人工知能学会研究会資料
SIG-FIN-013-10

企業の決算短信 PDF の自動要約

Automatic Summarizing PDF Files of Summary of Financial Statements

瀬戸孟[1]　酒井浩之[1]　坂地泰紀[1]

Takeshi Seto[1], Hiroyuki Sakai[1], and Hiroki Sakaji[1]

[1]成蹊大学理工学部情報科学科
[1]Department of Computer and Information Science, Faculty of Science and Technology, Seikei University

Abstract: In this paper, we proposed a method that summarizes pdf files of summary of financial statements. Specifically, our method extracts contents of financial statements, causal information and future forecasts from pdf files of summary of financial statements. Then, the summary is generated by connecting them.

1. はじめに

近年，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おこなう技術の必要性が高まっている．そのため，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，例えば，膨大な金融情報を分析して投資判断の支援をする技術が注目されている．その一例として，日本銀行が毎月発行している「金融経済月報」や新聞記事をテキストマイニングの技術を用いて，経済市場を分析する研究などが盛んに行われている[1][2][3].

投資家にとって，企業の業績に関する情報を収集することが重要であり，現状の把握と企業の今後の業績推移について知る必要がある．そこで，本研究では, 企業が業績発表を行った直後に企業の Web サイト等で一般に公開されている企業の決算短信に着目する．企業の決算短信は PDF 形式で企業の Web サイト等で配布され，誰でも閲覧ができる．しかしながら，公開される決算短信は多くの情報を含み，文書量も多いため，個人投資家にとって多くの企業の決算短信を熟読することは困難である．そこで，本研究では, 決算短信 PDF を要約し，個人投資家でも決算内容を把握しやすい情報に変換することを目的とする．

ここで，公開される決算短信の中で，どのような情報が重要であるかを考える．投資家が業績を評価するうえで，最低限理解しておかなければいけない情報として，本研究では以下の 3 点に注目した．

- 業績内容：今期の業績内容．例えば「その結果、当連結会計年度において、売上高は３８億２７百万円（前年同期比１１．９％減)、営業利益は６７百万円（前年同期比８２．０％減)、経常利益は７４百万円（前年同期比７３．２％減)、当期純利益は７８百万円（前年同期比４７．９％減）となりました。」 など．
- 業績要因：今期の業績内容になった要因．例えば「ビル事業は、国内の昇降機新設及びリニューアル事業の増加や、中国を中心とした海外の昇降機新設事業の増加に加え、円安の影響もあり、受注・売上とも前年同期を上回りました。自動車機器事業は、北米の新車販売市場の回復に加え、円安の影響もあり、受注・売上とも前年同期を上回りました。電子システム事業は、電子事業の増加により受注・売上とも前年同期を上回りました。」など．
- 今後の予測：来期ではどのようなことに注力していくか，また，年度発表からの変更点がどれほどあるかを示すもの．例えば「これらにより、当連結会計年度は、連結売上高１３７億円（前年比０．６％増収)、連結営業利益７億４千万円（前期比５．３％減益)、連結経常利益７億円（前期比９．３％減益)、連結当期純利益３億５千万円（前期比４５．８％増益）を予想いたしております。」など．

決算短信 PDF には，上記で示した決算内容，業績要因，今後の予測に関する情報が含まれている．そこで, 本研究では, 企業が公開している決算短信 PDF を収集し，その決算短信 PDF から決算内容，業績要因, 今後の予測の情報を含む文を自動的に抽出する．

そして，それらを連結して，決算短信 PDF の要約を生成する手法を提案する．

２．関連研究

関連研究として，酒井らは，企業の業績発表記事から業績要因を抽出する手法を提案した[4]．例えば，業績発表記事から「ソフト販売の収益が寄与する」といった文を抽出する．また，酒井らの手法[4]では，1 つの業績発表記事からは複数の業績要因が抽出される．その中から抽出される業績要因の中には，例えば「独禁法関連費用を営業外費用に計上」のような，事業と関連性の低い，それほど重要ではない業績要因も存在している．そのため，酒井らは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[3].

