言語処理技術を用いた有価証券報告書の分析

アシュアランス・イノベーション本部　AIラボ　佐野智久

▸ Tomohisa Sano, Ph.D.

2017年、当法人入所。仕訳の異常検知システムEY Helix GLADや不正会計予測システムDolphin等の開発・構築に従事。監査業務における、機械学習を用いたテキスト情報の活用に関する研究開発に取り組んでいる。博士（工学）。

Ⅰ　はじめに

有価証券報告書の記述情報は財務情報を補完する役割を持ち、企業の状況を理解するための情報や財務情報を理解するための文脈の情報等が記載されています。この記述情報を充実させることで、投資家のより適切な投資判断を促し、投資家と企業との間の建設的な対話が促進されます。近年では、コーポレートガバナンス情報の記載や経営戦略、リスク情報等に関する記載の充実が求められており、有価証券報告書の記述情報の見直しを目的とした開示府令の改正も行われています。従来の有価証券報告書の記述情報はボイラープレート化しているという批判もありましたが、これらの取り組みにより、有価証券報告書の記述情報は今後さらに充実していくことが予想され、より重要な意味を持つことになります。

当法人では、リスク評価の観点でAIの活用に取り組んでいます※1。有価証券報告書の記述情報や特定の記述項目の有無の情報を用いた分析は主にリスク評価の監査の計画段階においてさまざまな応用が考えられます。例えば、同業他社と大きく異なるリスク認識の発見や監査人によるビジネス理解との相違の確認、有価証券報告書が将来的に訂正される可能性の予測、訂正の可能性の高い有価証券報告書に含まれる表現の識別等が期待される応用例として挙げられます。本稿では、XBRL形式で公開されている有価証券報告書のデータを分析した取り組みについて紹介します。

Ⅱ　XBRL形式の有価証券報告書の取得

金融庁の電子開示システムEDINET（Electronic Disclosure for Investors' NETwork）からXBRL（eXtensible Business Reporting Language）で記述された有価証券報告書や四半期報告書等のデータを取得することが可能です。XBRLはXMLと呼ばれるデータ表現方法を財務報告用に特化させたものです。XMLは1998年に登場し、レイアウトやデザインに関する情報を持たずにそのデータの意味を記述することができることから、コンピューター関連のドキュメントフォーマットのデファクトスタンダードとなりました。XML形式のドキュメントは複数のシステム間で容易に連携することができるという特徴があり、プログラムで扱うことに適しているといえます。

XBRL形式の有価証券報告書はEDINET APIを利用して取得することが可能です。このAPIを利用することで、EDINETのウェブページを表示することなく、プログラムを介してXBRL形式の有価証券報告書のデータを効率的に取得することができます。EDINET APIでは、特定の日付に提出された書類の一覧を取得したり、特定の文書IDを持つ書類をXBRL形式やPDF形式で取得したりすることができます。

