人工知能学会研究会資料
SIG-FIN-015-01

決算短信 PDF からの業績予測文の抽出

Extraction of Sentences of Forecasted Business Performance from PDF files of Summary of Financial Statements of Companies

北森詩織 [1]　酒井浩之 [1]　坂地泰紀 [1]

Shiori Kitamori[1],　Hiroyuki Sakai [1],　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 propose a method for extracting sentences of forecasted business performance from PDF files of summary of financial statements of companies. Specifically, our method extracts sentences of forecasted business performance containing causal information by using clue expressions. We evaluated our method and confirmed that it attained 82% precision and 56% recall .

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．逆に，現在の業績が黒字であったとしても，世界景気の減速が予測されるため，今後の業績が芳しくないことが企業側から示されれば，株価は下落するからである．

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 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，Web ページ上などで配布され，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．例えば，以下のような文を決算短信 PDF から抽出する．

「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」

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．このような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ことを目的とする．上記の業績要因は一時的な増益ではないため，投資の際に有用な業績予測の情報であると考えられる．よって，上記のような業績要因を含む業績予測を示している企業に投資を行うほうが，投資家にとって有益であると考えられる．このことから，投資の際，業績要因を含む業績予測を知ることの重要さが分かる．

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，自動要約を行った．瀬戸らが抽出した業績予測文の正解例では，下記のような文が挙げられている．

「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.」

上記の文では，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た[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．

ここで，業績予測に修正が必要である場合について述べる．JPX（日本取引所）は，上場企業の個別決算内容が前年比で一定以上変動した場合の取り扱いについて，以下の基準を定めている[4]．

1) 当事業年度の個別売上高が，直前に公表された当事業年度に係る予想個別売上高と比較して 10％以上増減しているとき
2) 当事業年度の個別経常利益が，直前に公表された当事業年度に係る予想個別経常利益と比較して 30％以上増減しており，かつ，増減額が前事業年度の個別純資産額又は資本金の額のいずれか大きい方と比較して５％以上であるとき
3) 当事業年度の個別当期純利益が，直前に公表された当事業年度に係る予想個別当期純利益と比較して 30％以上増減しており，かつ，増減額が前事業年度の個別純資産額又は資本金の額のいずれか大きい方と比較して 2.5％以上であるとき（ただし，直近予想が０の場合は抵触したものとする）

上記のいずれかの基準を満たしたとき，業績予想の修正を行う必要があり，「業績予想からの修正の有無：有」「連結業績予想数値の当四半期における修正の有無　有」のいずれの文が決算短信 PDF に記述される．これを利用して，この２文のどちらかを含む決算短信 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，107,251 ファイル(3,821 社)中，8,213 ファイル(2,067 社)を，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 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２で示す表現 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　表現 A

予想につきましては 見通しにつきましては

表２　表現 B

業績，業績の， 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．上記のＡとＢの組み合わせにより，計 10 個の手がかり表現を作成した．この手がかり表現を含む文を業績予測文として抽出した．

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．酒井らは，決算短信 PDF から抽出した手がかり表現と企業キーワードを使用して，決算短信 PDF から業績要因の抽出を行った．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，上記の手法では，業績予測文が獲得されない決算短信 PDF が多く存在した．それを確かめるために，簡易的な評価を行い，精度，網羅率を求める．ここで，精度は本手法により得られた業績予測文から無作為に選別した 10 個の業績予測文に対して，人手にて正解かどうかを判断した．網羅率は，今回収集した業績予測に修正のある決算短信 PDF ファイル 8213 個中，今回得られた業績予測文のファイル数が 337 個であったことより算出した． 評価の結果，精度は 100%であったが，網羅率は 4．1%であった．

結果を見ると，精度は高くなるが，網羅率が低く，良い結果とはいえない．網羅率が低い理由は，手がかり表現の種類が少ないからであり，新たな手がかり表現を獲得する必要がある．

