ファナック株式会社，富士通株式会社，日本電信電話株式会社，ＳＯＭＰＯホールディングス株式会社，京セラ株式会社，パナソニックホールディングス株式会社，株式会社マキタ，西日本旅客鉄道株式会社，ＫＤＤＩ株式会社，ニデック株式会社，ＭＳ＆ＡＤインシュアランスグループホールディングス株式会社，オムロン株式会社，野村ホールディングス株式会社，株式会社キーエンス，ＡＮＡホールディングス株式会社

操作画面および機械判読が困難な表の例

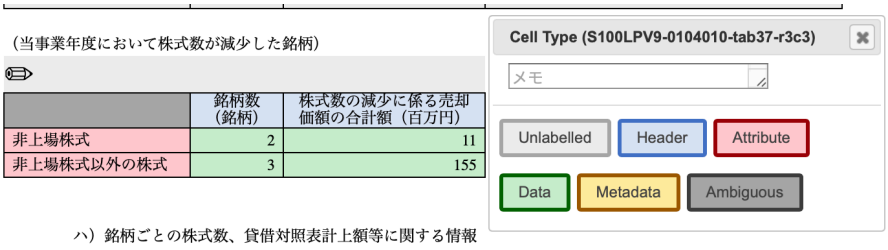

図 5　アノテーションツールの操作画面

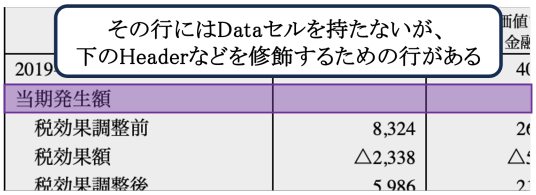

図 6　小見出し行を含む表の例

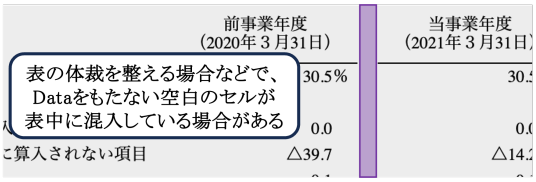

図 7　空白セルを含む表の例

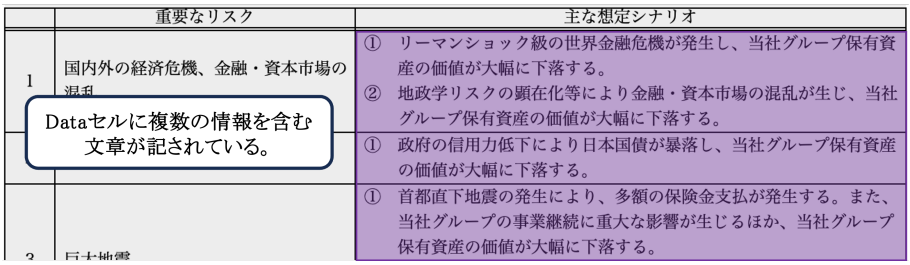

図 8　非スカラ値のセルを含む表の例

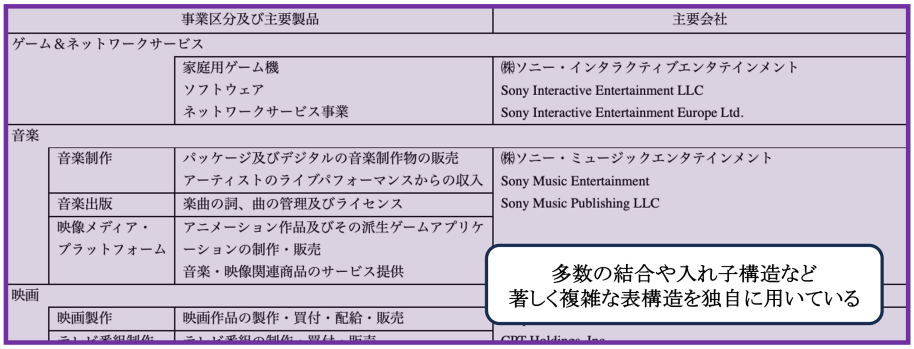

図 9　特殊な形の表の例