StruAP を用いた金融分野の開示文書からの情報抽出

柳井 孝介　佐藤 美沙　十河 泰弘　山脇 功一　渋谷 淳
日立製作所 研究開発グループ　QUICK ナレッジ開発本部
{kohsuke.yanai.cs, misa.sato.mw, yasuhiro.sogawa.tp}@hitachi.com
{koichi.yamawaki, jun.shibuya}@quick.jp

1 はじめに

本稿では、有価証券報告書や有価証券届出書等の金融ドメインの開示文書から、投資家が必要とする情報を抽出する手法について述べる。有価証券報告書は年に 3,000 本以上、有価証券届出書は 12,000 本以上が開示される。また投資家が必要とする情報は、1 つの文書あたり、50 から 80 項目ほどある。そのため、これらの文書をすべて人が読み、人手で情報を抽出するのはコストがかかる。

これらの文書は XBRL で公開されており、ある程度の大きさのテキストのブロックには分割されているものの、抽出すべき情報はテキストブロックの中に自然言語で書かれているため、XBRL のタグ構造をたどるだけでは必要な情報は抽出できず、自然言語解析を行う必要がある。また、1 つのテキストブロックの中にも章節構造が含まれている場合が存在するため、実用的な情報抽出システムの実現のためには、自然言語処理に加えて、章節構造の解析が必要となる。

例えば投資家が必要とする情報の 1 つである「ベンチマーク」に関しては、有価証券届出書の下記の 3 つの節・パラグラフのテキストが情報の抽出元となる。

- 第二部【ファンド情報】/第 1【ファンドの状況】/1【ファンドの性格】/(1)【ファンドの目的および基本的性格】/・ファンドの特色
- 第二部【ファンド情報】/第 1【ファンドの状況】/2【投資方針】/(1)【投資方針】
- 第二部【ファンド情報】/第 1【ファンドの状況】/3【投資リスク】

XBRL においては、PurposesAndBasicFeaturesOfFundTextBlock や InvestmentPolicyTextBlock などのタグの中に該当のテキストがあり、このテキスト中からそのファンドがベンチマークとしている指数を抽出する。言語的な表現としては、「○○を運用上のベンチマークとして」や「○○を対象指数として」など、50 を超えるパターンがある。

```
{
  "name":          "ベンチマーク",
  "xbrl_tag":      "PurposesAndBasicFeaturesOfFundTextBlock",
  "paragraph":     "ファンド.+目的.+基本的性格",
  "tree_pattern": "benchmark",
  "target":         "a2",
  "extracted_expression": [
    ["(対象(株価)?(指数|指標)|中長期的)(.*である|.*(){P<value>.+)
(の|を|に|と|に対しての)",
      "%(value)s"],
    [ "(?P<value>.+)(の|を|に|と|に対しての)",
      "%(value)s"],
    [ "(?P<value>.+(インデックス|[^対象]指数|)|¥¥d{3}))",
      "%(value)s"]
  ]
},
```

図 1: 有価証券届出書から「ベンチマーク」の項目を抽出する場合の抽出要件の記載例

著者らは関係抽出のツールとして、構文木の木構造のパターンマッチを行う StruAP（Structure-based Abstract Pattern）を提案している [3]。本稿では、2 節において StruAP を用いた抽出手法について述べ、3 節で有価証券届出書からファンド情報を抽出した結果について述べる。4 節では有価証券報告書から企業が目的としている経営指標を抽出した結果を報告する。5 節で考察を行い、6 節で本稿を締めくくる。