西沢らは，酒井らの手法を適用して決算短信 PDF から業績要因を抽出する手法を提案している[5]．本研究では，決算短信 PDF から業績内容を含む文，業績要因を含む文，今後の予測を含む文を抽出して要約を生成するが，業績要因を含む文の抽出は，西沢らの手法を使用する．そのため，本稿では，業績内容を含む文と，今後の予測を含む文を抽出する手法について述べる．

2. 企業 Web ページからの決算短信 PDF ファイルの取得

本章では，本研究で使用する決算短信 PDF ファイルを取得する方法について述べる．本研究では，企業の Web ページの中から決算短信 PDF ファイルがダウンロードできる IR 情報ページを識別し，そのページの PDF ファイルをダウンロードする．そして，ダウンロードした PDF ファイルから決算短信 PDF を選別する．以下に手法を示す．

Step 1: 企業 Web ページを収集し，企業 Web ページにおいて決算短信 PDF ファイルがダウンロードできる IR 情報ページを人手にて判定する．

Step 2: Step 1 で判定した企業 Web ページにおける IR 情報ページを学習データの正例とし，それ以外の企業 Web ページを負例とし，Web ページに含まれている名詞を素性としてサポートベクトルマシンで分類器を生成し，全ての企業 Web ページを IR 情報ページとそうでないページに分類する．

Step 3: Step 2 で識別した IR 情報ページの URL から，よく出現する URL の文字列を抽出する．

Step 4: Step 2 で識別した IR 情報ページの URL に加え，Step 3 で抽出した文字列を含む URL の Web ページから，ダウンロードできる PDF ファイルを全てダウンロードする．

Step 5: ダウンロードした PDF ファイルのテキスト情報を抽出し，最初から 10 行以内に「決算短信」という文字列を含むものを，決算短信 PDF ファイルとして選別する．

Step 1 においては，3,821 社の企業 Web ページを収集した．また，Step 1 にて，Step 2 で使用する学習用データとして，決算短信 PDF ファイルがダウンロードできる IR 情報ページを，155 の企業から人手で 200 ページ集めた．この 200 ページをサポートベクトルマシンにおける学習用データの正例とする．負例には，IR 情報ページ以外のページから，正例と同数の 200 ページを集めた．そして，Step 2 で識別した IR 情報ページの URL において，頻度が高い文字列を抽出し，その文字列を含んでいる URL からダウンロードできる PDF ファイルを全てダウンロードする．以下に，Step 3 の手法で取得した，IR 情報ページの URL によく出現する文字列をいくつか示す．

ir, library, company, investor, calendar.html, financial,library.html, calendar, investors, finance, IR, report,tanshin.html, irinfo, kessan

上記の手法にて，最終的に 107,251 個の決算短信 PDF ファイルを取得した．

3. 決算短信 PDF からの業績内容抽出

本章では，3,821 社の企業 Web ページから取得した 107,251 個の決算短信 PDF から，業績内容を含む文を自動的に抽出する手法について述べる．

まず，決算短信 PDF を pdftotxt を使用して，テキスト情報に変換して文を抽出する．そして，抽出された文に対して，業績内容を含む文を抽出するために，3.1 節から 3.4 節の規則を順に適用する．

3.1 キーワードに関する規則

業績内容を含む文を抽出するため手法を以下に示す．まずは，6 章で後述する人手で判定した決算内容を含む文に，共通して出現しているキーワードを人手にて抽出する．キーワード集合を以下に示す．

キーワード集合 1	千，万，億
キーワード集合 2	売上高，営業，経常，当期利益，当期損益，当期損失，純利益，純損益，純損失，純資産
キーワード集合 3	期，年

ここで決算短信 PDF を構成する文において，作成したキーワード集合 1 ，キーワード集合 2 ，キーワード集合 3 の各要素を 1 つ以上, 含んでいる文を抽出する. しかしながら, 上記の方法では以下のような, 業績内容を含んでいない文を抽出してしまった.

例：営業活動によるキャッシュ・フロー当連結会計年度の営業活動により獲得した資金は、２億８２百万円（５９．９％減）となりました。

そこで，抽出された文の中から，決算内容を含んでいない文を人手にて判定し，判定された文において共通して出現しているキーワード集合(キーワード集合 4)を人手にて作成する.

キーワード集合 4	部門，分野，キャッシュ，当たり，あたり，予測，見込，純資産

そして，文集合において，キーワード集合 4 の要素を含まない文の集合を得る.

3.2 「～事業」に関する規則

これまでの手法では、以下の例のような事業ごとの決算内容を含む文が抽出された.

例：これらの結果、サービス事業の売上高は前年同期（３１１百万円）比２１．４％増の３７８百万円となりました。

そのような文を除去するために，文集合の要素である文に「事業」が含まれていた場合，係り受け解析を実行し，図 1 のような文節の構造をもつ文を除去する.「事業」を含む文節が係っている文節に「は」が含まれていれば，その文を除去する.

