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

Extraction of Important Pages of Shareholder Convocation Notices Using Deep Learning by Automatic Generation of Training Data

高野 海斗
Kaito Takano

成蹊大学大学院 理工学研究科 理工学専攻
Graduate School of Science and Technology, Seikei University
ddl86201@cc.seikei.ac.jp

酒井 浩之
Hiroyuki Sakai

成蹊大学
Seikei University
h-sakai@st.seikei.ac.jp

中川 慧
Kei Nakagawa

野村アセットマネジメント株式会社
Nomura Asset Management Co., Ltd.
k-nakagawa@nomura-am.co.jp

keywords: deep learning, information extraction, text mining

Summary

A shareholder convocation notice is a letter that the company is obliged to send to shareholders when holding a shareholder's meeting. We can access it on corporate websites and acquire as PDF file. It contains a lot of useful information, such as company profile, major shareholders, and bills to be discussed. Therefore, institutional investors often use that information in their investment decisions. However, the following challenges exist for institutional investors to extract information that is likely to affect the stock price. The number of pages ranges from more than a dozen to more than a hundred. In addition, since they are issued before a shareholder's meetings, they are issued in large numbers in a particular month, i.e., thousands of company notices are issued in June, the most concentrated month. This is a significant burden for institutional investors.

The purpose of our research is to automatically extract pages that are likely to affect the stock price from shareholder convocation notices. To this end, we need to tag the pages to automatically extract what information is described on a page-by-page basis. In our research, we propose the following framework: We automatically create training data by a rule-based method and train the deep learning model that extracts important pages. We confirm the effectiveness of our framework for pages that cannot be extracted by the rule-based method.

1. はじめに

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应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 [和泉 11, 蔵本 13, 酒井 13]。

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

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^{*1}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

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) 分類器を十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した。

2. 株主招集通知とは

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^{*2}。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およ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程度のものもあれば、100 ページを超えるものもある。

3. 関 連 研 究

本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

これらの点を踏まえながら関連研究を挙げ、本研究の新規性および有用性を示す。

株主招集通知を対象とした研究として [高野 18] がある。この研究では、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。本研究では、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、[高野 18] 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。

[高野 20] では、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っている。しかし、人間が確認

^{*1} 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。

^{*2} なお、株主招集通知は公開情報であり、株主だけでなく誰でも閲覧が可能である。

すればすぐにわかるような些細な違いでさえも、ルールベースに当てはまらないことで抽出が正しくできないため、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。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。

そこで本研究では、[高野 20] のルールベースを参考に、新たにルールベース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て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ベースが適用できないような株主招集通知に対しても、本研究で提案する手法は有効である。

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 19] における決算短信からの業績要因の抽出や、[高野 19] における有価証券報告書からの業績要因・業績結果の抽出がある。また、決算短信などの金融テキストからの情報抽出の研究としては、[北森 17a, Kitamori 17b] における決算短信からの業績予測文の抽出や、[坂地 15, Sakaji 17] における決算短信からの原因・結果表現の抽出がある。しかし、これらの研究で対象としているのは、どれも一文単位での分類である。それに対して本研究では、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。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。これまでの自然言語処理の研究のほとんどは、単語、フレーズ、1 文を抽出の対象としているが、ページ単位での抽出に関する研究はなかった。もちろん可能であるならば、ページ単位よりもさらに限定し、重要文の抽出ができることが好ましいが、より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は、重要な箇所の見落としがあつてはならない業務においては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、本研究ではページ単位での抽出を行っている。本研究に限らずページ単位での抽出には需要があり、研究としての価値がある。

これらのことから、本研究はこれらの研究に対して、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、いくつかの評価実験は、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。

4. 提案手法

[高野 20] 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 (4.1 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 (4.2 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 (4.3 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

4.1 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 情報をクロールし、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、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^{*3}」を用いた。「k2pdfopt」を用いることで、PDF ファイルから図 1 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。

4.2 Step2: 学習データの自動生成

[高野 20] 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、ルールベース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、このようなルールベースに当てはま

