決算短信 PDF からの業績予測文の抽出

北森詩織[1] 酒井浩之[1] 坂地泰紀[1]

[1]成蹊大学 理工学部 情報科学科

1 はじめに

近年, 個人投資家の数が増加している. そのため, 金融市場における個人投資家に対する支援を行うための研究が盛んに行われている. 投資の際, 投資にとって重要なのは, 企業の今後の業績予測を知ることである. なぜなら, 例えば, たとえ現在の業績が赤字であったとして, 不振事業の整理が完了し, 今後の業績が回復することが企業側から示されれば, 株価は上昇するからである. 企業が示す今後の業績予測を知る手段として, 決算短信 PDF[4]を閲覧することが一般的である. 決算短信 PDF には, 業績情報や業績要因, 今後の業績予測など, 投資に有用だと思われるテキスト情報が多く含まれている. さらに, 企業 Web サイトなどで配布され, 誰でも閲覧可能である. しかし, 決算短信 PDF は文章量が多く, さらに多くの専門用語が含まれるため, 投資に関する知識をあまりもたない個人投資家にとって難解なものである. また, 個人投資家にとって, 多くの企業の決算短信 PDF を読み, 投資に重要な「今後の業績予測」の記述を見つけることは多大な労力を要する. そこで, 本研究では, 株式投資に関する知識をあまりもたない個人投資家支援のための手法として, 決算短信 PDF から企業の「業績要因を含む今後の業績予測」の抽出を行うことを目的とする. 例えば, キヤノンの決算短信 PDF から「IT インフラサービス事業は, アウトソージングサービス等が拡大することにより, 前年を上回る見込みであります」のような文を抽出する. 以降, 抽出する「業績要因を含む今後の業績予測」を含む文を「業績予測文」と定義する. 本研究によって業績予測文を抽出することにより, ある企業において予想よりも好調な事業や不振な事業が分かり, 個人投資家でも簡単に企業の業績予測を把握できる.

2 関連研究

関連研究として, 瀬戸らは決算短信 PDF を自動要約する手法を提案した[1]. 瀬戸らは, 決算短信 PDF から投資家が業績を評価する際に最低限理解しておかなければいけない情報として, 「業績内容」「業績要因」「業績予測」の 3 点に注目し, それらの情報を抽出して結合することで自動要約を行った. しかし, 瀬戸らが抽出した「業績予測」では売上高や利益の具体的な値が含まれているが, 業績予測の要因が含まれていない. それに対して, 我々は業績要因を含む今後の予測を抽出しており, どの事業が予想より好調であるか, あるいは不振であるかがわかる.

酒井らは, 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[2]. 坂地らは, 決算短信 PDF から原因・結果表現の抽出を行っている[3]. 坂地らは, 業績要因に対して, 例えば, 原因「猛暑」, 結果「冷房需要の盛り上がり」といった表現を自動抽出する手法を提案している. 酒井らや坂地らが抽出した業績要因文や原因・結果表現は, すでに確定済みの業績における情報から抽出しているのに対し, 本研究では, 今後の業績予測に関する業績要因を抽出している.

3 業績予測に修正のある決算短信 PDF の抽出

本研究における決算短信 PDF は, 文献[2]の手法に基づき収集した. この中から, まず, 業績予測に修正のある決算短信 PDF を抽出し, 抽出した「修正有」の決算短信から業績予測文を抽出する. 業績予想の修正があった決算短信PDFには「業績予想からの修正の有無:有」の文が決算短信 PDF に記述される. この記述を含む決算短信 PDF ファイルを, 業績予測に修正のある決算短信 PDF として抽出した. その結果, 収集した 107,251 個の決算短信PDF(3,821 社)中, 8,213 個(2,067 社)が業績予測に修正のある決算短信 PDF として抽出した.

4 業績予測文の抽出

4.1 文頭手がかり表現の獲得

決算短信 PDF から業績予測文を抽出するための手がかりとなる表現(以降, 手がかり表現と定義する)を調査した. その結果, 表 1 で示す表現 A と表現 B による組み合せにより構成される手がかり表現を人手により作成した.

