有価証券報告書からのリスク文抽出の試み

An attempt to extract sentences concerning contents of risks from securities reports

藤井 元雅[1]　坂地 泰紀[2]　佐々木 一[3]　増山 繁[1]
Motomasa Fujii[1]　Hiroki Sakaji[2]　Hajime Sasaki[3]　Shigeru Masuyama[1]
[1]東京理科大学経営学部
[1]School of Management, Tokyo University of Science
[2]東京大学大学院工学系研究科
[2]Graduate School of Engineering, The University of Tokyo
[3]東京大学未来ビジョン研究センター
[3]Institute for Future Initiatives, The University of Tokyo

Abstract: Recognizing risks that are descriptive information is important to judge how companies manage their risks, that is, the consistency of the strategy when evaluating companies. So, in this research, risk is defined as "uncertainty of future results that can be obtained by taking actions", and we tried to extract the details of the risks recognized by each company from the text of the securities report. We also tried to classify the extracted risk based on their respective contents. The extraction method is investigating and classifying the expression patterns (possibly, etc.) used when expressing the content of risk, and using them as clues.

1. はじめに

今や環境・経済・社会が事業に与える影響は多くの企業にとって無視できないリスク要因であり，SDGs や ESG に寄与する活動に取り組まないこと自体もまたリスクになり得ると指摘されている[1]．また，Covid-19 の世界的流行に対する事業継続計画(BCP)の策定内容が企業価値評価を行う投資家の間で重視され始めている．このような事業リスクをいかに把握・対処しているかを理解することは，企業活動や経営戦略の整合性の判断及び，企業価値評価の際に重要な役割を果たすと言える．例えば，上場企業が発行を義務づけられている有価証券報告書においては財務諸表だけでなく，業績要因やその企業の潜在的なリスク等多岐にわたる項目が記載されている．またその量は年々増加しており，この会計書類のみで様々な情報を入手することが可能になっている．現在，国や企業はこれらの情報公開に積極的になりつつある[2].

一方で，実際の企業価値評価に際しては過年度や同業他社のもの等複数の有価証券報告書の確認が必要であることや前述のような記載内容の増加に伴い，処理しなければならない情報は膨大なものとなっている．すなわち，テキストマイニング等の人工知能分野の手法により情報を効率的に処理する必要性が高まっていると言える．

そこで本研究では，企業が想定するリスク内容を抽出・分類し企業価値評価に役立てる前段として，有価証券報告書内の「事業等のリスク」の項目において，企業毎に異なるリスク表現の方法を分析し，企業が想定するリスク内容を含む文を抽出する．対象として有価証券報告書を用いた理由としては，金融商品取引法で規定された開示資料であり，規格が同一であるためである．リスクの定義については，

リスクの内容が企業の行動の前提にあることに焦点を当てた定義が必要であると考え，今回の判断基準を採用した．投資家は社会的責任を含めた，その企業理念やスタンスが一貫しているかを重視しており，その判断材料には具体的な企業活動の情報が必要不可欠であること，またその活動がどのような不確実性を伴っているのか，ということも同様であると判断したためである．

先行研究として，キーワードを設定することによりどのようなリスクが記載されているかを確認する研究[2]は存在するが，あくまで全体像を把握するためのものであった．よって本研究においては，最終的な目的である多様化するリスクについての評価を行う際に具体的な内容についての把握が難しいと判断し，様々な手法を用いて抽出を試みた．

2. 関連研究

関連研究としては，キーワード検索を実行して，Scopus と Engineering Village のデータベースから関連記事を取得し，機械学習手法を利用して出版物の構造化されたレビューを提示し，エンジニアリングリスク評価を支援する研究[3]や，予期せぬ事態に対処し，リスク管理を可能にするための適切なサポートを提供するために，機械学習，特に深層ニューラルネットワーク(DNN)モデルに基づくリスクアセスメントアプローチの提案を行っている研究[5]が挙げられる．財務諸表を用いたものとしては，テキストマイニング手法を用い財務諸表に報告されたテキストリスク開示からエネルギー企業のリスク要因を総合的に特定し，またそこからリスク要因階層システムの構築が試みられている[6].