*3 <https://www.willus.com/k2pdfopt/>

図 1 PDF データからテキストの抽出例

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 1 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 1 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。

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 4 に示す。表紙には、株主総会の開催日時や会場所、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 2 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 3 に示す。

表紙や目次となるページを、同一の Tag である Tag1 と

表 1 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

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 4 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 5 に示す。

Tag4 は大株主に関するページであり、例を図 6 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 7 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^{*4}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 [Dai 15, Lample 16, Ratinov 09, Sang 99]。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 2 に示す。

表 2 に示した通り、「IO」方式と「IOB2」方式の違いは、

表 2 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

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 2 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、その出力に対して、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

^{*4} 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。

Tag を区別するメリットの 1 つ目である。

また、4.3.5 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 2 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ため、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すると、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事ができる。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 ページが「I-resolution」が誤って付与される可能性をも下げる事が見込める。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 2 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

図 2 目次ページ (Tag1) の例

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 は付けられ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 [高野 20] 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。

§1 表紙や目次に関するページ候補の抽出

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 *5」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

*5 n は対象とする株主招集通知の中で一番大きい番号。

図3 目次ページが2ページにわたって記載されている例

図4 表紙となるページ (Tag1) の例

図5 議事項について記載されたページ (Tag2) の例

図6 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	地位および所属	重要な職務状況
吉 賀 信 行	取締役会長 社長兼会長（委員長） 取締役（委員長）	野村證券株式会社取締役（＊１） 野村證券株式会社代表取締役社長
永 井 清 二	取締役 代表取締役社長 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 代表取締役副社長	野村證券株式会社取締役（＊１、＊２）
木 村 宏	社外取締役 独立役員	日本たばこ産業株式会社社長 株式会社日本たばこ製薬
石 村 和 彦	社外取締役 独立役員	AGC株式会社取締役会長 TDK株式会社取締役 株式会社日本たばこ製薬
島 崎 聖 明	社外取締役 監査委員（委員長）	株式会社インヴェンティジャパン社外取締役 野村證券株式会社取締役（＊１）
廣 村 マ リ	社外取締役 監査委員	該当なし
室 下 真 人	取締役 監査委員（審判）	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） 野村フィナンシャル・アスリートグループホールディングス株式会社取締役（＊１）
Michael Lim Choo San [マイケル・リン・ツウ・サン]	社外取締役	Fulbright Healthcare Corporation Limited / エズベクティブ・フュアマン Nomura Singapore Ltd. / シ・エズベクティブ・フュアマン
Laura Simona Anger [ローラ・シンガー]	社外取締役	CIT Group, Inc. / インディペンデント・ディレクター Ontario 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）

[illegible]

図7 会社役員に関するページ (Tag5) の例

うかで判定した.

§ 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。

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 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

§ 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

§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。

§ 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 4.2.7 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。

§ 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する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 *6」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 8 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、「第 1 号議案」から「第 $n-1$ 号議案」までの Tag2 や Tag3 は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と、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

*6 n は対象とする株主招集通知の中で一番大きい番号。

2. 変更の内容 (下線...は変更部分)

現行定款	変更案
<p>(編纂) 第5条 当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。</p> <p>(1) 取締役会</p> <p>(2) 指名委員会、監査委員会および報酬委員会</p> <p>(3) 会計監査人</p> <p>(取締役の責任軽減) 第33条 (略)</p> <p>2. 当社は、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られ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。</p> <p>(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。</p> <p>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。</p> <p>3. (略)</p>	<p>(編纂) 第5条 当社は、株主総会および取締役のほか、次の機関を置く。</p> <p>(1) 取締役会</p> <p>(2) 指名委員会、監査委員会および報酬委員会</p> <p>(3) 会計監査人</p> <p>(取締役の責任軽減) 第33条 (現行どおり)</p> <p>2. 当社は、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られ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。</p> <p>(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。</p> <p>2. (現行どおり)</p> <p>3. (現行どおり)</p>

図 8 「以上」という文字列のみが行に出現するページの例

かは、5 章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で、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、4.2.6 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, [高野 20] 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ルすることで, 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の質の高い学習データを自動生成することが可能となった.