表1 文頭手がかり表現

表現 A	予想につきましては, 見通しにつきましては
表現 B	業績, 業績の, 今後の, 通期の, 今後の, 通期の, 通期

例えば, 表現 B の「今後の」と表現 A の「予想につきましては」を組み合わせ, 「今後の予想につきましては」という手がかり表現を得る. これらの手がかり表現は, 業績予測文の文頭によく出現する表現であるため, 文頭手がかり表現と定義する. 上記の業現 A と表現 B の組み合わせにより, 計 10 個の文頭手がかり表現を作成した.

4.2 業績要因を含む業績予測文の抽出

4. 1 節で業績予測文として抽出した文には, 業績要因が含まれていないものも多く含まれていた. そこで, 4. 1 節で抽出された文に対して, 酒井らの決算短信 PDF からの業績要因抽出手法[2]を使用して業績要因を抽出し, 業績要因が含まれている文を業績予測文として抽出した. 以下に, 抽出された業績予測文の例を示す.

連結業績予想につきましては，電線線材事業を中心に売上高が想定を上回ることが見込まれるため，売上高は前回予想を上回る見込みです

しかし，文頭手がかり表現のみでは業績予測文が獲得されない決算短信 PDF が多く存在した．より多くの業績予測文を抽出するために，新たな手がかり表現を獲得する必要がある．

4.3 文末手がかり表現の獲得

新たな手がかり表現の獲得する手法を以下に述べる．ここで，4.2 節で得られた業績予測文を分析すると，文末に手がかりとなる表現が多く出現していることが分かる．例えば，4.2 節で示した業績予測文の文末は「見込みです」であり，このような文末表現を含む文で，かつ，業績要因が含まれていれば，業績予測文である可能性が高い．しかし，業績予測文を抽出するのに有効な文末表現（以降，文末手がかり表現と定義する．）の種類は数多く，人手にて全て獲得することは困難である．そのため，文末手がかり表現を半自動的に獲得する．

まず，4．2 節で得られた業績予測文の文末に多く出現する文節を得る．以下に業績予測文の集合に 3 回以上出現する文末文節の例を示す．

修正いたしました，なりました，見込みであります，修正いたします，見込みです，見通しです，予想されます，いたしました，思われます，上方修正いたします，あります，しております

得られた文末文節から，例えば「修正いたしました」，「見込みです」のように，文末手がかり表現として有効な表現もある．しかしながら「なりました」，「いたしました」のように，文末手がかり表現として不適切な文節も含まれる．ただし，このような文節の場合は，この文節に係る文節列との組み合わせを考えると，有効な文末手がかり表現となる場合がある．例えば，「なりました」に係る「下回る見込みと」と組み合わせ，「下回る見込みとなりました」とすれば，有効な文末手がかり表現となる．しかしながら，文末文節と，それに係る文節列との組み合わせは膨大な数となる．そこで，文末文節とそれに係る文節列との組み合わせを絞り込む．

図1に，文節列の取得の例として「なりました」に係る文節列の例を示す．次に，文末文節 c に係る文節列 p に対して以下の式でスコアを求め，このスコアScore (p,c) がある閾値を上回る文節列のみを抽出する．