また，企業が想定するリスクの内容については，リスクアプローチの洞察と予期せぬ事象に対する実用的なガイドラインを提示している文献[7]や，リスク評価が実際にリスクに及ぼす影響について分析している文献[8]がある．

3. リスクの定義

リスクの定義について，リスクマネジメントの国際規格「ISO31000:2018　リスクマネジメントー指針」[9]において，「effect of uncertainty on objectives(目的に対する不確実性の影響)」とされている．

一方，有価証券報告書の記載内容を定めている「企業内容等の開示に関する内閣府令」の 2019 年の改正[10]においては，記載すべきリスクは「連結会社の経営成績等の状況の異常な変動，特定の取引先・製品・技術等への依存，特有の法的規制・取引慣行・経営方針，重要な訴訟事件等の発生，役員・大株主・関係会社等に関する重要事項等，投資者の判断に重要な影響を及ぼす可能性のある事項」とされている．すなわち有価証券報告書においては，企業の目的ではなく，それを実現するための行動(事業活動)によって生じる不確実性により，焦点が当てられていると言える．

例えば，2018 年 12 月期のヤマハ発動機株式会社(以下，ヤマハ発動機)においては「当社グループの日本における主力製造拠点は，予想される南海トラフ巨大地震の震源域近傍に集中している」といった文言が確認できる．これは，ヤマハ発動機は目的を達成するために生産活動を行っていることは明白であるが，製造拠点の確保という現在の企業活動こそが直面しているリスクを表していると見ることができる．

以上の点や文献[11]を参考とし，本研究では，リスクを**「行動したことにより獲得できる将来の結果の不確実性のこと」**と定義する．なお，ここでの結果とは企業の業績のことであり，不確実性とは，どのような状態や結果が出現するかはわかっているが，状態や結果の出現確率がわからない状況，すなわち，意思決定環境に応じた不確実性の分類における「不確実性下の曖昧性下」にあたる状況[12]としている．これは，前述の巨大地震の例を始めとした企業が直面する(有価証券報告書に記載される)多くの事象の出現確率の測定が困難であるためである．言い換えると，企業は現在の企業活動を踏まえた上で業績を予測しているが，それらは当然実際とは異なる場合があり，それらの要因となるものをリスクと定義しているということである．

4. リスク文へのタグ付け

評価データには，日経 225 に指定されている企業からランダムに 70 社を選択し，その有価証券報告書を用いた．有価証券報告書は PDF ファイルをテキスト化し，「事業等のリスク」の項目を抽出し，文単位に分割した．該当する文は計 5,007 文であり人手でタグを付与した結果，リスクの内容を含む文(リスク文と定義する)は 2,588 文，リスク文でない文は 2,419 文であった．このうち 7 割にあたる 3,507 文を学習データとして用いた．

5. 分類手法

本節では，本研究で用いる分類手法について紹介する．以下に本研究で用いた分類手法を示す．

パターンマッチング

複数の文字列や図形等を比較し同一あるいは類似したものであるかどうか，またどこに出現するのか

を調査する．詳細については 5.1 節で後述する．

SVM[1]

機械学習の一種であり，データ空間上に識別超平面を構築し，その際には分類がきわどいデータのみを上手く選んで，それらと識別超平面がなるべく離れるように学習を行う方法である．

Logistic Regression

定性的データの従属変数を予測・説明するために用いられる多変量解析手法であり，目的変数に対する影響の大きさを調べることができる．

Random Forest

Random 重複を許すランダムサンプリングによって多数の決定木を作成し，各木の予測結果の多数決をとることで最終予測値を決定する．

Bidirectional LSTM

系列データをうまく扱うために開発された Recurrent Neutral Network(RNN)の一種である Long Short-Term Memory(LSTM)[13]を用いる．本研究では，文頭から文末までと，文末から文頭までの双方向の情報を利用したいことから，Bidirectional LSTM(BiLSTM)を採用した．これは，語の曖昧さ解消など前方の文脈だけでなく後方の文脈が必要な場合があり，双方向の方が性能は高い場合が多いためである．特に，有価証券報告書は一文が長い傾向にあるため，前・後方の文脈の両方を考慮する必要があり，片方向だけでは情報が落ちてしまう可能性がある．加えて，要素間の関係性や全体のコンテキストを考慮する Attention 機構を適用している．