本稿の分析で利用しているデータは2016年から2019年までに提出された有価証券報告書をEDINET APIを用いて取得したものです。

※1　本誌2020年10月号「リスク評価におけるAI活用について」

Ⅲ　有価証券報告書に出現する語句の出現頻度の変化の度合いに関する分析

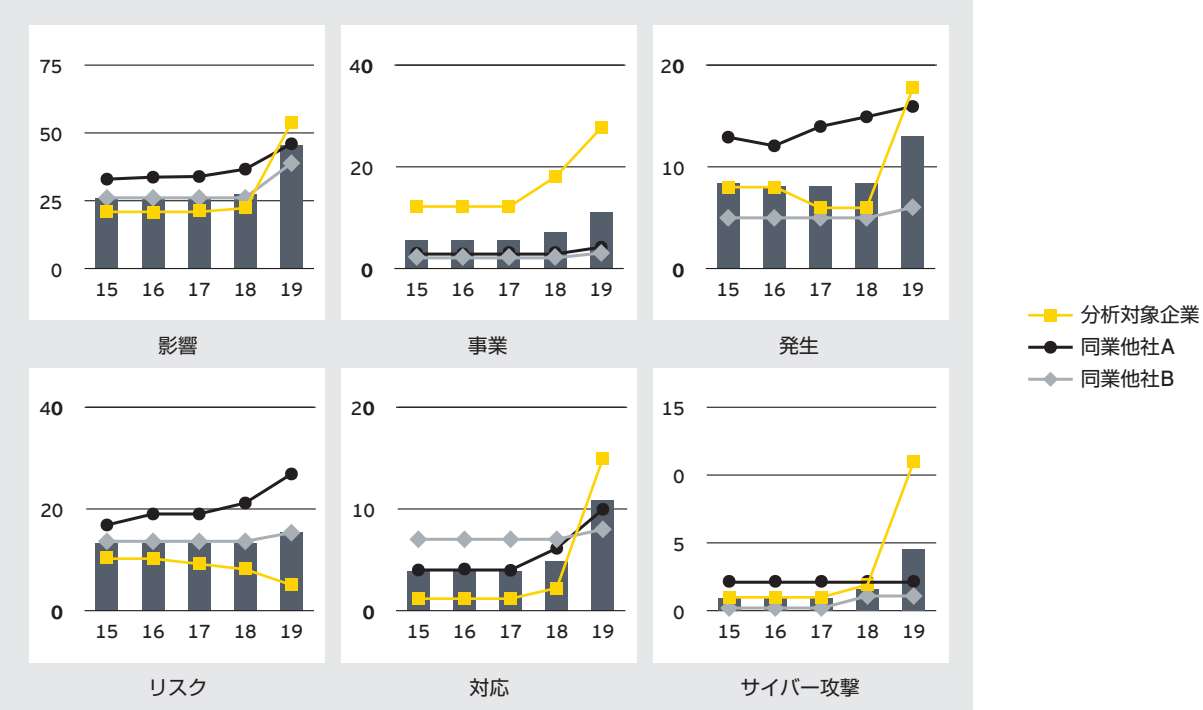
ある企業の複数年度にわたる有価証券報告書を比較することで、その企業のリスク認識の経年変化を捉えることができます。さらに、その変化を同業他社における変化と比較すれば、その企業が考えるリスク認識とその業種の一般的なリスク認識との間の相違の有無を判断することも可能です。

ここで、この変化の度合いを、対象となる企業の有価証券報告書での語句の出現回数と比較対象の企業グループでの語句の平均出現回数の差分の大きさを一定期間内で集計したものと定義します。ある語句に着目したときに、その変化の度合いが高く、さらに上昇傾向にある場合には、その語句に関連する出来事がその企業に関するホットトピックであると考えることができます。＜図1＞は、ある企業の有価証券報告書を対象として、その同業他社の有価証券報告書と比較した場合に変化の度合いが大きな一部の語句について、時系列的な変化を図示したものです。＜図1＞の「サイバー攻撃」のグラフから、この企業で2018年度以降にサイバー攻撃に対するリスクの認識が高まっていることが読み取れます。この企業は2017年にランサムウェアの被害にあっており、サイバー攻撃への対応や対策を強化しています。このように語句の出現頻度の変化の度合いに基づいて可視化を行うことで、その企業に対するビジネス理解やリスク認識を深めることができるようになります。

Ⅳ　訂正される有価証券報告書に出現しやすい語句の抽出

有価証券報告書の「事業等のリスク」の項目には、企業の財政状態、経営成績およびキャッシュ・フローの状況等に重要な影響を与える可能性があると経営者が認識している主要なリスクが記載されています。リスクに対する認識が異なれば記載内容が異なるため、記載されているテキストに含まれる語句の分布も有価証券報告書ごとに異なります。ある企業のグループの有価証券報告書における語句の分布と他のグループにおける語句の分布を比較することで、有価証券報告書のグループ間の特徴を見つけ出すことができます。例えば、過去に有価証券報告書を訂正したことがある企業のグループと一度も訂正したことがないグループとに分けた時、それぞれの有価証券報告書のテキストに現れる語句の傾向を分析することで、有価証券報告書の訂正の可能性の有無を予測することができる見込みがあります。ただし、有価証券報告書に記載されているテキストの長さはまちまちであることから、単純に語句の出現回数を比較することはできません。そこで、語句の出現回数ではなく、相対的な出現頻度の比率に着目します。語句tに対する相対的な出現頻度の比率$\mathrm{rtf}(t)$は、次式のように定義することができます。