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

$$P(p,c) = f(p,c)\,/N(c)$$

ただし，取得した業績予測文の集合において，
$P(p,c)$: c から取得される文節列pの出現確率，

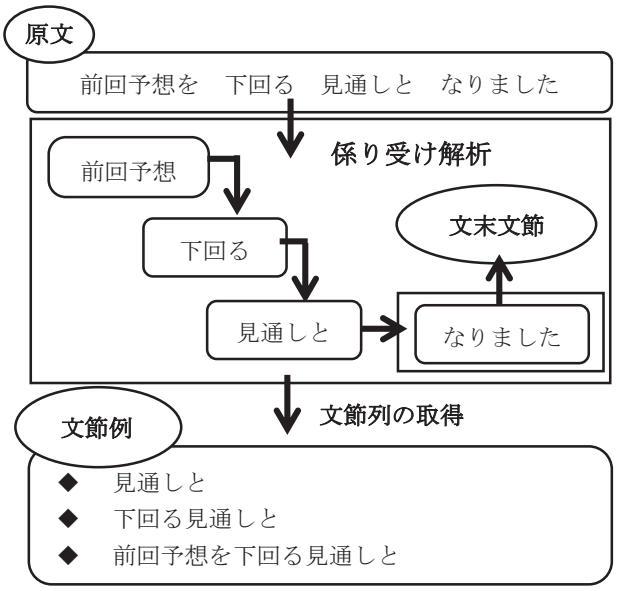

図１　文節列の取得例

$f(p,c)$: c から取得される文節列pの取得回数，
$N(c)$: 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，

上記の処理によって，99 個の文節列と文末文節の組み合わせを得た．この中から，人手により文末手がかり表現を選択し，88 個の文末手がかり表現を獲得した．それにより，例えば「下回る見込みとなりました」，「下回る見通しとなりました」などが文末手がかり表現として獲得された．

4.4 文末手がかり表現を使用した業績予測文の抽出

4.3 節で得られた文末手がかり表現を使用して業績予測文を抽出する．ここで，酒井らが業績要因を抽出するために決算短信 PDF から抽出した企業キーワード[2]を使用する．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り，例えば「東芝」の場合では「電子デバイス」や「フラッシュメモリ」などが企業キーワードとなる．本手法では，文末手がかり表現を文末に含み，かつ，その企業の企業キーワードが含まれている文を業績予測文として抽出した．以下に「雪国まいたけ」の決算短信PDFから抽出した業績予測文の例を示す． 例の太字は企業キーワード，下線は文末手がかり表現を示す．

業績予想につきましては，**まいたけ**・えりんぎを含む茸全般について，**最需要期**の9月から10月半ばにかけての気温が高い状態で推移したことや，デフレ下の需要低迷という市況悪化による販売単価・販売数量の落ち込み等により，売上高は大きく計画を下回ったことにより，予想値を下回る見込みであります

5　業績予測文の抽出の拡張

4章までの手法によって抽出した業績予測文を評価したところ，精度は良好であったが再現率が低い結果となった（7章「評価」を参照）．そこで，新たな文頭手がかり表現と文末手がかり表現を獲得することで，より多くの業績予測文を抽出するできるように手法を拡張する．

5.1 新たな文頭手がかり表現の獲得

4.1節で作成した文頭手がかり表現（以降，第一文頭手がかり表現と定義）を使用して抽出した業績予測文の精度は良好であったことから，第一文頭手がかり表現は業績予測文を抽出することに有効であるといえる．このことから，第一文頭手がかり表現に含まれる5つの単語「予想」「見通し」「業績」「今後」「通期」を含む文頭の表現は有効であると考えられる．また，第一文頭手がかり表現は，例えば「今後の予想につきましは」のように，10個の全て第一文頭手がかり表現において文末が係助詞「は」となっていることにも着目する．よって，抽出した業績予測文を形態素解析し，上記の5つの単語のいずれかを1つ含み，かつ，係助詞「は」を含んだ表現を「は」までで区切り抽出することで，新たな文頭手がかり表現（第二文頭手がかり表現と定義）を獲得した．獲得した第二文頭手がかり表現の例を以下に示す．

> 個別業績予想につきましては，今後の業績見通しにつきましては，通期業績予想については，通期の見通しにつきましては

しかし，獲得された第二文頭手がかり表現は1355個と多くなり，また，不適切な表現も含まれていた．よって，業績予測文の抽出に有効であると考えられる表現を絞り込む必要がある．

