金融テキストを対象とした有益情報抽出に関する研究

Proposal of methods for extracting useful information from financial texts for analysis

高野　海斗
Kaito Takano

2021 年 1 月

概要

近年，通信技術の発達に伴い大量の情報が Web 上に溢れるようになったが，人が一日に触れることができる情報には限界がある．金融業界でも人工知能分野の手法や技術を金融市場における様々な場面に応用することが期待されており，膨大な金融情報を分析し投資判断を支援する技術にも注目が集まっている．その中でも特に，金融テキストから投資判断において有益な情報を抽出し，抽出された情報と市場変動の関係性を発見し，市場分析に応用する研究は金融テキストマイニングと呼ばれている．投資判断において有益な情報とは，具体的には株価に影響を与えるような情報であり，営業利益のような業績結果だけでなく，そのような業績になった要因，さらには，役員人事，配当の実施の有無など，様々な情報が該当する．また，金融テキストマイニングで対象としているテキストは多岐にわたり，企業が公開している決算短信，有価証券報告書などの文書や，金融の専門家が発信する経済レポートはもちろんのこと，不特定多数の人が投稿している SNS や掲示板の情報なども，抽出や分析の対象である．

このような背景の下，本論文の前半では，金融テキストである株主招集通知を対象にした，投資判断において有益な情報の抽出に関する研究について述べる．株主招集通知とは，株主総会の開催前に株主へ送付される文書であり，その記載内容としては，株主総会で議論される決議事項，大株主情報，役員情報などである．その中には，配当の実施，役員人事など，株価に影響を与える可能性のある情報が多く記載されているが，ページ数が百ページを超えるものも珍しくない．したがって，投資判断に有用な情報が多く記載されているが，発行時期が株主総会開催時期に集中することもあり，資産運用業務で企業分析を行う部署では，株主招集通知から投資判断に有用な情報を確認する作業に膨大な労力を割いている．このような課題を解決するために，本論文では投資判断に有用な情報が何ページから何ページに記載されているかの推定を，ページ単位で実現する方法論の検討を行う．本研究により，確認したい情報が何ページから何ページまで記載されているのかを自動で推定することが可能になることにより，大幅な作業の効率化が実現可能になる．

本研究が従来の研究と大きく異なる点は，情報抽出のために文単位で区切り位置を推定するのではなく，ページ単位で区切り位置を推定する点と，モデルを学習するための学習データを自動生成している点が挙げられる．ページ単位で抽出を行うことにより，文単位での抽出よりも学習データの自動生成は容易になるが，ページの途中で次の記載内容が始まるといった独自の問題がいくつか存在する．また，自動生成で得られる学習データは，人手で作成した場合に比べ，精度やデータの偏り（バイアス）の問題が存在するため，それらの特性を考慮した上で学習させるモデルを選択する必要がある．そこで本研究では，人手で作成された学習データを用いた上で，該当ページに記載されている決議事項である議案の分類が可能であるかどうかの検討を行い，良好な精度で分類が可能であることを示した上で，ページ単位での分類や抽出にどのような問題があるのかの考察を行った．さらに，決議事項である議案だけでなく，大株主情報や役員情報などの情報が何ページから何ページに記載してあるかの推定を自動生成した学習データを用いて行うことにより，学習データの自動生

成によってどのような問題が生じるのかを議論した上で，それらの問題を軽減できるモデルの検討を行った．

本手法のルールベースによる学習データ生成により，人手による学習データ作成では生成することができない大量の学習データの生成が可能となったが，その反面，学習データに偏りが生じていることが明らかになった．そして，この偏りにより，ページに対しての分類を行う系列ラベリング問題において，CRF などの従来研究で良好な結果が得られる手法が必ずしも最良の結果が得られる保証がない．そこで本研究では，いくつかの従来モデルとの検討も行った上で，本手法による BiLSTM モデルがミクロ F1 値 0.970 と最も良好な結果であることを示し，従来の研究で有効とされている CRF 層などの追加が，自動生成した偏りのある学習データを用いる場合には，過学習の原因になっていることを示した．

本論文の後半では，同じく金融テキストである有価証券報告書を対象にした，投資判断において有益な情報の抽出に関する研究について述べる．有価証券報告書は，事業年度ごとに作成する企業内容の外部への開示資料であり，企業の概況，事業の状況，設備の状況など多くの内容が記載されている．有価証券報告書は，PDF だけでなく XBRL 形式のデータで提供されており，XBRL 形式のタグを解析することで，章ごとのテキスト情報を抽出することが可能である．しかし，章に限定して抽出しても，そのテキスト情報は依然として膨大である．そこで，本論文では有価証券報告書の 2 章「事業の状況」を対象に，投資判断において特に有益な情報である，企業の業績に関する要因について書かれた文（業績要因文）と，どれだけの売上高や経常利益だったのかについて書かれた文（業績結果文）の抽出を行い，抽出した業績要因文，業績結果文が，対象企業のどの事業セグメントに関するものであるかを自動付与する方法論について述べる．

上記を実現することにより，有価証券報告書に含まれる膨大なテキスト情報から投資判断に有用な情報のみを抽出することが可能になり，抽出した各文に対して事業セグメントが付与されていることで，業績要因文のみの抽出では把握することができないことを把握することが可能になる．例えば，自動付与した事業セグメントの情報から，業績要因文に対応する業績結果文が得られるため，売上高から事業規模を把握することが可能であり，経常利益の前年度比からセグメント別の事業の成長度合いを把握することも可能になる．

業績要因文の抽出は，決算短信を対象に学習データを自動生成し，分類モデルを学習させることで抽出が可能であることが先行研究によって示されているため，その従来手法を本論文で対象としている有価証券報告書に適用し，業績結果文の抽出はルールベースで行った．そして，有価証券報告書の特徴を利用することで，事業セグメントもルールベースで抽出し，抽出した業績要因文と業績結果文が，どの事業セグメントに対する記述であるかを，文の出現順序に着目することで自動付与した．評価実験の結果，事業セグメントの付与が正しく，業績要因文判定が正しいものの適合率は 0.693，再現率 0.725 であり，事業セグメントの付与が正しく，業績結果文判定が正しいものの適合率は 0.788，再現率 0.911 であった．

最後に，全体の結論と今後の研究展望について述べる．

目次

1　はじめに

1.1　背景

近年，機械学習などの手法が注目を集め，様々な分野への応用研究が活発に行われている．特に，データが紙媒体から電子媒体に移行したことにより，これまで必要だった多くの業務の自動化や半自動化が今後より進むと予測される．金融業界でも人工知能分野の手法や技術を金融市場における様々な場面に応用することが期待されており，膨大な金融情報を分析して投資判断を支援する技術にも注目が集まっている．さらに，最近では証券市場における個人投資家の比重が増大しており，個人投資家に対して投資判断の支援を行う技術の必要性が高まっている [1, 2].

また，計算機の発展に伴い，煩雑な計算が可能になったことで，複雑なモデルを学習させることが可能となった．これまでデータといえば数値データが中心であったが，複雑なモデルを学習させることが可能になったことにより，画像，音声，テキストなど様々なデータを扱うことが可能になり，現在ではそれらのデータを組み合わせた研究が盛んに行われている [3]．金融業界でも，これまで株価の予測などは過去の株価や企業の業績などをモデルの入力に利用してきたが，近年では，投資戦略にテキスト情報から得られたスコアを組み込むことの検証 [4] や，テキストデータを含むオルタナティブデータ（Alternative Data）を使用した運用に関する研究も盛んに行われている [5, 6]．しかし，未だに機械では判断できない領域は残っており，人手による投資判断が必要なことの方が多いのが現実である．

人手による投資判断において問題となるのが，判断のためのテキスト情報が氾濫していることが挙げられる．一昔前であれば，テキスト情報をいかにして集めるかが重要であったが，現在では様々な情報を大量に取得することが可能である．しかし，時間あたりに目を通すことができるテキスト情報には限界があることから，膨大なテキスト情報の中から有益な情報の取捨選択が課題となっている．

膨大なテキスト情報から自然言語処理技術と機械学習を用いて有益な情報を抽出することをテキストマイニングという．特に，機械学習を用いたテキストマイニング手法によって，金融に関連する膨大なテキストから投資判断に有益な情報を識別，抽出し，抽出された情報と市場変動の関係性を発見したり市場分析に応用する研究は，金融テキストマイニングと呼ばれている [1]．金融テキストマイニングで対象としているテキストは，企業が公開している情報や，金融の専門家が発信する経済レポートはもちろんのこと，不特定多

数の人が投稿している SNS や掲示板の情報も対象である．企業が公開している情報としては，決算短信，有価証券報告書，株主招集通知，統合報告書，プレスリリースなどが挙げられる．その他には，企業が所有している特許情報や，Web ページの情報なども有益な情報源である．

1.2 研究目的

このような背景を踏まえた上で，本論文では金融テキストマイニングの一環として，企業が公開している膨大な金融テキストデータから，投資判断に有益な情報を抽出するのに必要な各種の要素技術の開発を行い，その有効性を検証する．本論文における投資判断に有益な情報の例としては，企業の役員情報，大株主情報，業績情報などが挙げられる．また，本論文で研究の対象とする金融テキストは，株主招集通知と有価証券報告書である．

株主招集通知は，株主総会の開催日の二週間前までに株主への送付と公開が義務付けられており，誰でも企業 Web サイトから PDF ファイルとしてダウンロードすることが可能である．その記載内容としては，株主総会で議論される決議事項，大株主情報，役員情報などであるが，その中には，配当の実施，役員人事など，株価に影響を与える可能性のある情報が多く記載されている．しかし，記載内容は多岐にわたり，ページ数が百ページを超えるものも珍しくない．

そこで本論文の 2 章および 3 章では，この株主招集通知を対象に，投資判断に有益な情報が，何ページから何ページに記載があるのかを推定することで抽出することを目的とする．本研究では，ある程度まとまった文集合を用いて投資判断に有益な情報を抽出することで，高い精度の達成を目指す．ここで，PDF ファイルは，ページごとにテキストデータに変換することが可能であり，ページごとのテキスト情報をまとまった文集合とし，ページに記載されている内容の推定を行うことで，ページ単位での情報抽出を行う．このページ単位での情報抽出に関する研究は，先行研究が存在しないが，多くの企業が抱える業務を効率化するのに有効な手段である可能性があり，従来のテキストマイニング手法をページ単位に拡張した応用研究である．

有価証券報告書は，事業年度ごとに作成する企業内容の外部への開示資料であり，企業の概況，事業の状況，設備の状況など多くの内容が記載されている．また，有価証券報告書のデータは，金融庁が運営している Web サイトである EDINET に開示されているため，誰でも入手が可能である．有価証券報告書は，PDF だけでなく XBRL 形式のデータで提供されており，非常に複雑なタグで構成されているが，このタグを利用することで，章ごとのテキスト情報を抽出することが可能である．したがって，多くの情報が掲載され

ているものの，その中から必要な章に限定してテキスト情報が抽出できることが，XBRL形式の特徴である．しかし，章に限定して抽出しても，そのテキスト情報は依然として膨大である．そこで，本論文の 4 章では，有価証券報告書の 2 章「事業の状況」を対象に，投資判断において特に有益な情報である，企業の業績に関する要因について書かれた文（業績要因文）と，どれだけの売上高や経常利益だったのかについて書かれた文（業績結果文）の抽出を行い，抽出した業績要因文，業績結果文が，対象企業のどの事業セグメントに関するものであるかを自動付与する方法論について述べる．

上記を実現することにより，有価証券報告書に含まれる膨大なテキスト情報から投資判断に有用な情報のみを抽出することが可能になり，抽出した各文に対して事業セグメントが付与されていることで，業績要因文のみの抽出では把握することができないことを把握することが可能になる．例えば，自動付与した事業セグメントの情報から，業績要因文に対応する業績結果文が得られるため，売上高から事業規模を把握することが可能であり，経常利益の前年度比からセグメント別の事業の成長度合いを把握することも可能になる．

1.3 本研究の貢献

本論文の自然言語処理分野への学問的な貢献は，大きく分けると以下の 2 点である．

1. 金融テキストにおけるページ単位での情報抽出を自動生成した学習データを用いて行うための方法論の提案
2. 金融テキストにおける文単位での情報抽出として，事業セグメントが付与された業績要因文と業績結果文を抽出する方法論の提案

本論文で述べる研究の位置づけを明確化するために，次節 1.4 節でこれまでの従来手法に関する問題点や，従来手法との相違点などを列挙しつつ，本論文の有効性と新規性について述べる．

1.4 関連研究

本論文は，自然言語処理の分野の中でも，特に金融テキストマイニングと呼ばれる分野に含まれる．したがって，提案手法の多くは，テキストマイニングに関する研究をヒントに手法の提案，検討を行っているため，まず 1.4.1 節でテキストマイニングに関するこれまでの研究を踏まえた上で，本論文の新規性と有効性について述べる．また，1.4.2 節では，学習データの自動生成に関する研究の必要性を，いくつかの関連研究を示しつつ述べ

る．最後に，1.4.3 節では，金融テキストマイニングに関する研究において，どのような背景の下で，どのようなテキストデータを対象に，テキストマイニングが行われているかを述べた上で，本論文で抽出の対象とする情報の新規性および重要性について述べる．

1.4.1 テキストマイニング

本論文での抽出対象は，投資判断に有益な情報の抽出と一言でまとめたが，テキストマイニングにおける抽出の単位は様々であり，抽出単位によって手法が異なる．

一番小さい単位での抽出は，文字列単位（Word）での抽出である．例えば，「コロナウィルス」のような，最新のトレンドとなるような文字列を抽出することや，企業ごとの特徴的な語を抽出することが挙げられる．鳥海らは，Twitter に投稿されたテキストデータを用いて，新型コロナ禍における大きなイベントの発生と，ユーザの投稿にあらわれる感情の変化を抽出した特徴的な単語を用いて分析している [7]．

しかし，解決したい課題によっては，文字列単位での抽出では不十分である．例えば，企業分析をするためには，「コロナウィルス」によって，業績がどのように変化したのかを把握する必要がある．そのためには，「コロナウィルスによる外出自粛要請や渡航制限措置により、観光と出張の両面で需要が減少し、売上が大幅に減少した。」や「コロナウィルスによる巣ごもり需要の増加により、インスタント食品やレトルト食品、冷凍食品など、自宅で手軽に調理できる食品の売上が好調に推移しました。」のように，文単位（Sentence）の抽出が有効である．

文章から重要な情報を持った文を抽出する重要文抽出技術は，文章要約技術の 1 つである．平尾らは，SVM（Support Vector Machine）を用いた重要文抽出手法を提案しており，TSC（Text Summarization Challenge）のデータを用いて評価実験を行い，提案手法は Lead 手法などの従来手法と比較し，優れていることを実証している [8]．金融テキストを対象にした研究としては，業績予測文 [9, 10]，業績要因文 [11, 12] を抽出する研究などがあり，これらは 1.4.3 節で改めて詳細を述べる．

さらに大きい単位での抽出は，段落の抽出，記事の抽出などの文の集合である文章の抽出である．文章単位での抽出は様々な場面において必要となる．まず挙げられるのは，膨大なデータから分析対象となるデータを絞りこむために，文章を分類し，抽出する必要がある．例えば，コロナウィルスに関連するテキストデータを集めるためには，インターネット上のニュース記事に対して，コロナウィルスに関連するニュースであるかどうかの 2 値分類を行い，抽出する必要がある．また，データの絞り込みは，前述した単語単位での抽出や，文単位での抽出精度の改善に有効である可能性もある．例えば，川口らは，SVM を用いて Weblog 記事から主観的な意見を含むレビュー記事を抽出し，抽出したレ

ビュー記事を対象に，新聞記事から抽出した辞書に基づいて意見文を抽出する二段階抽出手法を提案している [13]．レビュー記事を最初に主観的な意見文が含まれるかどうか分類することにより，非レビュー記事から抽出されたしまう文を 46% 削減することに成功している．

また，本来は文単位での重要文抽出を行いたい課題に対しても，扱うデータや条件によって文章単位での抽出を選択することもある．例えば，PDF データをテキストデータに変換する場合には，文が崩れてしまうことによって，抽出対象となる文が抽出できない場合などがある．このような文の抽出失敗を極力避けたい，再現率を重視する場合には，抽出したい文が記載されている文章を抽出するモデルを作成することで，仮に抽出対象の文の形式が崩れてしまった場合や，学習データには出現していなかった単語の多い文だとしても，文章全体のテキスト情報を用いることで，抽出対象の文章を抽出することが可能になる．特に，「販促領域では新型コロナウイルス感染症の影響発現が主に 2020 年 3 月以降だったため、業績に与える影響は限定的でした。」，「しかし、外出自粛要請等により旅行分野及び飲食分野では、売上収益への影響がありました。」のような連続して出現する 2 文が本来の抽出対象であるとき，2 文目の抽出の有無によって，この文章の印象は大きく変わることが予想される．したがって，中途半端に文を抽出するよりは，文章単位でまとめて抽出した方がいいケースも多いため，文章単位での抽出は，自然言語処理の分野で重要な技術として研究されている．

本論文では，4 章の有価証券報告書を対象とした研究で，文単位の抽出について述べ，2 章と 3 章の株主招集通知を対象とした研究で，ページ単位での抽出について述べるが，本論文の新規性は，PDF ファイルである株主招集通知からページ単位の抽出を行っている点が挙げられる．ページを文の集合である一つの文章と捉えることで，ページ単位の抽出は，文章の分類問題と定義することが可能である．例えば，本論文の 3 章では，「議案が記載されているページ」，「大株主情報の記載があるページ」，「重要な事項が記載されていないページ」のように，ページに対して分類を行うことで重要なページの抽出を行っているが，これはニュース記事を，「政治」，「スポーツ」，「経済」といったトピックに分類する問題と非常に類似した特徴を持ち合わせる．このような，文章の分類問題を解くための手法に関する先行研究は数多く存在し，SVM や深層学習などの機械学習手法を用いることで，分類を行うことが可能であることが示されている [14, 15, 16]．本論文では，ページを一つの文章と捉え，従来の SVM や深層学習などの機械学習手法を用いることで，分類を行うことが可能であるかの検討を 2 章と 3 章を中心に議論している．

ただし，ページ単位での情報抽出が従来の文章分類問題と大きく異なる点は，文章の順に意味がある点であり，日本語で記載された PDF データを対象に，ページ単位での抽出

を行っている先行研究は存在しない．特に，ページ単位での抽出は，内容がページで区切られていない可能性があることや，前後のページの情報を用いることが有効である可能性があるなど，一文単位での分類やニュース記事の分類問題などとは異なる点が多い．このような順序のあるデータに対して分類を行う問題は，系列ラベリング問題（Sequential Labeling）と捉えることもできる [17, 18, 19, 20, 21]．自然言語処理であれば，文を単語の列と捉え，単語に対して品詞割当てや固有表現判定を行う問題 [22] や，対話における発話に対して，発話タイプの判定などを行う問題 [23] などが系列ラベリング問題に該当する．ページ単位での分類において，その前後のページの情報は分類の有効な特徴量になる可能性が高いため，単一ページを用いた分類よりも，前後のページの情報も用いた分類の方が，良好な結果を得られることが期待される．よって本研究では，これらの系列ラベリング問題で有効であることが示されている LSTM[24] や CRF[21] などの手法が，ページ単位での情報抽出においても有効であると考え，これらの手法を適用し評価を行うことで，有効性の検討を行っている．

また，ページを文の集合である一つの文章とし，文書全体を文章の列と捉えるのではなく，文書全体を文の列によって構成されていると捉えることで，従来研究の段落部分を判定する問題や，対話における対話内容のトピック変換点や，現在どのようなトピックが話し合われているのかを推定する研究などの手法を利用することも考えられる [25, 26, 27]．これらの研究は，入力が文単位であるのに対して，本研究で提案しているのは，入力がページ単位である点が相違点である．本研究でも，いくつかの条件を満たすことができれば，株主招集通知のテキスト情報を 1 文ずつ入力とすることで，上記に上げたような従来研究の手法を適用することができると考えるが，以下に挙げるように，その条件は厳しく，上記の従来研究を，本研究の目的のためにそのまま使用することは容易ではない．

まず，PDF データをテキストデータに変換を行う必要があるが，表や図などが多く含まれている株主招集通知のような PDF データは，テキストデータ変換時に思うように変換できないことが多々ある．また，ページとページの間には，ページ番号や表示されていないテキスト情報，左右にインデックス情報が縦書きされているなど，文に直す際には多くのノイズが含まれる傾向を持つ．また，文単位ではなく，行単位で変換が行われるため，行を連結処理したあとに，文への分割処理をする必要がある．さらに，注意事項，その他の補足，表の数値テキスト情報などが，複数行にわたって出現することも多いため，系列ラベリングのような問題と捉えて解く場合，LSTM などの再帰的なモデルは，比較的近い位置の情報を加味することに特化しているため，近年の attention 機構などを用いて，遠く離れた情報を加味することが可能なモデルを検討する必要もある．

次に，仮に上記の PDF からのテキスト変換が問題のないレベルで行うことができた場

合を考えるが，学習データ生成が容易でないことが挙げられる．2 章の研究では，人手にて従来作業でデータベースに記録されたデータがあるが，これらのデータは，ある株主招集通知の各議案がどの議案分類に該当するのかと，その議案が何ページから始まるかのデータでしかないため，ページの 1 行目から，その議案の記載が始まっている保証がなく，ページ前半にはひとつ前の議案についての記載があり，ページの後半分から，該当の議案についての記載が始まっている可能性がある．また，同一ページに複数の議案が出現する可能性もある．このようなデータに対して，何ページの何行目から何ページの何行目までが，該当の議案であるかの判定を，ルールベースで行い学習データを作ることも可能であったが，その場合，どの行から議案が始まっているのかを，学習データになるレベルの質で作成可能なルールを設定する必要があり，このようなルールの作成は非常に難しい．また，3 章の研究では，そのような従来の業務で蓄えたデータもない状態から学習データを生成する必要があったため，文単位での学習データの自動生成は不可能であった．

最後の条件としては，実用化を考えた場合に処理速度の問題が残る．文単位での開始位置推定を行うためには，まず入力される PDF をテキストに変換し，文単位に分割するための前処理を行う必要ある．そして，各文を分散表現で表し，一文一文を入力とするモデルに，数十から数百ページに出現する大量の文を入力し，推定を行う必要がある．これに対して，本研究でのページ単位での抽出であれば，各ページをテキストデータに変換し，出現する素性となる語の頻度のベクトルを生成するだけであるため，高速に処理することが可能であり，モデルへの入力もページ数であるため文に対して非常に少なくなる．

これらの理由から，本研究ではページ単位での抽出を行う方法論の検討を行っている．ページ単位での抽出を行うことにより，メリットもあるが，ページ単位での抽出独自のデメリットもいくつか存在し，それらに対する考察も本論文の貢献の一つである．

1.4.2　学習データの自動生成

テキストデータの分類に関わらず，分類器を学習させるためには，学習データが必要である．近年，学習データを用いた教師あり学習における分類問題の精度向上は目覚ましく，学習データが十分な量と質で存在するのであれば解決できる問題は非常に多い．これまで人手で分類し，抽出してきたデータを学習データとすることで，それらの分類，抽出の自動化に成功したといった研究は多く存在し，近年では，教師無しデータを大量に用いた事前学習済みモデルを用いることで，大量の教師ありデータがなくても少量のデータをファインチューニング（Fine Tuning）することで良好な結果を得ることが可能になりつつある [28, 29, 30, 31].

しかし，これまで人手で抽出してこなかった情報を抽出する場合には，学習に使用する

ためのデータが存在しないという問題がある．学習データが存在しない場合，一番シンプルな解決法としては学習データを人手で作成することであるが，学習データの人手による作成には非常に多くの労力と時間がかかる．特に金融テキストマイニングの場合は，その金融テキストの内容を理解することにある程度の専門的な知識が必要となり，学習データの作成にもその分野に対する専門的な知識が必要になる場合も多い．また，テキストデータは時間とともに特徴が変化していく傾向があるが，その中でも特に金融テキストは，これまでなかった商品，サービス，職種の登場などにより変化が大きく，過去のデータを用いた学習が，現在のデータにあわず，予測がうまくできない可能性がある．したがって，たとえ学習データがもともとあったとしても，時間経過に伴い，その都度，学習データを作り直す必要がある．これらの理由から，人手による学習データの作成をし続けることは実用上困難である．

そこで，本論文では学習データの自動生成方法を提案する．分類問題の学習データに関する研究は盛んに行われている．少量の学習データが存在する場合においては，学習データの水増しを行う方法論の検討も行われており，例えば，Arazo らは，半教師あり学習における Pseudo Labeling による影響を分析している [32]．画像分類の分野であれば，画像を左右反転させるなどのシンプルな処理をすることで，学習データを増やすことができ，Shorten と Khoshgoftaar の研究では，画像データの様々な学習データの水増し方法を検討し，評価実験によりその有効性を示している [33]．テキスト分類の分野であれば，Wei と Zou は，ランダムに，1. 同義語置換，2. 同義語挿入，3. 単語移動，4. 単語削除，を利用可能な学習データの 50% に適用することで，利用可能な学習データ全てを使用した結果と同様の結果が得られることを示している [34]．また，データ分析力や機械学習のモデリング力を競い合うコンペティションプラットフォームである Kaggle[*1] で開催されたテキスト分類のコンペティションでは，再翻訳を行いデータを水増しする方法が，精度向上に有効であることを，優勝者がディスカッションにて公開している[*2]．

これらのデータ水増し方法は，少量の元でとなるデータが必要であるが，どのような問題を扱うかによるが数百から数千程度のデータが必要である．それに対して，0 から学習データを生成する場合，自然言語処理分野では，正例となるテキストに対してのみ出現するような特徴的な語がいくつもあれば，正例を作ることが可能であるが，そのような語があるのであれば学習データを生成するまでもなく分類が可能である．特に，抽出の対象が文である場合には，多種多様な文が存在することから，単純な方法での学習データ生成は

[*1] https://www.kaggle.com/

[*2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/discussion/52557

難しい．例えば，近藤らは，計測対象文の集合から，景気関連文の抽出を行うためのモデルを学習させるための正例として，景気ウォッチャー調査の景気判断理由集を学習データに活用している [35]．近藤らの研究のように，抽出対象となる文集合とは別のデータを活用することで学習データを準備する研究は多いが，それ以外の方法で 0 から学習データを生成するための研究は非常に少なく，本論文で扱うページ単位での抽出を行うための学習データを生成する従来研究は存在しない．また，文に対しての学習データ生成も，単純なルールベースの手法では多種多様な文に対応することはできず，タスクに特有の手がかり表現などを使用することになる．

そこで本論文では，0 から学習データを生成する研究として，3 章では，ページ単位での抽出を行うモデルを学習させるための学習データをルールベースで生成する方法について述べ，4 章では，文単位での抽出を行うモデルを学習させるための学習データを，手がかり表現や企業キーワードのような特徴語を用いて生成する方法について述べる．

学習データを自動で生成することが可能になれば，データの特徴が変化したときや，抽出したい対象が変わったときも，多大な労力を割くことなく，モデルの再学習が可能となる．しかし，自動生成で得られる学習データは，人手で作成した場合に比べ，精度やデータの偏り（バイアス）の問題が存在するため，それらの特性を考慮した上で学習させるモデル選択をする必要がある．学習データの自動生成によってどのような問題が生じるのかを議論した上で，それらの問題を軽減できるモデルの検討を 3 章で行う．

1.4.3　金融テキストマイニング

本論文では，金融テキストとして，有価証券報告書や株主招集通知を対象としており，金融テキストマイニングと呼ばれる研究に含まれる．したがって，金融テキストマイニングに関するこれまでの研究が，どのような背景の下で，行われているかを述べた上で，本論文で抽出の対象とする情報の新規性および重要性について述べる．

和泉らは，日銀が発行している「金融経済月報」を用い，国債市場の分析を行っている [36]．この研究では，共起解析（co-occurrence analysis），主成分分析（principal component analysis），回帰分析（regression analysis）からなる CPR 法を提案し，テキストを解析することで国債市場の分析を行っている．さらに，蔵本らは和泉らの研究 [36] をさらに発展させ，入力テキストを新聞記事として，長期的な株式市場の分析を行っている [2]．Milea らは，欧州中央銀行が発行している報告書から Fuzzy Grammar Fragment を抽出し，それに基づき，MSCI ユーロ・インデックスを予測（上向き，もしくは，下向きに推移するかどうか）している [37]．Koppel らは，企業に関する記事が，その企業の株価に影響を与えるかどうかを判別する手法を提案している [14]．この研究では，一定以

上，株価が上がった記事と下がった記事を学習データとし，ある記事が Good News か，Bad News かを判別している．Lavrenko らは，企業に関する記事内容によって引き起こされる株価の動きを予測するための手法を提案している [38]．これらの研究では，日銀の金融経済月報，景気動向記事のような新聞記事といったテキスト情報を用いて，将来的な市場の予測を行っている．

金融テキストの分析を目的とした研究として，松田らは，日本銀行政策委員会金融政策決定会合議事要旨のテキストデータから，トピック抽出の研究を行っている [39]．また，竹内らは有価証券報告書を分析し，特徴語抽出により，倒産企業の特徴を明らかにしようとしている [40]．この研究では，文脈情報や係り受け情報などにも注目し，継続企業群と倒産企業群の特徴語の抽出を行っている．

これらの研究に対して，本論文の 2 章と 3 章の目的は，現在の技術では，市場や株価にどのような影響を与えるのか判断することが難しい，人手での判断が必要となる情報を抽出することである．したがって，直接的な分析を行う研究とは異なり，情報抽出に重きを置いた研究である．現状，金融テキストマイニングの研究は分析がメインであり，その分析を行うためのテキスト情報は，単語単位で抽出した情報や，係り受け情報などを利用した研究が中心である [39, 40]．そして 1.4.1 節で述べた通り，テキストマイニングにおいて，様々な抽出技術の研究が活発に行われているが，それらの技術の金融テキストへの応用は十分に進んでいない．その要因には，教師あり学習モデルを使用するためには大量の学習データが必要であることや，ドメインや金融テキストファイル独自の特徴があることが挙げられる．そこで 1.4.2 節の学習データの自動生成手法と 1.4.1 節のテキストマイニング手法を，金融テキストに応用するための方法論を本論文では検討する．また，金融テキストとして株主招集通知を対象とした研究は珍しく，不必要な情報などが多く記載されていることから，分析対象として敬遠されていたが，本論文の手法を用いることで，分析対象とするテキスト情報を絞り込むことが可能となるため，今後の株主招集通知の自然言語処理による分析への貢献が期待される．

本論文の 2 章と 3 章では，ページ単位での抽出を行っているが，それに対して，単語単位や文単位での抽出に重きを置いた金融テキストマイニングの研究は多い．単語単位での抽出に関する研究として，極性辞書の作成を目的とした研究がある．例えば，坂地らは，景気動向について述べた経済新聞記事を対象に，景気が上がるか下がるかの根拠表現を抽出し，抽出した根拠表現に対して極性（ポジティブ，ネガティブ）を付与する手法を提案している [41]．また，五島と高橋は，金融分野に特化した極性辞書の作成を目的とし，ニュースデータと株式価格データから極性辞書の作成を行っている [42]．

金融テキストを対象とした文単位の抽出の研究も数多く存在している．酒井らは，企業

の業績発表記事から業績要因を抽出し [43]，抽出した業績要因に対して業績に対する極性 (「ポジティブ」,「ネガティブ」) を付与する手法 [44] や，抽出される複数の業績要因から重要な業績要因を自動的に抽出する手法を提案している [45]．更に，酒井らは，上記の手法 [43] を決算短信 PDF に適用し，決算短信 PDF から，例えば「半導体製造装置の受注が好調でした」のような業績要因を含む文を抽出する手法を提案している [11]．これらの研究から派生して，坂地らは決算短信から原因・結果表現を抽出する手法を提案している [46]．例えば,「夏の記録的な猛暑の影響により冷房需要の増加がみられた」という文を抽出し,「夏の記録的な猛暑の影響」が原因,「冷房需要の増加がみられた」を結果として識別する．さらに，抽出された原因・結果表現の中から意外性のある組み合わせを識別する手法を提案している [47, 48]．坂地らの研究でも「により」のような原因と結果を繋げる手がかり表現を使用して文の抽出と原因・結果表現の識別を行っている．業績要因文抽出の既提案手法である酒井らの研究 [11] でも「が好調」「が不振」といった手がかり表現をブートストラップ的に自動獲得しているが，全ての手がかり表現を網羅することができず，再現率よりも適合率が高い結果となっている．また，北森らは決算短信から業績予測文を抽出する手法を提案している [9, 10]．例えばソニーの決算短信から「音楽分野音楽制作が好調であることなどにより，分野全体の売上高は１０月時点の想定を上回る見込みです」という文を抽出する．業績要因文と業績予測文の違いは，業績要因文は前期の業績の要因について述べている文であるのに対して，業績予測文は今期（あるいは来期）の業績予測の要因について述べている文である．そのため，抽出に必要な手がかり表現が異なり，北森らは「見込みです」などの文末や「予想につきましては」などの文頭に出現する表現を手がかり表現として抽出している．また，田中らは決算短信から抽出した業績要因文から複数企業に共通する要素を抽出し，さらに，抽出した共通する要素から新たな関連企業を推定する手法を提案している [49]．

このように，金融テキストからの業績要因などの文を抽出する研究は盛んに行われている．本論文の 4 章における業績要因文の抽出は，決算短信からの業績要因文の抽出を行っている酒井らの手法 [50] を適用し，業績要因文抽出のための学習データを有価証券報告書から自動的に生成し，深層学習を用いて行っている．酒井らの研究 [11, 50] との相違点は，酒井らの研究では，決算短信から業績要因文を抽出するのみであり，その業績要因文がどの事業セグメントに属するのかは判定しておらず，決算短信から業績結果文や事業セグメント名といった情報の抽出も行っていない．それに対して本論文では，有価証券報告書からその企業の事業セグメント名を獲得し，有価証券報告書から抽出した業績要因文，業績結果文に対して，事業セグメントの付与を行っている点で大きく異なる．

業績要因文に事業セグメントを付与する研究としては，村野らは決算短信から抽出した

業績要因文を事業セグメントに基づき k 近傍法にて分類する手法を提案している [51]．しかし，村野らの手法を行うためには，企業ごとに人手で事業セグメント名を収集する必要があり，さらに，大量の事業セグメント名のラベルが付いた正解データを作成する必要がある．証券市場の上場企業数は東京証券取引所の上場企業だけでも 2020 年 12 月 31 日現在，3,756 社も存在するため[*3]，人手で各企業の事業セグメントを収集し，学習データを作成するのには多大な労力が必要となる．また，企業の事業セグメント名は頻繁に変更され，事業セグメント名の変更が起こった場合，正解データを人手で作り直す必要がある．これは現実的にはコスト面などの問題から不可能であり有効な方法ではない．このような方法を取らざるを得なかった理由は，決算短信のフォーマットが企業や年度によって異なるためである．それに対し，本研究で提案する手法は，ある程度フォーマットが定まっている有価証券報告書を用いることで，事業セグメント名を自動的に獲得し，業績要因文抽出のための学習データをも自動で生成するため，人手による作業を必要としないものとなる．

これまで抽出方法が提案されてきた業績要因文と，本論文の 4 章によって抽出することが可能になる事業セグメントが付与された業績要因文と業績結果文の相違点は，より応用的な分析が可能になる点である．まず，企業の持つ事業セグメントは，売上や利益が事業セグメントごとに大きく異なる．例えば，「株式会社 SUBARU」は，「自動車セグメント」と「航空宇宙セグメント」を持っているが，その利益状況は，「自動車セグメント：397,657 百万円」に対して，「航空宇宙セグメント：9,102 百万円」と 45 倍程度の差がある．したがって，「航空宇宙セグメント」の業績が好調だとしても，メインの事業セグメントである「自動車セグメント」の業績が不調であれば，企業全体としての業績は不調である．事業セグメントを業績要因文に付与することが可能になることで，事業セグメントの規模によって，業績要因文の重要性を測ることが可能となる．また，事業セグメント別に，市場を分析することも可能となり，事業セグメント全体の傾向分析や，同一事業セグメント内で，好調な企業と不調な企業の分析なども可能となる．

さらに，事業セグメントを業績結果文とも紐づけることにより，業績要因文に対して，数値情報を付与することも本手法で可能となる．文や表現に対して極性を付与するためには，テキストとポジティブ・ネガティブを表す分類や数値情報が必要となる．これまでの多くの先行研究では，これらの数値情報に，企業の株価や，企業全体の売上高や営業利益などを用いているが，企業の株価や全体の売上高は，様々な要因が入り乱れてしまっており，対象としているテキストがどの程度の影響を与えているかを考慮しきれていない問題

[*3] https://www.jpx.co.jp/listing/co/index.html

がある．それに対して，本研究で抽出することができる各事業セグメントごとの業績要因文と業績結果文は，これまで先行研究で使用してきたデータよりも，テキストと業績結果の数値情報の関係が強いものとなっている．したがって，これらのデータを用いて極性付与を行うモデルを学習させることで，これまでの研究で得られた結果よりも良好な極性付与を行うことが可能になる見込みが高い．このようなモデルの学習を可能にするためには，業績要因文とそれに紐づく業績結果文を人手で大量に作成する必要があるが，それは非常に困難であり，そのようなデータを大量に生成することを可能にした．このように，これまで抽出してきた情報に，事業セグメントと業績結果文を紐づけることにより，分析の幅が広がることや，極性辞書作成のための言語資源となることが本論文の貢献の 1 つである．

1.5　本論文の構成

2 章では，自動生成した学習データを用いて，ページ単位での有益情報抽出を行うことが可能であるかどうかを検討するために，まず人手で作成した学習データを用いて分類器を学習し，分類対象となるページをルールベースで絞った上で，ページ単位での分類が可能であることを示す．具体的には，株主招集通知を対象に，議案が開始しているページをルールベースで推定し，そのページに記載されている議案が，どの議案分類に該当するかを分類する方法を検討しており，単一ページの入力に対して，従来の機械学習手法である SVM や深層学習モデルの MLP などを用いて分類を行うことで，F 値 0.930（SVM），0.937（MLP）と良好な結果を得ることが可能であることを示している．

2 章のページ単位の分類において，人手で学習したデータを用いて分類器を学習し，分類するページをルールベースで絞ることができれば高い結果を示すことができたため，続く 3 章では，ページ単位での有益情報抽出を行うために，自動生成した学習データを用いて，各ページを分類するモデルをいくつかの従来手法から提案し，比較検討を行っている．その結果，自動生成した学習データを用いて学習を行った BiLSTM モデルの分類器によって，ページに対して分類の Tag を付与することで，重要なページの抽出が可能であることを示している．具体的には，ルールベースを用いた方法で分類が行えない株主招集通知に対しても，どのモデルは有効な結果を示しており，特に，単一ページを入力としたモデルよりも，前後のページを加味できるモデルの方が有効であることを評価実験により示している．また，従来の系列ラベリング問題を解決するのに有効なモデルである CRF 層を追加したモデルが，学習データを自動生成した場合には，必ずしも最良なモデルとならないことを示している．

続く 4 章では，有価証券報告書を対象に，これまで決算短信から抽出してきた業績要因文に，事業セグメントを付与し，業績結果文を紐づける方法を提案する．有価証券報告書は，XBRL 形式のデータであることから，タグ情報からある程度，業績要因文などが記載されている箇所の推定は可能であるため，より詳細な文単位の抽出を行っている．業績要因文の抽出は，酒井らの手法 [50] を有価証券報告書のデータに適用しており，酒井らの手法が，決算短信以外のテキストにも有効であることを示している．そして，有価証券報告書のドメインに特化した方法を用いることで，これまで抽出することができなかった事業セグメントが付与された業績要因文と業績結果文の抽出が可能であることを示す．

最後に 5 章では得られた知見や成果を総括し，今後の研究展望について述べる．

2　株主招集通知における議案タイトルとその分類及び開始ページの推定

2.1　研究概要

本章では，人手で作成した学習データを用いて分類器を学習し，分類対象となるページをルールベースで絞った上で，ページ単位での分類が可能であることを示す．本研究では株主招集通知を対象として，開始ページの推定を行うが，開始位置に関する先行研究は少なく，開始位置の推定に関する関連研究としては，XML ドキュメントを対象とした研究はあるが [52]，株主招集通知を対象とし，議案ごとの開始ページを推定する研究はない．また，機械学習による分類手法の比較に関しては数多くの論文が存在するが [53, 54, 55, 56]，それらの論文のほぼすべてが手法の比較を行い，新規手法によって「適合率や再現率の改善が見られた」のような内容である．しかし，実際のシステムにおいては，一つの手法に依存するのは大変危険であるため（機械学習は過去のデータを学習データとすることから，新規データの変化に弱いため），複数の手法によって結果を出し，統合した確からしさ（信頼度）をユーザーに与える必要がある．個々の手法の適合率や再現率を示しても，ユーザーはどの結果を信頼するべきか迷うため，結果の統合は価値があると考えられる．この結果の統合による信頼度の算出が本章の特色の一つである．複数の分類手法の統合の関連研究としては，線形結合による分類手法の統合を行っているが [57]，本研究ではベイズを用いた確率的なアプローチを行っている．

金融系の調査会社では，各種データを収集し，様々なデータベースを構築している．データ処理にあたっては，たとえば XBRL 形式のように値に付与されたタグ等の付加情報を利用し，自動分類によるデータベース化を行っている．しかし，データ分類用付加情報が付与されているデータはまだ少数で，データベース構築の多くは自動分類化が進んでおらず，人手をかけた作業による分類が大半を占めている．また，手作業で必要な情報を抽出するには，専門的知識や経験が必要となる．そのような環境の中，2 章では「株主招集通知の議案別開始ページの推定」を研究課題として設定した[*4]．