2 手法

本稿では図 1 に示すような抽出要件に従って、以下の手順で抽出を行う。

Step 1 指定された XBRL のタグの中のテキストを取得する。

Step 2 指定された章節タイトルを持つテキストを取得する。

Step 3 CaboCha [1, 7] により係り受け解析を行う。

Step 4 StruAP により係り受け木の木構造のパターンマッチを行う。

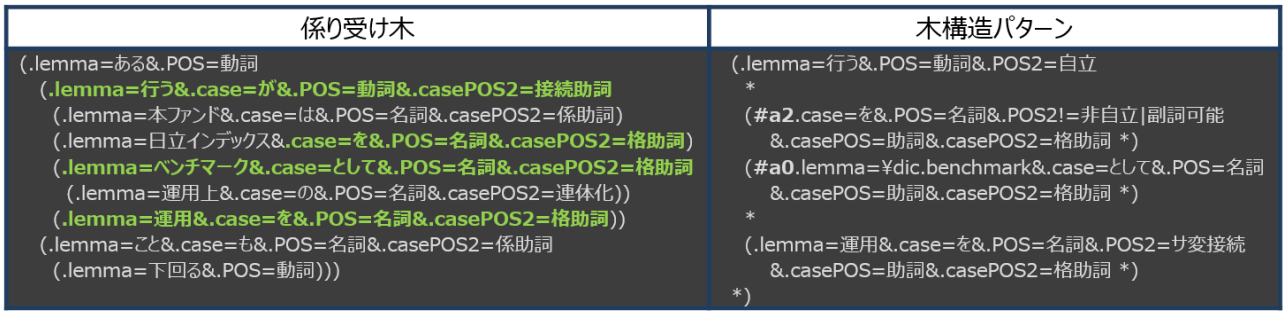

図 2: StruAP での木構造のパターンマッチ例

Step 5 正規表現により、トリミング・変換する。

まず図 1 の指定に基づいて、Step 1 と Step 2 で解析対象のテキストを絞り込む。章節構造解析は、文頭のパターン（「第〇部【」など）を正規表現でマッチングし、章節の始まりから終わりまでを検出するアルゴリズムを実装した。Step 3 の係り受け解析結果を受けて、Step 4 では、関係抽出ツールの StruAP を使って、ベンチマーク名称を含む文節を特定する。図 1 の指定では、benchmark という名称で木構造パターンが定義された関係の a2 のラベルの文節を対象としている。図 2 に StruAP での木構造パターンマッチの例を示す。図 2 の右に示すような、構文木を抽象化したパターンをあらかじめ定義しておき、このパターンにマッチした場合に、a0、a1、a2 などに該当する部分木を抽出することができる。図 2 の例では、a0 として「運用上のベンチマークとして」が、a2 として「日立インデックスを」が抽出される。`\dic.benchmark` の部分で、別の辞書ファイルに記述された語のセットを呼び出してパターンマッチに使うことができる。StruAP は、「日立インデックス」などのベンチマークとして抽出したいもの自体の辞書ではなく、木構造のパターンと関係を表す表現の辞書を使って抽出する。最後に Step 5 で抽出された文節に対して正規表現により、トリミング・変換を行い、所望の情報を抽出する。図 1 の例では、「中長期的に対象株価指数であるボベスパ指数を」が「ボベスパ指数」などに変換される。

3 有価証券届出書からのファンド情報抽出

有価証券届出書を対象として、文書で報告されているファンドのベンチマークや手数料の上限など、投資家が投資判断を必要とする 54 項目の情報の抽出を行った。この 54 項目の選定の際に、技術的な抽出可能性については考慮せず、実際のニーズのみに基づいて決定した。StruAP の木構造パターンに関しては、474 のパターンを定義した。StruAP の辞書に関しては、80 語を登録した。

表 1: 有価証券届出書からのファンド情報抽出結果

正しく抽出できた書類の割合	項目数
100%	7
95%以上 100%未満	12
90%以上 95%未満	2
75%以上 90%未満	9
50%以上 75%未満	8
25%以上 50%未満	4
0%以上 25%未満	12

表 1 に結果を示す。54 のそれぞれの項目に対して、379 の有価証券届出書のうち、どれだけ正しく抽出できたかを調べた。90%以上の文書で正しく抽出できた項目は、54 項目中 21（7+12+2）である。一方、75%未満は 24 項目もあった。今回の実験では、表の解析を行っていないが、文書によっては抽出対象の項目が表に記載されていることを確認している。そのため、我々の推算では表から情報抽出を行うことで、精度が 75%未満となる項目の数は 9 項目まで減らすことができる見通しである。

