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と極性付与

Extraction of Basis Information on Analyst's Forecasts and Assigning Polarity to Analyst Reports

小林和正[1]　酒井浩之[1]　坂地泰紀[2]　平松賢士[3]

Kazumasa Kobayashi[1], Hiroyuki Sakai[1], Hiroki Sakaji[2], Kenji Hiramatsu[3]

[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
[2]東京大学
[2] The University of Tokyo
[3]株式会社アイフィスジャパン
[3] IFIS JAPAN LTD

Abstract: In this paper, we proposed a method for supporting investors. Our method extracts basis information on analyst's forecasts from analyst reports and assigns polarity to the analyst reports. Analyst reports which are written about a company's performance or profitability by securities analysts are useful for investment but investors can only read it a little because many reports are published. Therefore, a system which judges investing by an artificial intelligence technique is required. By giving polarity to the analyst reports, the proposed method catches a slight change in performance. This ability of method is useful to judge whether investors need to read analyst reports carefully.

１．はじめに

近年，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り，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いる．例えば，決算短信から重要な情報を抽出して投資判断の支援を行うといった研究が行われている[3][4][5][6][7].

本研究において分析対象となるアナリストレポートは，証券アナリストが企業の経営状態や収益力などを調査してまとめたものである．業績予測や株価や事業の今後の展望などが記載されており，予想を元にレーティングが付与される．高度な専門知識をもつ証券アナリストによるレポートは，投資判断のための重要な情報源のひとつであり，株価の変動要因にもなりうる．多い時には 1 日に 1200 本以上ものアナリストレポートが発表されることもあるため，全てに目を通し，内容を把握することは困難であり，人工知能分野やテキストマイニングの手法を用いて投資判断を支援する技術が求められている．

そこで本研究は，学習データを自動生成し，深層学習を用いてアナリストレポートからアナリスト予想根拠情報を抽出する手法と，レーティングが変動しないアナリストレポートに対して，深層学習を用いて極性を付与する手法を提案する．ここで，アナリスト予想根拠情報とは，例えば，「世界経済の回復に加え、米ドル安や中国の供給削減期待などから市況は急回復し、同社の輸出も今後大きく改善する可能性が高い。」といった，投資判断やアナリストレポートの内容を把握するうえで重要な，アナリスト予想の根拠情報を含む文（以降，アナリスト予想根拠文とする）と定義される[1]．アナリスト予想根拠文の抽出や極性の付与により，アナリストレポートの内容把握に要する時間の削減や，レーティングが変動しない程度のわずかな業績変化を捉え，そのアナリストレポートを熟読するか判断するための情報となることが期待できる．

アナリストレポートからのアナリスト予想根拠文の抽出は，酒井らの先行研究が存在している[1]．しかし，酒井らの手法におけるアナリスト予想根拠文の再現率は 60%程度であり，それほど高いわけではない．酒井らはアナリスト予想根拠文を抽出するた

めの手がかりとなる表現を「手がかり表現」と定義し，そのような表現をブートストラップ的に獲得する。そして，獲得された手がかり表現を使用してアナリスト予想根拠文を抽出している．しかし，ブートストラップ的に手がかり表現を獲得する過程において，手がかり表現として不適切な表現を削除する必要があり，そのため，適切な手がかり表現であるにもかかわらず，獲得できない場合がある．そのため，酒井らの手法によるアナリスト予想根拠文の抽出は，比較的高い精度（75%程度）を達成しているものの，再現率は低いという結果となっている．

そのため，本研究では，深層学習を使用してアナリストレポートからアナリスト予想根拠文を抽出し，精度を落とさずに酒井らの手法よりも高い再現率を達成する手法を提案する．具体的には，酒井らの手法は高い精度を達成していることに着目し，酒井らの手法による抽出結果をさらに絞り込むことで，より高い精度のアナリスト予想根拠文の集合を作成する．そして，作成された高精度のアナリスト予想根拠文を深層学習の学習データとすることで学習データを自動生成し，その自動生成された学習データを使用して深層学習を行い，アナリスト予想根拠文を抽出する．

２．アナリスト予想根拠情報の抽出

本研究は，アナリストレポートからアナリスト根拠情報を抽出する酒井らの手法[1]によって抽出したアナリスト予想根拠文を深層学習の学習データとして使用する．以下，酒井らの手法[1]について簡単に述べる．

2.1　アナリストレポートからの手がかり表現の自動獲得

アナリスト予想根拠情報の抽出には，業績発表記事や決算短信から業績要因文を抽出した酒井らの手法[2][3]を適用し，アナリストレポートからアナリスト予想根拠情報を抽出する際に有効な手がかりとなる表現（以降,「手がかり表現」と定義）を獲得する．この手がかり表現からアナリスト予想根拠情報を含むアナリスト予想根拠文を抽出する．

まず,「予想する」,「考える」,「高い」の 3 つの手がかり表現を人手で与え,それに係る節を取得する．取得された節の集合において，頻繁に出現する表現を共通頻出表現として抽出する．ここで，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ういう仮定に基づき，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 1 で求める．この値が，式 2 で求めた閾値T_e以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s)\, log_2 P(e,s) \qquad (1)$$

ここで，アナリストレポートの集合において，
$S(e)$: 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.