まず，株主招集通知について述べる．企業が株主総会を開催する場合，企業は招集の手続きが必要になる．会社法では公開会社である株式会社が株主総会を招集する場合，株主総会の開催日の二週間前までに，株主に対してその通知を発しなければならないと定めて

[*4] 株主招集通知から推定するべき情報は人事案件など他にもあり，そのような他テーマへの応用も可能である．

いる（会社法第二百九十九条）．また，株式会社が取締役会設置会社である場合，その通知は書面で行わなければならない（会社法第二百九十九条第二項）．この株主総会に関する書面通知が株主招集通知である．

取締役会設置会社においては，定時株主総会の招集の通知に際し，取締役会の承認を受けた計算書類及び事業報告を提供しなければならない，株主総会の目的が役員等の選任，役員等の報酬等，定款の変更等に係る場合，当該事項に係る議案の概要を通知する必要がある等，会社法および株主招集通知にて通知する事項は会社法および会社法施行規則で定められている（会社法第二百九十八条，会社法施行規則第六十三条）．

一般的な株式公開会社の株主招集通知は，株主総会の日時・場所・目的事項（報告事項・決議事項）が記載される他，参考書類・添付書類として決議事項の議案概要，事業内容等の株式会社の現況に関する事項，株式に関する事項，会社役員に関する事項，会計監査人の状況，計算書類，監査報告書等が記載される．記載内容が法令で定められている株主招集通知だが，有価証券報告書等のように様式が定められておらず，その形式は記載順序や表現方法を含め各社で異なっており，ページ数も数ページのものから 100 ページを超えるものもある．

本章で述べる研究の対象は，株主招集通知に記載されている決議事項に関する議案である．議案については，その記載がどのページにあるか，何の事項の前後に記載されるかは各社各様であり，多様なパターンを識別するには株主招集通知を読み解く経験を積む必要があった．従来は抽出したい議案（「取締役選任」「剰余金処分」などの項目）が報告書のどのページに記載されているか人手により確認し，データを作成していたが，各社で報告書のページ数や議案数が異なるため，確認に時間を要していた．現状は，株主招集通知を紙で印刷するとともに，PDF ファイルで取得，人手にてデータベースに収録，校正リストの出力，チェックという流れで収録業務を行っている．ここで，抽出したい議案がその報告書にあるのか，どのページに記載されているのかが自動で推定できれば，時間の短縮やペーパーレス化などの業務の効率化につながる．したがって本研究の目標は，株主招集通知の各ページが議案の開始ページであるかそうでないかを判別し，さらに，開始ページであると判断されたページに記述されている議案が，どのような内容の議案であるかを自動的に分類することである[*5]．

株主招集通知の開示集中時期には，短時間に大量の処理を進めるため，臨時的に収録作業者を配置し，データ入力を行う．臨時作業者には，株主招集通知の理解から始まり，収

[*5] 上記のようなアプローチを採用した理由は，最初に議案の開始ページを推定することで，議案分類に使用する文書を絞り込むためである．

録定義に関する教育時間や練習時間が２日程度必要となる．この教育時間を経て，実際のデータ入力を始めると，慣れるまでは１社あたりのデータベース化に１時間半～２時間を要し，本研究で対象としている議案分類のみの作業でも慣れた作業者さえ数分かかる．特に議案など必要なページにたどりつくまでに株主招集通知を一枚一枚めくって探すこと，議案分類について議案タイトルやその詳細から対応する語を見つけ出すことに時間を要している．処理・判断が早くなるには，各社で異なる株主招集通知の構成を見極め，構造の特徴をつかむことが必要になる．しかし，これらの勘をつかむにはおよそ１週間程度かかっている．さらに，信頼性の高いデータ収録ができるようになるには３ヶ月以上を要している．

本研究によるシステムによって，これらの構造理解と勘の習得が不要になると共に，議案の開始ページの推定や議案内容が分類されることにより，当該部分の１社あたりにかかる処理時間の短縮が期待される．さらに，理解の十分でない作業者の判断ミスや判断の揺れが減少し，信頼度の高いデータ生成を支えることとなる．その結果，データベース収録に係る人件費の削減と，データベース化に伴うデータ収録の早期化をはかることが可能となる．また，一般的に株主は，株主招集通知に掲載されている議案を確認し，「この議案に賛成もしくは反対」の判断をしている．多数の企業の株式を保有している株主は，これに時間がかかることが推測される．株主招集通知に載っている議案が分類されれば，議案の内容をより早く把握することができ，判断に集中できるようになると考える．

ここで，本章で提案するシステムの全体像について述べる．本章で提案するシステムは，株主招集通知を入力として，表１に示すような結果を返すシステムである．この結果

表１　出力結果

開始ページ	議案分類
40	剰余金処分
43	定款変更
48	取締役選任
50	取締役選任
52	役員賞与
52	退職慰労金
53	役員報酬
54	役員報酬

を得るためには，まず議案が何ページから記載されているのかを推定する必要がある．そ

して推定したページに対して，議案がいくつ存在し，どの議案分類に分類されるかを自動で行う．

本章の 2.2 節では議案がある開始ページの推定について述べる．2.3 節から 2.5 節では各議案分類の手法について詳細な内容について述べる．2.6 節では各手法の評価を適合率，再現率，F 値を用いて述べる．2.7 節では 2.6 節の評価結果を踏まえて考察を述べる．2.8 節では応用システムについて述べる．

2.2 議案がある開始ページの推定

株主招集通知から議案を分類するために，議案が開始するページの推定を行う必要がある．開始ページの推定は，テキスト情報を用いてルールベースで行う．まず，株主招集通知の PDF を 1 ページごとに分割したうえで，テキストデータに変換する．変換には pdftotext[*6] を用いた．PDF データをテキストデータに変換するとき，以下の処理を加えて変換を行った．

処理１　半角を全角に変換する
処理２　空白は除去する
処理３　(　) が含まれる文は括弧ごと除去する
処理４　「の件」または「する件」を含んでいる場合，「件」の後に改行を加える
処理５　行頭が「第◯号議案」または「議案」と一致する場合，行頭に改行を加える
処理６　「。」は改行に置き換える

上記の処理によって PDF データは図 1 や図 2 のようなテキストデータに変換される．議案がある開始ページは以下の条件のもと推定した．ここで，「決議事項」または「目的事項」という表現が含まれているページを目次ページとし，「参考書類」，「議題及び参考事項」または「議題および参考事項」という表現が含まれているページを参考ページとする．

条件１　議案がある開始ページは，「議案」または「第 X 号議案」が先頭に含まれるページを対象とする
条件２　目次ページは議案がある開始ページの候補から除外する
条件３　参考ページから議案に関する記載が始まるため，参考ページ以降を対象とする

議案がある開始ページの推定は上記の条件に関わるページ番号の抽出の他に，議案数の

[*6] http://www.foolabs.com/xpdf/home.html

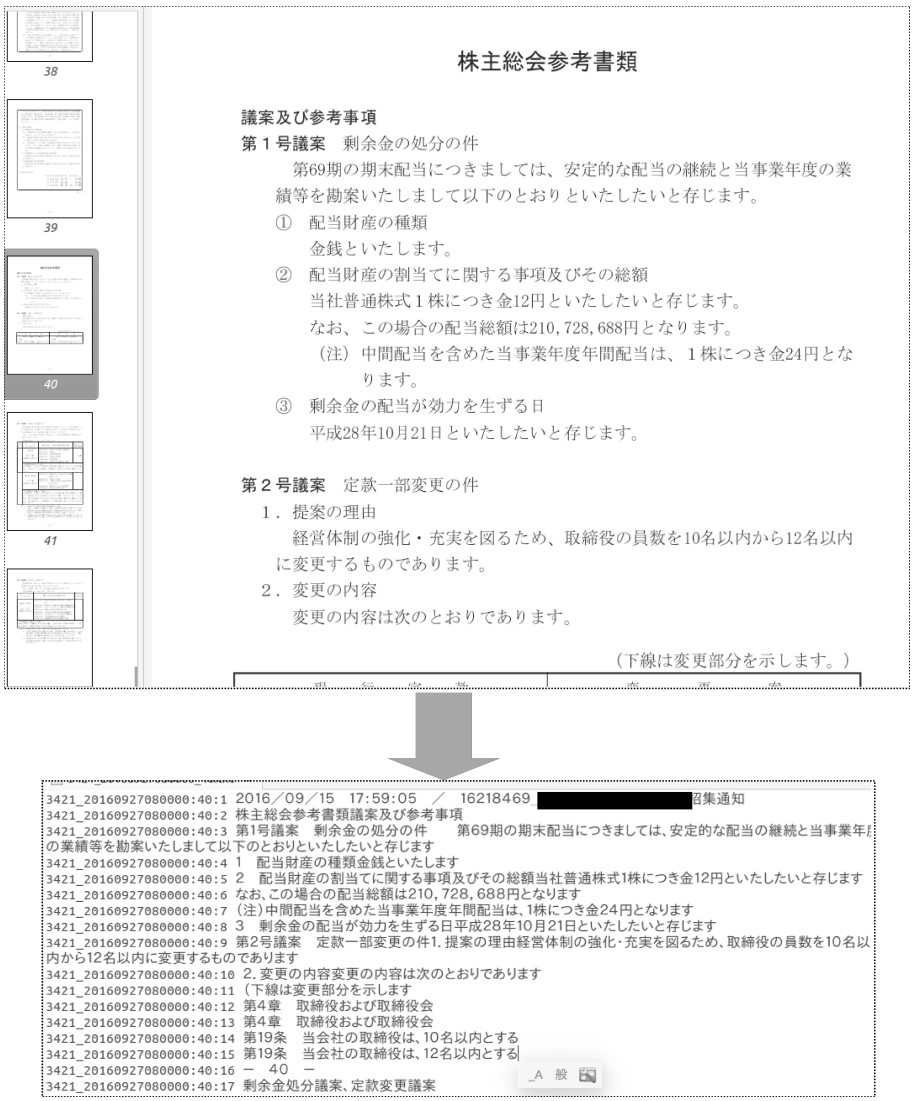


図１　株主招集通知のテキストデータ変換の例 1.

推定とページ数の抽出が必要となる．「第 X 号議案」という語が株主招集通知に含まれている場合，議案数は「X」の最大値であると推定できる．「第 X 号議案」という語が株主招集通知に含まれていない場合，議案という語が含まれていれば，議案数は 1 であると推定できる．ページ数の抽出は，PDF データを 1 ページずつテキストデータ化し，テキストデータが取得できた最後のページ番号をページ数として抽出する．

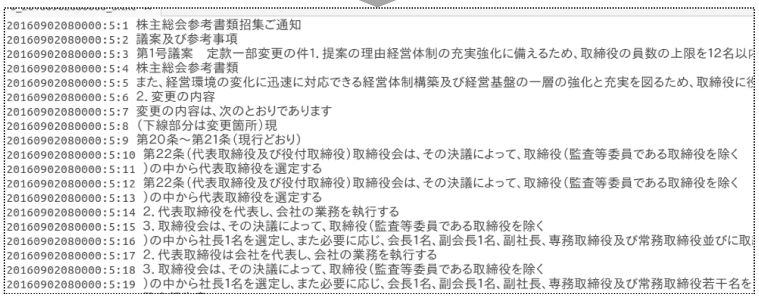


図２　株主招集通知のテキストデータ変換の例 2.

これらを踏まえて，変換されたテキストデータから必要な情報を抽出した例を表 2，表 3，表 4 に示す．例えば，図 1 は株主招集通知の P.40 の例であり，そのページから抽出される議案は「第 1 号議案」，「第 2 号議案」となる．これらの情報を基にページの推定を図 3 のフローチャートに従って行う．

表 2　抽出された情報 1

議案	ページ番号
第 1 号議案	P.1
第 2 号議案	P.1
第 3 号議案	P.1
第 4 号議案	P.1
第 5 号議案	P.1
第 6 号議案	P.1
第 7 号議案	P.1
第 8 号議案	P.1
第 1 号議案	P.40
第 2 号議案	P.40
第 3 号議案	P.48
第 4 号議案	P.50
第 5 号議案	P.52
第 6 号議案	P.52
第 7 号議案	P.53
第 8 号議案	P.53

表 3　抽出された情報 2

議案数	8 議案
ページ数	55 ページ

表 4　抽出された情報 3

参考ページ	P.1
参考ページ	P.40
参考ページ	P.42
目次ページ	P.1

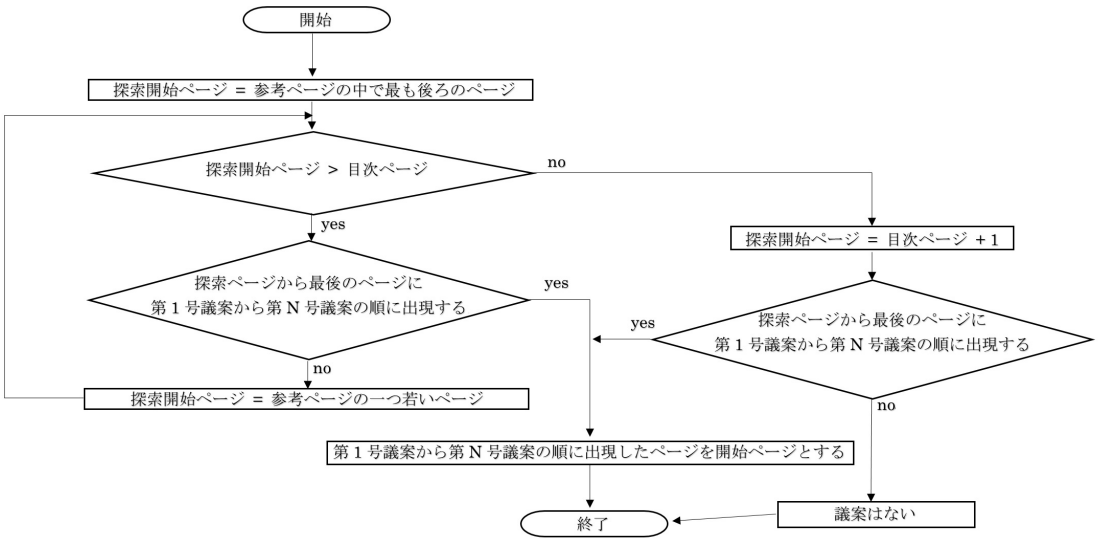


図 3　議案の開始ページ推定のためのフローチャート.

表 2 と表 4 の例の場合，まず参考ページの最も後ろのページである 42 ページが探索開始ページとなるが，第 1 号議案が 42 ページ以降に出現しないため，探索開始ページを 40 に更新して行うことで，第 1 号議案から第 8 号議案の開始ページの推定に成功する.

2.3　提案手法 1：特徴語による議案分類

本章では，特徴語による議案分類について説明する．この提案手法は以下のステップで議案の分類を行う．

Step 1:　議案分類ごとの特徴語の獲得
Step 2:　議案の分類

2.3.1　議案分類ごとの特徴語の獲得

2.3.1.1 特徴語候補の抽出

株主招集通知に出現する議案を分類するために，議案分類ごとの特徴語の抽出をする．例えば，「取締役選任」の特徴語として「現任取締役」のような語を抽出する．議案分類ごとの特徴語の獲得をするために，株主招集通知における議案別の開始ページとその議案の分類が記述されたデータ（6,444 件）を，株主招集通知の PDF を 1 ページごとに分割したうえでテキストデータに変換し，学習データとして使用する．特徴語の抽出は上記の学習データを形態素解析し，それから以下の条件のもと各議案分類の開始ページに 2 回以上出現する語を特徴語の候補とする．

条件 1　名詞を対象
条件 2　分割は N-gram 単位
条件 3　２５文字以上の長すぎる複合名詞は除外

例えば，名詞を対象に N-gram 単位の分割を「株主招集通知における議案タイトル」に対して行うと，「株主」，「招集」，「通知」，「株主招集」，「招集通知」，「株主招集通知」，「議案」，「タイトル」，「議案タイトル」のように分割される．

2.3.1.2 特徴語候補への重み付け

特徴語の候補 n_i に対して議案分類ごとに重み付けを行い，特徴語を選択する．重み付けの式には式 1 を用いる．

$$W(n_i, C(t)) = (0.5 + 0.5\frac{TF(n_i, C(t))}{max_{j=1,\ldots,m} TF(n_j, C(t))}) \times \ H(n_i, C(t)) \log_2 \frac{N}{df(n_i)} \quad (1)$$

ここで，学習データにおいて，

$C(t)$: 議案分類 t の開始ページの文書集合.

$TF(n, C(t))$: $C(t)$ において，名詞 n が出現する頻度.

$H(n, C(t))$: $C(t)$ の各文書である d に名詞 n が出現する確率 $P(n, d)$ に基づくエントロピー．以下の式 2 によって求める．

$$H(n, C(t)) = -\sum_{d \in C(t)} P(n, d) \log_2 P(n, d) \tag{2}$$

ここで，$P(n, d)$ は d に名詞 n が出現する確率である．

$df(n)$: 名詞 n を含む文書の数.

N: 学習データにおける文書の総数.

エントロピーを用いた理由は，各議案分類の文書集合中で多くの文書に分散して出現している語の方が少数の文書に集中して出現している語と比較してよりその議案分類の特徴を表し，特徴語としても有効であるという仮定に基づく．例として，議案分類「監査役選任」の特徴語候補への重み付けを，エントロピーを用いて行った結果を表 5 に示し，エントロピーを用いずに行った結果を表 6 に示す．

表 5　エントロピーを用いた重み付け結果

特徴語候補	重み
各監査役候補	19.84
各監査役候補者	19.84
補欠監査役選任議案	19.83
件監査役全員	19.76
補欠監査役選任	19.68
前任者	19.05

表 6　エントロピーなしの重み付け結果

特徴語候補	重み
志村	5.83
萩原	5.83
加々美	5.83
米久	5.83
深澤	5.83
大庭	5.83

表 6 に示した特徴語候補は上位 6 単語だが，上位 300 単語のほとんどが固有名詞であるため議案分類の特徴としては適切でなく，エントロピーを計算に用いることは有用と思われる．

2.3.1.3 特徴語の選択

各議案分類ごとの特徴語候補の重み付けの平均値を算出し，平均値よりも重みの高いものを特徴語として選択する．すなわち, 以下の条件が成り立つ語 n_i を特徴語として選択

する．

$$W(n_i, C(t)) > \frac{1}{m}\sum_{j=1}^{m} W(n_j, C(t)) \tag{3}$$

m:　議案分類 t の特徴語候補の総数．

例えば，「取締役選任」の特徴語の一部を表 7，「定款変更」の特徴語の一部を表 8 に示す．表 8 に注目すると，「定款変更」の項目なので「変更」，「施行」，「定款」などを含む特徴語が上位に来ることは想像できるが，それ以上に「下線部」のような単語に高い重みが付与されることは，重み付けのメリットである．実際，図 2 は「定款変更」に分類される議案の開始ページであるが，「下線部」という単語が出現している．

表 7　「取締役選任」の特徴語

特徴語	重み
リーダーシップ	18.13
当社代表取締役社長就任	18.00
現任取締役	17.85
指揮	17.67
在任取締役	17.49
ＣＥＯ	17.19

表 8　「定款変更」の特徴語

特徴語	重み
下線部	18.22
本定款変更	17.90
修正	16.95
日施行	16.73
改正	16.29
変更内容	15.86

2.3.2　議案の分類

2.3.1 節で得られた特徴語の重みと 2.2 節の手法で推定した議案の開始ページを用いて，開始ページごとの議案の分類を行う．議案分類 t の開始ページ j に対するスコア付与は式 4 を用いる．

$$score(j,t) = \frac{\boldsymbol{V(t)} \cdot \boldsymbol{V(j)}}{|\boldsymbol{V(t)}||\boldsymbol{V(j)}|} \tag{4}$$

ここで，

$\boldsymbol{V(t)}$:　議案分類 t の特徴語を要素，特徴語の重みを要素値とするベクトル
$\boldsymbol{V(j)}$:　開始ページ j の名詞 N-gram を要素，出現数を要素値とするベクトル

複数の議案が同ページに存在する場合，スコアが上位のものから順に選ばれるものとする．

2.4 提案手法２: 多層ニューラルネットワークによる議案分類

提案手法 1 では，学習データから各議案の特徴語を抽出し，それに基づいて議案を分類している．この学習データを使用すれば，機械学習手法に基づく手法でも議案分類が可能である．そこで，本研究では MLP（多層ニューラルネットワーク）を用いた議案分類も試みた．

2.4.1 素性選択

株主招集通知に記載されている議案の開始ページの議案分類を，MLP により行う．すなわち，議案の開始ページが，ある議案分類であるかそうでないかを判別する分類器を議案分類の数だけ生成し，テストデータとなる議案の開始ページがどの議案分類に属するかを判定する．したがって，例えば議案分類「取締役選任」を判別するための学習データは，「取締役選任」の開始ページが正例，それ以外の議案分類の開始ページが負例となる．また，テストデータは，学習データとして使用した株主招集通知を除き，株主招集通知を 1 ページごとに分割したうえで，2.2 節の手法を用いて推定されたページを対象とした．

まず，入力層の要素となる語（素性）を選択する．具体的には，学習データにおいて正例に含まれる内容語（名詞，動詞，形容詞）に対して，以下の式 5 にて重みを計算する．

$$W_p(t, S_p) = TF(t, S_p)H(t, S_p) \tag{5}$$

ただし，

S_p:　学習データにおいて正例に属する文の集合
$TF(t, S_p)$:　文集合 S_p において，語 t が出現する頻度
$H(t, S_p)$:　文集合 S_p における各文に含まれる語 t の出現確率に基づくエントロピー

$H(t, S_p)$ が高い語ほど，正例の文集合に均一に分布している語であることが分かる．$H(t, S_p)$ は次の式 6 で求める．

$$H(t, S_p) = -\sum_{s \in S_p} P(t, s) \log_2 P(t, s) \tag{6}$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \tag{7}$$

ここで，$P(t, s)$ は文 s における語 t の出現確率を表し，$tf(t, s)$ は文 s において語 t が出現する頻度を表す．

次に，負例に含まれる内容語（名詞，動詞，形容詞）に対しても，同様に重みを計算する．

$$W_n(t, S_n) = TF(t, S_n)H(t, S_n) \tag{8}$$

ただし，S_n は学習データにおいて負例に属する文の集合である．

ここで，ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ より大きければ，その語 t を素性として選択する．もしくは，語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の 2 倍より大きければ，その語 t を素性として選択する．すなわち，以下の条件のどちらかが成り立つ語 t を素性として選択する．

$$W_p(t, S_p) > W_n(t, S_n) \tag{9}$$
$$W_n(t, S_n) > 2W_p(t, S_p) \tag{10}$$

上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，ともによく出現するような一般的な語を素性から除去する．以下に議案分類「取締役選任」を判別するための学習データから選択された素性の一部を例示する．

取締役，監査，議案，配当，株主，社外，変更，事業，代表，現任，責任，部長，社長

上記の学習データでは，2,845 語が素性として選択された．

2.4.2 モデル

入力は，学習データから抽出された語（素性）を要素，語 t における $\log(W_p(t, S_p))$，もしくは，$\log(W_n(t, S_n))$ の大きいほうを要素値としたベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数（すなわち素性の数）と同じとし，隠れ層は，ノード数 1000 が 3 層，ノード数 500 が 3 層，ノード数 200 が 3 層，ノード数 100 が 3 層の計 12 層とする．出力層は 1 要素である．活性化関数はランプ関数（ReLU）[58] を使用し，出力層は sigmoid 関数を使用した．また，エポック数は 50 回とした．

機械学習による分類を 2 値で行った理由は，1 つのページに複数の議案が記載されており，複数の議案分類が割り当てられる可能性があるからである．例えば，「取締役選任」と「監査役選任」が同じページに記載されている場合があり，その場合は，そのページを「取締役選任」と「監査役選任」に分類する必要がある．そのため，ある議案分類とそれ以外を分類する分類器を議案分類数だけ生成するアプローチを採った．

2.5　提案手法３: 抽出した議案タイトルを用いた議案分類

本章では，議案タイトルの抽出手法と，抽出した議案タイトルを用いた議案分類手法について説明する．議案タイトルとは，図 1 の第 1 号議案の隣に記載されている「剰余金の処分の件」や，第 2 号議案の隣に記載されている「定款一部変更の件」のことを指す．議案タイトルを正確に抽出することは難しいが，正しく議案タイトルを抽出できれば議案分類の大きな手助けとなるとともに，議案タイトルを人手で入力する手間を省くことができる．

2.5.1　議案タイトルの抽出

PDF データを 2.2 節と同様の条件でテキストデータに変換し，図 4 のフローチャートの条件で議案タイトルの抽出を行う．

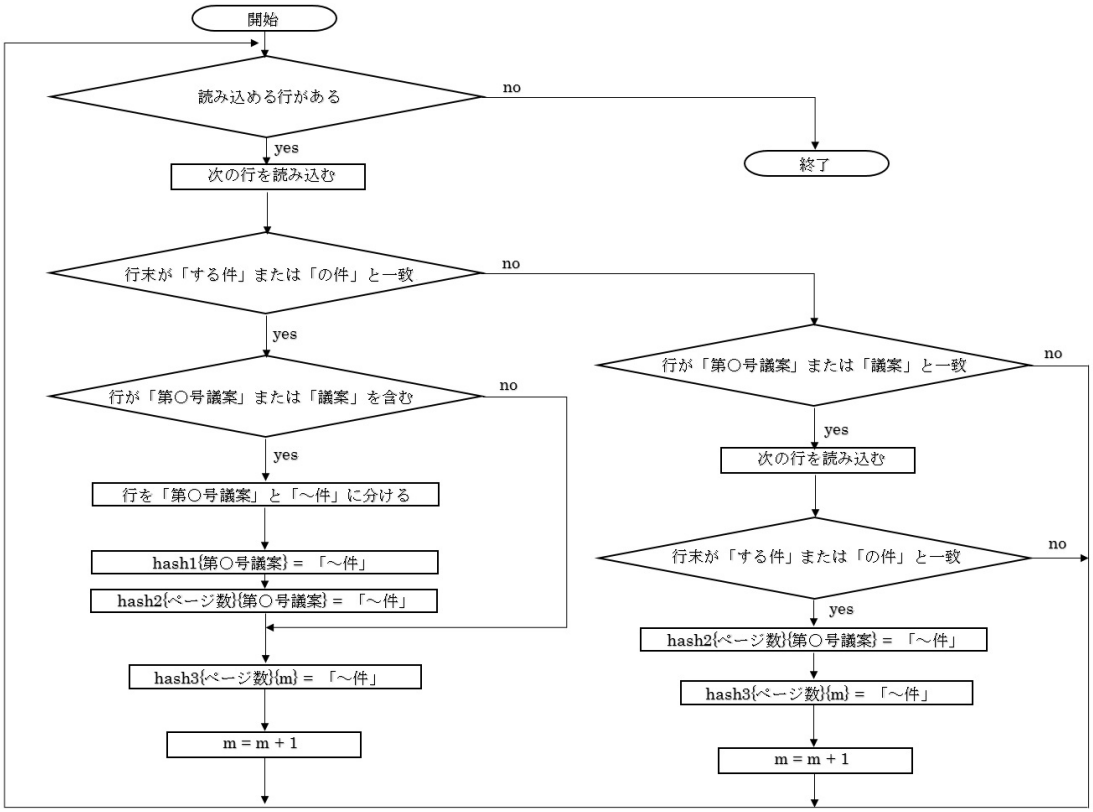


図 4　議案タイトル抽出のフローチャート．

議案タイトルは 2.2 節の手法を用いて推定された開始ページを用いて，ハッシュに保存された議案タイトルが選択される．以下の優先順位でハッシュは選ばれ，優先すべきハッシュに議案タイトルがない場合，次順位のハッシュに保存された議案タイトルを選択する．

$$hash1\{\text{”第○号議案”}\} > hash2\{\text{ ページ数 }\}\{\text{”第○号議案”}\} > hash3\{\text{ ページ数 }\}\{m\} \quad (1)$$

ここで，$hash3\{\text{ ページ数 }\}\{m\}$ はそのページに出現する m 番目の議案タイトルである．

2.5.2　議案タイトルを用いた議案分類手法

議案タイトルに含まれるキーワードを用いたルールベースによって議案分類を行う[*7]．図 5 に本手法である議案分類のフローチャートを示す．

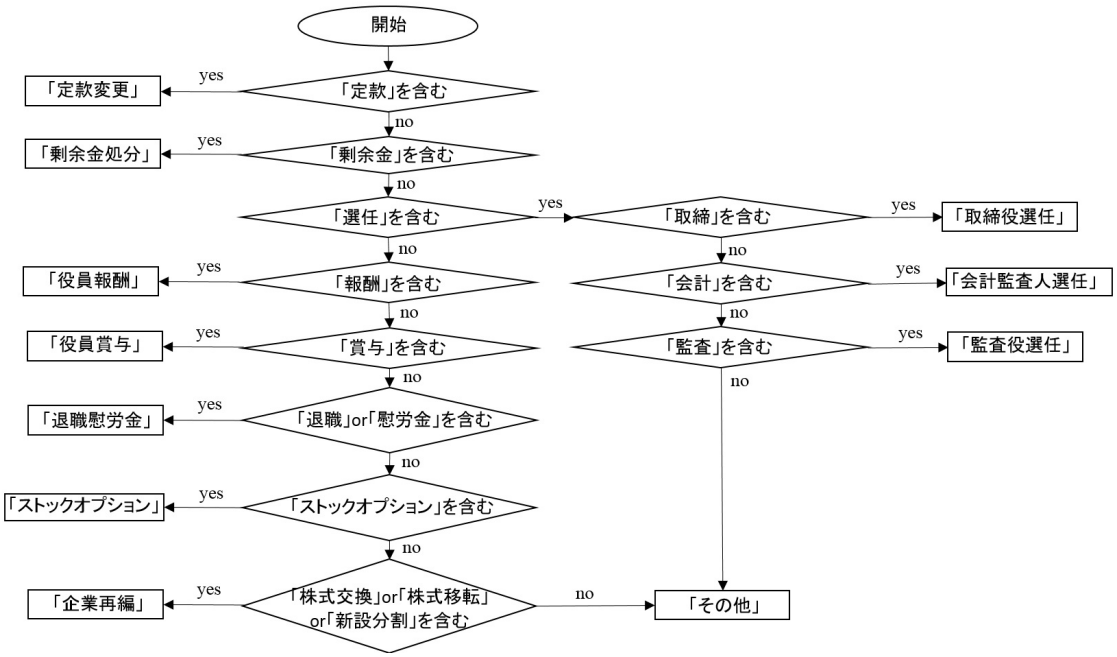


図 5　抽出した議案タイトルを用いた分類のフローチャート．

2.6　評価

本手法を実装した．学習データとして，2016 年 4 月から 8 月までの株主招集通知から人手にて議案開始ページとその分類を作成し使用した（学習データ数は 6,444 件）．実装

[*7] 作成したルールは，学習データとして用いられている株主招集通知を参考に作成した．

にあたり，形態素解析器として MeCab[*8]を使用した．深層学習においては TensorFlow[*9] を使用した．評価において，正解データとして，2016 年 9 月から 10 月までの株主招集通知から人手にて議案開始ページとその分類を作成した（正解データ数は 345 件）．議案分類ごとの学習データと正解データは以下の表 9 の通りである．次に，正解データと同じ 9

表 9　議案別データ数

議案分類	学習データ	正解データ
会計監査人選任	69	4
監査役選任	1140	51
企業再編	37	4
ストックオプション	90	13
退職慰労金	349	10
定款変更	858	38
取締役選任	1750	110
役員賞与	144	6
役員報酬	437	21
剰余金処分	1371	72
その他	199	16
合計	6444	345

月から 10 月までの株主招集通知から，各手法を用いて議案開始ページとその議案分類を推定した．表 10 は，ある企業の株主招集通知における正解データと提案手法 1 による議案開始ページと議案分類の推定結果を示す．そして，各手法の推定結果と正解データが一致すれば正解とし，議案ごとの適合率，再現率，F 値を算出した．また，比較手法として，SVM[59] による分類も行った[*10]．素性選択の方法は MLP と同様である．評価結果を表 11 に示す．表 11 の手法 1 は提案手法 1 の特徴語による議案分類手法，手法 2 は提案手法 2 の MLP による議案分類手法，手法 3 は提案手法 3 の抽出した議案タイトルを用いた議案分類手法の結果を示す．

MLP による提案手法 2 を評価するために，素性選択を行わなかった場合，MLP の隠れ層やエポック数を変化した場合，MLP ではなく SVM を使用した場合の評価を行った．

[*8] http://mecab.googlecode.com/ svn/trunk/mecab/doc/index.html
[*9] https://www.tensorflow.org/
[*10] SVM のカーネルとして線形カーネルを使用した．

表 10　提案手法 1 による議案開始ページと議案分類の推定結果とその正解データ

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
34	剰余金処分	34	剰余金処分
35	定款変更	35	定款変更
37	取締役選任	37	取締役選任
38	監査役選任	38	監査役選任
39	退職慰労金	39	退職慰労金

表 12 に各手法や設定を示す．ここで，比較手法 2 の「素性選択なし」とは，学習データにおける内容語を全て素性として使用し，重みとして情報利得を使用したものである．また，各手法や設定ごとに全ての議案分類の適合率，再現率を載せるのは数値が多すぎるため，各手法や設計ごとの全議案における適合率，再現率のみを表 13 に示す．

2.7　各手法に対する考察

各手法の評価の結果，提案手法 3，提案手法 2，手法 (SVM)，提案手法 1 の順に良好な結果が得られた．各手法には特徴があり，「ストックオプション」の F 値では提案手法 1 が最も高く，「剰余金処分」の F 値は提案手法 2 が最も高い．

提案手法 1 は「ストックオプション」の結果が他の手法と比較して良好な結果となった．これは特徴語の選択とその重み付けに起因している．得られた特徴語と重みを表 14 に示す．また，特に提案手法 3 と比較すると高い結果が得られているが，これは「ストックオプション」に該当する議案のタイトルは多種多様であることから，提案手法 3 のタイトルを用いた分類は難しいことが原因でもある．

提案手法 1 の分類推定が誤っていたものを確認したところ，誤分類は 46 件だった．その誤分類の詳細を確認したところ，「取締役選任」の項目が「監査役選任」の項目に誤分類されている件数が 15 件あった．これはどちらの項目も選任の件であり，分類が難しいことと，議案としての出現確率が高いことに起因している．また，分類には「その他」といった項目が存在するが，今回は「その他」への分類をしていないため，15 件が誤分類となった．「その他」の項目は，どの議案にも分類されない議案で構成されるため，特徴語となるものが抽出できない．そのため分類対象から除外した．

提案手法 1 は議案の分類をスコアが上位のものから順に割り当てるため，同一ページに

表 11　評価結果

議案分類	手法	適合率	再現率	F 値	手法	適合率	再現率	F 値
全議案	手法 1	0.866	0.881	0.873	手法 3	0.970	0.961	0.966
	手法 2	0.953	0.921	0.937	手法 (SVM)	0.970	0.894	0.930
会計監査人選任	手法 1	1.000	1.000	1.000	手法 3	1.000	1.000	1.000
	手法 2	1.000	1.000	1.000	手法 (SVM)	1.000	1.000	1.000
監査役選任	手法 1	0.727	0.960	0.828	手法 3	1.000	1.000	1.000
	手法 2	0.845	0.961	0.899	手法 (SVM)	0.860	0.961	0.907
企業再編	手法 1	1.000	0.500	0.667	手法 3	1.000	0.750	0.857
	手法 2	1.000	0.750	0.857	手法 (SVM)	1.000	0.750	0.857
ストック	手法 1	0.813	1.000	0.897	手法 3	1.000	0.538	0.700
オプション	手法 2	0.846	0.846	0.846	手法 (SVM)	0.917	0.846	0.880
退職慰労金	手法 1	0.526	1.000	0.690	手法 3	1.000	1.000	1.000
	手法 2	1.000	1.000	1.000	手法 (SVM)	1.000	0.900	0.947
定款変更	手法 1	0.947	0.947	0.947	手法 3	0.974	0.974	0.974
	手法 2	0.949	0.974	0.961	手法 (SVM)	1.000	0.895	0.944
取締役選任	手法 1	0.958	0.873	0.910	手法 3	1.000	0.982	0.991
	手法 2	0.990	0.87	0.928	手法 (SVM)	1.000	0.855	0.922
役員賞与	手法 1	0.833	0.833	0.833	手法 3	1.000	1.000	1.000
	手法 2	1.000	1.000	1.000	手法 (SVM)	1.000	0.667	0.800
役員報酬	手法 1	0.842	0.941	0.889	手法 3	0.809	1.000	0.895
	手法 2	1.000	0.714	0.833	手法 (SVM)	1.000	0.714	0.833
剰余金処分	手法 1	0.931	0.931	0.931	手法 3	0.960	1.000	0.980
	手法 2	0.986	1.000	0.993	手法 (SVM)	1.000	0.986	0.993
その他					手法 3	0.833	0.769	0.800

対し同じ議案分類を割り当てることができない．そのため，同じ議案分類が複数出てきた開始ページの推定に影響を与え，6 件の誤分類となった．また，その場合の推定は「退職慰労金」に分類される傾向にあり，それに起因して「退職慰労金」の分類の適合率が低くなってしまった．その際の正解データと提案手法 1 の出力結果を表 15 に示す．この例は，P.53 に「役員報酬」に関する議案が 2 件記載されているが，「役員報酬」の次にスコアが高い「退職慰労金」が議案分類として割り当てられてしまった結果である．

表 12　比較手法と設定

	ラベル	手法	素性選択	隠れ層	ノード数 1000	500	200	100	エポック数
提案手法 2	MM	MLP	あり	12 層	3 層	3 層	3 層	3 層	50
手法（SVM）	SVM	SVM	あり	-	-	-	-	-	-
比較手法 1	CM1	MLP	なし	12 層	3 層	3 層	3 層	3 層	50
比較手法 2	CM2	SVM	なし	-	-	-	-	-	-
設定 1	S1	MLP	あり	8 層	2 層	2 層	2 層	2 層	50
設定 2	S2	MLP	あり	4 層	1 層	1 層	1 層	1 層	50
設定 3	S3	MLP	あり	12 層	3 層	3 層	3 層	3 層	25
設定 4	S4	MLP	あり	12 層	3 層	3 層	3 層	3 層	10

表 13　比較評価結果

ラベル	適合率	再現率	F 値
MM	0.953	0.921	0.937
SVM	0.970	0.894	0.930
CM1	0.952	0.851	0.898
CM2	0.970	0.893	0.930
S1	0.952	0.914	0.932
S2	0.943	0.914	0.928
S3	0.946	0.920	0.932
S4	0.949	0.920	0.934

表 14　「ストックオプション」の特徴語

特徴語	重み
予約権	31.37
新株	30.91
オプション	22.66
調整	20.08
付与	18.61
報酬型ストック	18.18

提案手法 1 の分類推定を向上させるためには，「取締役選任」と「監査役選任」の項目の特徴語の選択をヒューリスティックに調整することが考えられる．また，「その他」の項目も，同様の手法で分類できるようにすることも考えられる．同じ議案が複数存在するページに関しては，「退職慰労金」の項目に分類されることが多いため，「退職慰労金」への分類に制約を与えることで，解消されると考えられる．

表 15　同じ議案が同一ページに複数出てきた例

正解データ		分類結果	
開始ページ	議案分類	開始ページ	議案分類
40	剰余金処分	40	剰余金処分
40	定款変更	40	定款変更
48	取締役選任	48	取締役選任
50	取締役選任	50	取締役選任
52	役員賞与	52	役員賞与
52	退職慰労金	52	退職慰労金
53	役員報酬	53	役員報酬
53	役員報酬	53	退職慰労金

提案手法 2 の結果は，議案があると推定されたページに対して各分類器を適用するため，そのページにある議案数とは異なった結果を返すことがある．提案手法 1 で述べた，同じ議案が複数存在するページに関しては，結果を余分に返さないメリットがある．しかし，提案手法 2 は 1 ページ毎に分類を行うため，議案が複数存在し，該当ページの後半以降から始まり次のページに跨ぐような議案に対しては，正しい分類結果を返せないデメリットが存在する．

提案手法 3 の抽出した議案タイトルを用いた議案分類は，高い適合率，再現率を示しているが，議案タイトルの抽出の結果に強く依存しているため，他の分類手法と照らし合わせる必要がある．他の分類手法はページに出現する特長語を用いて分類を行うため，議案タイトルが正しく抽出できない場合においても分類を行うことが可能である．

議案がある開始ページの推定は，議案があると推定された株主招集通知に対しては 100% 推定できていた．しかし，議案が本来は記載されているが議案がないと判断された株主招集通知が 2 件あった．原因を調べたところ，PDF からテキストデータへの変換のとき，「議案」といった表現が抽出されてないことが原因であった．その例を図 6 に示す．

図 6 の矢印で示した位置に，本来 PDF に記載されている「議案」という文字列がテキストに変換されるはずだができていない．これは文字列「議案」の PDF への入力が特殊であるために生じる．PDF を画像ファイルとして読み込み，文字認識をかけるなど解決策はあるが，本研究から内容が逸脱するため例外として処理した．

表 13 から，素性選択ありの提案手法 2 と，素性選択なしの比較手法 1 の比較により，

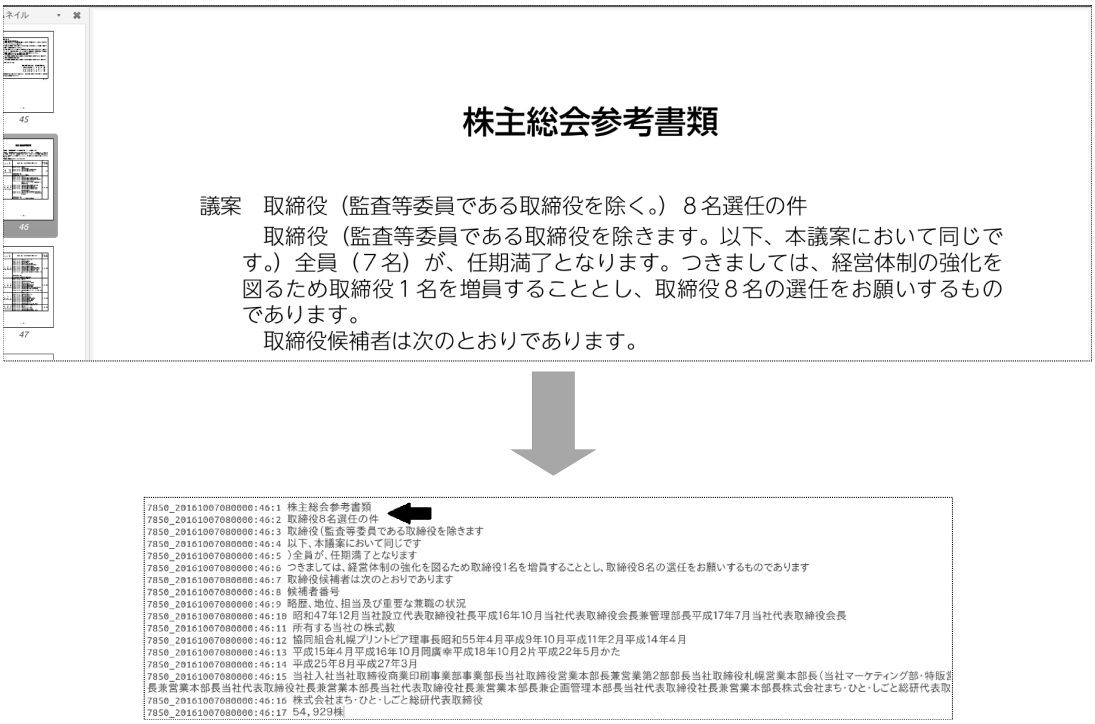


図 6　議案が正しく変換できなかったテキストデータの例.