▶図1　語句の出現頻度の時系列的変化



$$rtf(t)=\frac{P(t;x)}{P(t)}=\frac{C_x(t)}{N_x}\Big/\frac{C(t)}{N}$$

ここで、$C_x(t)$は語句tがある有価証券報告書のグループxに出現した回数、$C(t)$は語句tが全有価証券報告書に出現した回数、N_xは有価証券報告書のグループxに出現した全語句の出現回数、Nは全有価証券報告書内の全語句の総出現回数を表しています。有価証券報告書のグループxを訂正が行われた有価証券報告書のグループとした場合、相対的な出現頻度の比率は訂正が行われた有価証券報告書におけるその語句の出現しやすさを表すことになります。

訂正が行われた有価証券報告書に出現しやすい語句を抽出する手順は次のとおりです。はじめに、有価証券報告書に登場する語句から相対的な出現頻度の比率がある閾値（いきち）を超える語句を抽出します。相対的な出現頻度の比率が高くなると、固有名詞や業界用語等、一部の有価証券報告書のみに頻繁に出現する語句が抽出される可能性も高くなります。これらのノイズを軽減するために、特定の有価証券報告書に偏って出現しているものを除外します。この偏りは、有価証券報告書に対する語句の出現頻度の分布の標準偏差を計算することで測ることができます。偏りが大きければ、その語句に対する標準偏差が大きくなるため、標準偏差がある閾値を下回っているような語句に絞り込みます。最後に、その語句が含まれる有価証券報告書の数が多い順に並べ替えます。このような手順によって抽出した語句を＜表1＞に示します。

▶表1　訂正された有価証券報告書に出現しやすい語句

年度	出現しやすい語句
2016	遅延等、有利子負債依存度、暴動、新興国、需給
2017	インフラ、需給、事業資金、保持、経済動向
2018	暴動、インフラ、新興国、保持、帰属
2019	改廃、連結財務諸表作成、津波、点検、洪水等

これらの語句は訂正の可能性の高い有価証券報告書に出現しやすい語句といえますが、依然として、一次的な出来事に関連したものがノイズとして残っています。そこで、抽出された語句の中から複数年度で継続的に出現するようなもののみに絞り込むために、全ての期間において一定の相対的な出現頻度の比率を上回る語句だけを抽出します。絞り込んだ結果の上位20個の語句は＜表2＞のようになります。

▶表2　有価証券報告書を訂正している企業が継続的に使用している語句

利益、固定資産、軽減、購入、フロー、部品、会計基準、税制、要求、ウィルス、新設、有効、原油価格、インフラ、電力、出荷、全国、カバー、遅延等、暴動

これらの語句は、有価証券報告書を訂正している企業が継続的に使用しているものといえます。つまり、これらの語句を多用する有価証券報告書は将来的に訂正される可能性が通常よりも高いと考えられます。

最後に、抽出されたこれらの語句を用いて訂正の予測モデルを作成したときに、それぞれの語句がこのモデルにおいて説明力を有しているかどうかを確認します。これらの語句を用いてロジスティック回帰による訂正の有無の分析を「事業等のリスク」と「注記事項」のテキストに対して行った結果は＜表3＞のようになります。

▶表3　ロジスティック回帰による訂正の有無の予測モデル

事業等のリスク	
語句	p値
要求	0.035*2
原油価格	0.089*3
購入	0.185
遅延等	0.191
カバー	0.233

注記事項	
語句	p値
評価方法評価基準	0.001*1
海外連結子会社	0.009*1
リース資産自己所有	0.035*2
個人債務	0.058*3
当期純損失	0.113