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

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．

$$T_e = \alpha log_2 |N_s| \qquad (2)$$

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．先ほどと同様に，様々な手がかり表現に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，手がかり表現を選別する．

2.2　手がかり表現と共通頻出表現を使用したアナリスト予想根拠文の抽出

アナリストレポートから獲得した手がかり表現と共通頻出表現を用いて，アナリストレポートからアナリスト予想根拠文を抽出する．まず，アナリストレポートから手がかり表現を含む文を抽出して係り受け解析し，手がかり表現を含む文節を取得する．この文節にかかっている文節を取得して連結する．連結された文節列に共通頻出表現が含まれていれば,その文をアナリスト予想根拠文として抽出する．

2.3　文末手がかり表現の自動獲得

2.1 節で取得された手がかり表現は 1 文節のみだが，多分節で構成されている手がかり表現も多く存在する．アナリストレポートには文末に特徴的な表現が多く出現している傾向があることに着目し，2.2 節で抽出されたアナリスト予想根拠文の文末に出現しており,多文節で構成される手がかり表現(以降,文末手がかり表現と定義）を取得する．まず，2.2 節で抽出されたアナリスト予想根拠文の文末に出現する 1 文節（以降，文末文節と定義）を取得する．これらの文末文節は有効な手がかり表現ではないが，文末文節に係っている文節列を取得して文末文節と

組み合わせると，有効な文末手がかり表現となる可能性がある．そこで、文末手がかり表現を獲得するにあたり、文末文節に係っている文節列を取得するが，文末文節とそれに係る文節列の組み合わせは膨大な数になるため，組み合わせを絞り込む必要がある．文末文節cに係る文節列pに対して式 3 でスコアを求め、平均値を上回る文節列のみを抽出する．そして，抽出した文節列と文末文節とを連結した表現を，文末手がかり表現として獲得する．

$$Score(p,c) = -f(p,c)\sqrt{fp(p)}log_2 P(p,c) \quad (3)$$

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

ただし，アナリストレポートから取得したアナリスト予想根拠文の集合において，
$P(p,c)$: 文末文節cから取得される文字列pの出現確率.
$f(p,c)$: 文末文節cから取得される文字列pの取得回数.
$N(c)$: 文末文節cから取得される文字列の総数.

文末手がかり表現を使用したアナリスト予想根拠情報の抽出は，2.2 節で示した手法と同様である．

３．深層学習によるアナリスト予想根拠文の抽出

3.1　学習データの自動生成

酒井らの手法[1]により抽出されたアナリスト予想根拠文から学習データを生成する．しかし，抽出されたアナリスト予想根拠文の精度は 75%程度であり，それをそのまま学習データとすることはできない．そこで，以下の処理を行うことで，抽出されたアナリスト予想根拠文を絞り込み，その結果を学習データとする．

Step 1: 手がかり表現を 3 つ以上，または，1 つの文末手がかり表現を含んでおり，かつ，共通頻出表現を 3 つ以上含む文において，後述のキーワードに基づくスコアが 30 以上のアナリスト予想根拠文を正例とする．

Step 2: 共通頻出表現，手がかり表現，文末手がかり表現を含まない 20 文字以上の文を負例とする．

Step 3: Step 1 および Step 2 で抽出した学習データを利用し，深層学習でテストデータのアナリストレポートからアナリスト予想根拠文を抽出する．

Step 1 で使用するスコアはアナリスト予想根拠根文に含まれる企業にとって重要なキーワードのスコアの和である．このキーワードは対象企業の決算短信 PDF から抽出している．決算短信 PDF を採用した理由としては，上場企業が年 4 回，決算短信を発表しているのに対し，アナリストレポートは企業によって数にばらつきがあるためである．決算短信 PDF を使用したキーワードのスコア付与は，企業tの決算短信 PDF に含まれる名詞n に対して以下の式 4 で重み$W(s,S(t))$を算出する．

$$W(n_i,S(t)) = \left(0.5 + 0.5\frac{TF(n_i,S(t)}{max_{j=1,\ldots,m}TF\left(n_j,S(t)\right)}\right) \times H(n_i,S(t)) \times \log_2\frac{N}{df(n_i)} \quad (4)$$

ここで，
$S(t)$: ある企業tの決算 PDF の集合.
$TF(n,S(t))$: $S(t)$において，名詞nが出現する頻度.
$H(n,S(t))$: $S(t)$の各決算短信 PDF 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.