素性選択ありのほうが素性選択なしよりも高い F 値を達成した．従って，提案手法 2 における素性選択は本タスクにおいて有効であると考える．SVM においては，マージンを最大化する超平面を学習データから決める手法であり，その特徴から素性選択の有無に関わらず同程度の性能を得ることが可能であった．提案手法 2 と SVM との比較では，SVM のほうが適合率が高いが再現率が低いという結果となり，F 値はほぼ同じ結果となった．また，隠れ層の数やエポック数の変化でも本手法とほぼ同じ結果となっており，機械学習に基づく手法では，手法の如何にかかわらず，この結果が上限であると考える．

2.8　応用システム

評価の結果と考察を踏まえ，特徴語の重みによる議案分類と抽出した議案タイトルを用いた議案分類を実際に運用できるシステム構築を行った．システムへのインプットは株主招集通知の PDF，アウトプットには議案のあるページ，抽出した議案タイトル，各手法の議案分類の結果，分類の信頼度とする．図 7 にシステムの構造を示す．

ここで，信頼度とは，抽出したタイトルを用いた議案分類の結果の確からしさを示したものであり，値が 1 に近いほど議案分類の推定結果が正しいことを意味している．人が最

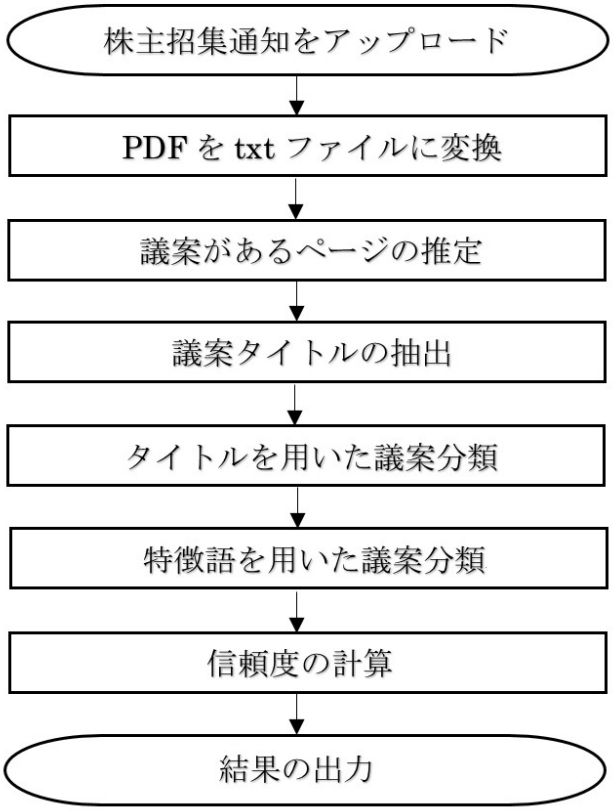


図 7　システム構造.

終確認をする際に確認対象の判断を行うため，抽出した議案タイトルを用いた議案分類の分類結果の信頼度を追加した．信頼度の計算はベイズを利用した [60].

2.8.1　ベイズ推定による信頼度の推定

ここで説明のため，議案集合を U とし，抽出した議案タイトルによる分類結果「t」が得られたときの事象を以下のように定義する.

$A(t)$:　議案分類は「t」である.
$B(t)$:　抽出した議案タイトルによる議案分類結果が「t」であった.
$C(m)$:　特徴語による議案分類が「m」であった[*11].

[*11] m は $m = t$ でも可．また，複数議案が選ばれることもある．例：t =「定款変更」，m =「定款変更」.

このとき求めたい信頼度は以下の式 12 で得られる．

$$P(A(t)|B(t),C(m)) = \frac{P(B(t),C(m)|A(t))P(A(t))}{\sum_{d\in U} P(B(t),C(m)|A(d))P(A(d))} \tag{12}$$

ベイズ推定では，$P(A(t)|B(t),C(m))$ を事後確率，$P(B(t),C(m)|A(t))$ を尤度，$P(A(t))$ を事前確率と呼ぶ．すなわち式 12 は，元々の議案の出現確率に尤度をかけることで，$B(t),C(m)$ という根拠を加味した条件付き確率を得ることができる．しかし，特徴語による議案分類は，同ページに議案が複数ある場合，複数の分類結果を返すため，全パターンの尤度の計算はデータ数の関係もあり困難である．よって，事象を以下のように変更した．

A:　議案分類は「t」である．
$\bar{A}$:　議案分類は「t」でない．
B:　抽出した議案タイトルによる議案分類結果が「t」であった．
C:　特徴語による議案分類が「t」を含む．
$\bar{C}$:　特徴語による議案分類が「t」を含まない．

よって先ほどの式 12 は式 13 のようになる[*12]．

$$P(A|B,C) = \frac{P(B,C|A)P(A)}{P(B,C|A)P(A) + P(B,C|\bar{A})P(\bar{A})} \tag{13}$$

また，ベイズ流のアプローチは条件をひとつずつ加え，事後確率を変化させていくのが基本となる．したがって，条件 B を加えたことによって得られた事後確率を，更新された事前確率として条件 C を加えたときに用いることで，最終的な事後確率を計算する[*13]．これは，事象 B と事象 C はマルコフ性のある確率過程であるという仮定に基づく．

2.8.2　条件 B を加えた事後確率 $P(A|B)$ の算出

まずは事前確率 $P(A)$ に，条件 B を加えた事後確率 $P(A|B)$ を求める必要がある．$P(A|B)$ はベイズの定理より以下の式 14 で求めることができる．

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \tag{14}$$

これは議案タイトル抽出の議案分類の適合率の計算と同様であるため，評価で得られた各議案の適合率を $P(A|B)$ として用いる．例えば，「定款変更」の事前確率 $P(A)$ は 0.133

[*12] $\bar{C}$ の場合もほぼ同様であるためここでは省略する．
[*13] これはベイズ更新と呼ばれる．

であるが，ベイズ更新によって事後確率 $P(A|B)$ は 0.9737 となる．

2.8.3 条件 C を加えた事後確率 $P(A|B,C)$ の算出

すでに条件 B を加えたので，更新された事前確率 $P(A|B)$ を用いて $P(A|B,C)$ を求める．$P(A|B,C)$ はベイズの定理より以下の式 15 で求めることができる．

$$P(A|B,C) = \frac{P(C|A,B)P(A|B)}{P(C|A,B)P(A|B) + P(C|\bar{A},B)P(\bar{A}|B)} \tag{15}$$

ここで，事象 B は議案タイトルを用いた分類による分類結果，事象 C は文中に出現した特徴語を用いた分類結果であるため，分類に用いる情報源が異なることから分類結果は独立であるという仮定の下，式 15 は以下の式 16 で表すことができる．

$$P(A|B,C) = \frac{P(C|A)P(A|B)}{P(C|A)P(A|B) + P(C|\bar{A})P(\bar{A}|B)} \tag{16}$$

ここで，

$P(A|B)$: 抽出した議案タイトルによる議案分類の適合率．
$P(\bar{A}|B)$: $1 - P(A|B)$．
$P(C|A)$: 特徴語による議案分類の再現率．
$P(C|\bar{A})$: 特徴語による議案分類の誤検知率[*14]．

同様に $P(A|B,\bar{C})$ も以下の式 17 で表すことができる．

$$P(A|B,\bar{C}) = \frac{P(\bar{C}|A)P(A|B)}{P(\bar{C}|A)P(A|B) + P(\bar{C}|\bar{A})P(\bar{A}|B)} \tag{17}$$

ここで，

$P(\bar{C}|A)$: $1 - P(C|A)$．
$P(\bar{C}|\bar{A})$: $1 - P(C|\bar{A})$．

2.8.4 システムの実行例

株主招集通知の PDF ファイルをアップロードすると，ページ数，議案番号，議案タイトル，抽出した議案タイトルを用いた議案分類の結果，特徴語を用いた議案分類の結果，

[*14] 再現率は真の値が 1 のものに対して 1 であると予測される比率であるため，誤検知率を真の値が 0 であるものに対して，1 であると予測される比率と定義する．例：真の分類が「定款変更」のものに対して，「定款変更」と予測されたものの比率が再現率であり，真の分類が「定款変更」でないものに対して，「定款変更」と予測されたものの比率が誤検知率である．

信頼度が画面に表示される．

システムの実行結果の例を，図 8 と図 9 に示す．左上の「～.pdf」のリンクをクリックすると，アップロードした PDF ファイルの閲覧が可能である．また，右の PDF のリンクをクリックすると，該当ページの PDF ファイルを閲覧することが可能である．

サムコ

PDF : 6387_1495247685.pdf

ページ数	議案番号	議案題目	議案分類（議案題目）	議案分類（特徴語）	信頼度	
33	第１号議案	剰余金の処分の件	剰余金処分	剰余金処分,	0.99	PDF
34	第２号議案	取締役７名選任の件	取締役選任	取締役選任,	0.99	PDF
37	第３号議案	監査役３名選任の件	監査役選任	監査役選任,	0.99	PDF
39	第４号議案	退任監査役に対し退職慰労金贈呈の件	退職慰労金	退職慰労金,	0.99	PDF

図 8　「サムコ」の株主招集通知をアップロードした出力結果.

テクノメディカ

PDF : 6678_1495247399.pdf

ページ数	議案番号	議案題目	議案分類（議案題目）	議案分類（特徴語）	信頼度	
38	第１号議案	定款一部変更の件	定款変更	定款変更,	0.99	PDF
39	第２号議案	剰余金処分の件	剰余金処分	剰余金処分,	0.99	PDF
40	第３号議案	取締役６名選任の件	取締役選任	取締役選任,	0.99	PDF
42	第４号議案	監査等委員である取締役１名選任の件	取締役選任	監査役選任,	0.96	PDF
43	第５号議案	会計監査人選任の件	会計監査人選任	会計監査人選任,	0.99	PDF

図 9　「テクノメディカ」の株主招集通知をアップロードした出力結果.

図 9 の第４号議案の結果に注目すると，抽出した議案タイトルを用いた分類では「取締役選任」となっているが，特徴語を用いた分類では「監査役選任」に分類結果となっている．ここで信頼度を見ると，抽出した議案タイトルによる分類結果である「取締役選任」は約 96% の信頼度である．実際に確認したところ，真の分類は「取締役選任」だった．これは，「取締役選任」の分類において，抽出した議案タイトルを用いた手法は高い適合率を誇る一方で，特徴語による議案分類は「取締役選任」の再現率が低いため，約 96% の信頼度という結果が得られた．これは実際に人手で確認するとき有用であると考えられる．

2.8.5　応用システムの評価

本応用システムによる業務の効率化を評価するために，実装した応用システムと従来の人手による作業の比較評価を行った．評価基準は実際の業務に基づいたデータ収録にかかった時間と適合率・再現率とする．応用システムは，2 つの議案分類手法を採用しているため 2 つの結果が得られるが，信頼度は抽出したタイトルを用いた議案分類の結果の確からしさを示したものであることから，適合率と再現率は抽出したタイトルを用いた議案分類の結果を用いた．

2.8.6　評価方法

学習データとテストデータとは別に，新たに 95 社分の株主招集通知を人手にて確認し，実際の業務と同じ手順によりデータベースに収録した．その後チェックを行ったものを正解データとする．この 95 社分のデータを 1 社ずつ時間を計り収録する作業を 12 人 の方に行ってもらい，12 人 から得られた各 1 社ずつの収録時間の平均を算出した．応用システムは 95 社分のデータをアップロードし，出力結果を csv ファイルでダウンロードし，データベースに収録するまでの時間を測定した．この時点までにかかった時間を測定し，収録されたデータがどのくらい正しく入力されているかを確認した．

2.8.7　収録時間

実装した応用システムの収録時間の結果と人手による収録時間の結果を表 16 に示す[*15]．

2.8.8　適合率・再現率・F 値

各作業者の適合率・再現率と応用システムの適合率・再現率・F 値を表 17 に示す．

2.8.9　考察

応用システムの実装によって，当該部分の 1 社あたりにかかる処理時間は約１０分の 1 となった．また，適合率・再現率から明らかなように，人手による作業よりも高い精度で開始ページの推定と議案分類が実現できた．

応用システムは例外に対応できないことや 100% の正解率で議案の分類ができないため，この後の作業工程であるチェックの際に，人手で作成したデータよりも時間がかかる

[*15] 企業 5〜企業 93 はスペースの関係上省略

表 16　収録にかかった時間の結果

	人手	応用システム
企業 1	173.36 秒	
企業 2	79.55 秒	
企業 3	289.09 秒	
企業 4	178.82 秒	
企業 94	175.70 秒	
企業 95	82.10 秒	
合計	13601.55 秒	1300 秒
1 社あたりの平均	143.17 秒	13.68 秒

と考えていた．しかし，人手にて作成したデータは勘違いによるミスや誤分類が多く，応用システムよりも質の低い結果となっているため，応用システム以上に時間をかける必要がある．また，この後のチェック段階において，人手によるデータには信頼度はないため，全収録データを複数人で確認する必要がある．しかし，応用システムの信頼度が 0.9 を越えるものと下回るものに対しての議案数と適合率は表 18 のようになっており，信頼度が 0.9 を下回るものに対しては複数人での確認が必要だが，信頼度が 0.9 を越えるものに対しては一人が確認を行えば十分であると思われる．これにより，さらなる業務の効率化が期待される．

2.9　本章のまとめ

本研究では，株主招集通知における議案の開始ページを推定し，その議案を分類する手法を提案した．本研究により，人手で株主招集通知から議案の開始ページを探しだし，分類をする作業時間を大幅に削減できた．議案の開始ページ推定は，議案がある開始ページには「議案」または「第 X 号議案」が先頭に含まれるといった規則に基づく．議案分類の推定は，議案ごとの特徴語を抽出し，その特徴語のスコアに基づき分類する手法，MLP を用いた分類，議案タイトルを用いたキーワードによる分類の 3 手法を用いた．評価の結果，特徴語による議案分類，深層学習による議案分類，抽出した議案タイトルを用いた議案分類，どの手法も良好な適合率，再現率を達成した．これらの結果を踏まえ，応用システムの実装により，株主招集通知の構造理解の習得が不要になると共に，議案の開始ページの推定や議案内容が分類されることにより，当該部分の 1 社あたりにかかる処理時間が

表 17　収録されたデータの適合率・再現率・F 値

	適合率	再現率	F 値
応用システム	0.950	0.926	0.938
作業者 1	0.428	0.404	0.415
作業者 2	0.467	0.470	0.468
作業者 3	0.513	0.500	0.506
作業者 4	0.539	0.510	0.524
作業者 5	0.489	0.489	0.489
作業者 6	0.490	0.490	0.490
作業者 7	0.745	0.736	0.741
作業者 8	0.716	0.671	0.693
作業者 9	0.710	0.710	0.710
作業者 10	0.795	0.798	0.797
作業者 11	0.789	0.786	0.787
作業者 12	0.822	0.778	0.799
作業者平均	0.625	0.612	0.618

表 18　信頼度別応用システムの適合率

	0.9 以上	0.9 未満
議案数	301	41
適合率	0.997	0.610

１０分の１程度に短縮された．また，信頼度により，その後の作業工程であるチェックの時間も短縮されることが見込まれる．

これらのことから，本研究によって，理解の十分でない作業者の判断ミスや判断の揺れが減少し，信頼度の高いデータ生成を支えることが可能となる．その結果，収録に係る人件費の削減と，データベース化に伴うデータ収録の早期化をはかることができるであろう．

本章の実験と評価によって，人手で作成した学習データを用いて分類器を学習し，分類対象となるページをルールベースで絞った上で，ページ単位での分類は，従来のテキスト分類手法である SVM や深層学習である MLP を用いることで，十分な精度で実現可能で

あることを示した．次章では，学習データを自動生成した上で，従来の単一ページの情報を入力とするモデルと，前後のページ情報を加味するモデルをいくつか従来手法から提案し，比較検討を行う．

3　学習データの自動生成による深層学習を用いた株主招集通知の重要ページ抽出

3.1　研究概要

2 章で述べた研究の結果より，人手で学習したデータを用いて分類器を学習し，分類するページをルールベースで絞ることができれば高い結果を示すことができた．本章では，ページ単位での有益情報抽出を行うために，自動生成した学習データを用いて各ページを分類し，重要ページの抽出を行う．

資産運用会社では，様々な部署で文書に目を通して重要な部分を確認することや，レポートを作成するなどの金融テキストに関わる業務を多く行っている．これらの業務を人工知能分野の技術を用いて効率化することにより時間を大幅に削減することができれば，コスト削減や生産性の向上が期待できる．本章で述べる研究では，そのような業務の一部で扱っている「株主招集通知」の確認作業を効率化することを目的とする．

株主招集通知は，企業が株主総会の開催前に株主に送付する義務があり，株主でなくとも企業の Web サイトから PDF ファイルとして閲覧，取得することが可能である．この通知には，会社のプロフィール，大株主情報，役員情報，決議事項である議案など，多くの情報が含まれている．したがって，機関投資家は投資判断だけでなく議決権行使においても，これらの情報を使用している．ただし，機関投資家が投資判断に影響を与える可能性のある情報を抽出するには，次のような課題がある．まず，株主招集通知のページ数は，数ページ程度のものもあれば 100 ページを超えるものまで様々である．また，株主総会前に発行されることから，特定の月に発行が集中する特徴を持つため，最も株主総会が集中する月である 6 月には全体で何千もの通知が発行される．そのため，多くの銘柄を扱う機関投資家にとって，株主招集通知の確認は大きな負担となっている．

本章の研究の目的は，株主招集通知から投資判断に影響を与える可能性の高いページ[*16]を，自動的に抽出することである．そのために，自動的に重要ページに Tag を付けることで，情報を抽出する必要がある．

本章の研究では，重要ページ抽出のために，次のようなフレームワークを提案する．

Step1　各企業の Web サイト上にある IR 情報ページから，企業が公開している株主招集通知の PDF ファイルを収集

[*16] 以下，確認が必要な投資判断へ影響を与える可能性が高いページを「重要ページ」とする．

Step2　Step1 で収集したファイルに対して，ルールベースで各ページに Tag を付与することで，学習データを自動生成

Step3　Step2 で作成した学習データを用いて，各ページの Tag を付与する分類器を学習

Step4　学習させた分類器を用いて，Tag を付与することで重要ページを抽出

このようなフレームワークを提案する背景として，(1) 分類器を充分に学習するために必要な学習データを人手で作成するには多くの時間と労力が必要であるため，また，(2) 本研究では株主招集通知を扱うが，株価に影響を与える可能性のある有益なテキスト情報源は他にも多数存在しており，他の金融テキストの重要ページ抽出をするために，その都度，そのテキストに精通している人材に学習データ作成の依頼をするのは非効率的であることが挙げられる．

本章の研究では，ルールベースによって自動生成される学習データの質を高めるために，ルールに少しでも当てはまらないファイルのデータは学習データから除かれているため，学習データは，想定される入力データ全体を網羅していないバイアスの強い学習データになっている特徴がある．このようなルールベースに当てはまる学習データを用いて，ルールベースでは抽出できないページを含むファイルに対してフレームワークの有効性を検討する．

3.2　先行研究

本章の研究の特徴は以下の通りである．

1. 対象が株主招集通知
2. 学習データの自動生成
3. バイアスがある学習データによる分類器の生成
4. ページ単位での抽出

まず前章で述べた研究との違いを述べる．前章で述べた研究では，まずルールベースの手法で決議事項である各議案が何ページから始まるかを推定し，そのページのテキスト情報などを用いて，記載されている議案がどの分類に属するかを自動で判定している．この分類には，実際の人手による業務によりデータベースに蓄えられたデータを学習データとして使用している．それに対して本章では，議案の分類は行っておらず，株主招集通知の各ページに対して確認すべき情報が記載されているページであるか，Tag 付けによる分類を行っている点で異なる．特に，前章の研究では開始ページの推定をルールベースにて

行っているが，本研究では最終的にはルールベースを用いることなく，開始ページを推定するだけでなく何ページまでそのことが記載されているかも推定することが可能である点で大きく異なる．

また，本章で扱う研究では当初，ルールベースを用いて株主招集通知から重要ページの抽出を行うことを検討していた [61]．しかし，人間が確認すればすぐにわかるような些細な違いでさえも，作成したルールに当てはまらない場合は抽出が正しくできない．そこで，重要ページである可能性の低いページも多く抽出することで再現率を高めている．しかしながら，企業によって記載の仕方に些細な違いがあることや，年度によって記載の仕方が変更になる可能性を考慮すると，ルールベースで対応し続けるには限界がある．

そこで本章では，文献 [61] のルールベースを参考に，新たにルールを作成することによって学習データを自動生成し，分類器を用いて抽出を行う．ルールベースを用いていることから，正しく抽出ができている可能性が極めて高いデータと，そうでないデータを区別することができる．ページ全体のテキスト情報を用いて重要ページを抽出するため，些細な違いでルールが適用できないような株主招集通知に対しても，本章で提案する手法は有効であることを示す．

また，学習データの自動生成による情報抽出の研究としては，酒井らは決算短信からの業績要因の抽出を行う研究 [12] や，次章で述べる有価証券報告書からの業績要因・業績結果の抽出の研究があるが，本章では，ページ単位での学習データを自動生成し，ページ単位で分類を行う点で大きく異なる．

もちろん可能であるならば，ページ単位よりもさらに詳細な情報である重要文の抽出ができることが好ましいが，ページ単位での抽出よりも質の高い学習データが必要であることや，一見重要ではない文に見えても，前後の文の情報を加味すると見落としてはいけない重要な文であったりすることがあるなど，現在の機械学習モデルでは，重要な箇所の見落としがあってはならない業務において，適用可能な領域に達していない場合も多い．さらに，本研究で扱うデータは，PDF をテキスト情報に変換したデータであり，文が崩れて抽出されることもある．このようなことから，ページ単位での抽出を行う方法を本章では提案している．

本章で，自動生成した学習データを用いてページ単位での Tag 付けによる分類をしている点で新規性があり，評価実験は，他のテキストデータに対しての応用を検討する際に有用である．学習データの自動生成は，ルールベースを用いているため，本課題に特化したものであるが，このルールベースを作成する以外の部分は，ページ抽出を行う上で汎用性の高い手法である．

3.3 提案手法

文献 [61] の提案する手法では，ルールベースによって株主招集通知から重要ページの抽出を行っているが，ルールから少しでも外れたりするものは正しく抽出ができない．

この問題を解決するために，本研究では以下の手法を提案する．

Step1　各企業の Web サイト上にある IR 情報ページから，企業が公開している株主招集通知の PDF ファイルを収集（3.3.1 節）

Step2　Step1 で収集したファイルに対して，ルールベースで各ページに Tag を付与することで，学習データを自動生成（3.3.2 節）

Step3　Step2 で作成した学習データを用いて，各ページの Tag を付与する分類器を学習（3.3.3 節）

Step4　学習させた分類器を用いて，Tag を付与することで重要ページを抽出

3.3.1 Step1: 企業 Web サイトからの株主招集通知の収集

企業の Web サイトでは決算短信などの金融に関する情報を IR 情報として公開している．株主招集通知も同様に IR 情報として公開されているため，企業 Web サイトの IR 情報をクローリングし，そこで公開されている PDF ファイルを自動でダウンロードすることで，多くの株主招集通知を含む PDF ファイルを収集することが可能である．ただし，取得した PDF ファイルには株主招集通知以外のファイルも数多く含まれていることから，自動的に株主招集通知を判別する必要がある．まず，2 章における研究と同様，PDF ファイルをテキストデータに変換する必要がある．PDF ファイルをテキストデータに変換するために「k2pdfopt[*17]」を用いた．「k2pdfopt」を用いることで，PDF ファイルから図 10 のようなテキストデータを各ページごとに取得することが可能である．このテキストデータを改行ごとに分割し，1 行の文字列に「株主総会」もしくは「株主招集」を含み，さらに，「通知」を含む文字列が出現している PDF ファイルを株主招集通知と判別する．

[*17] https://www.willus.com/k2pdfopt/

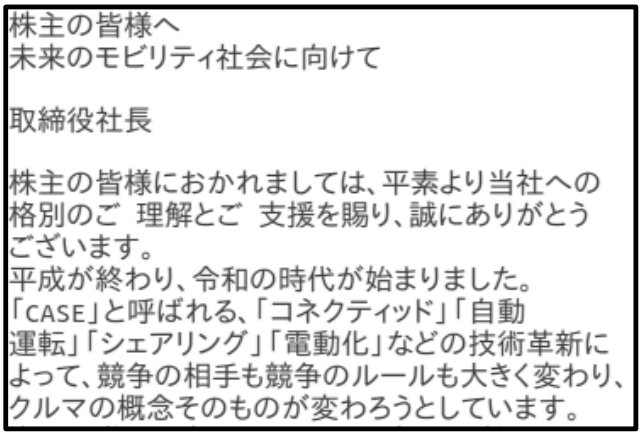


図 10　PDF データからテキストの抽出例

3.3.2　Step2: 学習データの自動生成

文献 [61] では，ルールベースを適用することで重要ページの抽出を行っているが，ルールに当てはまらない株主招集通知 PDF ファイルに対しては，余分に多くのページを抽出することで適合率を犠牲に再現率を高めている．そのため，重要ページの抽出には改善の余地がある．本研究では，Step1 で株主招集通知であると判定されたファイルを対象に，このようなルールに当てはまらない項目がある PDF ファイルと，抽出すべき重要ページが全て抽出できている可能性が極めて高い PDF ファイルに分け，後者を学習データとして扱う．学習データは各ページに対して以下のいずれかの Tag が付与される．

Tag1 (index)　表紙や目次となるページ
Tag2 (B-resolution)　決議事項である議案について記載されたページ（ 1 ページ目）
Tag3 (I-resolution)　決議事項である議案について記載されたページ（ 2 ページ目以降）
Tag4 (shareholder)　大株主に関するページ
Tag5 (B-officer)　会社役員に関するページ（ 1 ページ目）
Tag6 (I-officer)　会社役員に関するページ（ 2 ページ目以降）
Tag0 (O)　上記に当てはまらない印刷不要なページ

表 19 に，学習データの例を示す．各ページに対して，いずれかの Tag が付与された状態となっている．また，研究の目的は，Tag が付与されていない株主招集通知を分類器に入力することで，表 19 に示したような Tag を自動で付与するモデルを作成することである．Tag を付与することで，重要ページを抽出することが可能となる．

表 19　データへの Tag 付与の例

Page	Tag	Tag の意味
1	Tag1	表紙・目次
2	Tag0	印刷不要
3	Tag1	表紙・目次
4	Tag1	表紙・目次
5	Tag0	印刷不要
6	Tag0	印刷不要
7	Tag2	議案に関する記述の開始
8	Tag3	議案に関する記述の続き
9	Tag3	議案に関する記述の続き
10	Tag3	議案に関する記述の続き
11	Tag3	議案に関する記述の続き
12	Tag3	議案に関する記述の続き
13	Tag3	議案に関する記述の続き
14	Tag3	議案に関する記述の続き
15	Tag0	印刷不要
16	Tag0	印刷不要
17	Tag0	印刷不要
18	Tag0	印刷不要
19	Tag0	印刷不要
20	Tag0	印刷不要
21	Tag4	大株主情報
22	Tag0	印刷不要
23	Tag5	会社役員に関する記述の開始
24	Tag6	会社役員に関する記述の続き
25	Tag6	会社役員に関する記述の続き
26	Tag6	会社役員に関する記述の続き
27	Tag0	印刷不要
28	Tag0	印刷不要
29	Tag0	印刷不要
30	Tag0	印刷不要
31	Tag0	印刷不要
32	Tag0	印刷不要

各 Tag におけるページの抽出すべき重要性は以下の通りである．

Tag1 の表紙や目次となるページとして，表紙ページの例を図 13 に示す．表紙には，株主総会の開催日時や会場場所，決議事項である議案などの一覧が記載されている．表紙はたいてい先頭のページに出現する特徴がある．

目次ページの例を図 11 に示す．表紙ページとは別に記載されていることが多いが，企業によっては表紙ページを割愛し，目次ページを表紙ページとして 1 枚目に持ってくる企業も存在する．目次ページには決議事項である議案などの情報が記載されている．議案の

数は株主招集通知によって異なり，多い企業だと議案数が 10 を超える企業もあることや，株主の皆様への記載が多い企業などは，この目次ページにあたるページが複数ページにまたがる可能性もある．目次ページが複数ページにまたがる例を図 12 に示す．

表紙や目次となるページを同一の Tag である Tag1 とした理由はいくつかある．まず，記載されている確認したい重要な事項が，どちらか，もしくは両方に記載されているためである．そして，似たような情報が記載されることから，出現する文字列も類似したものが多く出現する．また，目次となるページは株主招集通知に必ず記載されているが，本研究で定義している図 13 に示したような表紙ページは PDF データでは省略されることがあるため，目次となるページから始まる株主招集通知も多い．そのため，表紙ページは他の重要ページに比べてデータ数が少ない．したがって，これらの理由から，表紙や目次となるページを同一の Tag である Tag1 に割り当てた．

Tag2 と Tag3 は，決議事項である議案について記載されたページに付与される Tag である．話し合われる議案の数が企業によって異なることや，議案の内容によって記載すべき必要な事項が異なり文字数や図表の量も変わるため，数ページのものから 30 ページを超えるものも存在する．特徴として，議案についてのページは連続して出現し，途中に他の記載（企業紹介や従業員数など）が出現しないため，議案についての記載が始まる開始ページには Tag2 を付与し，それ以降の議案について記載されたページには Tag3 を付与する．Tag2 のページ例を図 14 に示す．

Tag4 は大株主に関するページであり，例を図 15 に示す．株主招集通知では，大株主の情報として上位 10 名の大株主とその保有株数や保有率が記載されている．各企業の大株主情報は大きく変わることが少ないが，変更があったときには株価に影響を与える可能性があるため，重要ページである．たいていの企業は 1 ページにまとめているが，企業によっては記載内容の多さや表のフォーマットなどの理由から複数ページに記載している企業もある．

Tag5 と Tag6 は，会社役員に関するページに付与される Tag である．会社役員に関するページはたいてい数ページにわたって記載されているため，役員に関する記載が始まる開始ページの Tag5 と，それに続く Tag6 を付与することとした．Tag5 のページ例を図 16 に示す．

次に，1 ページ目と 2 ページ目以降で異なる Tag（Tag2 と Tag3，および Tag5 と Tag6）を一部の項目に対してのみ付与した理由を述べる．まず，このような各ページに対して Tag を付与する問題は，ある系列の各要素に適切なラベル列を付与する系列ラベリング問題（Sequential Labeling）と捉えることもできる．主な系列ラベリング問題として，文章を単語区切りに分割し，単語に対して名詞や形容詞などの品詞を割り当てる問題や，固有

表現に対して「IOB2」方式[*18]などの Tag を付与することで，固有表現を抽出する問題が挙げられる．これらの研究は Tag を付与する対象が単語列であるのに対して，本研究で Tag を付与する対象は各ページに対してである．また，本研究で抽出対象としている重要ページはその対象ごとに特徴が異なるため，その特徴や学習器を学習させることなどを加味した上で，抽出対象ごとに Tag 付けの方式を決定した．

系列ラベリング問題における Tag 付けの方式は，大きく分けると「IO」方式，「IOB2」方式，「IOE2」方式，「IOBES」方式などが存在する [17, 19, 20, 21]．本研究での Tag 付け方式は「IO」方式と「IOB2」方式を Tag ごとに使い分けている．本研究での Tag 付けの方式と各々の Tag 付け方式によって，学習データに Tag を付けた場合の例を表 20 に示す．

表 20 に示した通り，「IO」方式と「IOB2」方式の違いは，先頭かどうかを区別するかどうかである．重要事項が記載されるページの先頭は，書き出しは似ていることから区別しやすい（例えば，話し合われる議案には番号が付いているが，議案の記載は番号が若い順に記載されるため，「第 1 号議案」といった文字列が出現しやすいなど）のに対し，それらの記載が終了するページは，企業や年度，補足事項の有無によって多種多様である．そのため本研究では，終了を意味する E の Tag を用いる「IOE2」方式や「IOBES」方式ではなく，「IO」方式と「IOB2」方式に Tag 付けの方式を絞った．

本研究において，開始の Tag と続きの Tag を区別するメリットは 2 つある．株主招集通知には，必ず話し合われる議案や会社役員に関する記載があるが，どの程度の割合を占めているかは株主招集通知によって異なる．しかし，表 20 に示した例のように，これらの記載は連続して出現し，途中に他の記載が入ることはないため，Tag2，Tag5 となる開始のページは，1 つの株主招集通知に 1 ページずつしか存在しない．したがって，素性選択やモデルを検討する研究の初期段階において，それらの取捨選択をするのに有効である．例えば，あるモデルを学習させている途中に学習データにないデータを入力し，その出力において多くのページに対して Tag2 や Tag5 が付与されるモデルは学習がうまくできていないことが，正解データを人手で時間をかけて作らずとも判断することが可能となる．これが開始の Tag と続きの Tag を区別するメリットの 1 つ目である．

また，3.3.3.5 節で提案するようなモデルは，ラベル間の依存性を学習することができる特徴を持つ．例えば，本研究のデータであれば，議案の続きである「I-resolution」の Tag は，その前のページには「B-resolution」，「I-resolution」の Tag が出現しやすく，その後

[*18] 固有表現抽出であれば，文章を単語列に分け，固有表現の先頭に対して「Beginning」である B の Tag を付与し，固有表現が続く場合は「Inside」である I の Tag を付与する．そして，それ以外に対しては「Outside」である O の Tag を付与する方式．

表 20　Tag 付与方式ごとの Tag 付与の例

Page	本研究の方式	IO 方式	IOB2 方式
1	index	index	B-index
2	O	O	O
3	index	index	B-index
4	index	index	I-index
5	O	O	O
6	O	O	O
7	B-resolution	resolution	B-resolution
8	I-resolution	resolution	I-resolution
9	I-resolution	resolution	I-resolution
10	I-resolution	resolution	I-resolution
11	I-resolution	resolution	I-resolution
12	I-resolution	resolution	I-resolution
13	I-resolution	resolution	I-resolution
14	I-resolution	resolution	I-resolution
15	O	O	O
16	O	O	O
17	O	O	O
18	O	O	O
19	O	O	O
20	O	O	O
21	shareholder	shareholder	B-shareholder
22	O	O	O
23	B-officer	officer	B-officer
24	I-officer	officer	I-officer
25	I-officer	officer	I-officer
26	I-officer	officer	I-officer
27	O	O	O
28	O	O	O
29	O	O	O
30	O	O	O
31	O	O	O
32	O	O	O

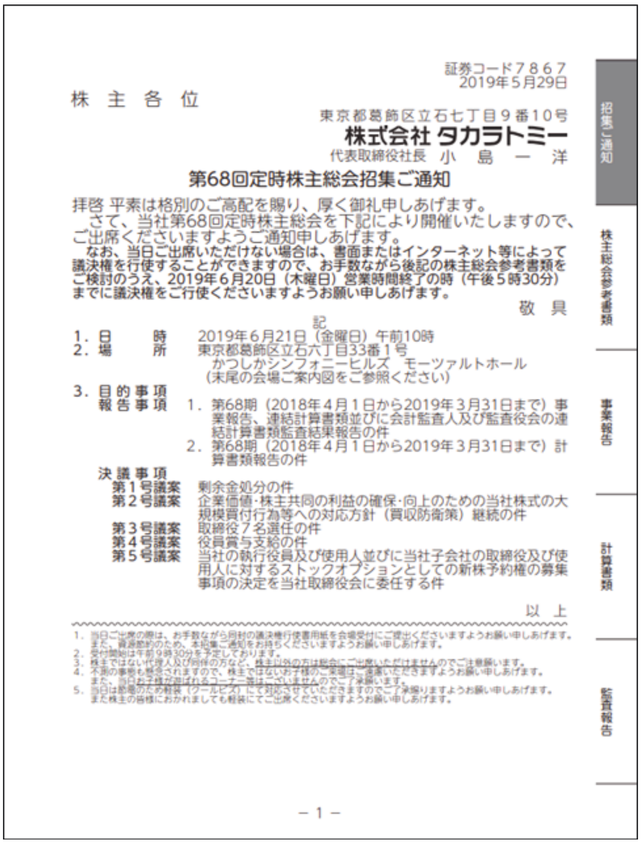
のページには「I-resolution」,「O」の Tag が出現しやすいことを学習することが可能である．それに対して「IO」方式での Tag 付けの場合には,「resolution」の Tag は，その前のページには「O」,「resolution」,「index」の Tag が出現しやすく，その後のページには,「resolution」,「O」の Tag が出現しやすいことを学習する．したがって,「IO」方式の場合，表 20 に示した例の 17 ページや 18 ページが，議案に関する記述に似たような文字列の出現するページだとすると，17 ページの前の 16 ページが「O」の Tag であったとしても「resolution」の Tag は「O」の次のページにも出現する可能性があるので，誤って

「resolution」の Tag を 17 ページや 18 ページに付与してしまう可能性がある．「IOB2」方式であれば，同様に 17 ページや 18 ページが議案に関する記述に似たような文字列の出現するページだとしても，17 ページが「I-resolution」のテキスト情報に似ている場合には前の 16 ページが「O」の Tag であるため，「I-resolution」が誤って付与される可能性を下げることができる．したがって，「第 1 号議案」などの議案の記載が開始するページにテキスト情報が似ていない限り，「B-resolution」が付与される可能性も低くなり，それに伴い，18 ページに「I-resolution」が誤って付与される可能性をも下げることが見込める．これが開始の Tag と続きの Tag を区別するメリットの 2 つ目である．

一部の項目に対してのみ開始のページと続きのページを区別する Tag 付与の方式を用いた理由について述べる．Tag1 (index) の表紙や目次となるページは，株主招集通知の若いページに出現する可能性が高いが，表紙と目次に同一の Tag を付与するため，連続して出現するとは限らない．例えば，表 20 に示した例だと，1 ページ目が表紙ページ (index)，2 ページ目に代表挨拶のページ (O)，3 ページ目と 4 ページ目が目次ページ (index) となる株主招集通知である．この場合，「IOB2」方式で開始位置かどうかで異なる Tag を付与するルールを適用すると，2 ページ目の代表挨拶のような「O」の Tag のページが間に入るかによって，目次ページの Tag が「B-index」から「I-index」に変わることになり，似たようなテキスト情報を持つページの Tag が変わるため，予測が難しくなることが推測できた．また，表紙ページが存在しない株主招集通知も多いため，表紙と目次ページの Tag を区別すると表紙の Tag が付与されたページが学習データとして少なくなってしまうため，同一の Tag を付与せざるを得なかった．これらの前提から，開始のページと続きのページを区別しない「IO」方式を用いた．Tag4 (shareholder) の大株主に関するページに関しては，2 ページに記載がまたがる株主招集通知が本手法での自動生成でほとんど生成することができなかった．そのため，開始のページと続きのページで異なる Tag を付与する「IOB2」方式を用いると，「I-shareholder」が付与されるページの学習データ数が非常に少なくなり，学習，予測をすることが難しいため，開始のページと続きのページを区別しない「IO」方式を用いた．特に，本研究で扱うデータは非常に Tag0 が多い不均衡なデータとなる．そのため，モデルの学習において，Tag の出現頻度で重み付けを行うが，そのときに出現頻度の極端に低い Tag は学習に悪影響を与えるため，少なくとも開始 Tag 以上の出現頻度がない Tag は付けない工夫が必要である．

これらの Tag を付与するためのルールを [61] を元に作成した．まず，重要ページを抽出するために，重要ページの候補となるページをテキストデータを用いて抽出する．

3.3.2.1 表紙や目次に関するページ候補の抽出


証券コード７８６７
2019年５月29日

株 主 各 位

東京都葛飾区立石七丁目９番10号
株式会社 タカラトミー
代表取締役社長 小 島 一 洋

第68回定時株主総会招集ご通知

拝啓 平素は格別のご高配を賜り、厚く御礼申しあげます。
　さて、当社第68回定時株主総会を下記により開催いたしますので、ご出席くださいますようご通知申しあげます。
　なお、当日ご出席いただけない場合は、書面またはインターネット等によって議決権を行使することができますので、お手数ながら後記の株主総会参考書類をご検討のうえ、2019年６月20日（木曜日）営業時間終了の時（午後５時30分）までに議決権をご行使くださいますようお願い申しあげます。