5.2 第二文頭手がかり表現の絞り込み

5.1節で獲得した第二文頭手がかり表現から，有効な文頭手がかり表現を絞り込む際に，4.3節で獲得した文末手がかり表現に着目する．獲得した文末手がかり表現を使用して抽出された業績予測文の精度は高いことから，この文末手がかり表現を含む業績予測文の文頭には，有効な文頭手がかり表現が含まれていることが予想される．このことから，多くの文末手がかり表現に共通して頻出する文頭手がかり表現は有効であると仮定できる．この仮定に基づき，以下の式で文頭手がかり表現が文末手がかり表現と同時に出現する確率に基づくエントロピーを求め，後述の基準に基づき，第二文頭手がかり表現の絞込みを行う．

$$H(n) = -\sum_{i \in TC(n)} P(n,i) \log_2 P(n,i)$$

$TC(n)$：文頭手がかり表現nを含む文において出現する文末手がかり表現iの集合

$P(n,i)$：文頭手がかり表現nが文末手がかり表現iと同時に出現する確率．以下の式で求める．

$$P(n,i) = f(n,i) \Big/ \sum_{i \in TC(n)} f(n,i)$$

$f(n,i)$：文頭手がかり表現nを含む業績予測文の集合において，文末手がかり表現iが出現する回数．

エントロピー$H(n)$は，文頭手がかり表現nが様々な文末手がかり表現と均一の確率で同時に出現している場合に高い値をとる．ここで，閾値を1として得られた文頭手がかり表現は85個であり，第一文頭手がかり表現の10個を足した合計95個を獲得した．その中から，人手で文頭手がかり表現を選択し，最終的に79個の文頭手がかり表現を獲得した．

5.3 第二文末手がかり表現の獲得

5.2節で獲得した第二文頭手がかり表現を使用して，4.3節と同じ手法で，新たな文末手がかり表現（第二文末手がかり表現と定義）を取得した．その結果，得られた第二文末手がかり表現は122個であり，第一文末手がかり表現の総数である88個を足した合計200個獲得した．その中から，人手により，文末手がかり表現を選択し，最終的に166個の文末手がかり表現を獲得した．以下に第二文末手がかり表現の例を示す．

> 予想を上回る見込みとなりました，想定されます，厳しい状況で推移すると思われます，売上高は前回予想を下回る見通しです，懸念されます

5.4 新たな業績予測文の抽出

5.3節で獲得した文末手がかり表現166個（第一文末手がかり表現＋第二文末手がかり表現）と企業キーワードを含む文を業績予測文として抽出した．また，文頭手がかり表現79個（第一文頭手がかり表現＋第二文頭手がかり表現）を含む文も業績予測文として抽出した．以下に，「青山財産ネットワークス」の決算短信PDFから抽出した業績予測文の例を示す．

> 連結業績予想に関する定性的情報連結業績予想につきましては、分譲**マンション**の販売が好調に推移し、販売経費等が削減できる見通しから、通期の業績予想を修正いたしました

6　実装

本手法を実装して，企業Webページから取得した107,251個の決算短信 PDF から業績予測文を抽出した．実装にあたり，形態素解析器として MeCab ，係り受け解析器として Cabocha[5]を使用した．また，酒井らが開発した企業名を入力するとその企業の決算短信 PDF と業績要因文を検索するシステム[2][1]に，本手法で抽出された業績予測文を組み込み，業績予測に修正のある決算短信 PDF が検索されると，業績予測文が表示されるようにした．

[1] http://hawk.ci.seikei.ac.jp/cees/

7　評価

本手法の評価を以下の方法で行った．まず，業績予測に修正のある 50 個の決算短信 PDF ファイルを無作為に抽出し，その 50 個から人手にて抽出した業績予測文 69 文を正解データとして，精度，再現率，F 値を求めた．結果を表 2 に示す．ここで，本手法(1)は，第一文頭手がかり表現と第一文末手がかり表現を使用して抽出した業績予測文の評価結果，本手法(2)は第二文頭手がかり表現と第一文頭手がかり表現，第一文末手がかり表現と第二文末手がかり表現を使用して抽出した業績予測文の評価結果である．