Step1 で抽出された正例と，それに含まれる手がかり表現，共通頻出表現，文末手がかり表現をいくつか示す.

しかし、消費材は耐久財に比べて利用頻度が高いことからサイトのＰＶの向上等のメディア価値の向上に繋がると考える。

- 手がかり表現： 高い，繋がると，比べて，考える
- 共通頻出表現： 向上，利用頻度，価値
- 文末手がかり表現： 繋がると考える。

国内トラック需要は強く、新興国市場に強い同社は、本来は成長余地を大きく有している。

- 手がかり表現: 強く，大きく，強い
- 共通頻出表現: 市場，国内，需要
- 文末手がかり表現: なし

Step2 で抽出された負例をいくつか示す. 手がかり表現および共通頻出表現を含まず，文末手がかり表現

もないことがわかる．

ただし、いずれのニュースも非常に大きなインパクトを持つというわけではない。

事業内容Ｆｌａｓｈコンテンツを中心にしたオリジナル動画キャラクタの開発等を展開。

上記の手法により，正例として 20,824 文，負例として正例と同数の 20,824 文の学習データを自動的に生成した．

3.2 素性選択

学習データから入力層の要素となる語（素性）を選択する．自動生成された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，式 5 で重みを計算する．

$$W_p(t,S_p) = TF(t,S_p)H(t,S_p) \qquad (5)$$

ただし，

S_p: 学習データにおける正例のアナリスト予想根拠文の集合．

$TF(t,S_p)$: 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ことがわかる．$H(t,S_p)$は次の式 6 で求める．

$$H(t,S_p) = -\sum_{s\in S_p} P(t,s)log_2P(t,s) \qquad (6)$$

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

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，負例の文に含まれる内容語（名詞，動詞，形容詞）に対して，式 7 で重みを計算する．

$$W_n(t,S_n) = TF(t,S_n)H(t,S_n) \qquad (7)$$

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．

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を満たす語tを素性として選択する．

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

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

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．選択された素性の一部を以下に例示する．

事業，拡大，販売，収益，株価，可能，向け，改善，影響，増加，需要，増益，国内，製品

3.3 モデル

深層学習のモデルについて以下に述べる．入力は 41,648 文の学習データから抽出された 5,699 語を要素，語tにおける$Wp(t,Sp)$，もしくは，$Wn(t,Sn)$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と同じ 5,699 とし，隠れ層は，ノード数 1,000 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．また，エポック数は 50 回，活性化関数として，ReLU を使用した．上記のモデルを図 1 に示す．

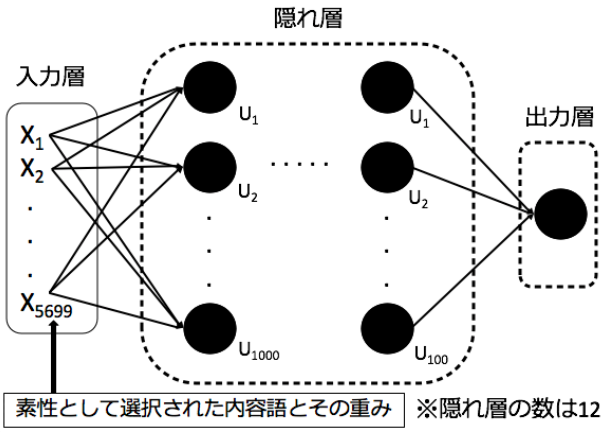

図 1: ニューラルネットワークのモデル

4 アナリストレポートに対する極性付与

アナリスト予想根拠情報の抽出により，アナリストレポートの内容把握に要する時間の削減が期待できるが，レーティングが変動しない程度のわずかな業績変化を捉え，そのアナリストレポートを熟読するか判断するための情報としては，アナリストレポートに対して極性を付与するほうが有効である．アナリストレポートにはレーティングが付与されており，レーティングが上がっていれば，そのアナリストレポートはポジティブな内容が記述されていることが予想できる．レーティングが下がっていればネガティブである．しかし，多くのアナリストレポートではレーティングの変更はないものの，記述されている内容をポジティブ，ネガティブに分類できる．