BERT

BERT は，Devlin ら[14]によって提案された大規模言語モデルで，様々なタスクにおいて優秀な成績を収めている

本研究では，fine-tuning することで，リスク文の抽出に利用する．

5.1. パターンマッチングによる抽出

本研究においては，有価証券報告書においてリスクを表現するパターンとして以下の 2 つに分類した．なお，以下の分類においては文献[15]を参考とした．

まず，論理文(複文の従属節の部分を前件，主節の部分を後件とするとき，前件の命題の真偽が後件の真偽に関連している関係(因果関係)を表す複文)に用いられる接続詞のうち，前件がまだ発生していない(前件が既に発生している場合，それは不確実なものではなくリスクの定義から外れる)場合に用いられるもの，あるいはこれに置き換えることができる表現．

次に，出来事の内容(命題)を可能性があるものとしている，あるいはこれに置き換えることができる表現．なお，このような条件を満たす場合でも，命題が「業績が低下する」「財政状態が悪化する」といった企業活動の最終的な成果を表している場合は除外される．理由としては，前述のような命題はリスクの定義における「結果」に該当するものであり，抽出すべき内容と異なっている(抽出すべきは結果が不確実な理由)ためである．

最後に，上記の表現には当てはまらないが，「～変動の影響を受ける」といった間接的な表現や「市場価格」といった不確実性が想起される語による表現．

上記 3 つの表現をそれぞれ，仮定表現，可能性表現とする．それぞれの表現のパターンの具体例は以下の通りである．

i.仮定表現を示すパターンの例

場合(には|は|においては) ときは すれば すると 局面では 計 7 例

ii.可能性表現を示すパターンの例

可能性(がある|生じる|否定できない|皆無ではない|ないとは言えない) リスク(がある|が考えられる|が存在する|が内在する|を有している|を含む|にさらされている|となり得る|が高まっている|を抱えている|を負っている|は避けられない|を伴う) ことがある 想定される かもしれない 考えられる 場合がある 恐れがある 懸念がある 計 25 例

6. 評価実験

どの分類手法が本タスクにおいて有効かを調べるために，評価実験を行った．本研究における後述の手法いずれにおいても学習の際には，Mecab[2]による形態素解析を用いて，形態素を抽出し，素性に利用している．また，係り受け解析器としては Cabocha[16]を用いた．Logistic Regression，Random Forest の実装には gensim を利用し，SVM の実装としては学習器には SVM-light を用いた．BiLSTM の実装には，Pytorch を利用し，BERT は東北大学乾研究室が公開している日本語 Wikipedia から学習したモデルを利用した．

なお，精度，再現率，F 値の算出にはいずれもマクロ平均を用いている．また，評価実験に用いたデータの詳細は表 1 に示す．

[1] http://svmlight.joachims.org/

[2] http://taku910.github.io/mecab/

	正例	負例	総数
学習データ	1,834	1,671	3,505
検証データ	498	503	1,001
テストデータ	256	245	501

表1　データ内容

	精度	再現率	F値	抽出数
可能性表現	0.44	0.95	0.6	2,470
仮定表現	0.75	1.00	0.85	1,590

表2　パターンマッチングの評価結果

	精度	再現率	F値
パターンマッチング	—	—	—
SVM	0.93	0.80	0.87
Logistic Regression	0.90	0.90	0.90
Random Forest	0.91	0.91	0.91
BiLSTM	1.00	1.00	1.00
BERT	0.99	0.99	0.99

表3　評価結果

7. 考察

実験結果を表 2・3 に示す．本手法においてはパターンマッチングに比べ，SVM や LR，RF が再現率を除き，高い性能を示した．これによりリスクを表現するとした語形がそうでない文においても多く用いられていると考えることができる．以下は，抽出が適切に行われなかった文の一例である．

①当社グループの財政状態及び経営成績に影響を及ぼす可能性のあるリスクには以下のようなものがあります.
(富士フイルムホールディングス株式会社)

②しかしながら，精密な故に荷役や輸送段階における軽微な衝撃等によって全損害となり，高価格化が故に損害が拡大するリスクをはらんでおります.
(キヤノン株式会社)

③当社グループの業績は上半期と下半期を比較した場合，下半期の業績がよくなる傾向にあります.
(横浜ゴム株式会社)

④このような急激な価格変動が長引かない，あるいは，これまでこのような変動がなかった市場で発生しないという保証はありません.
(スズキ株式会社)