敬 具

記

1. 日　　時　2019年６月21日（金曜日）午前10時
2. 場　　所　東京都葛飾区立石六丁目33番１号
　かつしかシンフォニーヒルズ　モーツァルトホール
　（末尾の会場ご案内図をご参照ください）
3. 目的事項
　報告事項
　1. 第68期（2018年４月１日から2019年３月31日まで）事業報告、連結計算書類並びに会計監査人及び監査役会の連結計算書類監査結果報告の件
　2. 第68期（2018年４月１日から2019年３月31日まで）計算書類報告の件
　決議事項
　第１号議案　剰余金処分の件
　第２号議案　企業価値・株主共同の利益の確保・向上のための当社株式の大規模買付行為等への対応方針（買収防衛策）継続の件
　第３号議案　取締役７名選任の件
　第４号議案　役員賞与支給の件
　第５号議案　当社の執行役員及び使用人並びに当社子会社の取締役及び使用人に対するストックオプションとしての新株予約権の募集事項の決定を当社取締役会に委任する件

以 上

1. 当日ご出席の際は、お手数ながら同封の議決権行使書用紙を会場受付にご提出くださいますようお願い申しあげます。また、資源節約のため、本招集ご通知をお持ちくださいますようお願い申しあげます。
2. 受付開始は午前９時30分を予定しております。
3. 株主ではない代理人及び同伴の方など、株主以外の方は総会にご出席いただけませんのでご注意願います。
4. 不測の事態も懸念されますので、株主ではないお子様のご来場はご遠慮いただきますようお願い申しあげます。また、当日お子様が遊ばれるコーナー等はございませんのでご了承願います。
5. 当日は節電のため軽装（クールビズ）にて対応させていただきますのでご了承賜りますようお願い申しあげます。また株主の皆様におかれましても軽装にてご出席くださいますようお願い申しあげます。

招集ご通知　株主総会参考書類　事業報告　計算書類　監査報告

－１－

図 11　目次ページ（Tag1）の例

表紙や目次には，特徴的な文字列が多く存在するので，「第 n 号議案[*19]」と下記に挙げる文字列が，対象のページにすべて含まれているかどうかで判定した．

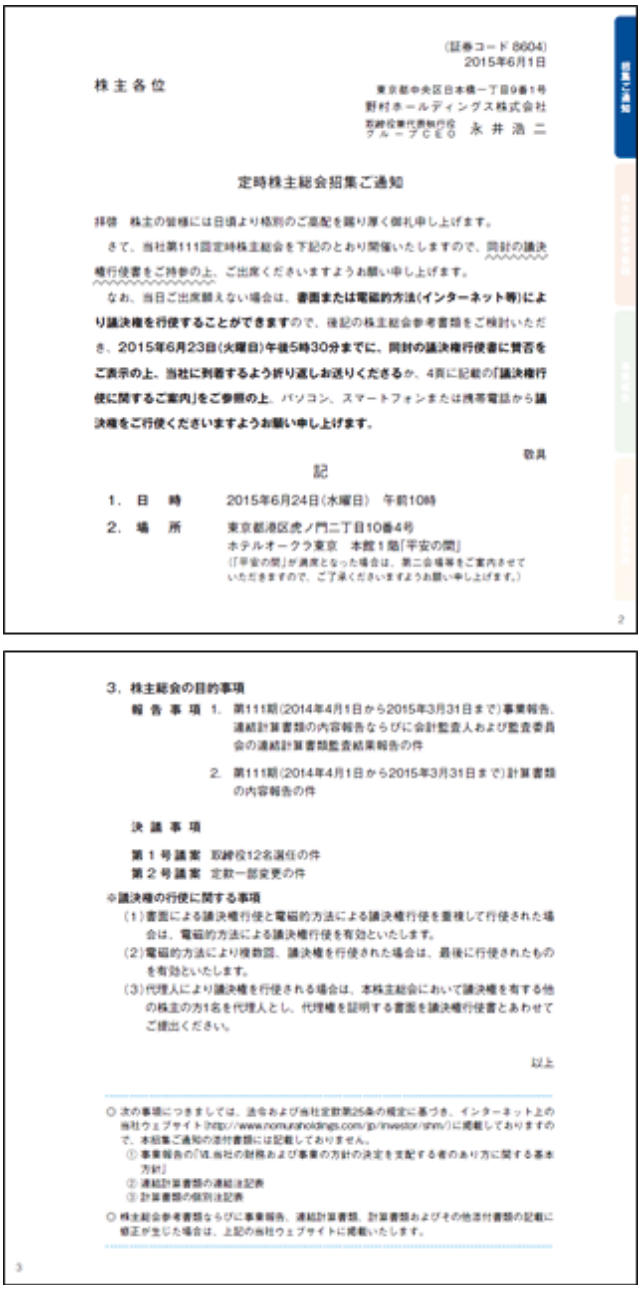
株主総会，日時，決議事項，通知，招集

目次のページは複数のページにまたがる可能性があるため，上記のルールに当てはまるページが存在しない場合には，新たに 2 ページごとにテキストデータをまとめたものを対象に，上記の文字列がすべて含まれているかどうかで判定した．

3.3.2.2 大株主に関するページ候補の抽出

大株主に関するページには，特徴的な文字列が多く存在するので，下記に挙げる文字列がすべて含まれているかどうかで判定した．

*19 n は対象とする株主招集通知の中で一番大きい番号．



(証券コード 8604)
2015年6月1日

株主各位

東京都中央区日本橋一丁目9番1号
野村ホールディングス株式会社
取締役兼代表執行役
グループCEO　永井浩二

定時株主総会招集ご通知

拝啓　株主の皆様には日頃より格別のご高配を賜り厚く御礼申し上げます。

さて、当社第111回定時株主総会を下記のとおり開催いたしますので、同封の議決権行使書をご持参の上、ご出席くださいますようお願い申し上げます。

なお、当日ご出席願えない場合は、**書面または電磁的方法(インターネット等)により議決権を行使することができます**ので、後記の株主総会参考書類をご検討いただき、**2015年6月23日(火曜日)午後5時30分までに、同封の議決権行使書に賛否をご表示の上、当社に到着するよう折り返しお送りくださる**か、4頁に記載の**「議決権行使に関するご案内」をご参照の上**、パソコン、スマートフォンまたは携帯電話から**議決権をご行使くださいますようお願い申し上げます。**

敬具

記

1．日　時　　2015年6月24日(水曜日)　午前10時

2．場　所　　東京都港区虎ノ門二丁目10番4号
ホテルオークラ東京　本館1階「平安の間」
(「平安の間」が満席となった場合は、第二会場等をご案内させていただきますので、ご了承くださいますようお願い申し上げます。)

2

3．株主総会の目的事項

報告事項　1．第111期(2014年4月1日から2015年3月31日まで)事業報告、連結計算書類の内容報告ならびに会計監査人および監査委員会の連結計算書類監査結果報告の件

2．第111期(2014年4月1日から2015年3月31日まで)計算書類の内容報告の件

決議事項

第1号議案　取締役12名選任の件
第2号議案　定款一部変更の件

※議決権の行使に関する事項

(1)書面による議決権行使と電磁的方法による議決権行使を重複して行使された場合は、電磁的方法による議決権行使を有効といたします。
(2)電磁的方法により複数回、議決権を行使された場合は、最後に行使されたものを有効といたします。
(3)代理人により議決権を行使される場合は、本株主総会において議決権を有する他の株主の方1名を代理人とし、代理権を証明する書面を議決権行使書とあわせてご提出ください。

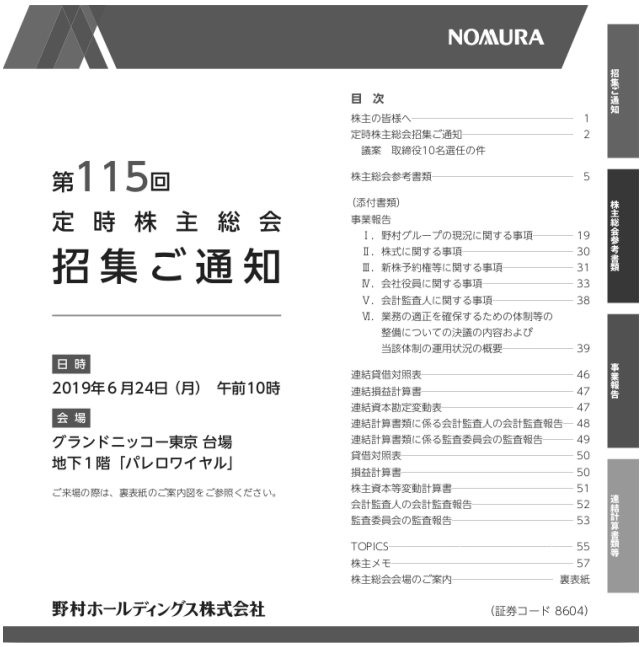
以上

○次の事項につきましては、法令および当社定款第25条の規定に基づき、インターネット上の当社ウェブサイト(http://www.nomuraholdings.com/jp/investor/shm/)に掲載しておりますので、本招集ご通知の添付書類には記載しておりません。
①事業報告の「Ⅵ.当社の財務および事業の方針の決定を支配する者のあり方に関する基本方針」
②連結計算書類の連結注記表
③計算書類の個別注記表

○株主総会参考書類ならびに事業報告、連結計算書類、計算書類およびその他添付書類の記載に修正が生じた場合は、上記の当社ウェブサイトに掲載いたします。

3

図 12　目次ページが 2 ページにわたって記載されている例



NOMURA

第115回
定時株主総会
招集ご通知

日時
2019年6月24日（月） 午前10時

会場
グランドニッコー東京 台場
地下1階「パレロワイヤル」

ご来場の際は、裏表紙のご案内図をご参照ください。

野村ホールディングス株式会社

（証券コード 8604）

目次

招集ご通知
株主総会参考書類
事業報告
連結計算書類等

図 13　表紙となるページ（Tag1）の例

株主総会参考書類

議案および参考事項

議案　取締役10名選任の件

本総会終結の時をもって、取締役10名全員が任期満了になります。つきましては、指名委員会の決定に基づき、取締役10名の選任をお願いしたいと存じます。

10名の候補者のうち、社外取締役候補者は6名であり、執行役を兼務する予定の取締役候補者は、永井浩二および永松昌一の2名であります。

取締役候補者は次のとおりです。

＜取締役候補者一覧＞

候補者番号	氏名	担当	取締役会への出席状況
1	古賀信行 重任 非業務執行取締役	取締役会長 指名委員（委員長） 報酬委員（委員長）	100%（10回/10回）
2	永井浩二 重任 執行役兼務	代表執行役社長 グループCEO	100%（10回/10回）
3	永松昌一 重任 執行役兼務	代表執行役副社長	100%（8回/8回）
4	宮下尚人 重任 非業務執行取締役	監査委員（常勤）	100%（10回/10回）
5	木村宏 重任 社外取締役・独立役員	指名委員 報酬委員	100%（10回/10回）
6	石村和彦 重任 社外取締役・独立役員	指名委員 報酬委員	100%（8回/8回）
7	島崎憲明 重任 社外取締役・独立役員	監査委員（委員長）	100%（10回/10回）
8	園マリ 重任 社外取締役・独立役員	監査委員	100%（10回/10回）
9	Michael Lim Choo San（マイケル・リム） 重任 社外取締役・独立役員		100%（10回/10回）
10	Laura Simone Unger（ローラ・アンガー） 重任 社外取締役・独立役員		100%（8回/8回）

5

図 14　議事項について記載されたページ（Tag2）の例

Ⅱ 株式に関する事項

1. 当社が発行できる株式の総数　6,000,000,000株

各種類の株式の発行可能種類株式総数は次のとおりです。

種　類	発行可能種類株式総数（株）
普通株式	6,000,000,000
第１種優先株式	200,000,000
第２種優先株式	200,000,000
第３種優先株式	200,000,000
第４種優先株式	200,000,000

2. 発行済株式総数　普通株式　3,493,562,601株

（注）2018年12月17日付で実施した自己株式の消却により、発行済株式の総数は、前期末に比べ、150,000,000株減少しております。

3. 株主数　371,292名

4. 上位10名の株主

株　主　名	持株数および持株比率	
	千株	%
日本マスタートラスト信託銀行株式会社（信託口）	180,391	5.44
日本トラスティ・サービス信託銀行株式会社（信託口）	160,284	4.84
日本トラスティ・サービス信託銀行株式会社（信託口５）	68,101	2.05
NORTHERN TRUST CO.（AVFC） RE SILCHESTER INTERNATIONAL INVESTORS INTERNATIONAL VALUE EQUITY TRUST	64,983	1.96
JP MORGAN CHASE BANK 385151	62,963	1.90
STATE STREET BANK WEST CLIENT-TREATY 505234	54,126	1.63
日本トラスティ・サービス信託銀行株式会社（信託口７）	46,435	1.40
NORTHERN TRUST CO.（AVFC） RE U.S. TAX EXEMPTED PENTION FUNDS	46,059	1.39
日本トラスティ・サービス信託銀行株式会社（信託口１）	45,498	1.37
SSBTC CLIENT OMNIBUS ACCOUNT	42,902	1.29

（注）1. 当社は、2019年３月31日現在、自己株式を182,411千株保有しておりますが、上位10名の株主から除外しております。
2. 持株数は千株未満を切り捨て、持株比率は自己株式を控除して計算しております。

5. 自己株式の取得、処分等および保有の状況

（1）取得した株式
普通株式　100,020,867株
取得価額の総額　51,713,634千円
うち、取締役会決議により買い受けた株式
普通株式　100,000,000株
取得価額の総額　51,702,989千円
買受けを必要とした理由
資本効率の向上および機動的かつ柔軟な資本政策の実施を可能とし、また株式報酬として交付する株式へ充当するため。

（2）処分した株式
普通株式　17,894,180株
処分価額の総額　10,817,001千円

（3）消却した株式
普通株式　150,000,000株
消却価額の総額　89,915,970千円

（4）当事業年度末日における保有株式
普通株式　182,411,802株

図 15　大株主に関するページ（Tag4）の例

Ⅳ 会社役員に関する事項

1．取締役の状況（2019年３月31日現在）

氏　名	地位および担当	重要な兼職状況
古　賀　信　行	取締役会長 指名委員（委員長） 報酬委員（委員長）	野村證券株式会社取締役（＊１） 神奈川開発観光株式会社代表取締役社長
永　井　浩　二	取締役 代表執行役社長 グループCEO	野村證券株式会社取締役会長（＊１）
永　松　昌　一	取締役 代表執行役副社長	野村證券株式会社取締役（＊１、＊２）
木　村　　宏	社外取締役 指名委員 報酬委員	日本たばこ産業株式会社社友 株式会社ＩＨＩ社外取締役
石　村　和　彦	社外取締役 指名委員 報酬委員	AGC株式会社取締役会長 TDK株式会社社外取締役 株式会社ＩＨＩ社外取締役
鳥　崎　憲　明	社外取締役 監査委員（委員長）	株式会社ロジネットジャパン社外取締役 野村證券株式会社取締役（＊１）
園　　マ　リ	社外取締役 監査委員	該当なし
宮　下　尚　人	取締役 監査委員（常勤）	野村アセットマネジメント株式会社取締役（＊１、＊２） 野村信託銀行株式会社取締役（＊１、＊２） 野村ファイナンシャル・プロダクツ・サービシズ株式会社監査役（＊１）
Michael Lim Choo San［マイケル・リム］	社外取締役	Fullerton Healthcare Corporation Limited ノン・エグゼクティブ・チェアマン Nomura Singapore Ltd. ノン・エグゼクティブ・チェアマン（＊１）
Laura Simone Unger［ローラ・アンガー］	社外取締役	CIT Group Inc. インディペンデント・ディレクター Navient Corporation インディペンデント・ディレクター Nomura Securities International, Inc. インディペンデント・ディレクター（＊１）

（注）1. 取締役 木村宏、石村和彦、鳥崎憲明、園マリ、Michael Lim Choo SanおよびLaura Simone Ungerは会社法第２条第15号に定める社外取締役であり、株式会社東京証券取引所の有価証券上場規程第436条の２に定める独立役員であります。
2. 監査委員（委員長）である取締役 鳥崎憲明は米国公認会計士であり、また、監査委員である取締役 園マリは公認会計士であり、それぞれ財務および会計に関する相当程度の知見を有するものであります。
3. 監査委員会による監査がより実効的に行われることを期待し、野村グループの業務に精通した取締役 宮下尚人を常勤の監査委員として選定しております。
4. ＊１の記載がある会社は当社の100%子会社（間接所有を含む）です。
5. ＊２の記載のある役職は、当事業年度の終了後、本事業報告作成日現在までの間に退任したもの、または本事業報告作成日現在において退任が予定されているものです。なお、取締役 木村宏は2019年３月28日付でAGC株式会社を、取締役 鳥崎憲明は2019年３月30日付で株式会社UKCホールディングスを退任いたしました。
6. 社外取締役の兼職先（＊１を除く）と当社との間には、いずれも特別な関係はありません。
7. 当社は、取締役 木村宏、石村和彦、鳥崎憲明、園マリ、宮下尚人、Michael Lim Choo SanおよびLaura Simone Ungerと会社法第423条第１項の損害賠償責任を限定する契約を締結しております。当該契約に基づく責任の限度額は、2,000万円または法令が規定する額のいずれか高い額となります。

33

図 16　会社役員に関するページ（Tag5）の例

株主, 株式, %, 持株, 発行, 総数

大株主に関するページも複数のページにまたがる可能性があるため，上記のルールに当てはまるページが存在しない場合には，新たに 2 ページごとにテキストデータをまとめたものを対象に，上記の文字列がすべて含まれているかどうかで判定した．

3.3.2.3 会社役員に関する開始ページ候補の抽出

会社役員に関する開始ページには，下記に挙げる文字列のいずれかが含まれているかどうかで判定した．

役員に関する事項，役員に関する状況，役員の状況

3.3.2.4 会計監査人に関する開始ページ候補の抽出

会計監査人に関する情報は，重要ページに該当しないが，高い可能性で会社役員に関する情報の次の項目として記載されているため，会社役員に関する記載が終了するページの予測に有効であるため抽出を行った．会計監査人に関する開始ページには，下記に挙げる文字列のいずれかが含まれているかどうかで判定した．

会計監査人に関する事項，会計監査人の状況，　　会計監査人に関する状況

3.3.2.5 各議案ごとの開始ページ候補の抽出

各議案ごとの開始ページ候補抽出のために，各議案に対応するタイトルを抽出した．各議案タイトルの抽出は，全ページを対象に文字列マッチングで以下のいずれかに当てはまるものを抽出した．

第 N 号議案～件，第 N 号議案～について

ただし，N は 1 から 99 までの数字とする．上記の文字列マッチングを適用することで抽出できる例を以下に示す．

第 1 号議案
剰余金の処分の件
第 2 号議案
定款一部変更の件
第 3 号議案
取締役（監査等委員である取締役を除く。）
９名選任の件
第 4 号議案
監査等委員である取締役 3 名選任の件
第 5 号議案
補欠の監査等委員である取締役 1 名選任の件

抽出した各議案のタイトルと議案番号を用いて，議案ごとの開始ページ候補の抽出を行った．すでに抽出した表紙や目次となる候補ページよりも後ろのページを対象に，第 1 号議案から順に，議案タイトルと「第 N 号議案」という文字列が出現するページをその議案の開始ページとして抽出した．

3.3.2.6 Tag0 が付与されるページの抽出

これまでに抽出した情報以外に，「連結計算書類」，「貸借対照表」，「損益計算書」，「メモ」，「新株予約権」に関する事項が記載されている可能性が高いページの抽出も行った．

これらのページは，本研究において重要ページではないと判断した．「連結計算書類」，「貸借対照表」，「損益計算書」については，本研究の対象となる上場企業の場合には決算短信にてあらかじめ公開されるため，株主招集通知で確認する必要はない．また，「新株予約権」の重要な内容については，重要ページと定義した株主招集通知の決議事項として記載されるため，重要性は低い．そのため Tag0 を付与する．しかし，これらのページを抽出しておく目的は，Tag0 の学習データを生成することではなく，ルールによる Tag 付与に誤りが生じている PDF ファイルを学習データから除くことができるようにするために抽出している．具体的には次項 3.3.2.7 節の抽出した情報を用いた Tag の付与において，これらの Tag0 が付与されたページに対して Tag0 以外の Tag が付与される場合，ある 1 ページに対して複数の Tag を付与しようとしていることから，ルールによる Tag の付与に誤りが生じていることになる．そのため，ルールの適用が失敗しているとみなし，学習データから除くことができる．

そのため，学習データの自動生成において，必ず抽出するべきページではないため，抽出のルールはシンプルに，対象のページのテキストデータで，「連結計算書類」，「貸借対照表」，「損益計算書」，「メモ」，「新株予約権」という文字列のみが行に出現しているページを抽出の対象とした．

3.3.2.7 抽出した情報を用いた Tag の付与

これまでに抽出した候補となるページの情報と，議案ごとのタイトル情報を用いて，Tag 付けを自動で行う．

表紙や目次に関する候補ページに対しては，Tag1 を付与する．次に，それよりも後ろのページに対して，決議事項である議案に関しての記載があるページ群に対して Tag2 と Tag3 を付与する．Tag2 は，「第 1 号議案」の開始ページに対して付与し，その次のページ以降には「第 n 号議案[*20]」の開始ページまで Tag3 を付与していく．「第 n 号議案」の記載が終わるページは「第 n 号議案」の開始ページから 5 ページ以内に「以上」という文字列のみが含まれる行が出現するページとし，そのページまでを Tag3 として付与する．「以上」という文字列のみが行に出現するページの例を図 17 に示す．また，「第 n 号議案」の記載が終わるページが特定できない場合は学習データから除く．

5 ページ以内の「5」という数字について，どのように決定したか以下に述べる．収集した株主招集通知をランダムに 20 件ほど抽出し，「第 n 号議案」が記載されているページ数を人手で確認し平均したところ 2.6 ページであり，5 ページのものが 1 件，5 ページを超えるものは 3 件であった．理由として考えられるのは，議案は話し合われる順に記載されていることが多く，大事な内容ほど前半に話し合われる特徴から，前半の大事な内容ほど細かく多くの記載がある傾向にあり，後半の議案は記載が少ない傾向にある．5 ページ以内という 5 をもっと大きい値に設定することで学習データから除くものは少なくなるが，終了ページに「以上」が記載されておらず，他の理由で「以上」が記載されたページまでに Tag3 を付与し，多くの誤った Tag3 が付与されたページを自動生成してしまう可能性があるため，5 ページ以内とした．5 ページ以内とすることで，ランダムに選んだ 20 件のうち 4 件の PDF ファイルはルールベースで終了ページを決めることができず，学習データにすることができないため，学習データの量を犠牲に質を重視するための選択である．5 ページ以内であれば「第 n 号議案」に関する内容が 2 ページで終わり，5 ページ目に「以上」という文字列のみが行に出現するページがあったとしても，Tag3 を $\frac{2}{5}$ は正しく付与することができる．「第 1 号議案」から「第 $n-1$ 号議案」までの Tag2 や Tag3 は

[*20] n は対象とする株主招集通知の中で一番大きい番号．

証券コード 8604
野村ホールディングス

2. 変更の内容

（下線＿は変更部分）

現行定款	変更案
（機関） 第5条　当会社は委員会設置会社として、株主総会および取締役のほか、次の機関を置く。 (1)取締役会 (2)指名委員会、監査委員会および報酬委員会 (3)会計監査人	（機関） 第5条　当会社は、株主総会および取締役のほか、次の機関を置く。 (1)取締役会 (2)指名委員会、監査委員会および報酬委員会 (3)会計監査人
（取締役の責任軽減） 第33条　（略） 2. 当会社は、会社法第427条第1項の規定により、社外取締役（会社法第2条第15号に規定する社外取締役をいう。）との間に、任務を怠ったことによる損害賠償責任を限定する契約を締結することができる。ただし、当該契約に基づく責任の限度額は、2,000万円以上であらかじめ定めた金額または法令が規定する額のいずれか高い額とする。	（取締役の責任軽減） 第33条　（現行どおり） 2. 当会社は、会社法第427条第1項の規定により、取締役（業務執行取締役等である者を除く。）との間に、任務を怠ったことによる損害賠償責任を限定する契約を締結することができる。ただし、当該契約に基づく責任の限度額は、2,000万円以上であらかじめ定めた金額または法令が規定する額のいずれか高い額とする。
（剰余金の配当の基準日） 第44条　当会社の剰余金の配当の基準日は、毎年6月30日、9月30日、12月31日、3月31日とする。 2. 前項のほか、基準日を定めて剰余金の配当をすることができる。 3. （略）	（剰余金の配当の基準日） 第44条　当会社の剰余金の配当の基準日は、毎年9月30日、3月31日とする。 2. （現行どおり） 3. （現行どおり）

以上

株主総会参考書類

20

図 17　「以上」という文字列のみが行に出現するページの例

かなり高い精度で正しく付与することができることを考慮すると，学習データに加える価値が大きいため，平均した 2.6 という数字の約 2 倍である 5 を選択した．もちろん，学習データから「第 n 号議案」の記載が 5 ページを超えるものが含まれなくなるため，このようなデータに対して正しく Tag 付けができるモデルになっているかどうかは，3.4 節の評価と考察にて後述する．

Tag4 に関しては，Tag1 よりも後ろのページであり，Tag2 や Tag3 が付与されていないページにおいて，抽出した大株主に関するページ候補の中で一番若いページに対して付与する．Tag4 の付与されたページの次のページが大株主に関するページ候補である場合は，そのページに対しても Tag4 を付与する．

Tag5 と Tag6 の会社役員に関するページは，ほとんどの株主招集通知で大株主に関するページよりも後ろに記載されるため，Tag4 が付与されたページよりも後ろを対象に，抽出した会社役員に関する開始ページ候補の中で一番若いページに Tag5 を付与する．Tag5 が付与されたページの次のページから，抽出した会計監査人に関する開始ページ候補のページまで Tag6 を付与する．会計監査人に関する開始ページは，会社役員に関するページの記載が終了するページから改ページされて始まることが多いが，改ページされることなくページの半ばから会計監査人に関する記載が始まることもあるため，重要ページが抽出されないことを防ぐために，会計監査人に関する開始ページまでを重要ページの対象として Tag6 を付与する．したがって，会計監査人に関する開始ページ候補が Tag5 を

付与したページ以降に出現しない場合，会社役員に関する記載が終わるページが特定できないため，学習データから除く．

また，Tag1 から Tag6 を付与しようとしたページに対して，3.3.2.6 節により，すでに Tag0 が付与されている場合には，学習データから除く．

ここまでのルールを適用し，Tag1，Tag2，Tag4，Tag5 が少なくとも一つずつ付与することができていれば，Tag が正しく自動で付与できているとみなし，まだ Tag が付与されていないページに対して Tag0 を付与し，学習データに追加する．

これらのルールを適用して，Tag が期待通りに付与できない場合，[61] ではさらに細かいルールを適用し，多くのページを重要ページの候補として抽出することで実用的な再現率を実現している．しかし，本研究で提案する学習データの自動生成においては，学習データの質を高くするために，これらの細かいルールの適用が必要なファイルに関してはルールベースが適用できないファイルとし，学習データから除いた．本研究で提案するフレームワークでは，自動でデータを収集することで多くのデータを収集することが可能であることから，ルールを厳しくした上でも多くの学習データの生成が可能である．

データを自動でクローリングして取得することでデータの量の問題を解決し，元々の収集したデータの分母が大きいためルールベースを厳しくすることが可能であり，人手で作成するには難しい量，かつ，質が高い学習データを自動生成することが可能となった．

3.3.3 Step3: 分類器の学習

まず，分類器にテキストの情報を入力するために，ページごとのテキスト情報を数値情報に変換する必要がある．そのため，分類に有効である可能性が高い素性を選択する方法を述べる．

そして，どのような分類器が本研究で扱うようなタスクにおいて有効であるのかを比較し，検討するために 4 つのモデルを提案する．

3.3.3.1 素性選択

学習データのページごとにテキストを形態素解析し[*21]，入力層の要素となる語（素性）を選択する．ただし，形態素解析を行う際に，本研究での分類において「第 n 号議案」という文字列はルールベースにおいても重要な役割があるため，「第 n 号議案」を一つの語として，分かち書きを行った．

そして，学習データにおける語 w を含むページの数を $df(w)$，学習データにおけるペー

[*21] http://mecab.googlecode.com/

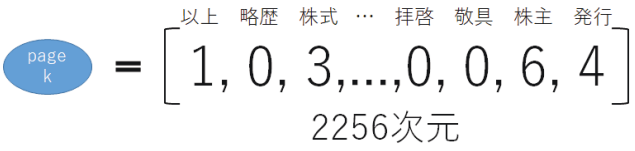


図 18　入力ベクトルの例

ジの総数を N とし，$\frac{df(w)}{N}$ が 0.005 未満の出現数が少ない語を除き，0.005 以上の語を素性として選択した．ただし，「第 n 号議案」という文字列は，出現数に関係なく素性として選択した．本手法では閾値として 0.005 を提案しているが，実験段階において，この閾値を極端に小さくした場合においての予備実験も行った結果，ほとんど同様の結果を得ることができた．この閾値の決め方に関しては 3.4 節の評価と考察にて詳しく後述する．

この素性選択によって，2,256 の素性を選択した．学習データから選択される素性の一部を例示する．

第 1 号議案，株式，略歴，持ち直し，拝啓，敬具，償却，解消，不足，取組，制度，手形，運営

3.3.3.2 モデル 1

入力を株主招集通知のあるページとし，そのページがどの Tag であるかを判定する分類器として，多層パーセプトロン（MLP）をモデルとして選択した．入力は，学習データから抽出された語（素性）を要素，頻度を要素値としたベクトルとする（Bag of Words）．イメージ図を図 18 に示す．

モデルの入力層のノード数を入力ベクトルの次元数（すなわち素性の数）と同じとし，隠れ層は，ノード数 1,000 が 3 層，ノード数 500 が 3 層の計 6 層とした．出力層は Tag の種類である 7 要素とした．モデルの簡略図を図 19 に示す．活性化関数はランプ関数 (ReLU) を使用し，出力層はソフトマックス関数 (softmax) を使用した．本研究のデータは，Tag によって出現頻度が異なる不均衡データであるため，誤分類した場合のペナルティに対して，Tag の出現頻度を用いて重み付けして対処を行った．

3.3.3.3 モデル 2

入力を株主招集通知のあるページとその前後 1 ページとし，そのページがどの Tag であるかを判定する分類器として，多層パーセプトロン（MLP）をモデルとして選択した．

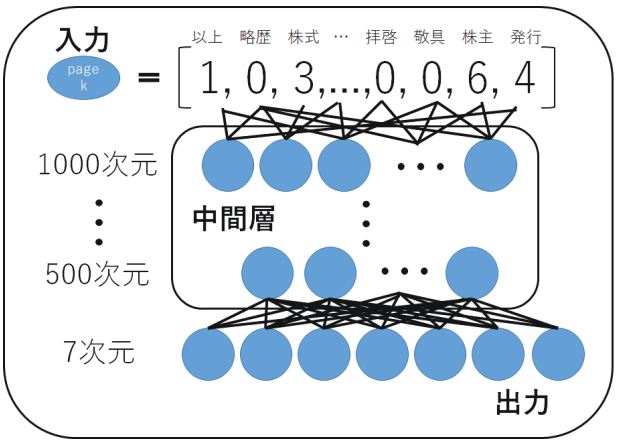


図 19　モデル 1 簡略図

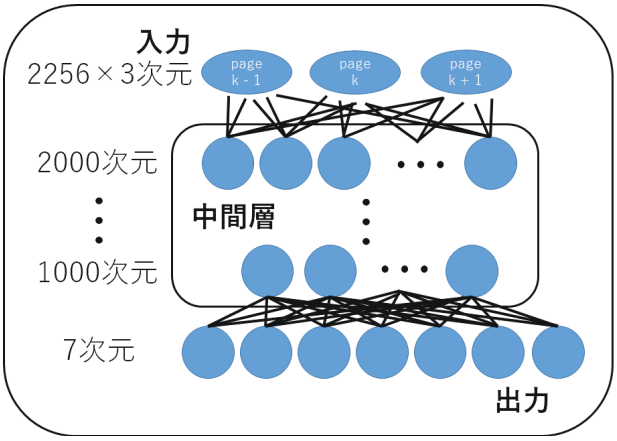


図 20　モデル 2 簡略図

入力は，ページごとに素性を要素，頻度を要素値としたベクトル（Bag of Words）を 3 ページ分を結合したベクトルとする．モデルの入力層のノード数を入力ベクトルの次元数 × 3 とし，隠れ層は，ノード数 2,000 が 3 層，ノード数 1,000 が 3 層の計 6 層とした．出力層は Tag の種類である 7 要素とした．

活性化関数はランプ関数 (ReLU) を使用し，出力層はソフトマックス関数 (softmax) を使用した．本研究のデータは，Tag によって出現頻度が異なる不均衡データであるため，誤分類した場合のペナルティに対して Tag の出現頻度を用いて重み付けして対処を行った．モデルの簡略図を図 20 に示す．

3.3.3.4 モデル 3

入力を株主招集通知ファイルとし，各ページがどの Tag であるかを前後のページの情

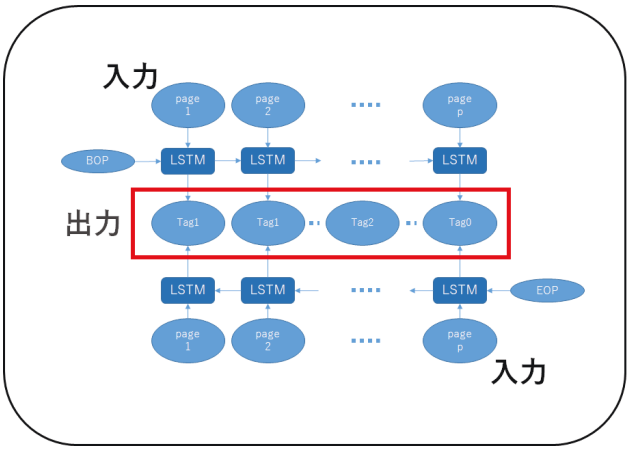


図 21　モデル 3 簡略図

報も用いて判定する分類器として，双方向 LSTM（BiLSTM）をモデルとして選択した．各ページごとに素性を要素とし，頻度を要素値としたベクトルをページ順に入力する．モデルの入力層のノード数を入力ベクトルの次元数（すなわち素性の数）と同じとし，隠れ状態ベクトルの次元数を 1,000 とした．前から順に入力することによって得られる各ページの hidden state vector 1,000 次元と，後ろから順に入力することによって得られる各ページの hidden state vector 1,000 次元を合わせた 2,000 次元のベクトルに線形変換を行うことで 7 次元にし，ソフトマックス関数 (softmax) を適用することで Tag を予測するモデルとする．モデルの簡略図を図 21 に示す．モデル 1 同様，Tag によって出現頻度が異なる不均衡データであるため，誤分類した場合のペナルティに対して，Tag の出現頻度を用いて重み付けして対処を行った．

3.3.3.5 モデル 4

最後のモデルは，モデル 3 の出力層に CRF（条件付き確率場: Conditional Random Field）層を追加した BiLSTM-CRF を選択した．

本研究で扱うような各ページに Tag を付与する問題は，系列ラベリング問題として定式化でき，これらの問題は CRF を用いることでラベル間の依存性を学習することが可能である．例えば，代表的な系列ラベリング問題である固有表現抽出の研究では，まず文章を単語区切りに分割し，固有表現に対して Tag を付与した学習データを用いることで学習器を学習させる．そして，学習させた分類器に単語列を入力し，各単語に対して Tag を付与することで固有表現を抽出することが可能となる．この固有表現抽出の研究における分類器は，BiLSTM-CRF を用いることで良好な結果を得ることが Lample らの研究で示

されている [21]．本研究は，入力が単語区切りになった文章ではないが，系列ラベリング問題としての特徴を持つため，BiLSTM-CRF を分類器として用いることを提案した．

本研究では，とある n ページの入力に対して，各ページの双方向 LSTM の出力を $\mathbf{H} = (\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_n})$ とし，各ページの予測される Tag を $\mathbf{t} = (t_1, t_2, ..., t_n)$ と定義することで，以下の条件付き確率，式 18 を最大化するように，双方向 LSTM のパラメータと Tag の遷移スコア $\mathbf{W}$ を更新することで学習を行う．

$$p(\mathbf{t}|\mathbf{H}) = \frac{exp(score(\mathbf{H}, \mathbf{t}))}{\sum_{t'} exp(score(\mathbf{H}, \mathbf{t}'))} \tag{18}$$

$$score(\mathbf{h}, \mathbf{H}) = \sum_{i=0}^{n} W_{t_i, t_{i+1}} + \sum_{i=1}^{n} P_{i, t_i} \tag{19}$$

ただし，

t_0:　始まりを表す TagB が割り当てられる
t_{n+1}:　終わりを表す TagE が割り当てられる
$\mathbf{W}$:　Tag の遷移スコアを表す正方行列 (9×9)[*22]
$\mathbf{W}_{t_i, t_j}$:　Tag t_i から Tag t_j へ遷移するためのスコア
P:　双方向 LSTM の出力をまとめた行列 $(n \times 7)$
t':　$\mathbf{t}$ が取り得るすべてのパターンの Tag の組み合わせ

式 19 の第一項は CRF が担保する Tag のつながりの正しさをスコア化し，第二項は LSTM による Tag の予測をスコア化している．例えば，学習データから，Tag2 の次は Tag3 の可能性は高いが，Tag6 の可能性は限りなく低いといったことを第一項は学習することが可能である．

実装では，式 18 の対数 $log(p(\mathbf{t}|\mathbf{H}))$ を最大化させるようにパラメータを更新する．学習したモデルを用いて予測を行うときは，スコアを最大化する $\mathbf{t}$ を選択することで，Tag を予測することが可能となる．

3.3.4　Step4: 分類器を用いた重要ページの抽出

Step3 で学習させた分類器に株主招集通知を入力することで，ページごとに Tag を割り当てる．そして，割り当てた Tag が Tag0 以外であれば抽出ページとすることで，重要ページを抽出することが可能となる．

[*22] Tag0，Tag1，Tag2，Tag3，Tag4，Tag5，Tag6，TagB，TagE の合計が 9 個のため．

3.3.4 節で 4 つのモデルを提案したが，3.4 節にて，どのモデルが本研究の目的に適しているのかをテストデータをもとに評価し，考察する．

3.4 評価と考察

提案手法を実装し，評価を行った．まず，Step1（3.3.1 節）では，3,601 の企業 Web サイトにおける IR 情報ページより 554,544 の PDF ファイルを取得し，テキストデータに変換後，このテキストデータを改行ごとに分割し，1 行の文字列に「株主総会」もしくは「株主招集」を含み，さらに，「通知」を含む文字列が出現している PDF ファイルを株主招集通知と判別することで，収集した 554,544 の PDF ファイルから 11,050 の株主招集通知である PDF ファイルを取得した．

次に，Step2（3.3.2 節）では，Step1 で収集した PDF ファイルに対してルールベースを適用し，各ページに Tag を付与することで，学習データを自動生成した．Step1 で収集した 11,050 の株主招集通知である PDF ファイルにルールベースを適用することで，2,795 の PDF ファイルから 125,746 ページの学習データを自動生成することに成功した（Tag1 が 4,063，Tag2 が 2,795，Tag3 が 13,216，Tag4 が 2,837，Tag5 が 2,795，Tag6 が 8,161，Tag0 が 91,879）．

ここで，本手法で自動生成された学習データの評価を行った．自動生成された学習データは，2,795 の PDF ファイル（125,746 ページ）と膨大であるため，ランダムに 10 件の PDF ファイルを選び，人手で確認し Tag 付けを行うことで，自動生成された学習データを定量的に評価した．ランダムに選んだ 10 件の PDF ファイルの合計ページ数は 483 ページであった．自動生成した Tag と正解の Tag の関係がわかるように，クロス集計した結果を表 21 に示し，評価指標として，Tag ごとの適合率・再現率・F1 スコアを表 22 に示す．クロス集計の行が自動で付与した Tag であり，列が人手にて付与した正解の Tag である．

収集した 11,050 の株主招集通知から，2,795 件の学習データを自動生成することができたが，その適合率・再現率は表 22 より良好な結果である．特に本研究では，最終的に重要ページとして Tag 付けしたページを抽出し印刷するため，重要ページが抽出されないようなことが極力起こらないようにするべきである．よって，一部のデータによる評価ではあるが，Tag1 から Tag6 の再現率が 1.000 を満たしているこの自動生成されたデータは，目標を達成するために有効な学習データであるといえる．また，Tag6 の適合率が低いのは，学習データの自動生成方法に起因している．3.3.2.7 節で述べた通り，会社役員に関するページの 2 ページ目以降である Tag6 は，会社役員に関するページの開始ペー

表 21　学習データのクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	342	0	0	0	0	0	0
Tag1	0	17	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	0	0	0	65	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	0
Tag6	9	0	0	0	0	0	20

表 22　Tag ごとの適合率・再現率・F1 スコア

Tag	適合率	再現率	F1 スコア
Tag0	1.000	0.974	0.987
Tag1	1.000	1.000	1.000
Tag2	1.000	1.000	1.000
Tag3	1.000	1.000	1.000
Tag4	1.000	1.000	1.000
Tag5	1.000	1.000	1.000
Tag6	0.690	1.000	0.816