4 有価証券報告書からの経営指標抽出

有価証券報告書を対象として、経営指標の抽出を行った。例えば、

表 2: 有価証券報告書からの経営指標抽出結果

抽出できなかった企業の数	6	（全 30 社中）
誤抽出された経営指標の数	3	（全体の抽出数は 41）

> 経営目標の達成を判断するための指標として、当社は連結 ROE を公表しております。(中略) 当連結会計年度における連結 ROE は 8.0 %であり、前年度と比べて 0.3 ポイント上昇いたしました。

から「連結 ROE」のようなその企業が目標としている経営指標と、「8.0%」などのその指標に関連する現状値や目標値を抽出する。金融庁の政策オープンラボの実証実験においても、有価証券報告書の中での経営指標の記載のされ方に注目している [6]。StruAP の木構造パターンに関しては、62 のパターンを定義した。StruAP の辞書に関しては、94 語を登録した。この 94 語に関しては、関係を表す表現であり、経営指標を表す語は含まれていない。

本実験では、定量的な指標のみを抽出対象とした。企業が目標として設定していても、定性的な目標は達成を定量的に測れないため、今回の抽出対象からは除外した。有価証券報告書から上記で作成した StruAP の木構造パターンを用いて、経営指標の候補を自動で抽出し、それらを人手で名寄せして整理することで、経営指標の辞書を作成した。結果、有価証券報告書の実際の記載に基づいて、367 語、239 種類の経営指標を含む辞書を整備でき、これを抽出結果に対するフィルタリングに用いた。

表 3 に結果を示す。30 社の有価証券報告書に対して、どれだけ正しく抽出できたかを調べた。まず 30 社のうち 6 社分の有価証券報告書に関しては、有価証券報告書の中に経営指標が記載されているにもかかわらず、経営指標の抽出ができなかった。すなわち企業ごとに算出した再現率は 80%といえる。また 30 社全体で、41 の経営指標が抽出され、そのうち 3 つは、文字列としては記載があるものの、その企業が経営目標の KPI として挙げていないものであった。すなわち抽出データに関する適合率は 93.7%といえる。

なお抽出した経営指標の項目に関して、記載している企業が多い順に並び替えた結果を表 3 に示す。なお本結果は、2018 年 3 月期に提出された 2,688 社分の有価証券報告書に対して経営指標を抽出して集計した結果である。表 3 から、売上、利益、キャッシュフローなどに加えて、自己資本利益率も重視されていることがわかる。

表 3: 抽出された経営指標

項目	企業数
売上高（既存店売上高、総売上高など含む）	2,064
純利益	1,933
営業利益（調整後営業利益など含む）	1,933
キャッシュフロー	1,070
純資産	1,039
営業キャッシュフロー	967
セグメント利益	615
ROE（株主資本当期純利益率など含む）	615
自己資本比率	604
売上総利益	507

5 考察

本稿では、構文木の木構造のパターンマッチにより情報抽出を行った。本稿で対象としたような情報抽出において、キーワードの辞書によるマッチングは、あらかじめ抽出対象となるものを網羅的に整理しておく必要があり実用的ではない。同様に、先にエンティティを同定して、その周辺の語を特徴量として抽出対象かどうかを判別する手法も現実的ではないと思われる。

有価証券届出書からの抽出対象とした 54 の項目には、既存の DB に登録されている属性名と属性値のペアを利用した Distant supervision に基づく情報抽出手法 [2] が有用と考えられる項目もある。今回の研究では、金融分野に関する既存の大規模 DB がなかったこと、また数量表現である項目もあることから、係り受け規則による抽出を行っている。数量表現抽出に係り受け規則を用いるものとしては、藤林らの研究 [5] が挙げられるが、有価証券届出書においては、数値と属性が同じ文に書かれておらず、章題から属性を特定する必要がある場合もある。高野らは有価証券報告書からその企業の事業セグメントごとの業績要因文と業績結果を抽出しており [4]、業績結果として、売上、営業利益、経常利益、当期純利益を対象としている。しかし、企業が目標とする経営指標はその企業の方向性を示すとも考えられ、企業ごとに異なるものであり、この 4 つ以外を目標とする企業も存在する。本稿の 4 節の実験では企業の経営指標の抽出を目的とし、239 種類の経営指標が抽出できることを確認した。