5．文末手がかり表現の獲得

4．2 節で得られた業績予測文より，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．

5.1　文末リストの作成

まず，4．2 節で得られた業績予測文から，文末に 3 回以上出現する文節のリストを作成する．得られた文末リストの例を以下に示す．

> 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

得られた文末リストから，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．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，文末の文節と，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，文末の文節とそれに係る文節列との組み合わせを絞り込む．

5.2　文末の文節とそれに係る文節列との組み合わせの絞り込み

5．1 節で得られた文末リストに示した文節から，その前に係る文節列を取得し，文末の文節とそれに係る文節列との組み合わせで，新たな文末手がかり表現を抽出することを考える．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．

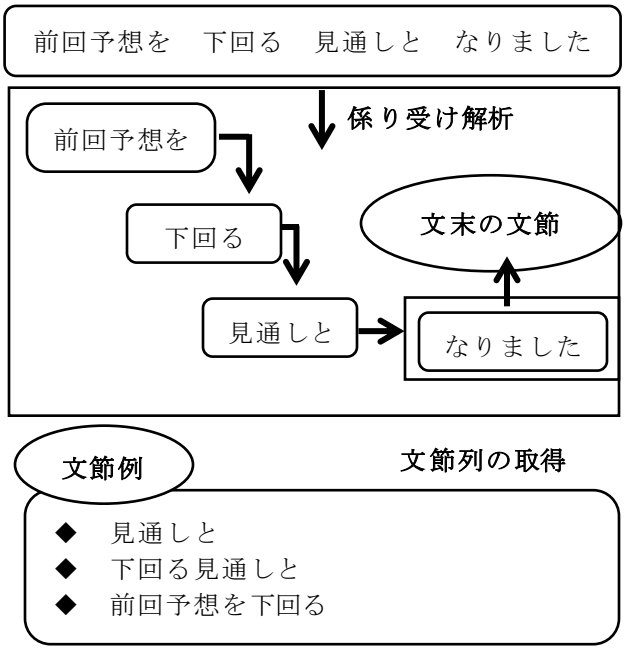

図 1　文節列の取得例

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分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，4. 2 節の手法により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,
$f(p,c)$：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に含まれる分節の数,

ここで，スコアScore (p,c)が高い文節列と文末文節の組み合わせの例を示す.

下回る見込みとなりました
前回予想を下回る見込みとなりました
下回る見込みとなりました
下回る見通しとなりました
見通しとなりました

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.

6. 文末手がかり表現を使用した業績予測文の抽出

5 章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，文末手がかり表現を含む文を抽出するが，それだけでは業績要因を含んでいない文を抽出する可能性がある．そこで，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．以下に抽出した業績予測文の例をいつくか示す.

会社名：雪国まいたけ
企業キーワード：まいたけ, 最需要期
文末てがかり表現：下回る見込みであります
【業績予測文】
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

図 2　業績予測文の例 1

会社名：KDDI
企業キーワード：**端末販売収入，スマートフォンシフト，データ通信料収入**
文末手がかり表現：上方修正いたしました
【業績予測文】
営業収益，営業利益，経常利益については，ＬＴＥ対応端末発売に伴う**端末販売収入**の増加や，**スマートフォンシフト**に伴う**データ通信料収入**の増加が当初予想を上回る見込みとなったため，上方修正いたしました

図 3 業績予測文の例 2

7. 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[1]，係り受け解析器として Cabocha[5]を使用した．

企業名を入力すると，その企業の決算短信 PDF を検索するシステム[2]に，本手法で抽出された業績予測文を組み込んだ．図 4 に「ソニー」で検索した場合に表示される業績予測文を示す．