ジ (Tag5) の次のページから，抽出した会計監査人に関する開始ページ候補のページまでを Tag6 として付与する．そのため，会計監査人に関する開始ページは，本来抽出すべき対象のページでないことから誤判定となり，適合率が低くなっている．このようなルールを設定したのは，あるページの途中まで会社役員に関する記述があり (Tag6)，さらに，ページの途中から会計監査人に関する記述が始まっているページを抽出できるようにするためである．表 21 に示すように，正解の Tag は Tag0 であるが自動で付与した Tag が Tag6 になってしまった誤った付与の数が 9 となっているのは，学習データの自動生成の評価に使用した株主招集通知の 1 つにあるページの途中まで会社役員に関する記述があり (Tag6)，さらに，そのページの途中から会計監査人に関する記述が始まっているページが存在したためである．サンプリング数が少ないので詳細な割合ではないが，10 件のうち 1 件の株主招集通知で，このようなページが存在するため，学習データの適合率は多少下がるが，このようなルールを設定している．

ルールベースにより，収集した PDF ファイルの約 25% には高い適合率・再現率で Tag を付与できるが，残りの約 75% の PDF ファイルは，ルールベースだけでは適切な Tag

を付与することはできない．これは，ページの一部の文字が画像として PDF に記載されていることがあることや，PDF からテキストデータの変換で形式が崩れてしまうなど，PDF データであることによる原因と，企業や年度ごとに書き方などが異なる多様性が原因である．しかし，重要ページに出現する単語頻度などの傾向は，多少のテキストデータへの変換ができていないことや，企業や年度ごとに書き方が異なることによって大きくは変わらないため，これらのデータを学習データとし，分類器を学習させることで，ルールから少しでも外れたりする株主招集通知に対しても Tag が付与できると考え，本手法を提案した．

また，今回の学習データ生成における候補ページ抽出の文字列の選択や，Tag 付けのルールは，質と量のトレードオフの関係になっている．例えば，文字列の選択に関して，すべて含まれるルールを採用している表紙や目次に関するページ候補 (Tag1) や大株主に関するページ候補 (Tag4) では，文字列を増やすことで学習データの質を上げることはできるかもしれないが，その分自動生成できるデータ数は少なくなる可能性が高い．また，文字列のいずれかが含まれるルールを採用している会社役員に関する開始ページ候補の抽出 (Tag5) などであれば，文字列を増やすことで学習データの量を増やすことは可能となるが，質は悪くなる可能性がある．そのため，他のデータに適用する場合には，収集できたデータの分母や，実際に学習データ自動生成によってどの程度の量の学習データが生成できたかによって，調整が必要である．

Step3（3.3.3 節）では，Step2 で作成した学習データを用いて，各ページの Tag を付与する分類器を学習させた．自動生成した学習データから素性選択を行い，$2,256$ の素性を選択した．この素性の頻度を入力とし，各モデルを学習させた．

本手法では 3.3.3.1 節で述べた通り，学習データにおける語 w を含むページの数を $df(w)$，学習データにおけるページの総数を N とし，$\frac{df(w)}{N}$ が 0.005 未満の出現数が少ない語を除き，0.005 以上の語を素性として選択している．この 0.005 の値をどのように決めたかをまず述べる．本研究での学習データは，Tag の出現頻度が不均衡なデータであることを考慮し，$DF(\mathrm{Tag}i)$ を Tag$i$ が付与されたページの数とすることで，閾値の上限を以下の式 20 で求める．

$$v = \frac{min(DF(\mathrm{Tag}i))}{N} \tag{20}$$

学習データにおけるページの総数が $N = 125,746$ であり，$min(DF(\mathrm{Tag}i)) = 2,795$ であることから，閾値の上限は $v = 0.022$ となる．この値を閾値の上限とする理由は，例えば，$DF(\mathrm{Tag2}) = 2,795$ であるが，仮に Tag2 にのみ出現する特有の語 w があるとすると，この語は分類に大きく貢献することが考えられる．しかし，閾値を 0.023 にした場合，

Tag2 で必ず出現したとしても他の Tag で出現しないため，その場合 $df(w) = 2,795$ となり，素性として選択することができないため，閾値の上限は式 20 と定めることができる．

もちろん，ある Tag にのみ出現し，その Tag で必ず出現する語などは存在しないため（もし，存在する場合ルールベースのみで分類が可能），あとは出現頻度の低い Tag$i$（本研究では Tag2，Tag4，Tag5) ごとに語の df 値を計算し，Tag$i$ に特徴的な語（例えば，学習データ自動生成に使用した語など）が少なくとも，Tag$i$ が付与されたページの中で，k ページに 1 回以上は出現しているかどうかなどを参考に，閾値を決めるのが有効である．本研究では，Tag2 や Tag5 が付与されたページに対して，少なくとも 4 ページに 1 回くらいのペースで，そのページを表す特徴的な語が出現すると仮定し，$k = 4$ とすることで，閾値を以下の式 21 にて求めた．

$$v = \frac{min(DF(\text{Tag}i))}{N \times \ k} \tag{21}$$

この閾値の小数第四位以下を切り捨てた結果が $v = 0.005$ となったため，閾値をこの値に定めた．本研究では，この閾値で良好な結果を得ることができたが，もし他の文書に対して本手法を適用する時，最初に定めた閾値で学習がうまくいかない場合には，k の値を徐々に大きくして，閾値を下げていく方針をとることが有効であると考える．なぜなら，閾値を下げれば下げるほどノイズとなる素性が増え，単純に次元数も増加することから，前処理や学習にかかる時間が増加する．したがって，学習コストが少ない閾値の設定が高いものから試していくことが有効である．また，$v = 0.005$ より大きい値での実験は，重要な素性が選択できない可能性が高まるため試していないが，閾値を極端に低い値として $v = 7.952 \times \ 10^{-5}$（$df(w)$ が 10 以上）で素性選択を行い，24,220 の素性を用いて同様の実験を行った結果，この後に示す結果とほとんど同様の結果を得ることができた．このことから，本研究におけるデータにおいては，$v = 0.005$ より小さい値であれば，結果に大きな差はなかった．

次に，各モデルの有効性を検討するために，ルールベースを適用した結果，ルールに当てはまらず学習データにならなかった PDF ファイルからランダムに 10 件選び，人手で確認し Tag 付けを行った．人手にてページごとに Tag 付けを行うことにより，10 件の PDF ファイルから 507 ページ分の正解 Tag が付与されたテストデータを作成した．

まず，各モデルの重要ページの Tag がどの程度精確に付与されているかと，モデルごとの特徴を考察するために，提案した各モデルによる Tag 付けの結果と，正解データの Tag の関係をクロス集計した．行が予測した Tag であり，列が人手にて付与した正解の Tag である．モデルごとにクロス集計した結果を，表 23，表 24，表 25，表 26 に示す．

また，定量的評価を行うために，各モデルの評価指標として，多値分類問題でよく使われ

表 23　モデル 1 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	298	2	0	2	0	0	1
Tag1	3	18	0	0	0	0	0
Tag2	0	0	10	4	0	0	0
Tag3	5	0	0	108	0	0	2
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	2
Tag6	9	0	0	0	0	0	23

る評価指標である micro 平均[*23], macro 平均, weighted 平均の適合率･再現率･F1 スコアをクロス集計表から算出した．その結果を表 27 に示す．各指標の算出方法は以下の通りである．micro 平均 F1 スコアは, micro 平均正解率（accuracy）とも等しいため, 各 Tag の正しく付与できているページ数の和を，テストデータの総ページ数で割った値である．例えば，モデル 1 の結果である表 23 であれば，$(298+18+10+108+10+10+23)/508 = 0.941$ である．macro 平均は，まず各 Tag ごとに適合率・再現率を計算し，その結果を相加平均した値である．例えば，モデル 1 の結果である表 23 の適合率であれば，$(0.983+0.857+0.714+0.939+1.000+0.833+0.719)/7 = 0.864$ である．weighted 平均は macro 平均と同様に, まず各 Tag ごとに適合率･再現率を計算し, その結果を正解 Tag の出現頻度で重み付けをした加重平均の値である．正解 Tag の出現頻度で重み付けを行うが，正解 Tag の出現頻度はクロス集計表の各列ごとの値を集計することで求めることが可能であり，モデル 1 の結果である表 23 の Tag0 であれば，$298+3+0+5+0+0+9 = 315$ である．

そして，実用上の評価を行うために，Tag 付けを行ったテストデータに対して，学習した各モデルによる重要ページ抽出の適合率，再現率を求め，既存研究である文献 [61] の手法と提案手法との比較評価を行った．なお，重要ページ抽出の適合率，再現率を評価するために，Tag0 が付与されたものを重要ページでないものとし，Tag0 以外が付与されたページを重要ページとみなして，ファイルごとに評価を行った．具体的な計算方法としては，Tag1 から Tag6 を Tag7 とし，Tag7 と Tag0 の二値分類として適合率・再現率を算出する．

重要ページ抽出の適合率・再現率と，全ページを印刷するのに比べてどのくらいのページを減らすことができたのかの指標となるページ圧縮率を表 28，表 29，表 30 に示す．

[*23] micro 平均の適合率，再現率，F1 スコアは同一の値となるため，F1 スコアのみを結果に示す．

表 24　モデル 2 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	305	4	0	4	0	0	1
Tag1	0	16	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	2	0	0	110	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	1	0	0	0	0	8	0
Tag6	7	0	0	0	0	2	27

表 25　モデル 3 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	304	3	0	1	0	0	0
Tag1	0	17	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	4	0	0	113	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	0
Tag6	7	0	0	0	0	0	28

表 26　モデル 4 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	306	4	0	1	0	0	2
Tag1	0	16	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	2	0	0	113	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	0
Tag6	7	0	0	0	0	0	26

表 28，表 29 からわかるように，文献 [61] で提案しているルールベースが当てはまらない PDF ファイルに対して，提案手法は高い適合率・再現率で，重要ページの抽出に成功した．ルールベースにおいて再現率が低かった証券コード 2899，3371 のような PDF ファイルに対しては再現率が向上し，再現率は高いがその分不要なページを重要ページとして多く抽出していた証券コード 2109，4228，5715，6302，8367 のような PDF ファイルに対しては，余分な抽出をしないことで適合率が向上した．

表 23 から表 26 に関しては，行が予測した Tag であり，列が人手にて付与した正解の

表 27　評価指標による評価結果

	micro	macro			weighted		
手法	F1 スコア	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
モデル 1	0.941	0.864	0.945	0.899	0.946	0.941	0.942
モデル 2	0.959	0.942	0.928	0.931	0.962	0.959	0.959
モデル 3	0.970	0.965	0.972	0.966	0.973	0.970	0.971
モデル 4	0.968	0.964	0.956	0.958	0.970	0.968	0.969

表 28　文献 [61] のルールベースとモデル 1 の評価結果

証券コード	ルールベース			モデル 1		
（全ページ数）	適合率	再現率	圧縮率	適合率	再現率	圧縮率
1605 (71)	32 / 49 = 0.653	32 / 32 = 1.000	69%	31 / 32 = 0.969	31 / 32 = 0.969	45%
2109 (40)	14 / 24 = 0.583	14 / 16 = 0.875	60%	15 / 17 = 0.882	15 / 16 = 0.938	43%
2899 (40)	6 / 7 = 0.857	6 / 15 = 0.400	18%	15 / 16 = 0.938	15 / 15 = 1.000	40%
3371 (44)	4 / 5 = 0.800	4 / 10 = 0.400	11%	10 / 11 = 0.909	10 / 10 = 1.000	25%
4228 (36)	13 / 22 = 0.591	13 / 13 = 1.000	61%	13 / 14 = 0.929	13 / 13 = 1.000	39%
4620 (56)	22 / 26 = 0.846	22 / 23 = 0.957	46%	23 / 24 = 0.958	23 / 23 = 1.000	43%
5715 (52)	20 / 39 = 0.513	20 / 20 = 1.000	75%	20 / 21 = 0.952	20 / 20 = 1.000	40%
6302 (64)	21 / 41 = 0.512	21 / 23 = 0.913	64%	22 / 23 = 0.957	22 / 23 = 0.957	36%
8334 (56)	18 / 18 = 1.000	18 / 21 = 0.857	32%	20 / 22 = 0.909	20 / 21 = 0.952	39%
8367 (48)	19 / 32 = 0.594	19 / 19 = 1.000	67%	18 / 19 = 0.947	18 / 19 = 0.947	40%

表 29　モデル 2 とモデル 3 の評価結果

証券コード	モデル 2			モデル 3		
（全ページ数）	適合率	再現率	圧縮率	適合率	再現率	圧縮率
1605 (71)	31 / 31 = 1.000	31 / 32 = 0.969	44%	32 / 32 = 1.000	32 / 32 = 1.000	45%
2109 (40)	14 / 15 = 0.933	14 / 16 = 0.875	38%	15 / 16 = 0.938	15 / 16 = 0.938	40%
2899 (40)	15 / 16 = 0.938	15 / 15 = 1.000	40%	15 / 16 = 0.938	15 / 15 = 1.000	40%
3371 (44)	10 / 11 = 0.909	10 / 10 = 1.000	25%	10 / 11 = 0.909	10 / 10 = 1.000	25%
4228 (36)	13 / 14 = 0.929	13 / 13 = 1.000	39%	13 / 14 = 0.929	13 / 13 = 1.000	39%
4620 (56)	21 / 22 = 0.955	21 / 23 = 0.913	39%	23 / 24 = 0.958	23 / 23 = 1.000	43%
5715 (52)	20 / 22 = 0.909	20 / 20 = 1.000	42%	20 / 21 = 0.952	20 / 20 = 1.000	40%
6302 (64)	21 / 24 = 0.875	21 / 23 = 0.913	38%	21 / 24 = 0.875	21 / 23 = 0.913	38%
8334 (56)	19 / 19 = 1.000	19 / 21 = 0.905	34%	20 / 22 = 0.909	20 / 21 = 0.952	39%
8367 (48)	19 / 19 = 1.000	19 / 19 = 1.000	40%	18 / 19 = 0.947	18 / 19 = 0.947	40%

Tag であるが，重要ページの抽出漏れを避けるために，Tag1 から Tag6 の再現率が高くなることが求められる．そのためには，予測が Tag0 で，正解が Tag1 から Tag6 のものが少ないことが望ましい．

まず，どのモデルにも共通していたのは，証券コード 8334 の Tag3 のページに対しての予測が Tag0 となってしまったことである．これは，「第 3 号議案 : 取締役に対する譲渡制限付株式及び業績連動型株式報酬制度に係る報酬決定の件」について記載された 4 ページ目となる最後のページであるが，該当議案についての中心的な話題がすでに 3 ページに

表 30　モデル 4 の評価結果

証券コード（全ページ数）	モデル 4 適合率	再現率	圧縮率
1605 (71)	32 / 33 = 0.970	32 / 32 = 1.000	46%
2109 (40)	14 / 15 = 0.933	14 / 16 = 0.875	38%
2899 (40)	15 / 16 = 0.938	15 / 15 = 1.000	40%
3371 (44)	10 / 11 = 0.909	10 / 10 = 1.000	25%
4228 (36)	13 / 14 = 0.929	13 / 13 = 1.000	39%
4620 (56)	23 / 24 = 0.958	23 / 23 = 1.000	43%
5715 (52)	20 / 21 = 0.952	20 / 20 = 1.000	40%
6302 (64)	21 / 22 = 0.955	21 / 23 = 0.913	34%
8334 (56)	19 / 19 = 1.000	19 / 21 = 0.905	34%
8367 (48)	18 / 19 = 0.947	18 / 19 = 0.947	40%

わたって記載されており，最後の 4 ページ目の記載は補足の説明であることや，文字数もページ半分に満たないため，誤った予測結果となってしまっていた．このように，議案について記載された最後のページは，予測が他の重要ページに比べて難しい．このようなページに対しては，考察の後半で説明する追加のルールを適用することで，適合率を多少下げることになるが抽出することが可能である．

また，学習データの自動生成で，一番最後に記載される議案が 5 ページ以内に収まるものしか学習データには存在しない．しかし，テストデータにおける証券コード 2109 や 4228 では，一番最後に記載されている議案が両方ともに 7 ページ続くにもかかわらず，議案について記載されたページである Tag3 を付与することができている．これは，一番最後に記載されている議案以外にも Tag3 が付与されており，それらには 10 ページ以上続くような議案も含まれているため，長く続く議案も学習することができているからと考えられる．

各モデルの考察として，提案モデル 1 は，単一ページを入力としたモデルを提案したが，表 23 からわかる通り，正解が Tag3 のものに対して Tag2 を付与する間違いや，Tag6 のものに対して Tag5 を付与する間違いが起こっている．しかし，これらの Tag 付けの間違いは許容することが可能である（Tag2 と Tag3 はどちらも「決議事項である議案について記載されたページ」の Tag であり，Tag5 と Tag6 はどちらも「会社役員に関するページ」の Tag であり，同じ項目であることは予測できているため）．問題があるのは，正解が Tag6 であるものに対して Tag3 を付与してしまったものや，Tag0 を付与してしまったものがあることである．議案で取り上げる話題には会社役員に関することも含まれていることがあるため，会社役員に関してのページに対して Tag3 が付与されてしまっている．また，会社役員に関するページも複数のページにわたって記載されているため，どうしても最後のページの記載内容が話題の補足であったり，文字数が少なくなってしまうことから，Tag0 が付与されてしまう可能性がある．これは，モデル 1 が単一ページを入

力とする特徴から，前後の情報を利用することができないために起こる．このようなデメリットはあるが，メリットとしては，やはり学習データを自動生成していることによる学習データのバイアスの影響が少ないことである．

モデル 2 では前後のページの情報を活用することで予測精度を向上させようとしているが，モデル 1 とは異なった Tag 付けの誤りが表 24 から見て取れる．モデル 2 では，モデル 1 で見られた，正解が Tag3 であるものに対して Tag2 を付与してしまう間違いや，正解が Tag6 であるものに対して Tag3 や Tag5 を付与する間違いもない．これは，ページの前後情報を入力として受け取っているため，前のページの情報から開始ページでないことが予測できることや，前後のページが会社役員に関するページであるため，議案の Tag ではないことが予測できるためである．このようなメリットはあるが，デメリットも存在する．デメリットは，自動生成された学習データのバイアスをも学習してしまう点である．本研究では，学習データの自動生成をルールベースに当てはまるものに限定しているため，学習データには，大株主情報，会社役員に関する情報の順で記載される場合か，大株主情報，新株予約権の情報，会社役員に関する情報の順で記載されたものしか存在しないが，企業によってはこの記載の順番が異なることがある．現状，異なった順に記載をしている企業はかなり少ないが，年度によって記載の順番が変化することなどを考えると，あまりこのようなページの出現順序までを学習するのは適切でない．本評価でも，証券コード 8334 にて，会社役員に関する情報，大株主情報の順で記載が出現しており，学習データにおけるすべてのデータにおいて大株主に関するページの前は Tag0 であることから，正解が Tag6 であるページの予測が Tag0 になってしまった．

自動生成した学習データの Tag がどのような特徴を持っているかを捉えることができるように，あるページの Tag から次のページの Tag への遷移行列を表 31 に示す．行があるページの Tag を表し，列が次のページの Tag を表す．表の見方としては，例えば，前のページが Tag1 の行に着目すると，Tag1 の次のページにその他の重要ページでないものが来ることは 3,071 回あり，続けて Tag1 が来ることは 820 回あり，議案についての記載が始まる Tag2 が来ることは 172 回あることが読み取れる．

まず，表 31 から見て取れるように，学習データにおいて Tag4 から遷移する可能性があるのは，大株主の情報が続く Tag4，会社役員に関する記述が始まる Tag5，それ以外の重要ページではないページ Tag0 のいずれかである．これは学習データの自動生成において，大株主の情報よりも後ろに会社役員に関するページが出現するというルールを設定しているためである．そのため，Tag6 から遷移する可能性があるのは，会社役員に関する記述が続く Tag6 か，それ以外の重要ページではないページ Tag0 のいずれかであり，Tag6 から大株主情報が記載されている Tag4 へ遷移するデータが学習データに含まれて

表 31　学習データにおける Tag から Tag への遷移行列

		ページ $i+1$						
	Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
ページ i	Tag0	82218	448	2623	0	2795	1021	0
	Tag1	3071	820	172	0	0	0	0
	Tag2	127	0	0	2667	0	0	0
	Tag3	2647	0	0	10549	0	0	0
	Tag4	1021	0	0	0	42	1774	0
	Tag5	7	0	0	0	0	0	2788
	Tag6	2788	0	0	0	0	0	5373

いない．これがモデル 2 のデメリットを引き起こしている原因である．

同様に学習データのバイアスを強く学習してしまったのが，モデル 3 の双方向 LSTM に CRF 層を追加したモデル 4 である．本来，系列ラベリング問題においては，ラベル間の依存性を学習することができる CRF を用いることが有効であることが多い．しかし，表 31 に示した通り，学習データを自動生成していることにより，学習データにバイアスが生じているため，CRF を用いるとモデル 2 のようにバイアスを強く学習してしまうことから，人手にて Tag 付けを行っていない本研究のようなデータにおいては最適なモデルとはならなかった．

それに対して，双方向 LSTM を採用したモデル 3 であるが，LSTM が長期的な依存関係の学習を目的としたモデルである特徴から，対象となるページよりも前全体のページ情報を加味することができ，双方向にすることにより，対象となるページよりも後ろ全体のページ情報を加味して Tag の予測が可能である．この特徴からモデル 2 では自動生成した学習データのバイアスを強く学習してしまっていたが，証券コード 8334 にて，会社役員に関する情報，大株主情報の順で記載が来た場合においても，前からの情報として，Tag5 や Tag6 などの会社役員に関する記載が続く可能性があるといった情報を引き継ぐことができ，後ろ側からの情報として，大株主の情報に関する記載のページの前であるといった情報だけでなく，それまでのページに会社役員に関する記載がまだ出現していないといった情報を引き継いだ状態で対象となるページの予測を行うことが期待でき，実際の Tag の予測も成功した．また，表 27 から明らかなようにモデル 3 の手法は，どの評価指標も一番高い値を示していることから，今回提案した 4 つのモデルの中で一番適したモデルである．

表 32　モデル 3 の各 Tag ごとの適合率・再現率・F1 スコア

Tag	適合率	再現率	F1 スコア
Tag0	0.987	0.965	0.976
Tag1	1.000	0.850	0.919
Tag2	1.000	1.000	1.000
Tag3	0.966	0.991	0.978
Tag4	1.000	1.000	1.000
Tag5	1.000	1.000	1.000
Tag6	0.800	1.000	0.889

表 33　重要ページ抽出を重視したモデル 1 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	275	1	0	1	0	0	1
Tag1	17	19	0	0	0	0	0
Tag2	0	0	10	4	0	0	0
Tag3	14	0	0	109	0	0	2
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	2
Tag6	9	0	0	0	0	0	23

最後に実用に向けて，モデル 3 の各 Tag ごとの適合率・再現率・F1 スコアを表 32 に示す．

表 23 から表 26 のクロス集計表の考察で一度述べた通り，重要ページのみを抽出し，印刷して人が読むことを考慮すると，重要ページの抽出漏れは極力防ぎたいため，Tag1 から Tag6 の再現率が高くなることが求められる．そのために，適合率は多少落ちてしまうが，Tag1，Tag3 と予測された次のページの Tag が Tag0 の場合，前ページと同様の Tag を付与（1 ページ余分に抽出）するルールを最後に適用することで，重要ページの抽出漏れをより防ぐことができ，重要ページ抽出の再現率を 1.000 に近づけることが可能となる．実際にこのルールを適用したモデルごとのクロス集計の結果を表 33 から表 36 に示し，表 27 と同様の評価指標を用いて計算した結果を表 37 に示す．

表 37 からわかるように，このルールを適用した上でも，各評価指標からモデル 3 が予測の優れていることが見て取れる．特にモデル 1，モデル 2，モデル 4 に関して，Tag6 も同様に追加のルールとして，Tag6 の次のページが Tag0 であった場合に Tag6 を 1 ページ多く付与することで再現率を上げることが可能である．しかし，学習データの自動生成

表 34　重要ページ抽出を重視したモデル 2 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	283	3	0	0	0	0	1
Tag1	14	17	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	10	0	0	114	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	1	0	0	0	0	8	0
Tag6	7	0	0	0	0	2	27

表 35　重要ページ抽出を重視したモデル 3 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	281	1	0	0	0	0	0
Tag1	14	19	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	13	0	0	114	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	0
Tag6	7	0	0	0	0	0	28

表 36　重要ページ抽出を重視したモデル 4 のクロス集計結果

	正解 Tag						
Tag	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6
Tag0	283	3	0	0	0	0	2
Tag1	14	17	0	0	0	0	0
Tag2	0	0	10	0	0	0	0
Tag3	11	0	0	114	0	0	0
Tag4	0	0	0	0	10	0	0
Tag5	0	0	0	0	0	10	0
Tag6	7	0	0	0	0	0	26

において再現率を高めるために，会計監査人に関する開始ページにも Tag6 をすでに付与しているため，追加のルールを適用することで，ほとんどのページに対して 2 ページ分を無駄に抽出することになる．そのため，このルールの追加は避けることが望ましい．モデル 3 では，モデル 1，モデル 2，モデル 4 と比較して，正解が Tag6 であるものに対して，すべて Tag6 が付与できていることからも，本研究の目的において有効なモデルである．

このルールを追加した場合における，モデル 3 の各 Tag ごとの適合率・再現率・F1 スコアを表 38 に示す．

表 37　重要ページ抽出を重視した評価指標による評価結果

	micro	macro			weighted		
手法	F1 スコア	適合率	再現率	F1 スコア	適合率	再現率	F1 スコア
モデル 1	0.899	0.808	0.943	0.861	0.921	0.899	0.905
モデル 2	0.925	0.870	0.930	0.893	0.939	0.925	0.929
モデル 3	0.931	0.896	0.977	0.928	0.947	0.931	0.934
モデル 4	0.927	0.890	0.954	0.916	0.940	0.927	0.930

表 38　重要ページ抽出を重視したモデル 3 の各 Tag ごとの適合率・再現率・F1 スコア

Tag	適合率	再現率	F1 スコア
Tag0	0.996	0.892	0.941
Tag1	0.576	0.950	0.717
Tag2	1.000	1.000	1.000
Tag3	0.898	1.000	0.946
Tag4	1.000	1.000	1.000
Tag5	1.000	1.000	1.000
Tag6	0.800	1.000	0.889

もちろん，不要なページに Tag が付与されることから，適合率は多少落ちる．それでも，507 ページの内，282 ページを削減できる結果は実用的である．再現率も Tag2 から Tag6 は 1.000 である．Tag1 のみ 1 ページ抽出しきれていないが，Tag1 の表紙ページと目次ページで確認したい事項は両方に記載されていることが多く，実際に確認したところ，抽出できた Tag1 のページで確認事項は網羅できていたため，数値的評価だけでなく，実用においても十分有効な結果を得ることができた．

3.5　本章のまとめ

本章では，企業の Web サイト上で日々大量に公開されている IR 情報を，企業分析や投資判断のための情報として活用するための研究の一環として，IR 情報として公開されている金融テキストの一つである株主招集通知の重要ページを自動で抽出するフレームワークを提案した．提案したフレームワークは，自動で学習データを大量に生成することが可能である．学習データを自動生成したことにより，学習データにバイアスが生じているこ

とを踏まえて複数のモデルを提案し，評価を行うことでモデルごとにメリット・デメリットの考察を行った．本章で提案した全てのモデルにおいて，文献 [61] の提案した手法では対応できないファイルに対して，良好な結果を示すことができた．特に，本章における学習データにバイアスがあることや，ページに対して Tag 付けを行うといった特徴において，４つの提案したモデルの中で，提案モデル 3 の双方向 LSTM が最も有用なモデルであることを実験による評価と考察において示した．

また，本章で提案した，自動生成した学習データを用いたページ単位での情報抽出の方法は，株主招集通知に限らず，多くのデータに対して，応用が可能な汎用性の高いものである．まず，学習データがあるのであれば，前章でも述べた通り，良好な結果が得られることを示した．したがって，ページ単位での分類に関しても，良好な結果を得ることが可能である．ただし，学習データがない場合は，本章で示したように学習データの生成を行う必要があるが，この学習データ作成のルールは，本章の目的に特化したものであるため，扱うテキストデータと目的によって，独自に作る必要がある．一般的に，本章の提案した方法を適用するにはいくつかの条件を満たしている必要がある．まず，ある程度はルールベースで学習データが作成できるような，特徴があるページを分類の対象としている必要がある．シンプルなルールで，学習データが作成できないようなページは学習データを作ることもできないため，本手法を適用することは難しい．また，学習データを作るためのテキストデータの分母が非常に大きい必要がある．シンプルなルールを設定し，精度の高いデータを得るためには，学習データは厳しい条件を満たす必要があるが，もともとの分母が大きくなければ，十分な量の学習データを生成することができないため，大量のデータを集めることが不可能な場合には，本手法の適用は難しい．

学習データの量と質は，分類器の性能に大きく影響を与えるため，本章で示したように，そのルールベースによって生成された学習データの量の確認と，一部をランダムサンプリングして，学習データとして十分な精度のデータになっているかの確認が必要であり，これらを満たしているのであれば，本論文の結果と同等の結果が期待できる．

4　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

4.1　研究概要

2 章および 3 章の株主招集通知を対象とした研究では，PDF データを対象としていたため，ページ単位での有益情報の抽出方法を検討した．続く本章では，有価証券報告書から文単位での投資判断に有益な情報の抽出方法について述べる．有価証券報告書は，XBRL 形式でデータが公開されており，タグ情報を用いることで章ごとのテキスト情報を抽出することが可能である．しかし，章に絞ってテキストを抽出したとしても，目を通すには情報量が膨大である．そこで本章では，従来手法を応用して学習データを自動生成したうえで，有価証券報告書の文書特徴を利用することにより，事業セグメントが付与された業績要因文と業績結果文を抽出する方法についての述べる．

投資家が投資活動を行うにあたり，上場企業の業績情報の収集は必要不可欠である．また，業績情報の中でも特に業績要因が投資判断において重要である [43, 44, 45, 11]．なぜなら，業績回復の要因が，その企業の主力事業が好調であることであったならば株価への影響は大きいが，株式売却益の計上などの特別利益の計上などが要因であるならば株価への影響は軽微であるからである．本論文で扱う業績要因とは，業績結果（売上，利益）に影響を与えた事業の推移に関する情報と定義した．業績に影響を与えたことが確定していなければ業績要因とはみなせないので，例えば「顧客基盤の拡大に努めました」や「効率化の支援を進めました」のような，業績に影響を与えたかどうかが不確定な文は，業績要因文ではないとした．

この定義は，多少曖昧な部分があるため，業績要因文の例と業績要因文ではない例をいくつか挙げる．

業績要因文の例
・当連結会計年度は、物流センターが順調に推移し、入庫量・在庫量ともに増加したことにより、全体として保管料売上、荷役料売上が増加しました。
・安定した授業料収入に加え、呉服・和装小物等の販売も概ね堅調に推移いたしました。
・雑貨においては、キャラクターグッズなどが全般的に低調な動きとなりましたが、シーズン商品については好調に推移し、生活雑貨関連も安定した推移となりました。

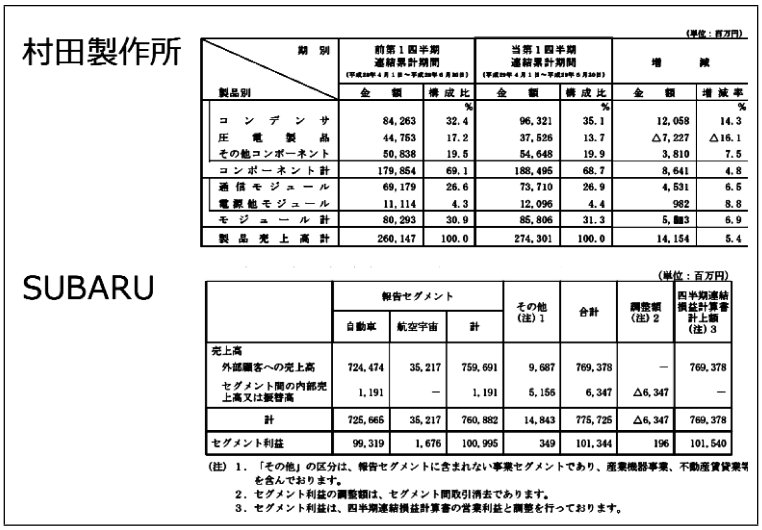
業績要因文ではない例
・販売面では、消費税増税前の駆け込み需要を取込むための積極的なポイント販促に加え、エブリデー・セイム・ロープライスをお客さまに認知していただくため、店頭に設置する「サツドラマンスリー」の内容を積極的に見直し、毎日安心してお買い物いただけることによる、新規顧客の増加とリピーターの確保に努めました。
・多様化するお客様のニーズを的確に捉え、グループ総合力を活かした総合的なソリューションサービスの提案を行い、新規案件とも連動して開発を強化し、事業拡大に努めてまいりました。

企業の業績情報の収集の情報源としては，企業の決算短信や後述の有価証券報告書が有用である．そのため，証券アナリストやファンドマネージャーは，多くの決算短信や有価証券報告書を読む必要があり，特に決算短信は決算期にもなると 1 日で数百と公開され，それらを素早く読み，内容を把握する必要がある．決算短信の全ての文を精読することは不可能であるため，重要な部分を読むことになるが，業績要因は決算短信の中で最も重要な内容の 1 つである．なぜなら，投資対象としている企業のどの事業セグメントの業績が伸びていて，どの事業セグメントの業績が悪化しているのか，その要因は何かという情報を得るには，どうしてもその企業の事業セグメントごとの業績要因を読み取る必要があるからである．そして，たとえ企業の全体の業績は伸びていなくても，今後，その企業の主力事業となることが期待できる事業セグメントの業績が伸びていることが確認でき，その要因も妥当なものであれば，その企業を投資対象とすることを検討できる．

例えば，株式会社 SUBARU では事業セグメントが「自動車」と「航空宇宙」の 2 つある．それぞれのセグメントの 2018 年 3 月期利益は，「自動車セグメント」で 361,454 百万円，「航空宇宙セグメント」で 12,259 百万円となっており，大きな差がある．そのため，SUBARU にとって「自動車セグメント」の業績要因は「航空宇宙セグメント」の業績要因よりも重要であると考えられる．しかし，事業セグメントごとの業績要因と，それに対応する業績結果（売上，利益）の情報を決算短信から読み取るには決算短信の精読が必要であり，現在は人手にてこれらの情報を収集しているのが現状である．

これまで，決算短信から業績要因が記述された文（以降，業績要因文と定義）の自動抽出は行われてきた [11, 50]．酒井らの研究 [11] では，「が好調」や「が不振」といった業績要因文を抽出するのに有効な手がかり表現をブートストラップ的に自動獲得し，さらに，企業ごとの重要なキーワードと組み合わせることで，業績要因文を抽出している．また，酒井らの研究 [50] では，手がかり表現に文節を追加した拡張手がかり表現を使用して業

績要因文抽出のための学習データを自動的に獲得し，自動的に獲得された大量の学習データを使用して，深層学習にて，業績要因文を抽出している．しかし，これらの研究では，抽出された業績要因がどの事業セグメントに属しているのかは判定できていない．決算短信は，企業が決算発表時に提出する決算速報であるため，タイムリーな業績情報が得られる反面，企業ごとに文書フォーマットが異なっているため，その企業の事業セグメント名の抽出や，業績要因文がどの事業セグメントに属しているかを自動的に判定することは困難である．また，事業セグメントごとの業績結果は，図 22 に示すように企業ごとに表のフォーマットが異なるため，抽出が困難である．そのため，決算短信から抽出された業績要因文が属する事業セグメントを推定し，さらに，その事業セグメントの業績結果を自動的に得るのは，極めて困難であった．



村田製作所

（単位：百万円）

製品別 ＼ 期別	前第１四半期連結累計期間 金額	構成比	当第１四半期連結累計期間 金額	構成比	増減 金額	増減率
		%		%		%
コンデンサ	84,263	32.4	96,321	35.1	12,058	14.3
圧電製品	44,753	17.2	37,526	13.7	△7,227	△16.1
その他コンポーネント	50,838	19.5	54,648	19.9	3,810	7.5
コンポーネント計	179,854	69.1	188,495	68.7	8,641	4.8
通信モジュール	69,179	26.6	73,710	26.9	4,531	6.5
電源他モジュール	11,114	4.3	12,096	4.4	982	8.8
モジュール計	80,293	30.9	85,806	31.3	5,513	6.9
製品売上高計	260,147	100.0	274,301	100.0	14,154	5.4

SUBARU

（単位：百万円）

	報告セグメント 自動車	航空宇宙	計	その他（注）1	合計	調整額（注）2	四半期連結損益計算書計上額（注）3
売上高							
外部顧客への売上高	724,474	35,217	759,691	9,687	769,378	－	769,378
セグメント間の内部売上高又は振替高	1,191	－	1,191	5,156	6,347	△6,347	－
計	725,665	35,217	760,882	14,843	775,725	△6,347	769,378
セグメント利益	99,319	1,676	100,995	349	101,344	196	101,540

（注）１．「その他」の区分は、報告セグメントに含まれない事業セグメントであり、産業機器事業、不動産賃貸業等を含んでおります。
２．セグメント利益の調整額は、セグメント間取引消去であります。
３．セグメント利益は、四半期連結損益計算書の営業利益と調整を行っております。

図 22　株式会社村田製作所と株式会社 SUBARU の事業セグメントごとの業績情報のフォーマットが異なっている例

そこで本章では，有価証券報告書からその企業の事業セグメント名を抽出し，さらに，事業セグメントごとの業績要因文と業績結果の抽出を行うことを目的とする．

ここで，本論文での業績結果についての定義を述べる．企業の業績を測る指標としては，売上高，営業利益，経常利益，当期純利益の指標を用いることが多く，決算短信や有価証券報告書の第一部「企業情報」の「主要な経営指標等の推移」にも，上記の指標が記載されている企業が大半である．そのため，業績結果とは売上，営業利益，経常利益，当期純利益についてのいずれかの情報と定義することにした．

企業全体の業績結果（売上，営業利益，経常利益，当期純利益）は，決算短信や有価証券報告書に表形式にて記載があり，その表のフォーマットも一定であるため抽出は容易で

あるが，事業セグメントごとの業績（売上，営業利益，経常利益，当期純利益）の記載は企業ごとに異なり，どのような指標を採用するか，その表現，表のフォーマットまで，企業ごとに異なっているため，抽出は困難である．そのため本研究では，企業ごとに異なるフォーマットの表から業績結果の抽出をするのではなく，テキスト情報から事業セグメントごとの業績結果について言及した文（以降，業績結果文と定義）を抽出し，業績結果文が属する事業セグメント名を推定することで対応した．

例えば，株式会社三菱ケミカルホールディングスの有価証券報告書から，業績要因文として，「機能商品分野及び素材分野においては、原料価格が下落する中、石油化学関連製品の市況が堅調に推移し、また、ヘルスケア分野においては、薬剤費削減策の影響等があるものの、ロイヤルティー収入の増加等もあり、好調に推移しました。」，業績結果文として，「当セグメントの売上高は 5,540 億円となり、営業利益は 1,034 億円となりました。」のような文を抽出し，「ヘルスケア」という事業セグメントを付与する．

本章の 4.2 節では，有価証券報告書について述べる．4.3 節では，本手法の概要について述べる．4.4 節では，有価証券報告書からの業績要因文抽出について述べる．4.5 節では，有価証券報告書からの業績結果文の抽出について述べ，4.6 節では，事業セグメント名の抽出について述べる．4.7 節では，4.4 節と 4.5 節で抽出した文に対して，4.6 節で抽出した事業セグメント名を付与する方法について述べる．4.8 節では，実装について述べる．4.9 節では，手法の評価について述べ，4.10 節では，評価結果を考察する．

4.2 有価証券報告書について

有価証券報告書とは，金融商品取引法で規定されている，事業年度ごとに作成する企業内容の外部への開示資料であり，審査を通す必要があることから，公表までに時間がかかる反面，決算短信と比較すると文書フォーマットが定まっている．有価証券報告書の記載内容は，企業の概況，事業の状況，設備の状況などであるが，決算短信に比べ，有価証券報告書は記載されている内容が多く，業績要因文以外の内容も多く記載されている点が大きく異なっている．なお，有価証券報告書は本文や表が図 26 のような XBRL 形式のデータで提供されている．図 26 に示すとおり，XBRL は非常に複雑なタグで構成されているが，このタグを利用することで，テキスト情報を部分的に抽出することが可能となる．決算短信は，現在のところ，売上や営業利益などの指標は XBRL 形式で提供されているが，本文は提供されておらず，決算短信 PDF からテキスト抽出を行うことしかできない．したがって，多くの情報が掲載されているものの，その中から必要な部分に限定してテキスト情報が抽出できることが，XBRL 形式で本文が提供されている有価証券報告書の

特徴である．また有価証券報告書のデータは，金融庁が運営している web サイトである EDINET[*24]に開示されているため，全ての企業の有価証券報告書の収集が容易である．

本章の研究対象である有価証券報告書を分析対象とする研究はいくつか存在する [40]．しかし，有価証券報告書からの業績要因文，業績結果文の抽出に関する研究は行われておらず，有価証券報告書から事業セグメント名を獲得し，さらに，抽出された業績要因文，業績結果文の属する事業セグメント名を推定する研究も行われていない．

4.3 本手法の概要

事業セグメントごとの業績要因文と業績結果文を有価証券報告書から抽出するには，事業セグメント名，業績要因文，業績結果文の抽出が必要である．本研究で使用する有価証券報告書については 4.8 節の実装で述べる．