表 2　業績予測文の精度・再現率・F値

	精度	再現率	F 値
本手法(1)	86.05%	53.62%	0.66
本手法(2)	70.31%	65.22%	0.68

8　考察

第二文頭手がかり表現と第一文頭手がかり表現、第二文末手がかり表現と第一文末手がかり表現を使用して業績予測文を抽出した本手法(2)の再現率は 65.22%と，比較的良好な再現率を達成した．ここで，第一文頭手がかり表現と第一文末手がかり表現を使用して業績予測文を抽出した本手法(1)の再現率は 53.6%であった．これは，例えば第一文頭手がかり表現には存在しない「個別業績予想につきましては」のような新たな文頭手がかり表現が獲得されたことや，第一文末手がかり表現には存在しない「懸念されます」のような新たな文末手がかり表現を抽出したことで，文頭手がかり表現と文末手がかり表現の種類が増え，本手法(1)では抽出出来なかった業績予測文が抽出できたからである．

ここで，得られた業績予測文から新たな文頭手がかり表現と文末手がかり表現を追加することで，再現率の向上がみられた．そこで，5 章で抽出した業績予測文から第二文頭手がかり表現と第二文末手がかり表現を使用して，5 章の手法と同様に新たな文頭手がかり表現 161個(第三文頭手がかり表現と定義)と文末手がかり表現 102 個(第三文末手がかり表現と定義)を獲得し，新たな業績予測文を抽出した(本手法(3))．その結果を同じ正解データを用い，精度，再現率を算出した．その結果を表 3 に示す．

表 3　本手法(3)の精度・再現率・F値

	精度	再現率	F 値
本手法(3)	60.81%	65.22%	0.62

本手法(3)の再現率は本手法(2)の再現率と同じ結果となった．また，精度は60.81%であり，良好な精度を得られなかった．これは，第三文頭手がかり表現や第三文末手がかり表現に，業績予測文を抽出するうえで不適切な手がかり表現が含まれてしまったため，精度が低下したと考える．よって，文頭手がかり表現と文末手がかり表現の追加は，第二以降はしないほうが良いといえる．

本手法(2)の精度は 70.31%であり，比較的良好な精度を達成しているものの，本手法(1)の精度 86.05%と比較すると精度の低下が見られる．これも，本手法(3)の精度が低下した理由と同様に，第二文頭手がかり表現と第二文末手がかり表現に，業績予測文を抽出するうえで不適切な手がかり表現が含まれてしまったためであると考える．しかし，F値で比較すると，比較的良好な精度を保ちながら，再現率の向上がみられた本手法(2)の結果が良いといえる．今後は，新たな手がかり表現を獲得した際に，精度の向上のため，不適切な手がかり表現を除去する手法が必要である．

9　まとめ

本研究では，企業の決算短信 PDF から，業績要因を含む業績予測文を抽出する手法を提案した．業績予測文の抽出では，はじめに業績予測文を抽出するための手がかり表現を人手にて調査して作成し，例えば，「今後の予想につきましては」のような第一文頭手がかり表現を用いて，業績予測文の抽出を行った．さらに，文末の文節とそれに係る文節の組み合せをスコアの高いもので絞込みを行い，例えば，「下回る見込みとなりました」のような第一文末手がかり表現を得た．そして，得られた文末手がかり表現と企業キーワードを含む文を業績予測文として抽出した．さらに，得られた業績予測文から，例えば「個別業績予想につきましては」のような第二文頭手がかり手表現，および，第二文末手がかり表現を得ることで，より多くの業績予測文を抽出した．評価の結果，業績予測文の抽出精度は 70.31%，再現率は 65.22%となり，比較的良好な精度，再現率を得ることができた．

参考文献

[1]　瀬戸孟，酒井浩之，坂地泰紀: 企業の決算短信 PDF の自動要約，第 13 回　人工知能学会　金融情報学研究会，pp.50-55，(2014)

[2]　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.30，no.1，pp.172-182，(2015)

[3]　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)

[4]　東京証券取引: 所決算短信・四半期決算短信作成要領等(2015 年　3 月版)，東京証券取引所，(2015)

[5]　工藤拓，松本裕治: チャンキングの段階適用におけ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)