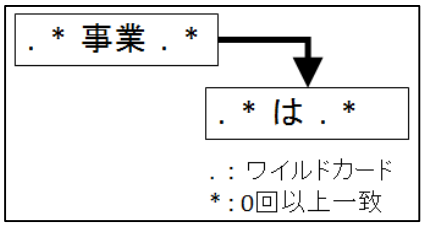

図 1 除去する文における文節構造

3.3 数字に関する規則

業績内容を含む文には，文中の括弧内に数字が多く含まれている傾向になる．まず括弧が文に含まれていた場合，括弧内の文字列を得る．その文字列を形態素解析して数字を得る．そして，文 s の括弧内の文字列において，数字が出現する割合を以下の式で得る.

$$\mathrm{P(s)} = \frac{n(s)}{f(s)}$$

f(s)： 文 s の括弧内に含まれる 2 文字以上で構成される形態素の数

n(s)：文 s の括弧内に含まれる数字の数

その割合が 0.05 未満であれば，その文を除去する.

3.4 複合語に関する規則

業績内容と関連のある複合語を多く含む文を，業績内容を含む文として抽出するために，以下の規則を適用する.

6 章で後述する人手で判定した決算内容を含む文集合の要素である文を形態素解析して複合語を抽出し，その中から業績内容と関連のある複合語を人手にて抽出する．以下に，抽出した業績内容と関連のある複合語の一部を示す.

売上高，経常利益，営業利益，増加，減少，四半期連結累計期間，四半期純利益，連結売上高，当連結会計期間

そして，文 s を形態素解析して複合語を抽出し，抽出した複合語において，上記のリストの複合語が含まれる割合を以下の式で求める.

$$\mathrm{P(s)} = \frac{fl(s)}{fr(s)}$$

$fr(s)$：文 s に含まれる複合語の数

$fl(s)$：文 s に含まれる上記のリストの複合語の数

この割合が 0.1 未満の場合，その文を除去する.

4. 決算短信 PDF から今後の予測抽出

本章では, 決算短信 PDF から今後の予測を含む文を自動的に抽出する手法について述べる.

4.1 手がかり表現の抽出

決算短信 PDF から今後の予測を含む文を抽出するために，例えば「見込み」や「見通し」などの手がかり表現を使用する．例えば，手がかり表現「見込み」が含まれている文からは，以下のような今後の予測を含む文が抽出できる.

例：以上により、次期の業績につきましては、売上高１６，７００百万円（当期比３．０％減）を想定しておりますが、不採算事業の縮小・撤退による固

定費の削減及び収益性の改善に加え、投資フェーズの事業の成長等を見込み、経常利益５８０百万円（当期比４８３百万円の利益の増加）、当期純利益１５０百万円（当期比１，０８５百万円の利益の増加）を予定しております。

しかし，今後の予測を含む文を抽出するのに有効な手がかり表現の種類は多く，人手で多くの手がかり表現を得ることは困難である．そのため，このような手がかり表現を決算短信 PDF から自動的に得ることを考える．具体的には，西沢らが決算短信 PDF から業績要因を含む文を抽出するための手がかり表現を得るための手法[5]を，今後の予測を含む文を抽出するための手がかり表現を得るために使用した．

ここで，手がかり表現を自動的に得る手法を示す．

Step 1: 決算短信 PDF から，「見込み」「見通し」を含む文を抽出する．

Step 2: 少数の手がかり表現（具体的には，「の見込み」，「なる見込み」「の見通し」「なる見通し」の 4 表現を用いる）を人手で与え，それに係る節を取得する．

Step 3: 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を共通頻出表現として，後述の基準を用いて抽出する．

Step 4: 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，後述の基準を用いて抽出する．

Step 5: 獲得した手がかり表現から，それに係る節を取得する．

Step 6: Step 3 から Step 5 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．

Step 2 では，手がかり表現に係る節の集合から，適切な共通頻出表現を選別する．具体的には，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以下の式で求め，その値が，ある閾値 Te 以上の共通頻出表現を選別する．

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

ここで，Step 1 で抽出された文集合において，
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．以下の式で求める．