本手法の概要を以下に示す．

Step 1: 有価証券報告書から自動生成された学習データを使用して深層学習のモデルを生成し，深層学習によって業績要因文を抽出する．

Step 2: 業績結果文をルールベースにて抽出する．

Step 3: 事業セグメント名候補を，有価証券報告書に記載されている「従業員の状況」からルールベースにて抽出する．

Step 4: 事業セグメント名候補を用いて，抽出した業績要因文，業績結果文に事業セグメントを付与する．

上記の処理の概要を図 23 に示す．業績要因文として，「ロイヤリティー収入の増加などもあり，好調に推移しました。」のような文を抽出し，抽出したセグメント名の 1 つである「ヘルスケア」を付与することで，事業セグメントごとの業績要因文の抽出が可能となる．

提案手法は，様々なテキスト情報に適用可能な汎用的な部分と，有価証券報告書に特化した部分に大きく分けることができる．

まず，汎用的な部分に関して述べる．本研究では，有価証券報告書からの業績要因文の抽出を深層学習で行うために，学習データを自動生成している．この部分は，酒井らが決算短信から業績要因文を抽出した手法 [50] を有価証券報告書に適用することで実施しているが，決算短信と有価証券報告書では業績要因文でも表現が異なることがあり，決算短

[*24] http://disclosure.edinet-fsa.go.jp/

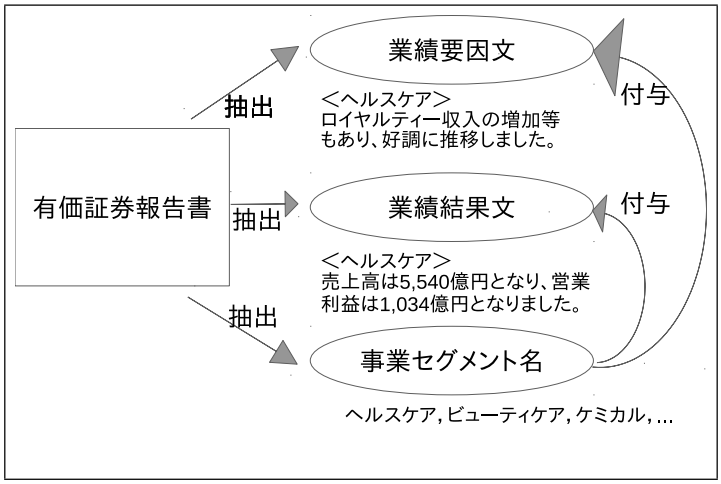


図 23　本手法の概要

信から自動生成した学習データをそのまま使用することは望ましくない．決算短信は主に投資家向けの文書であるため内容の速報性を重視しているが，有価証券報告書は金融庁への提出書類であるため，内容の正確性を重視している．そのため，業績要因文でも表現に違いが生じ，酒井らの手法をそのまま決算短信に適用して学習データを作成し，その学習データで学習したモデルを有価証券報告書に適用することができなかった．

本章では，文献 [11] および [50] を実装し，その手法を有価証券報告書に適用して学習データを生成しているため，決算短信における手法を，有価証券報告書からの業績要因抽出のために拡張したものである．この場合，手法のほぼ全てが共通であり，学習データ生成のための手がかり表現を獲得するための初期手がかり表現も，文献 [11] と同様の表現を用いた．この結果から，決算短信，有価証券報告書以外のテキストから業績要因を抽出する場合においては，酒井らの手法 [11, 50] を変更することなく，抽出が可能であると思われる．

なお，学習データ生成のための手がかり表現を獲得する手法 [11] は，既に決算短信以外にも様々なテキスト情報に適用されており，特許からの技術課題情報，新聞記事からの事故原因表現，アナリストレポートからの予想根拠情報などがある [62, 43, 12, 63, 64]．手法としての変更箇所は初期手がかり表現であり，どの手法も初期手がかり表現から獲得された手がかり表現と重要キーワードの組み合わせで，抽出対象の情報が含まれる文の抽出を行っている．したがって，業績情報以外の特別な情報抽出に拡張する場合には，初期手がかり表現を変更することで，目的の情報を抽出するための手がかり表現を獲得することができ，さらに，酒井らの手法 [50] を適用して拡張手がかり表現を生成した後に学習データを自動生成すれば，深層学習による判定にも適用が可能である．

業績要因文抽出の手法は，他テキストにおける（業績要因に限らず）情報抽出に適用できると考えられるが，本手法における，有価証券報告書からの事業セグメント名の抽出や，業績要因文に対する事業セグメント付与にあたる部分は，有価証券報告書に特化した手法となっている．

事業セグメント名の抽出は，有価証券報告書の「従業員の状況」から獲得しているが，この「従業員の状況」は決算短信にはない情報であり，本手法でのみ取得可能な情報である．また，業績要因文に対する事業セグメント付与は，有価証券報告書の「事業の状況」において，事業セグメント名と業績要因文，業績結果文の位置情報を利用して付与しており，有価証券報告書に特化した手法となっている．その結果，今までの決算短信を使用した既存研究では抽出できなかった情報が，本研究では抽出できるようになった．

4.4 有価証券報告書からの業績要因文の抽出

4.4.1 業績要因文の抽出手法の概要

酒井らの手法 [11] を適用して，有価証券報告書から手がかり表現，企業キーワードを抽出し，さらに酒井らの手法 [50] を適用して業績要因文抽出のための学習データを有価証券報告書から自動的に生成する．自動生成された学習データを用いて深層学習にて有価証券報告書から業績要因文を抽出する．業績要因文抽出の概要を以下に示す．

Step 1: 酒井らの手法 [11] を用いて，有価証券報告書から手がかり表現，企業キーワードを抽出する．

Step 2: 酒井らの手法 [50] を用いて，酒井らの手法によって抽出された手がかり表現の“拡張手がかり表現”を獲得する．

Step 3: 拡張手がかり表現を含み，かつ，スコア $W(n, S(t))$ が高い企業キーワードを含む業績要因文を正例，手がかり表現，企業キーワードをともに含まない文を負例として学習データを自動生成する．

Step 4: 自動生成された学習データを使用し，深層学習にて有価証券報告書から業績要因文を抽出する．

4.4.2 手がかり表現，企業キーワードの自動獲得

酒井らの手法 [11] は「が好調でした」のような手がかり表現を，企業 web サイトから収集した決算短信からブートストラップ的に自動的に獲得する．さらに，その企業にとって重要なキーワードを抽出する．例えば半導体製造装置のメーカーであれば，「半導体製

造装置」が重要なキーワードとなる．以降，その企業にとって重要なキーワードを「企業キーワード」と定義する．本研究では，この手法を有価証券報告書に対して適用した．

手がかり表現は以下の手法で獲得される．

Step 1: 少数の手がかり表現（具体的には，「が好調」，「が不振」の 2 表現を用いる）を人手で与え，それに係る節を取得する．

Step 2: 取得した節の集合から，その中で共通して頻繁に出現する表現（「売り上げ」など）を共通頻出表現として抽出する．

Step 3: 共通頻出表現が係る節を取得し，その中から新たな手がかり表現を抽出する．

Step 4: 獲得した手がかり表現から，それに係る節を取得する．

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなる，もしくは，予め定めた回数まで繰り返す（図 24 を参照）．

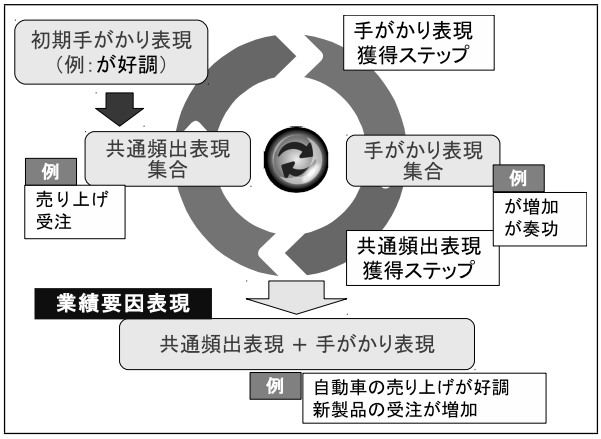


図 24　共通頻出表現・手がかり表現自動獲得手法の概要

Step 2 において，様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 22 で求め，その値が，ある閾値以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \tag{22}$$

ここで，有価証券報告書の集合において，

$S(e)$: 共通頻出表現 e が係る手がかり表現の集合．

$P(e,s)$: 共通頻出表現 e が手がかり表現 s に係る確率．

同様に，Step 3 において，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを求め，その値が，ある閾値以上の手がかり表現を選別する．

以上の手がかり表現，共通頻出表現の選別処理を行うことで，例えば以下のような適切な手がかり表現を獲得する．

> 占めており，回復し，増加し，続いており，上回る，推移しました。，推移するとともに，好転した，伸長した，堅調であった，低迷しました。，持ち直し

有価証券報告書から抽出された手がかり表現は 304 個であり，酒井らの決算短信から抽出できた手がかり表現 162 個よりも多く抽出された [11].

有価証券報告書からの企業キーワードの抽出は，企業 t の有価証券報告書における名詞 n に対して，以下の式 23 で重み $W(n, S(t))$ を計算することで行う．

$$W(n, S(t)) = (0.5 + 0.5 \times \frac{tf(n, S(t))}{\max tf(n, S(t))}) \times H(n, S(t)) \times \log_2 \frac{N}{df(n)} \qquad (23)$$

ここで，

$S(t)$: ある企業 t の有価証券報告書の集合．
$tf(n, S(t))$: $S(t)$ において，名詞 n が出現する頻度．
$H(n, S(t))$: $S(t)$ の各有価証券報告書である d に名詞 n が出現する確率 $P(n, d)$ に基づくエントロピー．以下の式 24 によって求める．

$$H(n, S(t)) = -\sum_{d \in S(t)} P(n, d) \log_2 P(n, d) \qquad (24)$$

$df(n)$: 名詞 n を含む有価証券報告書をもつ企業の数．
N: 有価証券報告書を収集した企業の数．

$W(n, S(t))$ は，情報検索で一般的な $tf \cdot idf$ 値を 1 つの企業の有価証券報告書集合を 1 つの文書とみなして求め，さらに，その企業の有価証券報告書集合においてまんべんなく出現している場合に高い値をとる尺度を組み合わせたものである．表 39 に，上記の手法によって，企業ごとの有価証券報告書から抽出された企業キーワードをいくつか示す．有価証券報告書から抽出された企業キーワードは，決算短信から抽出されたものとほぼ同様の結果が得られた．企業キーワードとは，その企業にとって重要なキーワードであるため，この結果は妥当である．

表 39　有価証券報告書から抽出された企業キーワードの例

企業名称	企業キーワード
三菱電機	産業メカトロニクス， 家庭電器
大日本印刷	エレクトロニクス， 印刷事業
カゴメ	野菜飲料， 野菜生活， 果美食品
エーザイ	医薬品， アリセプト， 医薬品事業
三菱商事	資源関連， 金融事業， ＬＮＧ

精度の高い業績要因文の集合を生成するために，手がかり表現にいくつかの文節を追加することで，“拡張手がかり表現”と定義する文節列（例えば「受注が好調でした」）を獲得する．これにより，例えば手がかり表現「好調でした」が「受注が好調でした」や「極めて好調でした」のような，より精度の高い手がかり表現へ拡張される．そして，例えば拡張手がかり表現「極めて好調でした」を含む業績要因文のみを抽出する．さらに，業績要因文に含まれる企業キーワードを使用して業績要因文にスコアを付与し，スコアの高い業績要因文のみを選別することで，学習データを自動生成する．

4.4.3　拡張手がかり表現の獲得

酒井らの手法 [50] により，有価証券報告書から抽出した手がかり表現に文節列を追加して，拡張手がかり表現を獲得する手法について簡単に述べる．具体的には，手がかり表現 c に係る文節列 p に対して以下の式 25 でスコアを求め，このスコアが，ある閾値を上回る文節列を抽出する．

$$Score(p,c) = -f(p,c)\sqrt{fp(p)}H(p)\log_2 P(p,c) \tag{25}$$

$$P(p,c) = \frac{f(p,c)}{N(c)} \tag{26}$$

ただし，有価証券報告書から取得した業績要因文の集合において，

$P(p,c)$:　手がかり表現 c から取得される文節列 p の出現確率．
$f(p,c)$:　手がかり表現 c から取得される文節列 p の取得回数．
$N(c)$:　手がかり表現 c から取得される文節列の総数．
$fp(p)$:　文節列 p に含まれる文節の数．
$H(p)$:　文節列 p がある企業の業績要因文に出現する確率に基づくエントロピー（後述）．

$H(p)$ は文節列 p がある企業の業績要因文に出現する確率に基づくエントロピーであ

り，$H(p)$ が高い文節列は多くの企業の業績要因文に出現している文節列であることが分かる．また，1 つの企業にのみ多く出現する文節列の $H(p)$ は 0 になるため，そのような文節列を除去できる．$H(p)$ は以下の式 27 で求める．

$$H(p) = -\sum_{s \in S(p)} P(p,s) \log_2 P(p,s) \tag{27}$$

ここで，$S(p)$ は文節列 p を含む業績要因文をもつ企業の集合，$P(p,s)$ は文節列 p が企業 s の業績要因文に出現する確率を表す．

$$P(e,s) = \frac{f(e,s)}{\sum_{s' \in S(e)} f(e,s')} \tag{28}$$

手がかり表現 c における $Score(p,c)$ の平均を求め，平均値より大きいスコアが付与された文節列 p を手がかり表現 c に追加し，それを拡張手がかり表現として獲得する．これにより，304 個の手がかり表現から 8,517 個の拡張手がかり表現を獲得した．表 40 に有価証券報告書から取得した手がかり表現「改善し」から取得された文節列の中で $Score(p,c)$ が高い文節列をいくつか示す．

表 40　「改善し」から取得された文節列

文節列	$Score(p,c)$	拡張手がかり表現
大幅に	4565.440	大幅に改善し
大きく	3068.250	大きく改善し
利益率が	1904.472	利益率が改善し
収益が	1618.658	収益が改善し
収益性が	1243.073	収益性が改善し

4.4.4　深層学習による業績要因文の抽出

拡張手がかり表現を含み，かつ，スコア $W(n,S(t))$ が高い企業キーワードを含む業績要因文を正例，手がかり表現，企業キーワードをともに含まない文を負例として学習データを自動生成する．上記の処理を行うことにより，4,525 企業の有価証券報告書から正例，負例，合計で 67,415 文の学習データを生成した．

自動生成された学習データを使用し，深層学習により業績要因文を抽出する．深層学習のモデルとしては，多層パーセプトロンを採用した．まず，入力層の要素となる語（素性）

を選択する．具体的には，自動生成された学習データにおいて正例の業績要因文に含まれる内容語（名詞，動詞，形容詞）に対して，以下の式 29 にて重みを計算する．

$$W_p(t, S_p) = TF(t, S_p) \times H(t, S_p) \tag{29}$$

ただし，

S_p:　学習データにおいて正例に属する業績要因文の集合．
$TF(t, S_p)$:　文集合 S_p において，語 t が出現する頻度．
$H(t, S_p)$:　文集合 S_p に語 t が出現する総数に対して，業績要因文 s にどのくらい語 t が出現しているかを表す確率に基づくエントロピー．

$H(t, S_p)$ が高い語ほど，正例の文集合に均一に分布している語であることが分かる．この指標を導入した理由は，正例の文集合中でも多くの文に分散して出現している語のほうが，少数の文に出現している語と比較してよりその文集合の特徴を表し，素性としても有効であるという仮定に基づく．$H(t, S_p)$ は次の式 30 で求める．

$$H(t, S_p) = -\sum_{s \in S_p} P(t, s) \log_2 P(t, s) \tag{30}$$

$$P(t, s) = \frac{tf(t, s)}{\sum_{s \in S_p} tf(t, s)} \tag{31}$$

ここで，$P(t, s)$ は業績要因文の集合に語 t が出現する総数に対して，業績要因文 s にどのくらい語 t が出現しているかを表す確率であり，$tf(t, s)$ は文 s において語 t が出現する頻度を表す．

次に，負例の文に含まれる内容語（名詞，動詞，形容詞）に対しても，同様に重みを計算する．

$$W_n(t, S_n) = TF(t, S_n) \times H(t, S_n) \tag{32}$$

ただし，S_n は学習データにおいて負例に属する文の集合である．

ここで，ある語 t の正例における重み $W_p(t, S_p)$ が負例における重み $W_n(t, S_n)$ の 2 倍より大きければ，その語 t を素性として選択する．もしくは，語 t の負例における重み $W_n(t, S_n)$ が正例における重み $W_p(t, S_p)$ の 2 倍より大きければ，その語 t を素性として選択する．すなわち，以下の条件のどちらかが成り立つ語 t を素性として選択する．

$$W_p(t, S_p) > 2W_n(t, S_n)$$
$$or$$
$$W_n(t, S_n) > 2W_p(t, S_p)$$

上記の条件を課すことで，正例，負例における特徴的な語のみを素性として選択し，正例，負例，ともによく出現するような一般的な語を素性から除去する．上記の手法により，67,415 文の学習データから 5,083 語が素性として選択された．以下に選択された素性の一部を示す．

・正例：事業，増加，推移，販売，減少，需要，堅調，好調，製品，増収，拡大，受注，回復，改善，向上
・負例：株主，支出，帰属，配当，取締役，現金，当たり，税金

選択された素性 5,083 語の正例と負例の内訳は，正例が 5,039 語であり，負例が 44 語である．負例における文章には，業績要因文を除く様々な文章があるため，特徴的な語というものは少なく，このような結果が得られる．

深層学習のモデルは多層パーセプトロンを採用し，入力は，67,415 文の学習データから抽出された 5,083 語を要素，語 t における $W_p(t, S_p)$，もしくは，$W_n(t, S_n)$ の大きいほうを要素値としたベクトルとする．中間層のノード数や活性化関数については，9 章「実装」において記述する．

学習されたモデルにより，有価証券報告書から業績要因文を抽出する．テストデータには，有価証券報告書に含まれる全ての文のうち，文の最後の文字が「た」で終わる文とした．これは，業績要因文は確定した業績の要因を述べている文であるので，過去形の文が多いからである．しかし，深層学習による抽出のみでは，例えば「営業利益は、前年度に比べ１３７億円増加し、４７８億円となりました」のような業績要因ではなく全体の業績のみが記述されている文も抽出されることがある．そのような文を除去するために，抽出された文に含まれている企業キーワードのスコア $W(n, S(t))$ の合計をその文のスコアとして付与し，スコアが高い文を業績要因文として抽出する．

抽出した業績要因文の例を以下に示す．

・中堅中小企業向けＳＩ分野では、食品業等の業種向けシステム販売が堅調に推移したほか、Ｗｉｎｄｏｗｓ１０への更新案件も増加いたしました。
・資材は、鋼管等の農業用ハウス関連資材の値上げ前の駆け込み需要、天候不順に伴う高機能液肥及び保温資材の需要増により、増収となりました。
・農材事業においては、新規薬剤の普及拡販や茎葉除草剤の伸長に加え、関東地区においては、土壌消毒剤も順調に推移いたしました。

4.5　有価証券報告書からの業績結果文の抽出

業績結果文は，業績要因文と比較して，表現が限定的であることからルールベースによる抽出を行った．ここで業績結果文とは，

定義 1：「〜円」と具体的な数値結果が記載されている．
定義 2：結果文であるため，文末が過去形の表現「〜た。」となっている．

を満たす文が対象となる．業績結果文によく含まれる表現としては，「利益」，「営業収益」，「経常収益」，「売上高」などの名詞がある．しかし、「利益」に関しては，「目標利益」，「資本利益」などの業績結果文らしくない表現もある．そこで，「〜利益」となる複合名詞を抽出し，ルール作成に使用した．表 41 は有価証券報告書から抽出した「〜利益」となる複合名詞の一部である．業績結果文の抽出のルールは，文末が「〜た。」であり，「万円」，

表 41　抽出した「〜利益」となる複合名詞

業績結果文に含まれやすい表現	業績結果文に含まれにくい表現
純利益	資本利益
営業利益	包括利益
経常利益	引前利益
セグメント利益	株主利益

「億円」，「千円」を含み，かつ，以下の表現を含んでいる文とした．

営業収益，売上高，経常収益，純利益，営業利益，経常利益，セグメント利益，総利益，当期利益，事業利益，業務利益，連結利益，帰属利益，最高利益，最終利益，販売利益

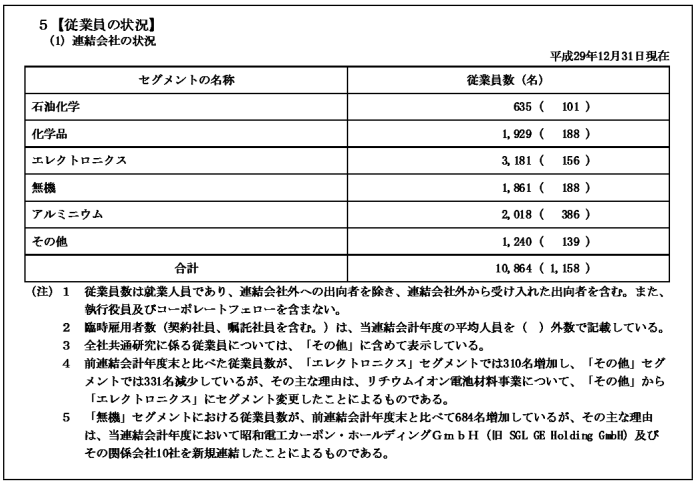
抽出した業績結果文の例を以下に示す．

・以上の結果、売上高は、6,390 億円となりました。
・上記の結果を受け、営業利益は 782 億円となりました。
・セグメント利益は 340,188 千円となりました。
・同事業の売上高は 30,935 百万円となりました。

このように，業績結果文に関しては，文脈を考慮しなければ，事業セグメントが付与できない文が多い．

4.6　有価証券報告書からの事業セグメント名の抽出

事業セグメント名の抽出は，有価証券報告書の 1.5. 節「従業員の状況」に事業セグメントごとの従業員数が記載されているため，この部分から事業セグメント名と思われる文字列を候補として抽出する．具体的には図 25 に記載されている表の中から「セグメントの名称」と「合計」の間にある文字列を事業セグメント名の候補とする．表 42 は図 25 のページをテキスト化した結果である*25．4 行目の「セグメントの名称」から，18 行目の「合計」の間にある文字列が事業セグメント名の候補として抽出される．昭和電工株式会社は「石油化学」,「化学品」,「エレクトロニクス」,「無機」,「アルミニウム」,「その他」が事業セグメント名であり，それらが候補として抽出される．



5【従業員の状況】
(1) 連結会社の状況

平成29年12月31日現在

セグメントの名称	従業員数（名）
石油化学	635（　101）
化学品	1,929（　188）
エレクトロニクス	3,181（　156）
無機	1,861（　188）
アルミニウム	2,018（　386）
その他	1,240（　139）
合計	10,864（1,158）

（注）1　従業員数は就業人員であり、連結会社外への出向者を除き、連結会社外から受け入れた出向者を含む。また、執行役員及びコーポレートフェローを含まない。
2　臨時雇用者数（契約社員、嘱託社員を含む。）は、当連結会計年度の平均人員を（　）外数で記載している。
3　全社共通研究に係る従業員については、「その他」に含めて表示している。
4　前連結会計年度末と比べた従業員数が、「エレクトロニクス」セグメントでは310名増加し、「その他」セグメントでは331名減少しているが、その主な理由は、リチウムイオン電池材料事業について、「その他」から「エレクトロニクス」にセグメント変更したことによるものである。
5　「無機」セグメントにおける従業員数が、前連結会計年度末と比べて684名増加しているが、その主な理由は、当連結会計年度において昭和電工カーボン・ホールディングＧｍｂＨ（旧 SGL GE Holding GmbH）及びその関係会社10社を新規連結したことによるものである。

図 25　昭和電工株式会社の有価証券報告書 1.5. 節「従業員の状況」PDF データ

昭和電工株式会社の例で示したような結果を抽出するために，以下のようなルールで抽出を行った．

ルール１:　スタートワードとエンドワードの間の文字列*26を対象とする．
ルール２:　ストップワードを設定し，それらの文字列を含む場合は，事業セグメント名の

*25　２０行目以下もテキスト化されているがスペースの関係上省略．
*26　ここでの文字列とは，2 文字以上のものであり，数字やかっこなどの記号でのみ構成されているものは除く．また，文字列であるため，最後が「。」で終わるものは文字列ではなく文章であるため，文字列ではないとする．例えば，昭和電工株式会社の例の 7 行目「635(101)」は文字列ではない．

表 42　昭和電工株式会社の有価証券報告書 1.5. 節「従業員の状況」のテキストデータ

行	本文
1	５【従業員の状況】
2	（１）連結会社の状況
3	平成２９年１２月３１日現在
4	セグメントの名称
5	従業員数（名）
6	石油化学
7	６３５（１０１）
8	化学品
9	１，９２９（１８８）
10	エレクトロニクス
11	３，１８１（１５６）
12	無機
13	１，８６１（１８８）
14	アルミニウム
15	２，０１８（３８６）
16	その他
17	１，２４０（１３９）
18	合計
19	１０，８６４（１，１５８）
20	（注）１従業員数は就業人員...

候補に加えない．

スタートワードとは，昭和電工株式会社の例では，4 行目の「セグメントの名称」のことを指す．スタートワードは以下の３つである[*27]．

スタートワード１:「セグメント～」
スタートワード２:「事業～」
スタートワード３:「部門～」

[*27] 「～」の部分には文字列が入る．

エンドワードは，昭和電工株式会社の例では，18 行目の「合計」のことを指す．エンドワードは以下の 2 つである．

エンドワード 1:「合計」
エンドワード 2:「報告セグメント計」

ストップワードとは，昭和電工株式会社の例では，5 行目の「従業員数（名）」のことを指す．ストップワードを以下に示す．

平均年数，平均勤続年数，平均年令，平均年間，平均年齢，増減，全社，従業員，使用人数，組合員数

これらにより，事業セグメント名の候補は抽出される．表 43 に企業ごとの事業セグメント名候補の抽出結果を示す．

表 43　抽出された事業セグメント名候補

企業名	セグメント名
東京急行電鉄株式会社	ビジネスサポート事業
	生活サービス事業
	ホテル・リゾート事業
	不動産事業
	交通事業
本田技研工業株式会社	二輪事業
	四輪事業
	汎用パワープロダクツ事業
	金融サービス事業
	その他の事業
旭硝子株式会社	ガラス
	電子
	化学品

4.7 業績要因文，業績結果文が属する事業セグメントの付与

抽出した業績要因文と業績結果文に対して，事業セグメント名の候補を用いて事業セグメントの付与を行う．付与の手法は以下の手順で行う．

Step 1: 有価証券報告書の 2. 章「事業の状況」を上から 1 文ごとに文を取得する．

Step 2: 「。」を含まない文であれば，事業セグメント名の候補を含んでいるかどうか確認する．

Step 3: 事業セグメント名候補を含んでいれば，そこから 5 行以内に出てきた文は，その事業セグメントの内容である文と判定する[*28]．

Step 4: 5 行以内に業績要因文や業績結果文と判定された文が取得された場合，そこから新たに 5 行以内に出てきた文は，その事業セグメントの内容であると判定する．

しかし，このルールだけだと，xbrl ファイルからテキストを抽出する際に，事業セグメント名と事業セグメントに対応する文が 1 つになってしまい，事業セグメントの付与ができない文が多く出てしまう．例えば，「アルミニウム」，「当セグメントでは、アルミ電解コンデンサ...」のように本来 2 つの文に分かれてテキストを抽出することを想定しているが，この抽出が「アルミニウム当セグメントでは、アルミ電解コンデンサ...」のように 1 つの文になって抽出してしまうことが起こる．そこで，文の先頭，1 文字削除した箇所，2 文字削除した箇所，3 文字削除した箇所に対して，事業セグメント名の候補があるかどうかでセグメントの付与を行った．もし，事業セグメント名候補を含んでいれば，この文を含めて 5 行以内に出てきた文は，その事業セグメントの内容であると判定する．先頭文字を削除した理由は，「1 アルミニウム当セグメント...」や「(1) アルミニウム当セグメント...」に対応するためである．

上記の処理の具体例を表 44 より述べる．要因に「○」が付いている文は，本手法によって業績要因文として判定された文であり，結果に「○」が付いている文は，本手法によって業績結果文として判定された文である[*29]．表 44 の例だと，26 行目の文が「。」を含まない文であり，「コンサルティング」が事業セグメント名候補として抽出できているため，以下 5 行の 31 行目までは「コンサルティング」事業の文であると判定する．31 行目の文

[*28] 一行に一文が対応する．

[*29] 表 44 の要因，結果の○は本手法が業績要因文，業績結果であると判定した文を○としいるため，○の文が本論文の定義による業績要因文，業績結果文であるということではない．29 行目，35 行目，38 行目は業績要因文ではなく，本手法の誤判定である．

表 44　野村総合研究所の業績要因文・業績結果文へのセグメント付与例

行数	要因	結果	本文
26			(コンサルティング)
27			当セグメントは、政策提言や戦略コンサルティング、業務改革をサポートする業務コンサルティング、IT マネジメント全般にわたるシステムコンサルティングを提供しています。
28			顧客の経営環境や IT 部門の環境が変化する中、経営・IT の両面でコンサルティングの需要が高まっています。
29	○		当社グループは、顧客のビジネス全般を支援する変革パートナーとなる体制を整えていくとともに、海外も含めた顧客基盤の拡大に努めました。
30	○		当年度は、企業収益の改善を受け、顧客業務の実行を支援する業務コンサルティングなどが増加したことに加え、グローバル関連では ASG GroupLimited がシステムコンサルティングの増加に寄与しました。
31		○	この結果、売上高 31,161 百万円 (前年度比 8.1% 増)、営業利益 5,853 百万円 (同 6.7% 増) となりました。
32			(金融 IT ソリューション)
33			当セグメントは、主に証券業や保険業、銀行業等の金融業顧客向けに、システムコンサルティング、システム開発及び運用サービスの提供、共同利用型システム等の IT ソリューションの提供を行っています。
34			事業領域の拡大に向け、業界標準ビジネスプラットフォームの生産革新を進めるとともに、IT と金融を融合した FinTech(フィンテック) 等を活用した新事業の開発に取り組んでいます。
35	○		既存事業の拡大に向けた取組みとして、業界標準ビジネスプラットフォームについては、制度改正への着実な対応を進めるとともに、顧客業務の高度化や効率化の支援を進めました。
36			リテール証券のバックオフィス業務をサポートする共同利用型システム「STAR - IV」については、災害時にシステム障害復旧を行う機能を追加したサービスを提供しています。
37			また、平成 30 年に実施予定の国債の決済期間短縮化や、証券保管振替機構の次期システムへの移行について、当社の共同利用型システムの対応を進めています。
38	○		また、資産運用領域の事業拡大を目的に、米国の CutterAssociates, LLC を子会社としました。
39	○	○	当年度の売上高は、証券業を中心にシステムコンサルティングが増加しましたが、前年度に大型の製品販売があった証券業向け開発・製品販売や、保険業向け開発・製品販売が減少し、248,188 百万円 (前年度比 2.2% 減) となりました。
40	○	○	子会社の (株) だいこう証券ビジネスにおいて業務体制見直しに向けた事業構造改善費用を計上したこともあり、営業利益は 26,461 百万円 (同 9.3% 減) となりました。

は業績結果文であるため，さらに以下 5 行の 36 行目までは「コンサルティング」事業の文であると判定される．しかし，32 行目の文が「。」を含まない文であり，「金融 IT ソリューション」が事業セグメント名候補として抽出できているため，32 行目から以下 5 行の 37 行目までは，「金融 IT ソリューション」事業の文であると判定が上書きされる．35 行目の文は業績要因文であるため，さらに以下 5 行の 40 行目までは「金融 IT ソリューション」事業の文であると判定される．

4.8　実装

本手法の評価を行うため，本手法を実装した．実装にあたり，形態素解析器として MeCab[*30]，係り受け解析器として CaboCha[65, 66, 67] を使用した．学習データは，2013 年から 2018 年 5 月までに提出された有価証券報告書を用いた．ここで，本研究で対象とする有価証券報告書の取得について述べる．

有価証券報告書の PDF データは企業の web サイトからでも取得可能であるが，企業の

[*30] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

web サイトから取得する方法だと収集が大変である．そこで，全ての企業の有価証券報告書のデータが取得可能な EDINET にて，データを収集した．取得したデータは，XBRL 形式の有価証券報告書データである．XBRL 形式のデータは図 26 に示すようなタグデータになっており，このタグを利用して，指定した章のテキストデータを抽出することが可能である．


```
<jpcrp_cor:InformationAboutEmployeesTextBlock contextRef="FilingDateInstant">
<p style="page-break-before:always; line-height:0.75pt; width:100%; font-size:0.75pt;"> </p> <h3 class="smt_head2">5【従業員の状況】</h3> <h4 class="smt_head3" style="padding-left:13.5pt; padding-right:0pt;margin-top:0pt;margin-bottom:0pt;text-justify:inter-ideograph;text-indent:0pt;font-family:&amp;apos;MS 明朝&amp;apos;;text-align:justify;letter-spacing:0pt;line-height:15pt;">(1) 連結会社の状況 </h4> <p class="smt_unitR" style="padding-left:368.3pt;line-height:10pt;">平成30年3月31日現在</p> <div class="tbld" style="margin-left:0pt;margin-right:0pt;margin-top:0pt;margin-bottom:0pt;"> <table cellspacing="0" cellpadding="0" style="border-collapse:collapse;border:solid 0pt #000000;width:468pt;"> <colgroup> <col style="width:234pt;min-width:234pt;"/> <col style="width:136.5pt;min-width:136.5pt;"/> <col style="width:97.5pt;min-width:97.5pt;"/> </colgroup> <tr style="height:22.5pt"> <td valign="middle" style="vertical-align:middle;border-top-style:solid;border-top-width:0.75pt;border-bottom-style:solid;border-bottom-width:0.75pt;border-left-style:solid;border-left-width:0.75pt;border-right-style:solid;border-right-width:0.75pt;"> <p class="smt_tblC">セグメントの名称</p> </td> <td valign="middle" colspan="2" style="vertical-align:middle;border-top-style:solid;border-top-width:0.75pt;border-bottom-style:solid;border-bottom-width:0.75pt;border-left-style:solid;border-left-width:0.75pt;border-right-style:solid;border-right-width:0.75pt;"> <p class="smt_tblC">従業員数(人)</p> </td> </tr> <tr style="height:22.5pt"> <td valign="middle" style="vertical-align:middle;border-top-style:solid;border-top-width:0.75pt;border-left-style:solid;border-left-width:0.75pt;border-right-style:solid;border-right-width:0.75pt;"> <p class="smt_tblL">子育て支援事業</p> </td> <td valign="middle" style="vertical-align:middle;border-top-style:solid;border-top-width:0.75pt;border-bottom-color:#a9a9a9;border-left-style:solid;border-left-width:0.75pt;border-right-color:#a9a9a9;"> <p class="smt_tblR" style="text-align:right;padding-left:0pt;padding-right:9pt;margin-top:0pt;margin-bottom:0pt;text-justify:inter-ideograph;text-indent:0pt;font-family:&amp;apos;MS 明朝&amp;apos;;letter-spacing:0pt;line-height:10pt;">3,051</p> </td> <td valign="middle" style="vertical-align:middle;border-top-style:solid;border-top-width:0.75pt; border-bottom-color:#a9a9a9;border-right-style:solid;border-right-width:0.75pt;"> <p class="smt_tblR" style="padding-right:66.4pt;">(2,209)</p> </td> </tr> <tr style="height:22.5pt"> <td
```


図 26　XBRL 形式のデータ例

図 26 の最初の行に書いてある「InformationAboutEmployeesTextBlock」が 1 章 5 節のタグであり，このタグの中のテキスト部分を抽出する．抽出した結果の例を図 27 に示す．本論文で扱うデータは，図 27 に示したような，XBRL 形式のデータから抽出された

```
1 5【従業員の状況】
2 (1)連結会社の状況
3 平成30年3月31日現在
4 セグメントの名称
5 従業員数(人)
6 子育て支援事業
7 3,051
8 (2,209)
9 全社(共通)
10 43
11 (4)
12 ШИИ
13 3,094
14 (2,213)
15 (注)1.従業員数は、就業人員であります。
16 2.臨時雇用者数につきましては、年間の平均人員を()外数で記載しております。
17 なお、臨時雇用者はパートタイマー、アルバイトを含み、派遣社員を除いております。
```

図 27　XBRL 形式のデータから抽出されたテキストデータ

テキストデータである．収集した有価証券報告書のデータ数は 21,326 件であり，対象と

表 45　昭和電工株式会社の事業セグメントが付与された業績要因文や業績結果文

要因 or 結果	事業セグメント	抽出文
要因	エレクトロニクス	リチウムイオン電池材料事業は、中国における電気自動車向け補助金政策の変更の影響を受け出荷が減少したため減収となった。
結果	エレクトロニクス	この結果、当セグメントの売上高は 1,230 億 64 百万円となり、営業利益は 219 億 25 百万円となった。
要因	石油化学	有機化学品事業は、原料価格低下を受け、酢酸ビニル、酢酸エチルの販売価格が低下し減収となった。
結果	石油化学	この結果、当セグメントの売上高は 1,857 億 83 百万円となり、営業利益は、206 億 90 百万円となった。
要因	化学品	情報電子化学品事業は、半導体・ディスプレイ業界の増産に伴い電子材料用高純度ガスの出荷が増加し増収となった。
結果	化学品	この結果、当セグメントの売上高は 1,487 億 58 百万円となり、営業利益は 164 億 74 百万円となった。
要因	アルミニウム	アルミ圧延品事業は産業機器・車載向けアルミ電解コンデンサー用高純度箔の出荷が増加し増収となった。
結果	アルミニウム	この結果、当セグメントの売上高は 1,054 億 39 百万円となり、営業利益は 66 億 97 百万円となった。

する企業の数は 4,525 企業である[*31].