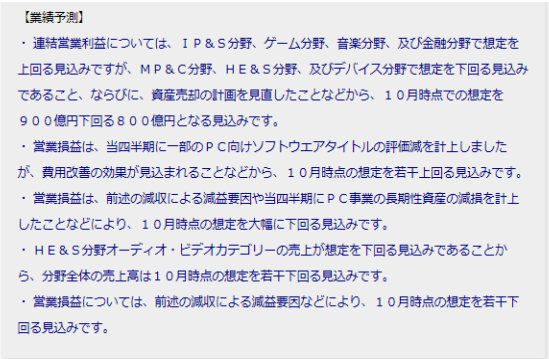
【業績予測】
・連結営業利益については、ＩＰ＆Ｓ分野、ゲーム分野、音楽分野、及び金融分野で想定を上回る見込みですが、ＭＰ＆Ｃ分野、ＨＥ＆Ｓ分野、及びデバイス分野で想定を下回る見込みであること、ならびに、資産売却の計画を見直したことなどから、１０月時点での想定を９００億円下回る８００億円となる見込みです。
・営業損益は、当四半期に一部のＰＣ向けソフトウエアタイトルの評価減を計上しましたが、費用改善の効果が見込まれることなどから、１０月時点の想定を若干上回る見込みです。
・営業損益は、前述の減収による減益要因や当四半期にＰＣ事業の長期性資産の減損を計上したことなどにより、１０月時点の想定を大幅に下回る見込みです。
・ＨＥ＆Ｓ分野オーディオ・ビデオカテゴリーの売上が想定を下回る見込みであることから、分野全体の売上高は１０月時点の想定を若干下回る見込みです。
・営業損益については、前述の減収による減益要因などにより、１０月時点の想定を若干下回る見込みです。

図 4「ソニー」の検索結果

8. 評価

本手法の評価を以下の方法で行った．まず，本手法により抽出された業績予測文のなかから無作為に 100 個選別し，投資歴 10 年以上の個人投資家に評価してもらった．その結果，82 個の業績予測文が正解であった．本手法の再現率の評価は以下の方法で行った．まず，業績予測に修正のある決算短信 PDF ファイルを無作為に 50 個選別し，人手にて業績予測文を抽出し正解データとした．そして，この 50 個の決算短信 PDF に対して本手法を適用し，本手法にて得られた業績予測文を含む決算短信 PDF ファイルと，正解データの決算短信 PDF ファイルが一致すれば正解とし，再現率を算出した．それらの評価結果を表 1 に示す．

表 1 本手法による業績予測文の精度，再現率

精度	再現率
82%	56%

比較手法として，５．１節で示した文末文節のみで業績予測文を抽出する手法の評価結果を示す．ここで，企業キーワードは本手法と同一である．

表 2 文末文節と企業キーワードで抽出した業績予測文の精度，再現率

精度	再現率
35%	94%

9. 考察

本手法の精度は 82%であり，比較的，良好な精度を達成した．ここで文末文節のみを使用した比較手法の精度は 35%であった．これは例えば以下のように，手がかり表現として「なりました」が使用された文を業績予測文として抽出されたからである．

固定負債は，全連結会計年度末と比べて３．０％増加し，１，４５７百万円となりました

それに対して本手法は，「なりました」に係っている文節列との組み合わせを手がかり表現とし「下回る見込みとなりました」を含む文を抽出しているため，上記のような文が抽出されることを防ぐことができた．

[1]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2]http://hawk.ci.seikei.ac.jp/cees/

本手法の再現率は 56%であり，まだ抽出できていない業績予測文が存在している．例えば，以下のような文が抽出されていない．

> 今後の当社グループを取り巻く事業環境は、国内外経済の下振れリスク、原燃料価格の上昇、円高による外需収益の圧迫など先行きの不透明感が懸念されます

上記のような文章を抽出するためには、文末手がかり表現の種類を増やすことが必要である．

10. 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」のような手がかり表現を用いて，業績予測文の抽出を行った．得られた業績予測文を，酒井らの先行研究を用いて，精度の高い業績要因を含む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」のような文末手がかり表現を得た。最終的に，文末手がかり表現と，酒井らの先行研究で得られた企業キーワードを使用して，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 82%，再現率は 56%となり，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] 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)

[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).