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

ここで，f(e,s) は，共通頻出表現 e が手がかり表現 s に係る回数である．

エントロピーH(e) は，共通頻出表現 e が業績発表集合において様々な手がかり表現に均一の確率で係っている場合に高い値をとる．

閾値 Te は，以下の式によって設定する．

$$\mathrm{Te} = \alpha \log_2 |Ns|$$

ただし，Ns は共通頻出表現を取得するのに使用した手がかり表現の集合，α は定数(0< α < 1) である．

Step 3 で共通頻出表現の抽出を行った後，Step 4 にて，その選別した共通頻出表現から新たな手がかり表現を獲得する．ここでも，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，手がかり表現の選別を行う．具体的には，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る．

$$\mathrm{H(s)} = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e)$$

ただし，Step 1 で抽出された文集合において，
E(s): 手がかり表現 s に係る共通頻出表現の集合．
P(s, e): 手がかり表現 s に対して共通頻出表現 e が係る確率．以下の式で求める．

$$\mathrm{P(s,e)} = \frac{f(s,e)}{\sum_{e' \in E(s)} f(s,e')}$$

ここで，f(s, e) は，手がかり表現 s に対して共通頻出表現 e が係る回数である．そして，ある閾値以上の候補を手がかり表現として抽出する．閾値は，共通頻出表現と同様に設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．

以下に，定数 α を 0.5 として得られた手がかり表現から人手で選別した手がかり表現の一部を示す．

計画しています，修正しておりません，推移しております，推進していきます，想定しています，変更ありません，見込まれます，見通しです，予想されます

4.2 手がかり表現を使用した今後の予測を含む文の抽出

4.1 節で得られた手がかり表現を使用して，決算短信 PDF から今後の予測を含む文を抽出する．具体的には，得られた手がかり表現を含む文を抽出する．ただし，以下に示す表現を含む文は除外した．

配当，サマリー，キャッシュ，グループ，上記，今後に，した。

上記の表現が含まれる文を除外した理由は，企業がどのようなことに注力するかを考えているので，お金の流れについての表現（配当，サマリー，キャッシュ）は文に含まれないようにした．また，“グループ”については，事業グループを示すものが多く，その企業についての説明がないものと判断した．そして，“上記”については，表で示されていることが多く，文に内容が書かれていない．最後に“した。”については，予測を抽出するので，語尾が過去形になるものは全て除外するようにした．

5. 実装

本手法を実装し，3,821 社の企業 Web ページから 107,251 個の決算短信 PDF ファイルを取得した．そして，それらから業績内容，業績要因，今後の予測を含む文を抽出し，それらを連結して要約を生成した．実装にあたり，決算短信 PDF からの形態素解析器として MeCab*[1]，係り受け解析器として CaboCha[6] を使用した．企業名を入力すると，その企業の決算短信 PDF の要約を生成するシステムを実装した．図 2 に実装したシステムによる「三菱電機」の決算短信 PDF から生成した要約を示す．

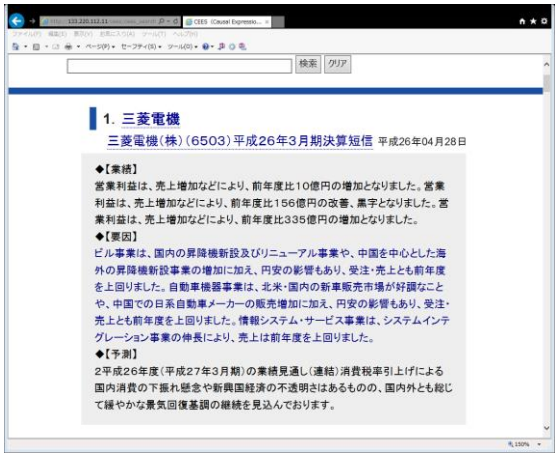

図 2　「三菱電機」の決算短信 PDF から生成した要約

また，以下に東芝の決算短信 PDF から生成した要約を示す．

【業績】売上高は，主要５セグメント全てで増収となった結果、前年同期比　２，３０９　億円増加し　１　兆　９，７０５　億円になりました。当期純損益は上記の資産価値の見直し、光学ドライブ事業の非継続化及び復興特別法人税廃止の影響等があり、前期比　２６６　億円減少の　５０８　億円になりました。（2）財政状態に関する分析１当期の財政状況・総資産は，２０１３年３月末に比べ１，４１６億円増加し、６兆２，４１６億円になりました。