4.3 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.

§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^{*7}, 入力層の要素となる語 (素性) を選択する. 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, この数字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 5 章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

$$\text{page}_k = \begin{matrix} \text{以上} & \text{略歴} & \text{株式} & \dots & \text{拝啓} & \text{敬具} & \text{株主} & \text{発行} \\ \left[1, 0, 3, \dots, 0, 0, 6, 4 \right] \\ 2256 \text{次元} \end{matrix}$$

図 9 入力ベクトルの例

図 10 モデル 1 簡略図

§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, 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 9 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 10 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 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. 入力, 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 $\times 3$ 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

^{*7} <http://mecab.googlecode.com/>

図 11 モデル2 簡略図

図 12 モデル3 簡略図

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 11 に示す。

§ 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 12 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。

§ 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 16] で示されている。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 $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n)$ とし、各ページの予測される Tag を $\mathbf{t} = (t_1, t_2, \dots, t_n)$ と定義することで、以下の条件付き確率、式 (1) 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。

$$p(\mathbf{t}|\mathbf{H}) = \frac{\exp(\text{score}(\mathbf{H}, \mathbf{t}))}{\sum_{\mathbf{t}'} \exp(\text{score}(\mathbf{H}, \mathbf{t}'))} \quad (1)$$

$$\text{score}(\mathbf{h}, \mathbf{H}) = \sum_{i=0}^n W_{t_i, t_{i+1}} + \sum_{i=1}^n P_{i, t_i} \quad (2)$$

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形行列 (9×9)*⁸

$\mathbf{W}_{t_i, t_j}$: 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 ($n \times 7$)

$\mathbf{t}$: $\mathbf{t}$ が取り得るすべてのパターンの Tag の組み合わせ

式 (2) 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 (1) 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。

4.4 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。

*8 Tag0, Tag1, Tag2, Tag3, Tag4, Tag5, Tag6, TagB, TagE の合計が 9 個のため。

表 3 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

4.4 節で 4 つのモデルを提案したが、5 章にて、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。

5. 評価と考察

提案手法を実装し、評価を行った。

まず、Step1 (4.1 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 (4.2 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)。

そして、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 3 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 4 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 4 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

表 4 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

が極力起こらないようにすべきである。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。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。4.2.7 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 3 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が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 (4.3 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 4.3.1 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 $df(w)$ 、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、 $DF(\text{Tag}i)$ を $\text{Tag}i$ 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 (3) で求める。

$$v = \frac{\min(DF(\text{Tag}i))}{N} \quad (3)$$

学習データにおけるページの総数が $N = 125,746$ であり、 $\min(DF(\text{Tag}i)) = 2,795$ 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、 $DF(\text{Tag}2) = 2,795$ であるが、仮に $\text{Tag}2$ 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、 $\text{Tag}2$ で必ず出現したとしても他の Tag で出現しないため、 $df(w) = 2,795$ 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 (3) 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 $\text{Tag}i$ (本研究では $\text{Tag}2$, $\text{Tag}4$, $\text{Tag}5$) ごとに語の df 値を計算し、 $\text{Tag}i$ 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、 $\text{Tag}i$ 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、 $\text{Tag}2$ や $\text{Tag}5$ 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 (4) にて求めた。

$$v = \frac{\min(DF(\text{Tag}i))}{N \times k} \quad (4)$$

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られる。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 5、表 6、表 7、表 8 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る評価指標である micro 平均^{*9}、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 9 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 5 であれば、 $(298 + 18 + 10 + 108 + 10 + 10 + 23)/508 = 0.941$ である。macro 平均は、

*9 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 5 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

表 6 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

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 5 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 5 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 [高野 20] 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 10, 表 11, 表 12 に示す。

表 10, 表 11 からわかるように、[高野 20] 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

表 7 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 8 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

ては、余分な抽出をしないことで適合率が向上した。

表 5 から表 8 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 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 となってしまうことである。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う。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 5 からわかる通り、