獲得できた手がかり表現は 304 個であり，そこから 8,517 個の拡張手がかり表現を獲得した．獲得した企業キーワードと拡張手がかり表現によって，4,525 企業の有価証券報告書から正例，負例，合計で 67,415 文の学習データを生成し，その 67,415 文の学習データから 5,083 語が素性として選択された．

業績要因文抽出の深層学習モデルの入力層のノード数を入力ベクトルの次元数と同じ 5,083 とし，隠れ層は，順にノード数 3,000 が 3 層，ノード数 1,000 が 3 層，ノード数 500 が 3 層，ノード数 100 が 3 層の計 12 層とした．出力層は 1 要素である．また，epoch 数は 20，活性化関数として ReLU を使用した．

表 45 に本手法により抽出され，事業セグメントが付与された業績要因文や業績結果文をいくつか示す．

本手法によって抽出された結果を検索対象とした有価証券報告書検索システムを実装した．検索システムでは，検索キーワードを入力すると，そのキーワードを含む業績要因文と，それが属する事業セグメントと，その事業セグメントの業績結果文が検索される．そ

[*31] 名称変更した企業や，上場廃止した企業も含む．

して，その事業セグメントを展開する企業が検索される．図 28 に，有価証券報告書検索システムにおいて「アルミ」で検索した場合の検索結果を示す．

図 28　有価証券報告書検索システム

本検索システムを使用することで，例えば「アルミ」を業績要因文にもつ有価証券報告書，その業績要因文の属する事業セグメントとその業績結果，および，企業が検索でき，あるキーワードに関連のある企業と，その企業における事業規模を検索するシステムとしても有効であると考える．なお，本有価証券報告書検索システムは一般公開される予定であり，誰でも利用可能である[*32]．

4.9　評価

本手法の評価を行った．評価用データとして，2018 年 7 月から 2018 年 11 月の 21 企業の有価証券報告書，4,691 文に対して人手にて業績要因文，業績結果文のラベル付けと，事業セグメントの付与を行い，正解データを作成した．評価用データは，学習データとして使用した有価証券報告書とは異なる期間の有価証券報告書を使用しているため，学習データとの重複はない．

対象企業の選定基準は，事業セグメントを 3 以上持つ企業から無作為に選択した．ラベ

[*32] http://hawk.ci.seikei.ac.jp/u-cees/

ル付けの結果，21 企業の 4,691 文に対して，事業セグメントに関する業績要因文は 200 文，事業セグメントに関する業績結果文は 135 文であった．この文の中には，「建材事業においては、ビル分野での短工期工事受注や住宅分野での販売網の拡充に努めたものの、新設住宅着工戸数など市況が前年比減で推移していることや競合環境が継続していること、アルミ地金など原材料価格の上昇影響などにより、売上高は１，９６９億４３百万円となりました。」のような，業績要因文かつ業績結果文と重複するものもある．作成した正解データを用いて，本手法の適合率（$precision$），再現率（$recall$），F 値（$F-measure$）を求めた．

評価は，事業セグメントが付与されている文を対象に，付与された事業セグメントが一致し，業績要因文であるかどうかの判定が合っていれば正解とし，どちらか一方でも異なっていたら不正解とした．例を表 46 に示す．表 46 の左側が人手によるラベルとセグメント付与の結果，右側が本手法によるラベルとセグメント付与の結果である．業績結果

表 46　正解・不正解の例

判定	要因	セグメント	要因	セグメント
正解	1	不動産	1	不動産
不正解	1	販売	1	不動産
不正解	0	不動産	1	不動産

文も同様の評価方法を用いた．

評価結果を表 47 と表 48 に示す． また，事業セグメント名の抽出の結果と事業セグメントの付与の結果を表 49 に示す．各企業の事業セグメントごとの適合率・再現率は，スペースの関係を考慮し，付録として表 59〜表 62 に示す．

最後に，事業セグメントの大きさを考慮した評価を行った．事業セグメントの規模を考慮したウェイトには売上高を用いた．売上高を用いた理由は，事業セグメントごとの利益に比べると年ごとの変動が少なく，事業セグメントの大きさを考慮する指標として妥当と考えたからである．ウェイトを用いて加重平均を計算し，比較対象として相加平均も計算し，評価を行った．評価対象は，適合率や再現率が 1.000 の企業以外から無作為に選択した企業で行った．その結果を表 50〜表 55 に示す．

表 47　業績要因文の評価結果

企業名	企業コード	$precision$	$recall$	$F-measure$
カネコ種苗株式会社	E00004	9 / 12 = 0.750	9 / 11 = 0.818	0.783
株式会社サカタのタネ	E00006	16 / 16 = 1.000	16 / 22 = 0.727	0.842
日本工営株式会社	E00078	2 / 8 = 0.250	2 / 10 = 0.200	0.222
コーセル株式会社	E01856	4 / 7 = 0.571	4 / 6 = 0.667	0.615
株式会社ＩＧポート	E02480	10 / 10 = 1.000	10 / 10 = 1.000	1.000
株式会社内田洋行	E02515	7 / 7 = 1.000	7 / 9 = 0.778	0.875
テーオーホールディングス	E03169	4 / 4 = 1.000	4 / 6 = 0.667	0.800
株式会社サンオータス	E03326	9 / 18 = 0.500	9 / 11 = 0.818	0.621
リベレステ株式会社	E03989	6 / 7 = 0.857	6 / 6 = 1.000	0.923
明豊エンタープライズ	E04024	4 / 7 = 0.571	4 / 8 = 0.500	0.533
株式会社ゼロ	E04230	5 / 7 = 0.714	5 / 7 = 0.714	0.714
東京博善株式会社	E04843	5 / 13 = 0.385	5 / 12 = 0.417	0.400
日本プロセス株式会社	E04873	16 / 18 = 0.889	16 / 19 = 0.842	0.865
株式会社ビューティ花壇	E05597	6 / 7 = 0.857	6 / 7 = 0.857	0.857
フリービット株式会社	E05680	4 / 7 = 0.571	4 / 4 = 1.000	0.727
株式会社インサイト	E05740	8 / 12 = 0.667	8 / 11 = 0.727	0.696
日本海洋掘削株式会社	E23800	6 / 19 = 0.316	6 / 6 = 1.000	0.480
株式会社ＴＨＥグローバル社	E24340	6 / 7 = 0.857	6 / 7 = 0.857	0.857
三協立山株式会社	E26831	4 / 5 = 0.800	4 / 7 = 0.571	0.667
タマホーム株式会社	E27305	11 / 11 = 1.000	11 / 16 = 0.688	0.815
ウエスコホールディングス	E30042	3 / 7 = 0.429	3 / 5 = 0.600	0.500
$TOTAL$		145/209 = 0.693	145/200 = 0.725	0.709

4.10　考察

表 47 と表 48 より，適合率および再現率ともに良好な結果が得られた．この学習データの中には，評価用データに選ばれた企業の過去の有価証券報告書が含まれており，過去に似たような内容の文が記載されているため結果が良くなってしまっている可能性も考慮し，学習データから評価用データに選ばれた企業の有価証券報告書を除いて学習を行い，再評価を行った．その結果，表 47 と表 48 と同様の結果を得ることができた．これは，学習データを自動で大量に生成することができたからである．また，この結果から，新規上場企業のような過去のデータがない企業に対しても，この手法を適用することが可能であると思われる．

表 48　業績結果文の評価結果

企業名	企業コード	$precision$	$recall$	$F-measure$
カネコ種苗株式会社	E00004	7 / 9 = 0.778	7 / 7 = 1.000	0.875
株式会社サカタのタネ	E00006	4 / 4 = 1.000	4 / 4 = 1.000	1.000
日本工営株式会社	E00078	6 / 6 = 1.000	6 / 6 = 1.000	1.000
コーセル株式会社	E01856	5 / 5 = 1.000	5 / 5 = 1.000	1.000
株式会社ＩＧポート	E02480	4 / 4 = 1.000	4 / 4 = 1.000	1.000
株式会社内田洋行	E02515	5 / 5 = 1.000	5 / 5 = 1.000	1.000
テーオーホールディングス	E03169	7 / 7 = 1.000	7 / 7 = 1.000	1.000
株式会社サンオータス	E03326	4 / 4 = 1.000	4 / 4 = 1.000	1.000
リベレステ株式会社	E03989	7 / 10 = 0.700	7 / 8 = 0.875	0.778
明豊エンタープライズ	E04024	5 / 11 = 0.455	5 / 5 = 1.000	0.625
株式会社ゼロ	E04230	3 / 4 = 0.750	3 / 3 = 1.000	0.857
東京博善株式会社	E04843	14 / 21 = 0.667	14 / 19 = 0.737	0.700
日本プロセス株式会社	E04873	6 / 10 = 0.600	6 / 6 = 1.000	0.750
株式会社ビューティ花壇	E05597	6 / 6 = 1.000	6 / 7 = 0.857	0.923
フリービット株式会社	E05680	5 / 5 = 1.000	5 / 5 = 1.000	1.000
株式会社インサイト	E05740	3 / 9 = 0.333	3 / 4 = 0.750	0.462
日本海洋掘削株式会社	E23800	4 / 6 = 0.667	4 / 7 = 0.571	0.615
株式会社ＴＨＥグローバル社	E24340	12 / 12 = 1.000	12 / 12 = 1.000	1.000
三協立山株式会社	E26831	6 / 8 = 0.750	6 / 7 = 0.857	0.800
タマホーム株式会社	E27305	5 / 5 = 1.000	5 / 5 = 1.000	1.000
ウエスコホールディングス	E30042	5 / 5 = 1.000	5 / 5 = 1.000	1.000
$TOTAL$		123/156 = 0.788	123/135 = 0.911	0.845

表 49　事業セグメント名の抽出と事業セグメントの付与の結果

	$precision$	$recall$	$F-measure$
抽出	103 / 106 = 0.972	103 / 104 = 0.990	0.981
付与	229 / 321 = 0.713	229 / 279 = 0.820	0.763

不正解になってしまった文は，大きく分けると，事業セグメントの付与に誤りがあるものと，業績要因文や業績結果文の抽出に誤りがあるものである．

事業セグメント付与の誤りとしては，有価証券報告書 1.5. 節「従業員の状況」に記載されている事業セグメントが，本文中の記載と微妙に異なっていたりすることが原因で起こる．例えば，企業コード E04843 の東京博善株式会社では，表 56 のように事業セグメントの付与が失敗しているが，これは，有価証券報告書 1.5. 節では「桐ケ谷斎場」と記載さ

表 50　E00004 の業績要因文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
種苗事業	78 億 55 百万円	0.134	5 / 6 = 0.833	5 / 5 = 1.000
花き事業	91 億 69 百万円	0.157	1 / 1 = 1.000	1 / 1 = 1.000
農材事業	267 億 65 百万円	0.457	2 / 2 = 1.000	2 / 3 = 0.667
施設材事業	147 億 40 百万円	0.252	1 / 3 = 0.333	1 / 2 = 0.500
		加重平均	0.873	0.769
		相加平均	0.792	0.792

表 51　E03326 の業績要因文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
エネルギー事業	9,358 百万円	0.310	3 / 6 = 0.500	3 / 3 = 1.000
カービジネス事業	20,239 百万円	0.670	4 / 7 = 0.571	4 / 5 = 0.800
ライフサポート事業	170 百万円	0.006	1 / 2 = 0.500	1 / 2 = 0.500
不動産関連事業	457 百万円	0.015	1 / 3 = 0.333	1 / 1 = 1.000
		加重平均	0.545	0.863
		相加平均	0.476	0.825

表 52　E03989 の業績要因文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
開発事業	1,550 百万円	0.268	2 / 2 = 1.000	2 / 2 = 1.000
建築事業	403 百万円	0.070	1 / 1 = 1.000	1 / 1 = 1.000
不動産販売事業	3,331 百万円	0.576	1 / 1 = 1.000	1 / 1 = 1.000
その他	502 百万円	0.087	2 / 2 = 1.000	2 / 2 = 1.000
		加重平均	0.712	1.000
		相加平均	0.875	1.000

れているが，190 行目では「桐ヶ谷斎場」と「ケ」が「ヶ」となっているため失敗してしまっていた．

それ以外にも，事業セグメント付与の誤りとしては，

表 53　E26831 の業績要因文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
建材事業	1,969 億 43 百万円	0.600	1 / 1 = 1.000	1 / 1 = 1.000
マテリアル事業	461 億 78 百万円	0.141	1 / 1 = 1.000	1 / 2 = 0.500
商業施設事業	385 億 84 百万円	0.118	1 / 2 = 0.500	1 / 2 = 0.500
国際事業	465 億 58 百万円	0.142	1 / 1 = 1.000	1 / 2 = 0.500
		加重平均	0.941	0.800
		相加平均	0.875	0.625

表 54　E03989 の業績結果文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
開発事業	1,550 百万円	0.268	1 / 1 = 1.000	1 / 2 = 0.500
建築事業	403 百万円	0.070	2 / 2 = 1.000	2 / 2 = 1.000
不動産販売事業	3,331 百万円	0.576	2 / 2 = 1.000	2 / 2 = 1.000
その他	502 百万円	0.087	2 / 5 = 0.400	2 / 2 = 1.000
		加重平均	0.948	0.866
		相加平均	0.850	0.875

表 55　E26831 の業績結果文の適合率と再現率の相加平均と加重平均の結果

事業セグメント	売上高	$Weight$	$precision$	$recall$
建材事業	1,969 億 43 百万円	0.600	1 / 1 = 1.000	1 / 1 = 1.000
マテリアル事業	461 億 78 百万円	0.141	2 / 2 = 1.000	2 / 2 = 1.000
商業施設事業	385 億 84 百万円	0.118	2 / 4 = 0.500	2 / 2 = 1.000
国際事業	465 億 58 百万円	0.142	1 / 1 = 1.000	1 / 2 = 0.500
		加重平均	0.941	0.859
		相加平均	0.875	0.750

・施設材事業においては、農業用フィルムの拡販と新規得意先開拓が功を奏したことや、小ロットや長尺な農業資材の配送にもタイムリーに対応できる当社配送体制の優位さが、運送物流事情悪化の影響でより鮮明となり、販売先の支持が得られたことなどから増収となり、農材事業においては、新規薬剤の普及拡販や茎葉除草剤の伸長に加え、関東地区においては、土壌消毒剤も順調に推移いたしました。

表 56　東京博善株式会社

行	本来付与したい事業セグメント	本手法によって付与された事業セグメント	本文
189	四ツ木斎場	四ツ木斎場	セグメント資産は、スポットライト新設工事等固定...
190	桐ヶ谷斎場	四ツ木斎場	5）桐ヶ谷斎場
191	桐ヶ谷斎場	四ツ木斎場	売上高は、四ツ木斎場の営業再開の影響を受け...
192	桐ヶ谷斎場	四ツ木斎場	営業利益は、売上高が減少したこと、屋上防水工事...

のような文があった．この文は「施設材事業」,「運送物流事情」,「農材事業」などのことを含む文である．今回このような文に対しては，企業全般のことが書いてあるとし，正解データに事業セグメント付与を行っていない．事業セグメント付与は，文の先頭に「施設材事業」があるため,「施設材事業」のセグメントが付与されるため誤りとなる．施設材事業のことについて言及していないわけではないため，正解として計算することも間違いではないと考えることもできるが，今回は厳しめに評価を行った．このように，業績要因文判定は合っているものの，事業セグメント付与が失敗している文は 6 文ほどあった．

また，このような事業セグメント付与の失敗により，その後に続く文に誤った事業セグメントは付与されてしまった．例えば，先ほどの例に続く以下の文,

> ・売上総利益については、利益率の高い種苗事業は順調に利益増に貢献したものの、花き事業の販売低迷、施設材事業においては養液栽培プラントの受注減、施設材事業と農材事業に共通した状況として競争激化が、いずれも利益率の低下要因となり、売上総利益は微増に終わりました。

に対しても,「施設材事業」が付与されてしまった．

業績結果文は，事業全体のことを言っている文に対して,「その他」の事業セグメントが付与されやすくなってしまっている．業績結果文の適合率計算で誤っている文 33 文のうち，19 文ほどはこれにあたる．しかし,「その他」に含まれるセグメントは，メインとなるセグメントと比較し，利益が極端に少ないため[*33],「その他」が失敗していることに大きな問題はないと考える．参考までに，事業セグメントが「その他」の文を除いた場合に適合率と再現率がどのような値になるかを表 57 に示す．「その他」で正解しているものも除いているため，再現率は多少下がるものの，適合率および F 値の値は向上しているのがわかる．

[*33] もし，利益が大きいセグメントが含まれるならば，その他扱いせずに，1 つのセグメントとして記載される．

表 57　事業セグメント「その他」を除いた全体の評価結果

	$precision$	$recall$	$F-measure$
業績要因文	132 / 177 = 0.746	132 / 187 = 0.706	0.725
業績結果文	112 / 126 = 0.888	112 / 124 = 0.903	0.896

むしろ企業全体の業績要因文に対して，「事業セグメント」が付与されているほうが問題となる．この問題に対しては，事業セグメントごとの業績情報の抽出とは異なり，決算短信には企業全体の業績情報が決められたフォーマットで記載されているため，企業全体の業績情報を抽出することは容易である．こうして抽出した業績情報と本手法で抽出した業績結果文に含まれる数値情報を比較することで，企業全体の業績結果文であるかどうかを判定することも可能であると考えている．

事業セグメント付与が成功しているが，業績要因文かどうかの判定を誤っている文に関しては，業績結果に直接影響したかどうかわからない表現を用いているもの，業績結果文を業績要因文と誤って抽出しているものに分類できる．

前者の例としては，以下のような文が挙げられる．

・保険部門につきましては、来店型保険ショップ『ほけんの窓口』を４店舗展開し、コンサルティング業務の質の向上に重点を置き、成約率の向上に努めました。
・このような情勢の中、ターゲット業界・顧客を絞り、新規プロジェクト獲得、新規顧客開拓に注力してまいりました。
・このような情勢の中、営業－開発部門の連携を強化し、新製品の拡販活動に注力するとともに、新規顧客の開拓、重点顧客の深堀活動に取り組んでまいりました。

「努めました。」「注力してまいりました。」「取り組んでまいりました。」「推進してまいりました。」などの，業績結果に直接影響しているかどうか不確かな文は，業績要因文ではないとし，正解データを作成した．それによって適合率が低くなってしまった．この問題に関しては，深層学習のモデル作成時の負例データにこのようなデータを入れることで，改善されると考える．

後者の例としては，以下のような文が挙げられる．

・生花祭壇事業の売上高は、３，３６６，５６５千円（前年同期比４．７％増）と、２期連続で過去最高を更新しました。
・不動産販売事業につきましては、一般不動産の販売による売上高が３，３３１百万円（前年同期比４１．４％減)、セグメント利益が６６７百万円（前年同期比３７．７％減）となりました。

これは，学習データの中に業績要因文かつ業績結果文となっているものが多く混じっていることが原因であると考えられる．このような文は，企業キーワードのスコア $W(n, S(t))$ の合計をその文のスコアとして，スコアが高い文を取ることで，ある程度除去できる．しかし，「生花祭壇」などの事業セグメント名が企業キーワードとして抽出されている場合には，除去しきれなかった．

また，E23800 の企業は，以下に示すように，事業セグメントごとにやったことを列挙していた．

・適正な来院者数を集客し継続的に維持することにより、業績の改善と採算を軌道に乗せ、当社グループ全体の収益安定化に貢献するよう取り組んでまいりました。
・その後、ロシア連邦共和国のサハリン島北東部沖に移動し、６月上旬から１０月中旬まで同国のＧａｚｐｒｏｍｎｅｆｔ－ＳａｋｈａｌｉｎＬＬＣの掘削工事に従事しました。
・同国アブダビ沖に移動し、平成３０年１月上旬から同国のＢｕｎｄｕｑＣｏｍｐａｎｙＬｉｍｉｔｅｄの掘削工事に従事しました。

これは業績要因文と捉えることもできなくないが，業績と直接の関係性が不確かであるため，業績要因文ではないとし，正解データを作成した．

4.11　本章のまとめ

本章では，有価証券報告書から業績要因文と業績結果文を抽出し，事業セグメントを付与する手法について述べた．具体的には，業績要因文を抽出するための手がかりとなる表現にいくつかの文節を追加することで，“拡張手がかり表現”と定義する文節列を獲得し，この“拡張手がかり表現”を使用することで業績要因文抽出のための学習データを自動的に生成した．そして，生成された学習データを使用して深層学習モデルの学習を行い，学習させたモデルを用いて，有価証券報告書から業績要因文を抽出した．業績結果文の抽出は，業績結果文に出現する表現を抽出し，ルールベースで行った．事業セグメントの付

与は，事業セグメント名の候補を「従業員の状況」から抽出し，ルールベースで行った．評価の結果，事業セグメントの付与が正しく，業績要因文判定が正しいものの適合率は 0.693，再現率 0.725 であり，事業セグメントの付与が正しく，業績結果文判定が正しいものの適合率は 0.788，再現率 0.911 であった．

5　結論

5.1　本論文の結論

本論文では全体を通して，金融テキスト，特に株主招集通知と有価証券報告書を対象にした金融テキストマイニングについて述べた．

2 章では，株主招集通知から開始ページを推定し，議案タイトルとその分類を推定する研究について述べた．人手で作成した学習データを用いて分類器を学習し，分類対象となるページをルールベースで絞った上で，ページ単位での分類が可能であることを示した．具体的には，議案が開始しているページをルールベースで推定し，そのページに記載されている議案が，どの議案分類に該当するかを分類する方法を検討し，単一ページの入力に対して，従来の機械学習手法である SVM や深層学習モデルの MLP などを用いて分類を行うことで，F 値 0.930（SVM），0.937（MLP）と良好な結果を得ることが可能であることを示した．また，本研究では，人手にて作成されたデータの一部をテストデータに使用し，その分類結果を利用し信頼度スコアを算出することで，実用的なシステムへの貢献を行った．さらに，株主招集通知の議案開始ページおよびそのページに記載されている議案がどの議案分類であるかの収録作業を人手で行う場合には，1 社あたり平均して 143.17 秒かかるのに対して，実装した応用システムは 13.68 秒で収録が可能であり，収録データがどの程度正しく収録されているかを示す F 値に関しても，人手での作業は平均して 0.618 であるのに対して，応用システムは 0.938 であり，応用システムが有効であることを示した．

3 章では，2 章と同じく株主招集通知を対象として，重要ページを抽出する研究について述べたが，抽出対象となるページが異なることや，学習データがもともと存在しない問題があった．学習データの生成は，単純なルールベースを用いて行ったが，学習データの質を高めるため，厳しいルールを設定した．その結果，学習データに含まれる株主招集通知は，株主招集通知をランダムサンプリングしたものを人手でラベル付けしたものに比べて，ルールに当てはまるものに限定されていることから，偏りが生じていた．この偏りのため，人手による学習データを用いた学習であれば，一番良好な結果が得られると考えられていた CRF 層を追加したモデルは，学習データの偏りを含めて学習（いわゆる過学習）した状態となり，提案したモデルの中で最良の結果を得ることができなかった．評価実験の結果，BiLSTM のモデルが唯一，学習データにないパターンの株主招集通知にも対応することが可能であった．この結果から，学習データの生成を機械的に行う場合には，従

来の研究で有効であったモデルが，必ずしも最良なモデルとはならないため，データ生成の特徴や学習データの偏りを考慮してモデルを検討する必要性を示した．

3 章で提案した，自動生成した学習データを用いたページ単位での情報抽出の方法は，株主招集通知に限らず，多くのデータに対して，応用が可能な汎用性の高いものである．2 章で示した通り，人手で作成した学習データがあれば，従来手法を用いることで十分な結果が得られる．しかし，学習データがない場合は，本研究で示したように学習データの生成を行う必要があるが，この学習データ作成のルールは，本研究の目的に特化したものであるため，扱うテキストデータと目的によって，独自に作る必要がある．一般的に，本研究の提案した方法を適用するにはいくつかの条件を満たしている必要がある．まず，ある程度はルールベースで学習データが作成できるような，特徴があるページを分類の対象としている必要がある．シンプルなルールで，学習データが作成できないようなページは学習データを作ることもできないため，本手法を適用することは難しい．また，学習データを作るためのテキストデータの分母が非常に大きい必要がある．シンプルなルールを設定し，精度の高いデータを得るためには，学習データには厳しい条件を満たす必要があるが，もともとの分母が大きくなければ，十分な量の学習データを生成することができないため，大量のデータを集めることが不可能な場合には，本手法の適用は難しい．

学習データの量と質は，分類器の性能に大きく影響を与えるため，3 章で示したように，そのルールベースによって生成された学習データの量の確認と，一部をランダムサンプリングして，学習データとして十分な精度のデータになっているかの確認が必要であり，これらを満たしているのであれば，3 章の結果と同等の結果が期待できる．

4 章では，先行研究である決算短信からの業績要因文の抽出を発展させ，事業セグメントを付与した業績要因文と業績結果文の抽出に関する研究について述べた．2 章と 3 章のページ単位の研究に比べて，文単位での抽出であることからタスクが難しく，学習データの質と量がより重要となる．そのため学習データの生成も単純なルールベースでは難しく，工夫が必要であった．先行研究の決算短信での研究 [50] が良好な結果であったため，有価証券報告書での業績要因文の抽出も良好な結果を得ることができた．

事業セグメントを付与した業績要因文と業績結果文の抽出は，事業セグメント，業績要因文，業績結果文をそれぞれ抽出する必要があるが，事業セグメントと業績結果文の抽出はルールベースにより行っており，ドメインに特化した手法となっているが，業績要因文の抽出は，学習データの自動生成を行っており，他の文の抽出にも有効な汎用性のある方法となっている．学習データを自動生成するために，抽出対象の文によく使用される表現である手がかり表現を最初にいくつか準備することで，酒井らの手法を用いてブートストラップ的に手がかり表現を自動で抽出することが可能であり，これらの手がかり表現を含

むものを正例，含まないものを負例とすることで，学習データを自動生成することが可能である．ただし，ブートストラップ的に自動獲得した手がかり表現を含む文を正例とする手法のみでは，学習データとしての精度に劣る．そのため，本研究では，ブートストラップ的に獲得した手がかり表現に適切な文節を追加することで手がかり表現を拡張し，その拡張手がかり表現を含む文を正例としている．これらの学習データ自動生成手法は汎用性が高い手法であり，金融テキストに限らず，他の文抽出タスクにも適用が可能な手法である．

本研究ではそれに加えて，企業キーワードという企業特有のキーワードを利用することや，業績要因文は過去の出来事に対しての記述であることから文末が「た。」で終わる文に正例を絞り込むなどの工夫を加えることで，学習データの精度を高めている．したがって，学習データの自動生成は汎用性があるが，より良好な結果を得るためには，抽出対象の文の特性を考慮して多少の工夫を加える必要がある．また，手がかり表現の獲得に関しては，文を係り受け解析する必要性があるため，係り受け解析が比較的機能する新聞記事のような文を用いる場合には有効である可能性が高いが，話し言葉や SNS の投稿などの係り受け解析が苦手とする文の場合には，望むような結果が得られない可能性が高いため，抽出の対象としている文に対して，係り受け解析が正しく機能するかどうかを，まず確認する必要がある．

この研究の貢献は，有価証券報告書を用いることで，事業セグメントごとの業績要因文と業績結果文を抽出することができるようになったことである．これにより，事業セグメント，業績要因文，業績結果（数値情報）の 3 つが紐づいたデータを抽出ことが可能となり，これまでよりもより発展的な分析が可能となる．

5.2　今後の展望

学習データに関しての課題は多く残っており，完全な自動生成は不可能である．少ない労力で学習データを生成できても，その学習データによって目的が達成できなければ意味がないため，今後も学習データの生成に関する研究は続ける必要があると考える．特に，3 章の研究では，文の抽出を文脈を考慮しない MLP での分類モデルを使用しているが，論文採択後，BiLSTM や BERT などのモデルで追加検証をしたところ，多少の精度の向上はあったが，期待するような精度の向上は見られなかった．学習データの分類は高い精度で可能だが，テストデータでの分類がそこまで高くないため，学習データを自動生成していることにより，学習データとテストデータの分布が一致しておらず，学習データに偏りが生じている可能性が高い．今後は，自動生成した学習データが出現する全データに対

してどのように分布しているか，文の分散表現などからの分析を本格的に行いたいと考えている．

また，これまで取り組んできた研究は有益な情報を抽出する段階に踏みとどまっている．特に，3 章や 4 章で抽出した情報は，最後に人が見ることで，投資判断等の意思決定に利用される．したがって，抽出した情報を用いて分析を行い，抽出情報に付加価値を付けることで意思決定をサポートすることが本来期待される．また，もともとのデータの分母が大きいため，抽出できる情報もまだまだ膨大である．したがって，抽出情報の重要性をスコア化することで，さらなる情報の取捨選択が可能となる方法の検討が必要である．

抽出したテキスト情報を分析，スコア化するためには，文の意味をコンピュータに解釈させる必要がある．特にテキストデータの分析は否定表現などの問題が多く残っている．また，業績要因文に関しては，例えば「コロナウィルスの感染拡大」によって飲食業界では負の影響があるが，ゲーム業界には正の影響があると考えられる．したがって，同じ業績要因だとしても業界や事業によって影響や極性が異なる．また，金融極性辞書などでは「事業の拡大」という表現で「拡大」が使われる傾向が強いことから，ポジティブな用語として辞書に登録されているが，「不良債権の拡大」ではネガティブな用語ともなり，前後関係によって極性が変化する．

このような課題を解決するために，対象企業の事業紹介などのテキスト情報と，抽出した文を入力にし，事業の業績結果や株価などを目的変数とすることで，入力文に対して極性や分散表現を出力できるようなモデルの検討を考えている．これらの極性付与の方法論の多くは，文や表現に対して，ポジティブ，ネガティブ度合いを示すために何らかの数値データを紐づける必要があるが，先行研究では，これらの数値情報に，企業の株価や，企業全体の売上高や営業利益などを用いている．しかし，企業の株価や全体の売上高は，様々な要因が入り乱れてしまっており，対象としている文がどの程度の影響を与えているかを考慮しきれていない問題がある．それに対して，本論文の 4 章で抽出することが可能となった，各事業セグメントごとの業績要因文と業績結果文は，これまで先行研究で使用してきたデータよりも，文と業績結果の数値情報の関係が強いものとなっている．したがって，これらのデータを用いて極性付与を行うモデルを学習させることで，これまでの研究で得られた結果よりも良好な極性付与を行うことが可能になる見込みが高い．このようなモデルの学習を可能にするためには，業績要因文とそれに紐づく業績結果文を人手で大量に作成する必要があるが，それは非常に困難であり，そのようなデータを大量に生成することを可能にした点が，本研究の分野への貢献の 1 つである．上記に上げたモデルを学習させ，利用することで，抽出したテキスト情報に対して，より金融に特化した分散表現を与えることができ，より高度な分析やスコア化が期待できる．

また，これまで抽出の対象としてこなかった金融テキストも多く存在しているが，それらにも有益な情報が多く記載されている．今後は，他の金融テキストからの情報抽出も試みたいと考えており，それらのテキスト情報が及ぼす市場への影響や，投資判断等の意思決定にどのような影響を与えているのかを，生涯研究し続けたいと考えている．

謝辞

本論文の執筆にあたり，多くの方々にご支援いただけたことに感謝申し上げます．本研究は成蹊大学理工学部情報科学科酒井浩之准教授のご指導の下で取り組みました．また，成蹊大学理工学部情報科学科 3 年次の後期から現在に至るまでの約 6 年半，横浜市立大学データサイエンス学部データサイエンス学科岩崎学教授にも，大変お世話になりました．先生方には，研究の取り組み方だけでなく，研究者としてのあるべき姿や，人材だけでなくコミュニティや，ひいては社会を育てる理念を学ばせていただきました．研究者としてだけでなく，人としても成長させていただいた先生方に深く御礼申し上げます．

また，本論文をまとめるにあたり，ご多忙にも関わらず丁寧なご意見いただきました，成蹊大学理工学部情報科学科中野有紀子教授，世木寛之教授，東京理科大学経営学部経営学科増山繁教授に対して心より感謝申し上げます．

最後に，阿部貴行准教授 (横浜市立大学データサイエンス学部データサイエンス学科) をはじめとする岩崎研究室 OB の皆様方，池上敦子教授 (成蹊大学理工学部情報科学科) や小森理准教授 (成蹊大学理工学部情報科学科) をはじめとする情報数理コースの先生方，共同研究やインターンで関わることになった皆様方にも，大変お世話になりました．専門分野問わず，身近な年長者である皆様の研究や姿勢は，参考とさせていただく点も多く，私にとってのよい刺激とさせていただきました．ここに感謝の意を申し上げます．

参考文献

[1] 和泉潔, 松井藤五郎. 金融市場における最新情報技術：8. 金融テキストマイニング研究の紹介. 情報処理, Vol. 53, No. 9, pp. 932–937, 2012.

[2] 藏本貴久, 和泉潔, 吉村忍, 石田智也, 中嶋啓浩, 松井藤五郎, 吉田稔, 中川裕志. 新聞記事のテキストマイニングによる長期市場動向の分析. 人工知能学会論文誌, Vol. 28, No. 3, pp. 291–296, 2013.

[3] Simeon Schüz and Sina Zarrieß. Knowledge supports visual language grounding: a case study on colour terms. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6536–6542, 2020.

[4] 山本零, 川代尚哉, 栗田昌孝. 決算短信と四季報テキスト情報の投資戦略への利用可能性検証. ジャフィー・ジャーナル, Vol. 18, pp. 46–62, 2020.

[5] 伊藤健, 佐藤広大. 金融イノベーション資産運用におけるオルタナティブ・データ活用の可能性と課題. 野村資本市場クォータリー, Vol. 23, No. 2, pp. 136–153, 2019.

[6] 吉野貴晶. オルタナティブデータを使った運用実務について: ストラテジストレポートの自然言語処理と環境指標. 資本市場, No. 412, pp. 58–69, 2019.

[7] 鳥海不二夫, 榊剛史, 吉田光男.

[8] 平尾努, 磯崎秀樹, 前田英作, 松本裕治. Support vector machine を用いた重要文抽出法. 情報処理学会論文誌, Vol. 44, No. 8, pp. 2230–2243, 2003.

[9] 北森詩織, 酒井浩之, 坂地泰紀. 決算短信 pdf からの業績予測文の抽出. 電子情報通信学会論文誌 D, Vol. J100-D, No. 2, pp. 150–161, 2017.

[10] Shiori Kitamori, Hiroyuki Sakai, and Hiroki Sakaji. Extraction of sentences concerning business performance forecast and economic forecast from summaries of financial statements by deep learning. In *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 67–73, 2017.

[11] 酒井浩之, 西沢裕子, 松並祥吾, 坂地泰紀. 企業の決算短信 pdf からの業績要因の抽出. 人工知能学会論文誌, Vol. J98-D, No. 5, pp. 172–182, 2015.

[12] 酒井浩之, 松下和暉, 北島良三. 学習データの自動生成による決算短信からの業績要因文の抽出. 日本知能情報ファジィ学会誌, Vol. 31, No. 2, pp. 653–661, 2019.

[13] 川口敏広, 松井藤五郎, 大和田勇人. 二段階アプローチによる weblog からの意見文抽出. 電子情報通信学会技術研究報告, pp. 49–54, 2007.

[14] Moshe Koppel and Itai Shtrimberg. Good news or bad news? let the market

decide. In *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pp. 86–88, 2004.

[15] Yuancheng Li, Xiangqian Nie, and Rong Huang. Web spam classification method based on deep belief networks. *Expert Systems with Applications*, Vol. 96, pp. 261–270, 2018.

[16] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 7370–7377, 2019.

[17] Erik F. Tjong Kim Sang and Jorn Veenstra. Representing text chunks. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 173–179, 1999.

[18] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, 2019.

[19] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*, pp. 147–155, 2009.

[20] Hong-Jie Dai, Po-Ting Lai, Yung-Chun Chang, and Richard Tzong-Han Tsai. Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *Journal of Cheminformatics*, Vol. 7(Suppl 1):S14, , 2015.

[21] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of NAACL 2016*, pp. 260–270, 2016.

[22] 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 44–53, 2002.

[23] 増村亮, 田中智大, 安藤厚志, 神山歩相名, 大庭隆伸, 青野裕司. 対話コンテキストを考慮したニューラル通話シーン分割. Technical report, 2019.

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

[25] 但馬康宏, 北出大蔵, 中野未知子, 中林智, 藤本浩司, 小谷善行. Hmm とテキスト分類器による対話の段落分割. 情報処理学会論文誌数理モデル化と応用（TOM）, Vol. 2,

No. 2, pp. 70–79, 2009.

[26] 但馬康宏. 分割位置を教師値としたテキストの段落分割. Technical Report 15, 岡山県立大学情報システム工学科, 2011.

[27] 泉春乃, 加藤昇平. One-versus-all と attention 機構を取り入れた rnn による対話行為推定. 電気学会論文誌C（電子・情報・システム部門誌）, Vol. 139, No. 12, pp. 1407–1414, 2019.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[29] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, Vol. 32, pp. 5753–5763, 2019.

[30] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: a lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[31] Masala Mihai, Ruseti Stefan, and Dascalu Mihai. RoBERT – a romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6626–6637, 2020.

[32] Eric Arazo, Diego Ortego, Paul Albert, Noel E.O' Connor, Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020.

[33] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, Vol. 6, No. 60, 2019.

[34] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6382–6388, 2019.

[35] 近藤浩史, 與五澤守, 成瀬道紀, 森正和. 金融機関のテキストデータを活用した景気セ

ンチメントの計測. 第 33 回人工知能学会全国大会, pp. 1P2–J–13–02, 2019.

[36] 和泉潔, 後藤卓, 松井藤五郎. 経済テキスト情報を用いた長期的な市場動向推定. 情報処理学会論文誌, Vol. 52, No. 12, pp. 3309–3315, 2011.

[37] Viorel Milea, Nurfadhlina Mohd Sharef, Rui J. Almeida, Uzay Kaymak, and Flavius Frasincer. Prediction of the msci euro index based on fuzzy grammar fragments extracted from european central bank statements. In *International Conference of Soft Computing and Pattern Recognition*, pp. 231–236, 2010.

[38] Victor Lavrenko, Matt Schmill, Dawn Lawrie, Paul Ogilvie, David Jensen, and James Allan. Mining of concurrent text and time series. In *Proceedings of the KDD 2000 Conference Text Mining Workshop*, pp. 37–44, 2000.

[39] 松田安咲子, 岡石一真, 白田由香利, 橋本隆子, 佐倉環. Lda 方式による金融経済月報からのトピック抽出 〜第 2 次安倍内閣の金融政策と経済動向分析 〜. 信学技報, Vol. 114, No. 204, pp. 79–84, 2014.

[40] 竹内広宜, 荻野紫穂, 渡辺日出雄. テキストマイニングによる倒産企業分析. 経営情報学会 2008 年春季全国研究発表大会, pp. 124–127, 2008.

[41] Hiroki Sakaji, Hiroyuki Sakai, and Shigeru Masuyama. Automatic extraction of basis expressions that indicate economic trends. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 977–984, 2008.

[42] 五島圭一, 高橋大志. 株式価格情報を用いた金融極性辞書の作成. 自然言語処理, Vol. 2, pp. 547–577, 2017.

[43] Hiroyuki Sakai and Shigeru Masuyama. Cause information extraction from financial articles concerning business performance. *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968, 2008.

[44] Hiroyuki Sakai and Shigeru Masuyama. Assigning polarity to causal information in financial articles on business performance of companies. *IEICE Trans. Information and Systems*, Vol. E92-D, No. 12, pp. 2341–2350, 2009.

[45] 酒井浩之, 増山繁. 企業の業績発表記事からの重要業績要因の抽出. 電子情報通信学会論文誌 D, Vol. J96-D, No. 11, pp. 2866–2870, 2013.

[46] 坂地泰紀, 酒井浩之, 増山繁. 決算短信 pdf からの原因・結果表現の抽出. 電子情報通信学会論文誌 D, Vol. J98-D, No. 5, pp. 811–822, 2015.

[47] 室野莉沙, 酒井浩之, 坂地泰紀, ベネットジェイスン. 決算短信から抽出した原因・結果表現の意外性の判定. 第 11 回テキストアナリティクス・シンポジウム, pp. 87–91, 2017.

[48] Hiroki Sakaji, Risa Murono, Hiroyuki Sakai, Jason Bennett, and Kiyoshi Izumi. Discovery of rare causal knowledge from financial statement summaries. In *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 602–608, 2017.

[49] 田中瑞竜, 酒井浩之, 坂地泰紀. 複数の顧客企業からの共通要素と新規関連企業の抽出. 言語処理学会第 23 回年次大会, pp. 1192–1195, 2017.

[50] 酒井浩之, 松下和暉. 決算短信からの業績要因文の抽出. 第 11 回テキストアナリティクス・シンポジウム, pp. 87–91, 2017.

[51] 村野壮人, 酒井浩之, 坂地泰紀, 江口潤一. 決算短信から抽出した業績要因文の事業セグメントに基づく分類と業績文の抽出. 第 19 回金融情報学研究会, pp. 59–64, 2017.

[52] Jaap Kamps, Marijn Koolen, and Mounia Lalmas. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 723–724, 2007.

[53] Ronen Feldman and James Sanger. *The Text Mining Handbook*. Cambridge University Press, 2007.

[54] Hao Chen and Tin K. Ho. Evaluation of decision forests on text categorization. In *Proceedings of the 7th SPIE Conference on Document Recognition and Retrieval*, pp. 191–199, 2000.

[55] Hang Li and Kenji Yamanishi. Text classification using ESC-based stochastic decision lists. *Information Proceedings and Management*, Vol. 38, No. 3, pp. 343–361, 2017.

[56] Jian Zhang and Yiming Yang. Robustness of regularized linear classification methods in text categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in informaion retrieval*, pp. 190–197, 2003.

[57] Giorgio Fumera and Fabio Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 6, pp. 942–956, 2005.

[58] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323, 2011.

[59] Vladimir Naumovich Vapnik. *Statistical Learning Theory*. Wiley, 1999.

[60] Emmanuel Lesaffre and Andrew B. Lawson. *Bayesian Biostatistics*. John Wiley & Sons, 2012.

[61] 高野海斗, 酒井浩之, 中川慧. テキストマイニングを用いた株主招集通知の重要ページ抽出. 言語処理学会第 26 回年次大会, pp. 565–568, 2020.

[62] 酒井浩之, 梅村祥之, 増山繁. 交通事故事例に含まれる事故原因表現の新聞記事からの抽出. 自然言語処理, Vol. 13, No. 2, pp. 99–123, 2006.

[63] 酒井浩之, 柴田宏樹, 平松賢士, 坂地泰紀. アナリストレポートからのアナリスト予想根拠情報の抽出. 第 17 回金融情報学研究会, pp. 25–30, 2016.

[64] 酒井浩之, 坂地泰紀, 和泉潔, 松井藤五郎, 入江圭太郎. 経済テキストからの市況分析コメントの自動生成. 第 20 回金融情報学研究会, pp. 44–49, 2018.

[65] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69, 2002.

[66] 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.

[67] Taku Kudo and Yuji Matsumoto. Fast methods for kernel-based text analysis. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, p. 24–31, 2003.

[68] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, No. 5, pp. 101–123, 2007.

[69] 伝康晴. 多様な目的に適した形態素解析システム用電子化辞書. 人工知能学会誌, Vol. 24, pp. 640–646, 2009.

[70] 岡照晃. Crf 素性テンプレートの見直しによるモデルサイズを軽量化した解析用 unidic ― unidic-cwj-2.2.0 と unidic-csj-2.2.0 ―. 言語資源活用ワークショップ 2017 発表予稿集, pp. 143–152, 2017.

[71] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2246–2249, 2018.

[72] Geoffrey E Hinton, James L McClelland, and David E Rumelhart. *Parallel Distributed Processing*, Vol. 1. MIT Press, 1986.

[73] 岡崎直観. 言語処理における分散表現学習のフロンティア. 人工知能, Vol. 31, No. 2,

pp. 189–201, 2016.

[74] John A. Bullinaria and Joseph P. Levy. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, Vol. 39, No. 3, pp. 510–526, 2007.

[75] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[76] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*, 2013.

[77] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[78] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[79] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, 2018.

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[81] David Meir Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, No. 1, p. 993–1022, 2003.

[82] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, Vol. 521, No. 7553, pp. 436–444, 2015.

[83] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 655–665, 2014.

[84] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks.

IEEE Transactions on Signal Processing, Vol. 45, No. 11, pp. 2673–2681, 1997.

[85] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, Vol. 4, pp. 2047–2052, 2005.

[86] Alex Graves, Navdeep Jaitly, and Abdel rahman Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 273–278, 2013.

[87] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 606–615, 2016.

研究業績

■ 査読付き学術論文

1. 高野 海斗, 酒井 浩之, 中川 慧, ”学習データの自動生成による深層学習を用いた株主招集通知の重要ページ抽出”, 人工知能学会論文誌, Vol.36, No.1, pp. WI2-06, 2021.
2. Kaito Takano, Miryu Tanaka, Hiroyuki Sakai, Ryozo Kitajima, Takahisa Ota, Chinatsu Tanabe, Hiroki Sakaji, ”Extracting characteristic terms from patent documents”, International Journal of Smart Computing and Artificial Intelligence, Vol.4, No.2, pp. 19-38, 2020.
3. 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, pp. wd-A_1-22, 2019.
4. 高野 海斗, 酒井 浩之, 坂地 泰紀, 和泉 潔, 岡田 奈奈, 水内 利和, ”株主招集通知における議案タイトルとその分類及び開始ページの推定システム”, 言語処理学会論文誌, Vol.1, No.25, pp.3-31, 2018.