＊1：1%有意　＊2：5%有意　＊3：10%有意

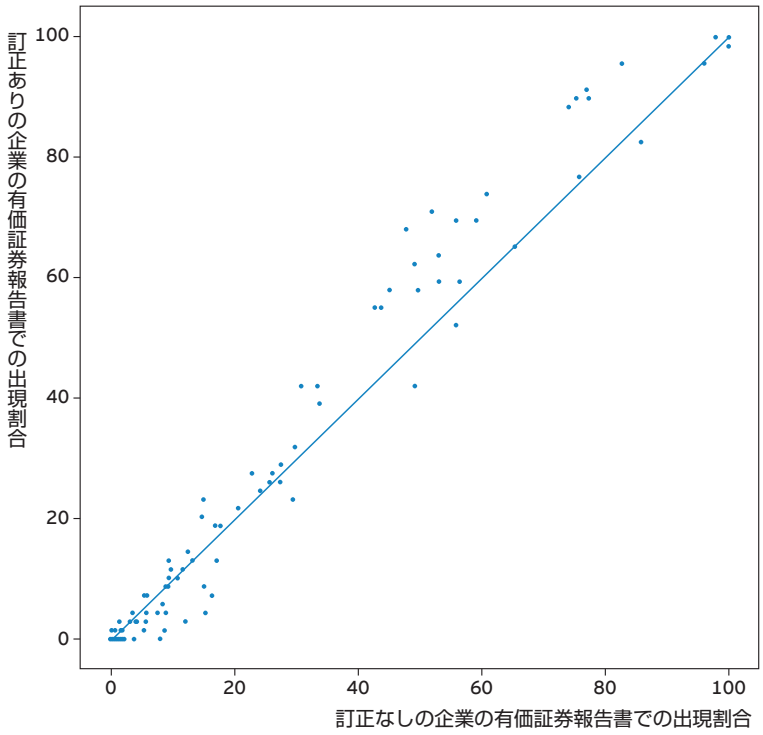
対象とするテキストによって違いはあるものの、p値として有意※2となる語句が幾つか識別されました。さらに高い予測の精度を達成するためには、事業等のリスク以外の記載内容の利用、PL項目に影響のある訂正に限定した分析、構造化された財務データと併せた分析等の取り組みが考えられます。

Ⅴ　注記事項の出現傾向の分析

重要な後発事象や会計方針の変更等の注記事項は、財務諸表に重要な影響を及ぼす可能性のある項目です。これらの注記事項は、XBRL形式ではそれぞれ異なるタグが付与されており、それぞれの注記事項の出現の有無を簡単に集計することができます。2016年に訂正された有価証券報告書に含まれる注記事項と、一度も訂正されたことのない企業の有価証券報告書に含まれ

※2　p値が有意水準を下回れば訂正の有無で語句の出現傾向に差がないという仮説が棄却され、訂正が行われた有価証券報告書で出現しやすい語句となる可能性が高いといえる。

▶図2　注記事項の出現割合



る注記事項の出現頻度を図示すると、＜図2＞のようになります。各プロットがそれぞれの注記事項の種類を表しています。

縦軸は訂正された有価証券報告書でそれぞれの注記事項が出現した割合を表し、横軸は一度も訂正されたことのない企業の有価証券報告書でそれぞれの注記事項が出現した割合を表しています。訂正の有無によって注記事項の出現傾向が変わらなければ、縦軸の値と横軸の値が同じとなり45度線上付近にプロットされるはずです。実際には、訂正された有価証券報告書に出現しやすい注記事項（45度線の左上側）が幾つかあることが分かります。

これらの注記事項の中から、訂正された有価証券報告書のグループと訂正が一度もなかった有価証券報告書のグループとの間で出現頻度に有意差があるものを識別するために、カイ二乗検定を利用します。分析の結果、「開示対象特別目的会社関係」「関連当事者情報」「表示方法の変更」「重要な後発事象」等の注記事項がグループ間で有意差を出し、訂正が一度もなかった有価証券報告書のグループよりも訂正された有価証券報告書で出現しやすいことが分かりました。しかし、訂正された有価証券報告書での出現確率は、一度も訂正されたことのない企業の有価証券報告書での出現確率の1.2倍から1.3倍程度であり、さらなる改善の余地が残されています。例えば、注記事項の内容のテキスト分析との組み合せやその他の財務データとの掛け合わせによって、訂正の有無の予測精度のさらなる向上が期待できます。

Ⅵ　おわりに

本稿では、有価証券報告書の記述情報や注記事項の項目データを分析する取り組みについて紹介しました。テキスト情報をはじめとする非構造化データは企業が保有するデータ全体の8割にものぼるといわれています。これらのデータの分析や活用により、財務数値からだけでは読み取ることのできないリスクを自動的に識別したり、不正の兆候を検知したりすることが実現されるでしょう。今後、有価証券報告書の記載内容が画一的なものでなくなり、さらに充実したものとなってくれば、リスク評価における重要な情報がより多く含まれるようになると期待されます。米国企業を対象に行われた先行研究での可読性やトーンの情報※3、監査上の主要な検討事項（KAM）のデータの分析等を取り入れることで、財務データの分析の補完や、予測を裏付ける要因に関連する情報の自動抽出の精度向上につながると考えます。

※3　本誌2019年5月号「テキスト分析と会計学研究」