【要因】国内経済は、日銀による異次元緩和や財政出動に加え消費税増税前の駆込み需要も寄与し、緩やかな景気回復が続きました。＜電力・社会インフラ部門＞：増収、減益国内の原子力発電システムが減収になったものの、電力流通システム、太陽光発電システム、鉄道向けシステム、自動車向け事業等の増収により社会インフラシステム事業全体が伸長しました。一方、火力・水力発電システムが好調を維持したものの減益になり、原子力発電システムが海外での一時的な費用や米国の原子力発電所の事業開発会社の資産価値を保守的に見直したこと等の一時的な影響により悪化しました。

【予測】来期は、海外では中国の不良債権問題、国内では消費税増税に伴う景気減速等の不安要因があり、欧州や新興経済地域の回復力が弱い状況にあるものの、世界経済全体としては当期を上回る成長が予想されています。

---

6. 評価

本手法の評価を以下の方法で行った．まず，無作為に選別した 70 個の決算短信 PDF の中から業績内容を含む文と今後の予測を含む文を人手で抽出した．その結果，業績内容に関する文 168 文，今後の予測に関する文 67 文を正解データとして得た．次に，選択した決算短信 PDF から，本手法にて業績内容を含む文と今後の予測を含む文を抽出する．それらが正解データの文と一致すれば正解とし，精度，再現率を算出した．それらの評価結果を表 1 ，表 2 に示す．

表 1 業績内容の再現率，精度

業績内容数	出力文数	正解文数	再現率	精度
143文	166文	168文	85%	86%

表 2 今後の予測の再現率，精度

今後の予測数	出力文数	正解文数	再現率	精度
47文	69	67	70%	68%

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

7. 考察

表１ より，業績内容を含む文においては，再現率 85% ，精度 86% と高い精度を出すことができた．その成功例を以下に挙げる．

「当連結会計年度においては、売上高は４３億４千３百万円（前年同期比３．８％減)、営業利益は３億７千４百万円（前年同期比２２．１％減)、経常利益は２億７千６百万円（前年同期比４５．５％減)、当期純利益は１億４千９百万円(前年同期比５３．５％減）となりました。」

失敗例を以下に挙げる．

「税金等調整前当期純利益は、７３４，１３９千円となり前連結会計年度に比べ３４１，５９５千円(８７．０％）増益となりました。」

失敗例における複合語は，“税金等調整前当期純利益”，“前連結会計年度”，“増益”で，そのうち 3.4 節で示した複合語のうち“前連結会計年度”，“増益”の 2 語が該当した．そのため，割合が 0.67 となり，文が除去されなかった．

一方，表 2 の今後の予測を含む文の抽出については，再現率，精度ともに改善する必要がある．その成功例を以下に挙げる．

「このような対応策を通して通期の業績予想を、連結売上高は１５，０００百万円(前期比０．１％減)、連結営業利益は１，０３０百万円（前期比１１．５％減)、連結経常利益は１，０５０百万円（前期比７．６％減)､連結当期純利益は４５０百万円(前期比４．５％減）を予想しております。」

失敗例を以下に挙げる．

「業績の見通しにつきましては、原材・素材価格の高騰や米国経済減速の影響が懸念され、わが国経済は不透明な状況下にあります。」

本手法では，企業についての情報を抽出する目的だったが、日本経済全体についての文が抽出されてしまった．その解決策としては，4.2 節で用いた表現を追加する必要がある．

8. まとめ

本研究では，企業の決算短信 PDF から業績内容，業績要因，今後の予測を抽出し，それらを連結して要約を生成する手法を提案した．その準備段階として，企業 Web ページから決算短信 PDF を取得する手法について述べた．決算短信 PDF から業績内容を含む文の抽出については，キーワードに関する規則，事業に関する規則，数字に関する規則，複合語に関する規則を順に適用することで行った．業績要因を含む文については，西沢らの先行研究を用いた．今後の予測を含む文については，手がかり表現を抽出し，それを含む文を抽出した．さらに，そこから関連がない語句を除外する手法を示した．これより，業績内容を含む文の抽出では，再現率，精度ともに 80%を超え，今後の予測を含む文の抽出では，再現率，精度，ともにおよそ 70%の結果となった．

参考文献

[1] 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)

[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] 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)

[4] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, IEICE Trans. Information and Systems, Vol. E91-D, No. 4, pp. 959–968, (2008)

[5] 西沢裕子，酒井浩之，企業の決算短信 PDF からの業績要因の自動抽出，第 3 回 テキストマイニング・シンポジウム, pp.67-72, (2013)

[6] 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)