表 9 評価指標による評価結果

手法	micro	macro			weighted		
	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 10 [高野 20] のルールベースとモデル 1 の評価結果

証券コード (全ページ数)	ルールベース			モデル 1		
	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%

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 6 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 13 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

表 11 モデル 2 とモデル 3 の評価結果

証券コード (全ページ数)	モデル 2			モデル 3		
	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%

表 12 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%

が来ることは 172 回あることが読み取れる。

まず、表 13 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 13 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 9 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 13 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

表 14 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 15 重要ページ抽出を重視したモデル 1 のクロス集計結果

Tag	正解 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 14 に示す。

表 5 から表 8 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 15 から表 18 に示し、表 9 と同様の評価指標を用いて計算した結果を表 19 に示す。

表 19 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

表 16 重要ページ抽出を重視したモデル 2 のクロス集計結果

Tag	正解 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 17 重要ページ抽出を重視したモデル 3 のクロス集計結果

Tag	正解 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

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、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 20 に示す。

もちろん、不要なページに Tag が付与されることから、

表 18 重要ページ抽出を重視したモデル 4 のクロス集計結果

Tag	正解 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

表 19 重要ページ抽出を重視した評価指標による評価結果

手法	micro	macro			weighted		
	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 20 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

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。

6. ま と め

本研究では、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。本研究で提案した全てのモデルにおいて、[高野 20] 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、本研究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。

◇ 参 考 文 献 ◇

- [Dai 15] Dai, H.-J., Lai, P.-T., Chang, Y.-C., and Tsai, R. T.-H.: 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)
- [和泉 11] 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)
- [北森 17a] 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)
- [Kitamori 17b] Kitamori, S., Sakai, H., and Sakaji, H.: 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)
- [蔵本 13] 蔵本 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)
- [Lample 16] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C.: Neural architectures for named entity recognition, in *Proceedings of NAACL 2016*, pp. 260–270 (2016)
- [Ratinov 09] Ratinov, L. and Roth, D.: 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)
- [酒井 13] 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)
- [酒井 19] 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)
- [坂地 15] 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)
- [Sakaji 17] Sakaji, H., Murono, R., Sakai, H., Bennett, J., and Izumi, K.: Discovery of Rare Causal Knowledge from Financial Statement Summaries, in *IEEE Symposium on Computational Intelligence for Financial Engineering & Economics*, pp. 602–608 (2017)
- [Sang 99] Sang, E. F. T. K. and Veenstra, J.: Representing text chunks, in *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 173–179 (1999)
- [高野 18] 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, 自然言語処理, Vol. 25, No. 1, pp. 3–32 (2018)
- [高野 19] 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. 1–22 (2019)
- [高野 20] 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)

[担当委員: 大塚 真吾]

2020 年 5 月 28 日 受理

著 者 紹 介

高野 海斗(学生会員)

2013 年成蹊大学理工学部情報科学科入学。2016 年同大学院理工学研究科理工学専攻博士前期課程入学。2018 年同博士後期課程入学。統計学，特に，欠測データ解析，集計データ解析の研究と，自然言語処理，特に，言語情報を用いたデータの自動分類の研究に従事。日本計算機統計学会学生会員。

酒井 浩之(正会員)

2005 年豊橋技術科学大学大学院工学研究科博士後期課程電子・情報工学専攻修了。博士（工学）。2005 年豊橋技術科学大学知識情報工学系助手。2012 年成蹊大学理工学部情報科学科講師。2014 年同准教授。自然言語処理，特に，テキストマイニング，テキスト自動要約の研究に従事。

中川 慧(正会員)

2012 年京都大学経済学部卒業。同年より資産運用会社にて定量分析業務に従事。2018 年より野村アセットマネジメント株式会社資産運用先端技術研究部クオンツアナリスト。2015 年筑波大学大学院ビジネス科学研究科博士前期課程修了。修士（経営学）。2020 年同博士後期課程修了。博士（経営学）金融工学および機械学習を用いた金融データ解析の研究に従事。