①は「可能性表現」を含むとして抽出されたが，これは記載するリスクについて説明する文章であり，リスクの内容を抽出できたとは言えない．逆に，②は「リスクがある」という言い方を変化させたものであり，これに限らず企業によって多くの異なる表現が確認できる．③は仮定表現として抽出されたものの，この「場合」は条件を説明するために用いられており，不確実な事象を示しているものではない．また，④のように「可能性表現」や「仮定表現」のいずれにも該当しないが「価格変動」といった語(変動自体が予測することができないというニュアンスを含んでいる)や「保証はない」という遠回しな言い方でリスクを表現している文はリスク文として抽出することができなかった．特に④のような事例を抽出するにあたっては，語形によるリスク内容の抽出には限界があると考えられる．

さらに，いずれの手法よりも BiLSTM や BERT の学習による性能が極めて高いことから，リスクの内容を含む文が定型的であり多様性があまりないこと，リスクを表現するのに際し文脈あるいは語句の順番が強い影響を与えていると考えることができる．なお，文が定型的であるということについては，分析対象が有価証券報告書という法令により形式が定められた書類であることも一因であると考えられる．

また，誤差だと考えられるが，BERT よりも BiLSTM の結果が上回った．このことから，対象言語データを有価証券報告書に絞る限りは，BERT ほどの複雑なモデルを用いる必要性が低いと考えられる．すなわち，文脈情報を利用できていれば，十分に分類可能なタスクであるということがわかった．

8. まとめ

本研究では，有価証券報告書からリスクの内容を含む文抽出を様々な手法を用いて行った．その中においては BiLSTM における学習が，精度・再現率・F 値のいずれも非常に高い性能を示した．しかし，文単位で抽出するとリスクと関係のない内容が混入する恐れがある．そのため今後，文単位ではなく表現単位での抽出を行う予定である．また，抽出したリスクの内容について分類を行う等，抽出された文の事業リスク把握に対する有効性について検証していきたい．

参考文献

[1] 経済産業省,「SDGs 経営ガイド」, 2019 年 5 月.

[2] 一般財団法人企業活力研究所,「新時代の非財務情報開示のあり方に関する調査研究報告書～多様なステークホルダーとのより良い関係構築に向けて～」, 2018 年 3 月.

[3] 張替一彰,「有価証券報告書事業リスク情報を活用したリスク IR の定量評価」,『証券アナリストジャーナル』, 第 46 巻第 4 号, pp. 32-44, 2008 年 4 月.

[4] Jeevith Hegde, Børge Rokseth, ” Applications of machine learning methods for engineering risk assessment – A review”, *Safety Science*, Volume 122, 2020.

[5] Lu Weiab, Guowen Liab, Xiaoqian Zhua, Xiaolei Suna, Jianping Liab, ” Developing a hierarchical system for energy corporate risk factors based on textual risk disclosures”, *Energy Economics*, Volume 80, 2019.

[6] Nicola Paltrinieria, Louise Comfortb, Genserik Renierscde, ” Learning about risk: Machine learning for risk assessment”, *Safety Science*, Volume 118, 475-486, 2019.

[7] Terje Avena, Bodil S.Krohnb, ” A new perspective on how to understand, assess and manage risk and the unforeseen”, *Reliability Engineering & System Safety*, Volume 121, 1-10, 2014.

[8] Graham D.Creedy, ” Quantitative risk assessment: How realistic are those frequency assumptions?”, *Journal of Loss Prevention in the Process Industries*, Volume 24, Issue 3, 203-207, 2011.

[9] International Organization for Standardization, ” ISO 31000:2018 Risk management — Guidelines”, 2018.

[10] 金融庁,「企業内容等の開示に関する内閣府令の一部を改正する内閣府令」, 2019 年.

[11] 伊藤邦雄,『新・企業価値評価』, 日本経済新聞出版社, 2014 年 4 月

[12] 竹村和久, 吉川肇子, 藤井聡,「不確実性とリスク評価－理論枠組みの提案－」,『社会技術研究論文集』, Vol.2, pp.12-20, 2004 年 10 月.

[13] Sepp Hochreiter and Jürgen Schmidhuber, Long short-term memory, *Neural computation*, Vol. 9, No. 8, 1735–1780, 1997.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, ”BERT: pretraining of deep bidirectional transformers for language understanding”, CoRR, 2018.

[15] 庵功雄,『新しい日本語学入門』, スリーエーネットワーク, 2012 年 4 月

[16] 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.