3 節の結果のエラー分析からわかったことを以下にまとめる。

- 自然言語で書かれたテキストを解析するのみならず、表からの情報抽出も必要である。これらの文書における表の構造は多種多様であり、単に HTML をパースして行列形式のデータに変化するだけでは不十分で、人が表から意味を読み取るやり方を一般化して、人と同様のやり方で表から情報を読み取る必要があると思われる。
- 1 つの書類に 2 つ以上のファンドの情報が記載されていることがある。これらの書類においては、ある項目に関しては共通で 1 つのみ情報が記載され、別の項目では異なる情報は並列に記載される。これらの情報を適切に抽出するためには、文書の章節構造のみならず、情報の並列性や共通性を認識して情報を抽出する必要がある。

また 4 節の結果のエラー分析からわかったことを以下にまとめる。

- 箇条書きの解析が必要である。自然言語の構文構造ほど複雑ではないが、箇条書きの構造を解析したり表現したりする標準的な方法やツールがない。表と同様に、自然言語で書かれた場合よりも、情報同士の関係が省略して書かれるため、意味の推測の必要性がより増す。
- 企業の経営指標と、セグメントごとの経営指標が別々に設定されていることがある。より正確かつ推測なしで抽出するためには、企業の構造と文書の構造のマッピングをとることまで実施する必要があると思われる。
- 経営指標が非財務指標である場合は、企業ごとに名称が異なるため、名寄せが困難である場合が存在する。例:「お客様数」「サービス別月次利用数」

6 おわりに

本稿では、有価証券報告書および有価証券届出書等から、投資家が必要とする情報を抽出した。本稿では、特に自然言語処理により情報を抽出する部分に注力し、構文木の木構造のパターンマッチを行う StruAP を使って、関係を表す表現に注目して情報を抽出した。文書の記載のされ方が多様であるにもかかわらず、効果的に抽出できていると思われる。

実務においては、多様な項目の抽出が求められ、高精度な抽出の実現のためには、自然言語処理以外の技術も必要である。本稿では、簡単な章節構造解析を用いたが、これ以外にも、複数のファンドの情報が並列に記載されている場合の文書構造を理解する手法や、表から情報を抽出する手法が必要である。特に、表に関しては、HTML で書かれているものの、同一の項目であっても文書によって情報の記載のされ方が多様であり、多様な構造の表に対して、意味的な構造を解析して、統一的な方法で情報を抽出できる手法が必要であると思われる。今後は、表の解析に注力し、抽出の精度を向上させる予定である。

参考文献

[1] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *Proceedings of CoNLL 2002 Post-Conference Workshops*, pp. 63–69, 2002.

[2] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP 2009*, pp. 1003–1011, 2009.

[3] Kohsuke Yanai, Misa Sato, Toshihiko Yanase, Kenzo Kurotsuchi, Yuta Koreeda, and Yoshiki Niwa. StruAP: A tool for bundling linguistic trees through structure-based abstract pattern. In *Proceedings of EMNLP 2017: System Demonstrations*, pp. 31–36, 2017.

[4] 高野海斗, 酒井浩之, 北島良三. 有価証券報告書からの事業セグメント付与された業績要因文・業績結果文の抽出. 人工知能学会論文誌, Vol. 34, No. 5, 2019.

[5] 藤畑勝之, 志賀正裕, 森辰則. 係り受けの制約と優先規則に基づく数量表現抽出. 情報処理学会研究報告. FI,. 情報学基礎研究会報告, Vol. 64, pp. 119–125, 2001.

[6] 金融庁政策オープンラボ. 有価証券報告書等の審査業務等における AI 等利用の検討-実証実験の結果概要について, 2019. `https://www.fsa.go.jp/news/r1/openlab/20190927/20190927.html`.

[7] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.