■ 国際会議論文

1. Kaito Takano, Miryu Tanaka, Hiroyuki Sakai, Ryozo Kitajima, Takahisa Ota, Chinatsu Tanabe, Hiroki Sakaji, ”Extraction of characteristic terms from patent documents for technical trend analysis”, 7th International Conference on Smart Computing and Artificial Intelligence (SCAI 2019), Japan, 2019.7.11.
2. Fumiya Sano, Kaito Takano, Shintaro Tomatsu, Manabu Iwasaki, ”Prediction by regression models with missing in covariates”, Conference of the International Federation of Classification Societies ,Tokai University, Japan, 2017.8.9.
3. Fumiya Sano, Kaito Takano, Shintaro Tomatsu, Manabu Iwasaki, ”Assessment of statistical power in relation to distance between propensity-score matched samples”, The 10th ICSA International Conference: Global Growth of Modern Statistics in the 21st Century, Shanghai Jiao Tong University, China, 2016.12.19.

■ 国内学会 口頭発表

1. 今井 康太, 酒井 浩之, 高野 海斗, 北島 良三, 末廣 徹, 稲垣 真太郎, 木村 柚里, ”債券市場における金融極性辞書の自動構築”, 第 25 回金融情報学研究会, 8, pp.38-43, オンライン, 2020.10.10.
2. 菅原 佑太, 高野 海斗, 酒井 浩之, ”大学医学部 Web サイトからの医療技術文の抽出”, 第 16 回テキストアナリティクス・シンポジウム, 2, pp.7-10, オンライン, 2020.9.10.
3. 高野 海斗, 酒井 浩之, 中川 慧, ”深層学習を用いた株主招集通知の重要ページ抽出”, 第 34 回人工知能学会全国大会, 1E4-GS-9-03, (pp.301-304), 熊本城ホール・熊本市民会館（オンライン）, 2020.6.9.
4. 高野 海斗, 酒井 浩之, 中川 慧, ”テキストマイニングを用いた株主招集通知の重要ページ抽出”, 言語処理学会第 26 回年次大会, B3-3, pp.565-568, 茨城大学（オンライン）, 2020.3.18.
5. 今井 康太, 高野 海斗, 酒井 浩之, ”業績要因を用いた決算短信のタイトル自動生成”, 言語処理学会第 26 回年次大会, B2-4, pp.417-420, 茨城大学（オンライン）, 2020.3.17.
6. 神田 裕輝, 高野 海斗, 酒井 浩之, 北島 良三, 中川 慧, ”新興市場を対象とした市況情報の抽出”, 第 24 回金融情報学研究会, 37, pp.219-225, 成蹊大学, 2020.3.15.
7. 關 涼介, 高野 海斗, 酒井 浩之, 北島 良三, ”判例テキストデータを用いた類似判例の自動抽出”, 第 15 回テキストアナリティクス・シンポジウム, pp.13-17, フューチャー株式会社, 2019.9.27.
8. 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ごとの業績要因・業績結果文の抽出”, 第 21 回金融情報学研究会, 13, pp.61-65, 東京大学, 2018.10.20.
9. 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ごとの業績要因文の抽出”, 第 13 回テキストアナリティクス・シンポジウム，pp.109-112, 成蹊大学，2018.9.7.
10. 高野 海斗, 岩崎 学, ”集計データの統計解析～選挙データへの適用～”, 日本計算機統計学会第 31 回シンポジウム，和歌山県立医科大学, 2017.11.16.
11. 岩崎 学, 関口 則子, 高野 海斗, ”集計データの統計解析 エコロジカル回帰の適用”, 日本行動計量学会第 45 回大会, 静岡県立大学, 2017.9.1.

12. 高野 海斗, 酒井 浩之, 坂地 泰紀, 和泉 潔, 岡田 奈奈, 水内 利和, ”株主招集通知における議案別の開始ページの推定”, 第 18 回金融情報学研究会, 09, pp.65-69, FinGate（茅場町一丁目平和ビル）, 2017.3.10.
13. 岩崎 学, 高野 海斗, 戸松 真太朗, ”説明変数に欠測を含む回帰モデルによる予測”, 日本行動計量学会 第 44 回大会, CD1-4, 札幌学院大学, 2016.9.1.

■ 国内学会 ポスター発表

1. 高野 海斗, 岩崎 学, ”標本調査における ecological data の活用”, 日本統計学会 第 12 回春季集会, 31, 早稲田大学, 2018.3.4.
2. 松岡 英佑, 戸松 真太朗, 高野 海斗, 岩崎 学, ”セイバーメトリクスによる 2016 年ベストナインの評価”, シンポジウム「スポーツアナリティクスと統計科学」第 7 回スポーツデータ解析コンペティション審査会, 前 12, 統計数理研究所, 2017.12.23.
3. 高野 海斗, 岩崎学, ”欠測インディケーター法の予測におけるパフォーマンスの評価”, 日本統計学会 第 11 回春季集会, 33, 政策研究大学院大学, 2017.3.5.

付録

本論文を読むための基礎知識

自然言語処理（Natural Language Processing）とは，人間が使う言語である「自然言語」をコンピュータに処理させる一連の技術である．「機械翻訳（Machine Translation）」，「質問応答（Question Answering）」，「対話エージェント（Dialogue Agent）」，「自動要約（Automatic Text Summarization）」，「情報検索（Information Retrieval）」，「情報抽出（Information Extraction）」，「感情分析（Emotion Analysis）」などが代表的な研究として挙げられる．自然言語処理の分野は，深層学習（Deep Learning）の発展により，近年急激な成長を遂げている．

本節では，本論文で扱う既存の自然言語処理技術や機械学習手法を，近年の有名な研究を交えつつ簡潔に紹介する．

形態素解析

形態素解析（Morphological Analysis）は，文法や辞書などの情報に基づき，形態素（Morpheme）の列に分割し，それぞれの形態素の品詞等を判別する技術である．例えば，「メロンパンを食べました。」という文であれば，表 58 のように分割する．テキストデータをコンピュータで扱うためには，テキストデータを数値データに変換する必要があるが，テキストをそのまま数値化することはできないため，いろいろな手法が提案されてきた．それらの数値化は，テキストをまず形態素に分割するところから始まるといっても過言ではないため，形態素解析は自然言語処理において重要な技術である．

表 58　「メロンパンを食べました。」を形態素解析した結果の例

メロン	パン	を	食べ	まし	た	。
名詞	名詞	助詞	動詞	助動詞	助動詞	記号

英語などの言語は，予め単語と単語が空白で区切られているため，単語分割の処理は比較的容易である．それに対して，日本語のような明確な単語の区切りがない言語は分割が難しく，辞書に基づく品詞情報などを加味することで，適切な分割や品詞を推定する必要がある．

形態素解析のツールは，フリーのものから商用のものまで多種多様であり，使用する形

態素解析器や辞書によって形態素解析の結果が異なる．そのため，どの形態素解析器を使うかによって後の結果に影響を与えるため，用途によって使い分ける必要がある．本論文で扱うデータは日本語のテキストデータであるため，いくつかの代表的な日本語の形態素解析器を紹介する．

本論文で使用している形態素解析器は「MeCab」[*34]である．MeCab の特徴は，システムと辞書が分かれていることや，他の形態素解析器に比べて高速に処理を行うことができる点である．特に近年，テキストデータが大量に手に入りやすくなったことから大規模なテキストデータを扱う必要があるため，処理速度が高速な形態素解析器が人気である．また，システムと辞書が分かれているため，辞書を用途によって使い分けることができる．辞書には，「IPA」，「NEologd」[*35]，「Unidic」[*36][68, 69, 70] などがあり，扱うテキスト情報に合わせて使い分けることが重要である．例えば，「NEologd」の辞書は，週 2 回以上更新されており新語に対応することが可能である．先ほどの「メロンパンを食べました。」の形態素解析の例では，「メロン」と「パン」が分割されていたが，「NEologd」の辞書を使用した場合には，「メロンパン」で一つの名詞として分割される．特に，固有名詞が豊富に含まれており，例えば，グループ名，漫画のタイトル，登場キャラクター，さらにはそれらの略語なども辞書に含まれているため，SNS などのテキストデータを分析するときによく使用される．

それ以外の形態素解析器としては，京都大学の黒橋研が開発した「JUMAN」[*37]，ワークスアプリケーションズ徳島人工知能 NLP 研究所が開発した「Sudachi」[*38][71] などがある．Sudachi は，分割の長さを A 単位（Unidic 相当），B 単位（IPAdic 相当），C 単位（NEologd 相当）から選択することが可能であり，表記を正規化する機能なども存在する．

係り受け解析

係り受け解析（Parse）は，文節間の修飾関係，つまり「修飾する（係る）」と「修飾される（受ける）」の関係を明らかにするための解析である．係り受け解析の例を図 29 に示す．

係り受け解析は，形態素解析で分割するだけでは得ることができない情報を追加で取得でき，その情報を用いることで様々な応用的手法が可能となる．テキストマイニングで

[*34] https://taku910.github.io/mecab/
[*35] https://github.com/neologd/mecab-ipadic-neologd
[*36] https://unidic.ninjal.ac.jp/download#unidic_bccwj
[*37] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[*38] https://github.com/WorksApplications/Sudachi

図 29　係り受け解析の例

は，ある情報を抽出するために有効は手がかりとなる表現（以降，手がかり表現と定義）を使用して，テキストから情報を抽出することが多いが，本論文では，この手がかり表現を獲得するために係り受け解析を利用している．

代表的な日本語の係り受け解析器（Parser）としては，「CaboCha」[*39][65, 66, 67] がある．CaboCha は，SVM[59] に基づく日本語係り受け解析器であり，形態素解析器は MeCab を使用している．それ以外の係り受け解析器には，形態素解析器に Sudachi を使用している「GiNZA」[*40]などがある．GiNZA は日本語 NLP ライブラリであり，係り受け解析以外にも，文章を文単位に分割，固有表現の抽出，文を分散表現に変換する機能などを備えている．

分散表現

形態素解析の節で述べた通り，テキストデータをコンピュータで扱うためには，テキスト情報を数値情報に変換する必要がある．分散表現（Distributed Representation）は，単語，文，文章をベクトル空間（Vector Space）で表現する技術である [72, 73]．例えば，単語を分散表現で表すことにより，単語と単語の類似度を計測することなどが可能となる．どのような分散表現を用いるかによって最終的な結果に大きな影響を与えるため，自然言語処理においては形態素解析に次いで重要な技術である．

分散表現の対となる表現方法が，ある要素のみが 1 でその他の要素が 0 である「one-hot 表現」であり，これは局所表現（Local Representation）と呼ばれている．各次元に 1 か 0 を設定することで「その単語か否か」を表すことができる．one-hot 表現の例を，図 30 に示す．

しかし，局所表現である one-hot 表現は多くのデメリットを抱えている．まず，単語と単語の類似度を測ることができず，似たような単語，例えば「りんご」と「梨」，「バナナ」と「BANANA」のような単語が，まったく別の単語として扱われる．また，ボキャブラ

[*39] https://taku910.github.io/cabocha/
[*40] https://megagonlabs.github.io/ginza/

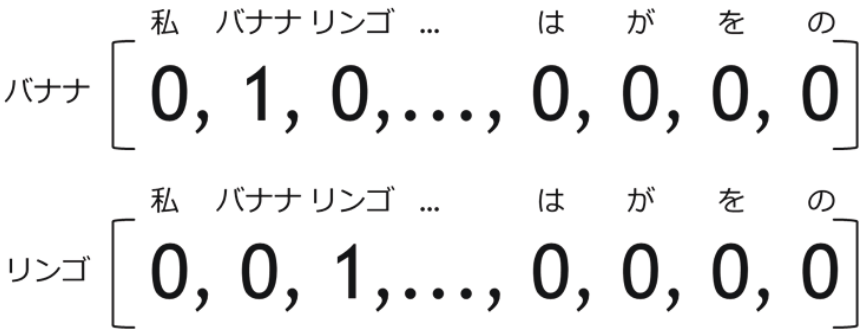


図 30　one-hot 表現の例

リーが増えるに伴って，次元数が非常に高次元になることや，1 要素を除いて残り全てが 0 なので非常にスパースなベクトルになるデメリットが存在する．そのため，この後に紹介するような様々な分散表現が提案されており，現在も多くの研究が行われている．本論文でも，様々な分散表現を使用しているため，いくつかの代表的な単語分散表現の獲得手法と，文や文章の分散表現の表現方法について紹介する．

簡単な方法として，単語の共起情報などを用いて単語文脈行列を作成することで分散表現を獲得する方法がある．各単語について，その周辺（例えば前後 k 語）に出現する単語を文脈語とし，共起出現頻度を用いて共起の強さを表現したものが，最も単純な単語文脈行列である．単語文脈行列の例を，図 31 に示す．

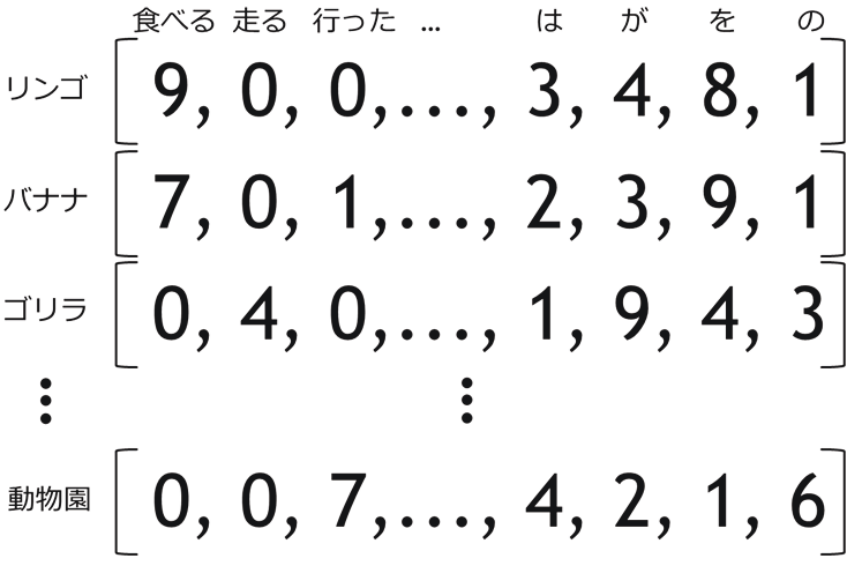


図 31　共起情報を用いた単語文脈行列の例

図 31 に示した単語文脈行列を用いることで，例えば，「りんご」であれば，[9, 0, 0,..., 3, 4, 8, 1] のような分散表現を得ることが可能となる．似た意味を持つ単語は，周辺に出現する文脈語が似たようなものになる傾向あるため，この分散表現を用いることで，単語間の距離を計測することなどが可能となる．

単語文脈行列は，様々な共起尺度が提案されてきた．例えば，正の自己相互情報量（PPMI）[74] は，頻出文脈語の影響を軽減することが可能である．しかし，これらの方法を用いて獲得できる分散表現の次元数は文脈語の数に等しく，非常に高次元であるというデメリットがある．そのため，単語文脈行列の情報をできるだけ保持しつつ，低次元密行列に圧縮する方法の研究も行われてきた．そのうちのひとつに，潜在的意味解析（Latent Semantic Analysis : LSA）[75] が挙げられる．LSA はまず，単語文脈行列に特異値分解（Singular Value Decomposition : SVD）を適用し，SVD で求まった特異値を影響度の少ない順に削っていくことにより次元の削減を行うことが可能である．

深層学習を用いた単語分散表現の研究が急激に盛んになるきっかけとなったのが，「word2vec[*41]」[76] である．word2vec は，CBOW（Continuous Bag-of-Word Model）と Skip-gram の二種類が提案されている．どちらも入力層，隠れ層，出力層の 3 層からなるニューラルネットワークで，隣接する層のノードは全結合となっている．

CBOW は，ターゲットとなる単語 $\{W_t\}$ を，その単語の周辺に出現する単語 $\{W_{t-c}, ..., W_{t+1}, W_{t-1}, ..., W_{t+c}\}$ を入力することで予測するモデル設計となっている．それに対して Skip-gram は，ある単語 $\{W_t\}$ を入力とし，その単語の周辺に出現する単語 $\{W_{t-c}, ..., W_{t+1}, W_{t-1}, ..., W_{t+c}\}$ を予測するモデル設計となっている．CBOW と Skip-gram のモデルを図 32 に示す．

どちらのモデルも，入力層から中間層への変換の重みを，その単語の分散表現として利用する．モデルを学習するためのデータは人手でのラベル付けが不要であるが，大量のテキストデータが必要である．また，どちらのモデルも context window（c）をハイパーパラメータとして決める必要がある．context window は，周辺の単語をどこまで学習に使用するか決めるためのものであり，context window が小さいほど，品詞情報を分散表現に反映させることができる．逆に，context window が大きいほど，トピックやドメイン情報が分散表現に反映されるため，用途に合わせて調整する必要がある．

word2vec は，ある単語に対して一意な分散表現を与えることしかできない，文脈依存なしの分散表現獲得方法である．例えば「バイト」のようなアルバイトの略語であり，かつ，情報量の単位でもあるような多義語である単語に対しても一意な分散表現で表すこと

[*41] https://code.google.com/archive/p/word2vec/

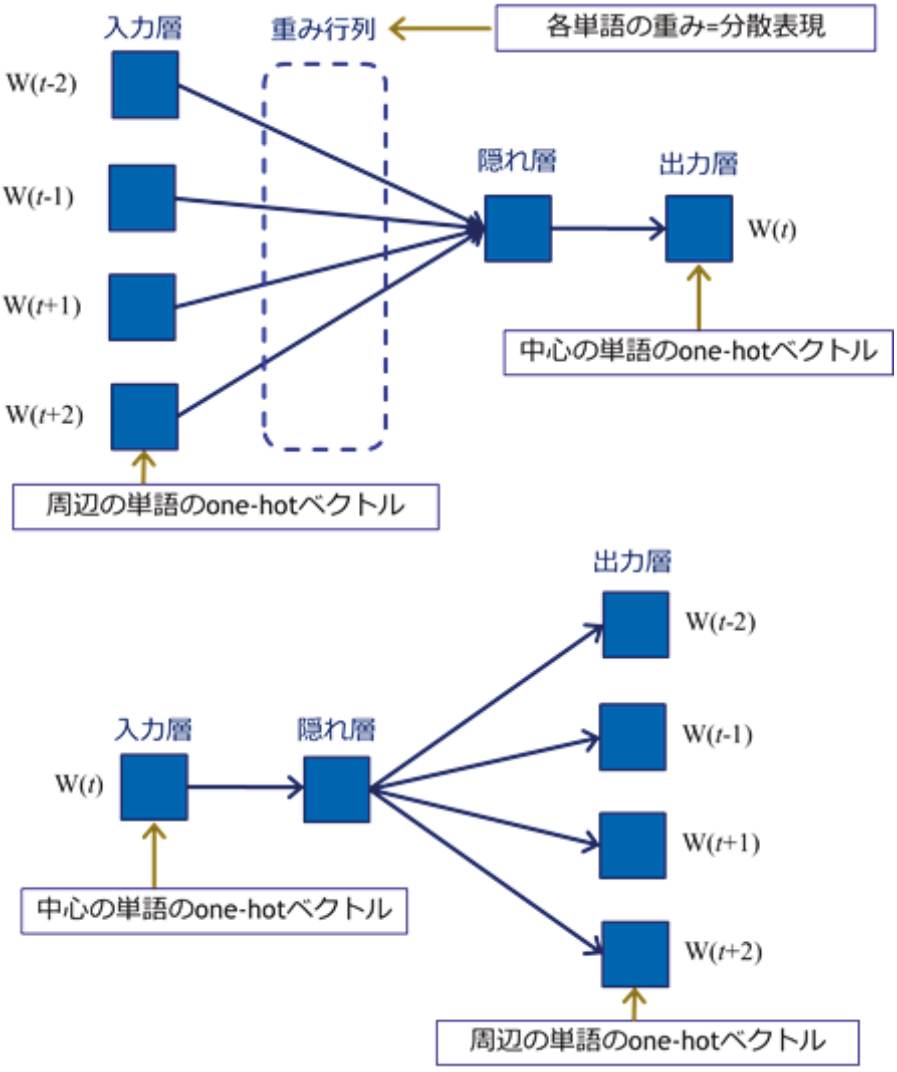


図 32　CBOW モデル（上）と Skip-gram モデル（下）

しかできないデメリットが存在する．word2vec 以降の文脈依存なしの分散表現獲得方法には，「GloVe（Global Vectors）」[77] や「FastText」[78] があるが，多義語の分散表現に問題が残る．

これらの問題を解決するために，文脈依存ありの手法の研究が，近年は盛んに行われている．文脈依存ありの手法で，最初に話題になったのは，「ELMo」[79] である．ELMo は，BiLSTM を用いることで，文字列ベースの特徴と，前後に出現する単語の情報を加味して分散表現を獲得する手法である．そして現在の自然言語処理の中心になっているのは，Attention[80] を使用したモデルをベースとした「BERT」[28] である．BERT は，翻訳や分類問題を解くためのアルゴリズムでもあるが，単語の分散表現としても利用可能で

ある．具体的には，BERT の中では Masked Language Model（MLM）という Mask された単語を推定するモデルがあり，その中で単語の分散表現を学習させている．これにより，単語の類似度を測るだけでなく，穴埋め問題を解くことや，固有表現抽出問題を解くことも可能である．

文や文章の分散表現方法は，単語の分散表現に対して多様な方法がある．一番シンプルなものは，その文中に出現する単語の頻度情報を要素とした分散表現である（Bag of Words）．他には，重要度の高い単語に高い重みを与えた TFidf 値などを要素とした分散表現，トピックモデル [81] によるトピック確率を用いた分散表現，出現単語の分散表現を平均（合計）した分散表現などがあり，目的によって使い分ける必要がある．

SVM

機械学習手法のひとつである SVM（Support Vector Machine）[59] は，1990 年代の終わり頃から各分野において爆発的に使われはじめた線形二値分類機である．SVM は，n 次元空間を二つに分離するための超平面をマージンが最大化するように決定する手法である．現在では，二値分類などは深層学習が主流であるが，係り受け解析で紹介した CaboCha は，SVM を使用して係り受け解析を実現している．

DNN, MLP

深層学習（deep learning）は，DNN（Deep Neural Network）や MLP（Multi Layer Perceptron）と呼ばれるニューラルネットワークモデルの中間層を増やしたモデルである [82]．モデル図を図 33 に示す．モデルの枠組みとしては，20 世紀のうちに開発されていたが，4 層以上の深層ニューラルネットについては局所最適解や勾配消失などの問題により実用化されていなかった．しかし，多層ニューラルネットワークの学習方法に関する研究，学習に必要な計算機の性能向上，および，インターネットの発展による学習データの増加などにより，抱えていた問題が解決され，現在のモデルの主流となっている．多層ニューラルネットワークの学習が可能になったことで，深層学習モデルを元にした多種多様なモデルが研究されており，例えば，画像の分野では CNN（Convolutional Neural Network）[83] のような畳み込みモデルがある．その他には，時系列や順序関係のあるデータを扱える RNN（Recurrent Neural Network）モデル [84]，Attention を組み込んだモデル [80] など，発展的な研究が現在も盛んに行われている．

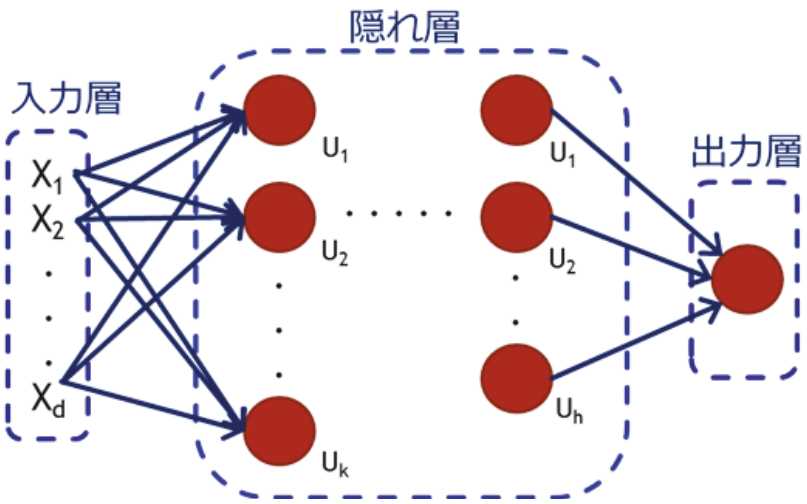


図 33　MLP のモデル図

LSTM

文に出現する単語の順序には，その順序に意味があるため，その出現順序を学習させることがより良い結果につながると考えられる．しかし，これまでの SVM や MLP などの入力には，単語の順序を考慮しない文の分散表現を入力にしていた．この単語の順序を考慮できるモデルとして，回帰型ニューラルネットワークモデルである LSTM（Long short-term memory）[24] がある．LSTM のモデル図を図 34 に示す．LSTM は入力と

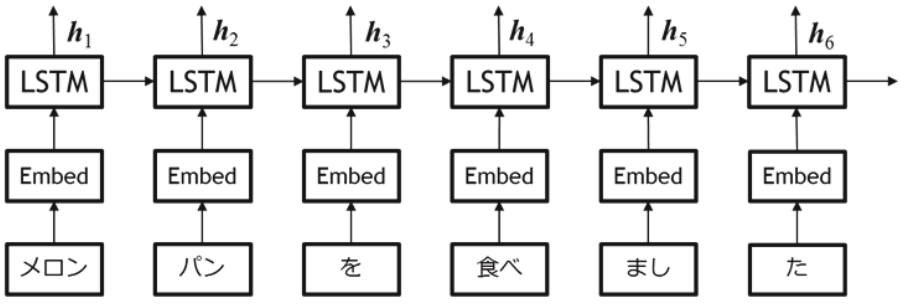


図 34　LSTM のモデル図

して単語の分散表現を順に受け取ることができるため，単語の出現順序によって出力層の結果が変わる特徴がある．また，単方向だけでなく逆方向の層を加えた双方向の LSTM モデルである BiLSTM（Bidirectional LSTM）[85, 86]，LSTM 層の出力に Attention を追加したモデル [87] など，様々な応用モデルも提案されている．

本論文では，3 章において，各ページの単語頻度を LSTM への入力とすることで，ページの分類を行うモデルを提案する．

4 章の詳細な評価結果

4 章の各企業の事業セグメントごとの適合率・再現率を，表 59 から表 62 に示す.

表 59　事業セグメントごとの業績要因文の適合率・再現率

企業名	企業コード	事業セグメント	$precision$	$recall$
カネコ種苗株式会社	E00004	種苗事業	5 / 6 = 0.833	5 / 5 = 1.000
		施設材事業	1 / 3 = 0.333	1 / 2 = 0.500
		農材事業	2 / 2 = 1.000	2 / 3 = 0.667
		花き事業	1 / 1 = 1.000	1 / 1 = 1.000
株式会社サカタのタネ	E00006	海外卸売事業	8 / 8 = 1.000	8 / 10 = 0.800
		国内卸売事業	3 / 3 = 1.000	3 / 5 = 0.600
		小売事業	4 / 4 = 1.000	4 / 6 = 0.667
		その他	1 / 1 = 1.000	1 / 1 = 1.000
日本工営株式会社	E00078	コンサルタント海外事業	1 / 2 = 0.500	1 / 2 = 0.500
		電力エンジニアリング事業	1 / 2 = 0.500	1 / 3 = 0.333
		都市空間事業	0 / 2 = 0.000	0 / 2 = 0.000
		コンサルタント国内事業	0 / 1 = 0.000	0 / 3 = 0.000
		エネルギー事業	0 / 1 = 0.000	0 / 0
コーセル株式会社	E01856	日本生産販売事業	1 / 2 = 0.500	1 / 1 = 1.000
		ヨーロッパ販売事業	1 / 2 = 0.500	1 / 1 = 1.000
		アジア販売事業	1 / 2 = 0.500	1 / 1 = 1.000
		中国生産事業	1 / 1 = 1.000	1 / 1 = 1.000
		北米販売事業	0 / 0	0 / 2 = 0.000
株式会社 IG ポート	E02480	出版事業	4 / 4 = 1.000	4 / 4 = 1.000
		映像制作事業	3 / 3 = 1.000	3 / 3 = 1.000
		版権事業	2 / 2 = 1.000	2 / 2 = 1.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
株式会社内田洋行	E02515	公共関連事業	3 / 3 = 1.000	3 / 3 = 1.000
		情報関連事業	2 / 2 = 1.000	2 / 3 = 0.667
		オフィス関連事業	1 / 1 = 1.000	1 / 2 = 0.500
		その他	1 / 1 = 1.000	1 / 1 = 1.000
テーオーホールディングス	E03169	自動車関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		流通事業	1 / 1 = 1.000	1 / 1 = 1.000
		住宅事業	1 / 1 = 1.000	1 / 1 = 1.000
		木材事業	1 / 1 = 1.000	1 / 2 = 0.500
		建設事業	0 / 0	0 / 1 = 0.000
株式会社サンオータス	E03326	カービジネス事業	4 / 7 = 0.571	4 / 5 = 0.800
		エネルギー事業	3 / 6 = 0.500	3 / 3 = 1.000
		不動産関連事業	1 / 3 = 0.333	1 / 1 = 1.000
		ライフサポート事業	1 / 2 = 0.500	1 / 2 = 0.500
リベレステ株式会社	E03989	開発事業	2 / 2 = 1.000	2 / 2 = 1.000
		不動産販売事業	1 / 2 = 0.500	1 / 1 = 1.000
		建築事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	2 / 2 = 1.000	2 / 2 = 1.000
明豊エンタープライズ	E04024	不動産分譲事業	2 / 3 = 0.667	2 / 3 = 0.667
		不動産賃貸事業	2 / 2 = 1.000	2 / 2 = 1.000
		不動産仲介事業	0 / 1 = 0.000	0 / 1 = 0.000
		請負事業	0 / 1 = 0.000	0 / 1 = 0.000
		その他	0 / 0	0 / 1 = 0.000
株式会社ゼロ	E04230	自動車関連事業	3 / 3 = 1.000	3 / 4 = 0.750
		ヒューマンリソース事業	1 / 2 = 0.500	1 / 2 = 0.500
		一般貨物事業	1 / 2 = 0.500	1 / 1 = 1.000

表 60　事業セグメントごとの業績要因文の適合率・再現率

企業名	企業コード	事業セグメント	$precision$	$recall$
東京博善株式会社	E04843	堀ノ内斎場	1 / 1 = 1.000	1 / 2 = 0.500
		町屋斎場	1 / 2 = 0.500	1 / 2 = 0.500
		代々幡斎場	1 / 2 = 0.500	1 / 2 = 0.500
		落合斎場	1 / 1 = 1.000	1 / 2 = 0.500
		四ツ木斎場	1 / 2 = 0.500	1 / 2 = 0.500
		桐ヶ谷斎場	0 / 0	0 / 2 = 0.000
		その他	0 / 5 = 0.000	0 / 0
日本プロセス株式会社	E04873	制御システム	4 / 4 = 1.000	4 / 5 = 0.800
		組込システム	2 / 2 = 1.000	2 / 3 = 0.667
		特定情報システム	2 / 4 = 0.500	2 / 2 = 1.000
		自動車システム	2 / 2 = 1.000	2 / 2 = 1.000
		産業・公共システム	3 / 3 = 1.000	3 / 4 = 0.750
		ＩＴサービス	3 / 3 = 1.000	3 / 3 = 1.000
株式会社ビューティ花壇	E05597	生花卸売事業	3 / 3 = 1.000	3 / 3 = 1.000
		ブライダル装花事業	2 / 2 = 1.000	2 / 2 = 1.000
		生花祭壇事業	0 / 1 = 0.000	0 / 1 = 0.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
フリービット株式会社	E05680	アドテクノロジー事業	1 / 2 = 0.500	1 / 1 = 1.000
		ヘルステック事業	1 / 1 = 1.000	1 / 1 = 1.000
		ブロードバンド事業	1 / 1 = 1.000	1 / 1 = 1.000
		モバイル事業	1 / 2 = 0.500	1 / 1 = 1.000
		クラウド事業	0 / 1 = 0.000	0 / 0
株式会社インサイト	E05740	介護福祉事業	3 / 3 = 1.000	3 / 4 = 0.750
		広告・マーケティング事業	3 / 4 = 0.750	3 / 3 = 1.000
		ケアサービス事業	2 / 4 = 0.500	2 / 2 = 1.000
		債権投資事業	0 / 0	0 / 2 = 0.000
		その他	0 / 1 = 0.000	0 / 0
日本海洋掘削株式会社	E23800	掘削技術	2 / 3 = 0.667	2 / 2 = 1.000
		運用・管理受託	2 / 3 = 0.667	2 / 2 = 1.000
		海洋掘削	2 / 10 = 0.200	2 / 2 = 1.000
		その他	0 / 3 = 0.000	0 / 0
株式会社ＴＨＥグローバル社	E24340	ホテル事業	2 / 2 = 1.000	2 / 3 = 0.667
		戸建事業	1 / 1 = 1.000	1 / 1 = 1.000
		販売代理事業	1 / 1 = 1.000	1 / 1 = 1.000
		マンション事業	1 / 2 = 0.500	1 / 1 = 1.000
		建物管理事業	1 / 1 = 1.000	1 / 1 = 1.000
三協立山株式会社	E26831	建材事業	1 / 1 = 1.000	1 / 1 = 1.000
		マテリアル事業	1 / 1 = 1.000	1 / 2 = 0.500
		国際事業	1 / 1 = 1.000	1 / 2 = 0.500
		商業施設事業	1 / 2 = 0.500	1 / 2 = 0.500
タマホーム株式会社	E27305	住宅事業	5 / 5 = 1.000	5 / 7 = 0.714
		不動産事業	2 / 2 = 1.000	2 / 2 = 1.000
		金融事業	2 / 2 = 1.000	2 / 4 = 0.500
		エネルギー事業	1 / 1 = 1.000	1 / 2 = 0.500
		その他	1 / 1 = 1.000	1 / 2 = 0.500
ウエスコホールディングス	E30042	複写製本事業	1 / 2 = 0.500	1 / 2 = 0.500
		スポーツ施設運営事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産事業	1 / 1 = 1.000	1 / 2 = 0.500
		指定管理事業	0 / 1 = 0.000	0 / 0
		総合建設コンサルタント事業	0 / 2 = 0.000	0 / 0

表 61　事業セグメントごとの業績結果文の適合率・再現率

企業名	企業コード	事業セグメント	$precision$	$recall$
カネコ種苗株式会社	E00004	種苗事業	1 / 1 = 1.000	1 / 1 = 1.000
		施設材事業	2 / 4 = 0.500	2 / 2 = 1.000
		農材事業	2 / 2 = 1.000	2 / 2 = 1.000
		花き事業	1 / 1 = 1.000	1 / 1 = 1.000
		造園事業	1 / 1 = 1.000	1 / 1 = 1.000
株式会社サカタのタネ	E00006	海外卸売事業	1 / 1 = 1.000	1 / 1 = 1.000
		国内卸売事業	1 / 1 = 1.000	1 / 1 = 1.000
		小売事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
日本工営株式会社	E00078	コンサルタント海外事業	1 / 1 = 1.000	1 / 1 = 1.000
		電力エンジニアリング事業	1 / 1 = 1.000	1 / 1 = 1.000
		都市空間事業	1 / 1 = 1.000	1 / 1 = 1.000
		コンサルタント国内事業	1 / 1 = 1.000	1 / 1 = 1.000
		エネルギー事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産賃貸事業	1 / 1 = 1.000	1 / 1 = 1.000
コーセル株式会社	E01856	日本生産販売事業	1 / 1 = 1.000	1 / 1 = 1.000
		ヨーロッパ販売事業	1 / 1 = 1.000	1 / 1 = 1.000
		アジア販売事業	1 / 1 = 1.000	1 / 1 = 1.000
		中国生産事業	1 / 1 = 1.000	1 / 1 = 1.000
		北米販売事業	1 / 1 = 1.000	1 / 1 = 1.000
株式会社 IG ポート	E02480	出版事業	1 / 1 = 1.000	1 / 1 = 1.000
		映像制作事業	1 / 1 = 1.000	1 / 1 = 1.000
		版権事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
株式会社内田洋行	E02515	公共関連事業	2 / 2 = 1.000	2 / 2 = 1.000
		情報関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		オフィス関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
テーオーホールディングス	E03169	自動車関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		流通事業	1 / 1 = 1.000	1 / 1 = 1.000
		住宅事業	1 / 1 = 1.000	1 / 1 = 1.000
		木材事業	1 / 1 = 1.000	1 / 1 = 1.000
		建設事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産賃貸事業	1 / 1 = 1.000	1 / 1 = 1.000
		スポーツクラブ事業	1 / 1 = 1.000	1 / 1 = 1.000
株式会社サンオータス	E03326	カービジネス事業	1 / 1 = 1.000	1 / 1 = 1.000
		エネルギー事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		ライフサポート事業	1 / 1 = 1.000	1 / 1 = 1.000
リベレステ株式会社	E03989	開発事業	1 / 1 = 1.000	1 / 2 = 0.500
		不動産販売事業	2 / 2 = 1.000	2 / 2 = 1.000
		建築事業	2 / 2 = 1.000	2 / 2 = 1.000
		その他	2 / 5 = 0.400	2 / 2 = 1.000
明豊エンタープライズ	E04024	不動産分譲事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産賃貸事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産仲介事業	1 / 1 = 1.000	1 / 1 = 1.000
		請負事業	1 / 2 = 0.500	1 / 1 = 1.000
		その他	1 / 6 = 0.167	1 / 1 = 1.000

表 62　事業セグメントごとの業績結果文の適合率・再現率

企業名	企業コード	事業セグメント	$precision$	$recall$
株式会社ゼロ	E04230	自動車関連事業	1 / 1 = 1.000	1 / 1 = 1.000
		ヒューマンリソース事業	1 / 1 = 1.000	1 / 1 = 1.000
		一般貨物事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	0 / 1 = 0.000	0 / 0
東京博善株式会社	E04843	堀ノ内斎場	3 / 3 = 1.000	3 / 3 = 1.000
		町屋斎場	3 / 3 = 1.000	3 / 3 = 1.000
		代々幡斎場	2 / 2 = 1.000	2 / 4 = 0.500
		落合斎場	3 / 3 = 1.000	3 / 3 = 1.000
		四ツ木斎場	3 / 6 = 0.500	3 / 3 = 1.000
		桐ヶ谷斎場	0 / 0	0 / 3 = 0.000
		その他	0 / 4 = 0	0 / 0
日本プロセス株式会社	E04873	制御システム	1 / 1 = 1.000	1 / 1 = 1.000
		組込システム	1 / 1 = 1.000	1 / 1 = 1.000
		特定情報システム	1 / 5 = 0.200	1 / 1 = 1.000
		自動車システム	1 / 1 = 1.000	1 / 1 = 1.000
		産業・公共システム	1 / 1 = 1.000	1 / 1 = 1.000
		ＩＴサービス	1 / 1 = 1.000	1 / 1 = 1.000
株式会社ビューティ花壇	E05597	生花卸売事業	2 / 2 = 1.000	2 / 2 = 1.000
		ブライダル装花事業	2 / 2 = 1.000	2 / 2 = 1.000
		生花祭壇事業	1 / 1 = 1.000	1 / 2 = 0.500
		その他	1 / 1 = 1.000	1 / 1 = 1.000
フリービット株式会社	E05680	アドテクノロジー事業	1 / 1 = 1.000	1 / 1 = 1.000
		ヘルステック事業	1 / 1 = 1.000	1 / 1 = 1.000
		ブロードバンド事業	1 / 1 = 1.000	1 / 1 = 1.000
		モバイル事業	1 / 1 = 1.000	1 / 1 = 1.000
		クラウド事業	1 / 1 = 1.000	1 / 1 = 1.000
株式会社インサイト	E05740	介護福祉事業	1 / 1 = 1.000	1 / 1 = 1.000
		広告・マーケティング事業	1 / 2 = 0.500	1 / 1 = 1.000
		ケアサービス事業	1 / 2 = 0.500	1 / 1 = 1.000
		債権投資事業	0 / 0	0 / 1 = 0.000
		その他	0 / 4 = 0.000	0 / 0
日本海洋掘削株式会社	E23800	掘削技術	1 / 1 = 1.000	1 / 2 = 0.500
		運用・管理受託	1 / 1 = 1.000	1 / 2 = 0.500
		海洋掘削	1 / 1 = 1.000	1 / 2 = 0.500
		その他	1 / 3 = 0.333	1 / 1 = 1.000
株式会社ＴＨＥグローバル社	E24340	ホテル事業	2 / 2 = 1.000	2 / 2 = 1.000
		戸建事業	2 / 2 = 1.000	2 / 2 = 1.000
		販売代理事業	2 / 2 = 1.000	2 / 2 = 1.000
		マンション事業	2 / 2 = 1.000	2 / 2 = 1.000
		建物管理事業	2 / 2 = 1.000	2 / 2 = 1.000
		その他	2 / 2 = 1.000	2 / 2 = 1.000
三協立山株式会社	E26831	建材事業	1 / 1 = 1.000	1 / 1 = 1.000
		マテリアル事業	2 / 2 = 1.000	2 / 2 = 1.000
		国際事業	1 / 1 = 1.000	1 / 2 = 0.500
		商業施設事業	2 / 4 = 0.500	2 / 2 = 1.000
タマホーム株式会社	E27305	住宅事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産事業	1 / 1 = 1.000	1 / 1 = 1.000
		金融事業	1 / 1 = 1.000	1 / 1 = 1.000
		エネルギー事業	1 / 1 = 1.000	1 / 1 = 1.000
		その他	1 / 1 = 1.000	1 / 1 = 1.000
ウエスコホールディングス	E30042	複写製本事業	1 / 1 = 1.000	1 / 1 = 1.000
		スポーツ施設運営事業	1 / 1 = 1.000	1 / 1 = 1.000
		不動産事業	1 / 1 = 1.000	1 / 1 = 1.000
		指定管理事業	1 / 1 = 1.000	1 / 1 = 1.000
		総合建設コンサルタント事業	1 / 1 = 1.000	1 / 1 = 1.000