そこで，本研究では，レーティングが上がったアナリストレポートを正例，レーティングが下がったアナリストレポートを負例とした学習データを生成し，その学習データによる深層学習にて，レーティングの変更がないアナリストレポートに対する極性（ポジティブ，ネガティブ）を試みた．

ここで，学習データからの素性選択は 3.2 節で述べた手法と同じであり，深層学習のモデルは 3.3 節のモデルと同じである．

5 評価

5.1 アナリスト予想根拠文抽出の評価

本手法を実装し，自動生成した学習データを使用して図 1 のモデルを用いて深層学習を行い，アナリストレポートからアナリスト予想根拠文を抽出した．抽出された例を以下に示す．

> 文： ４月からの薬価改定影響、ジェネリック促進策による市場拡大効果を踏まえ業績見通しを見直した。

> 文： 政府の薬価改定強化がＧｘ市場にも及んでいること、ブランド医薬品メーカーがＧｘ事業でのプレゼンスを高めていること、この２点が最大の懸念事項だ。

評価のための正解データは，手がかり表現と共通頻出表現を獲得したアナリストレポートを含まない，すなわち深層学習の学習データとして使用していないアナリストレポート集合から 12 個のアナリストレポートを無作為に選択し，その中の 468 文から人手でアナリスト予想根拠文を抽出して作成した．次に，選択したアナリストレポートから本手法にて抽出したアナリスト予想根拠文が正解データの文と一致すれば正解とし，精度，再現率，F 値を算出した．

本手法 1: 深層学習による手法の抽出結果と，手がかり表現と文末手がかり表現を使用した既提案手法[1]の抽出結果との和集合をとる手法

本手法 2: 深層学習による手法のみ

本手法 3: 深層学習による抽出結果と，既提案手法による抽出結果との積集合をとる手法

比較手法: 手がかり表現と文末手がかり表現を使用する既提案手法

評価結果を表２に示す．

表２： アナリスト予想根拠文抽出の評価結果

手法	精度(%)	再現率(%)	F 値
本手法 1	75.13	69.60	72.25
本手法 2	83.33	44.11	57.68
本手法 3	83.90	35.78	50.16
比較手法	75.00	61.76	67.73

5.2 アナリストレポートへの極性付与の評価

４章の手法を実装し，7,454 個のアナリストレポートが学習データとして自動生成された．そして，この学習データによる深層学習を用いて，評価用のレーティングが変動しなかった 155 個のアナリストレポートに対して極性を付与した．正解データを評価用と同じアナリストレポートを人手にて極性を付与することで作成し，本手法の精度を求めた．また，学習手法として SVM を使用した場合を比較手法とした．SVM の場合も，学習データ，および，素性は深層学習と同じである．評価結果を表３に示す．

表３： 極性付与の評価結果

手法	Positive 精度(%)	Negative 精度(%)	全体 精度(%)
深層学習	75.0 (84/112)	76.7 (33/43)	75.5 (117/155)
ＳＶＭ	79.5 (70/88)	64.2 (43/67)	72.9 (113/155)

6　考察

本手法 1 は比較手法と比較して，精度，再現率，ともに高くなっており，良好な結果と言える．深層学習と手がかり表現と文末手がかり表現を使用する手法を組み合わせて和集合をとった結果が，積集合をとった結果や深層学習の結果より向上している．これは，手がかり表現と文末手がかり表現を使用する手法で抽出できなかったアナリスト予想根拠文は深層学習で抽出できており，さらに，深層学習で抽出できなかったアナリスト予想根拠文は手がかり表現と文末手がかり表現を使用する手法で抽出できていることを示している．

深層学習のみを使用した本手法 2 において，精度が高く，再現率が低い原因としては，負例に正例に分類されるべき文が含まれていたからであると考えられる．手がかり表現，共通頻出表現，文末手がかり表現を含んでいないとしても，アナリスト予想根拠文である可能性があり，負例の抽出条件を再考する必要がある．

7　まとめ

本研究では，アナリストレポートの内容を把握するうえで重要な，アナリスト予想の根拠情報を含む文（アナリスト予想根拠文）を自動的に抽出する手法を提案した．アナリストレポートから抽出されたアナリスト予想根拠文のみを提示することで，アナリストレポートの内容把握に必要な時間を削減できたり，アナリスト予想根拠文の中からその企業の事業に関連する根拠情報のみを選別することで，その企業への投資判断やアナリストレポートを熟読するかどうかを判断するための情報となることが期待できる．

本手法は酒井らの手法[1]から学習データを自動生成し，深層学習を行うことで，酒井らの手法[1]で抽出できなかったアナリスト予想根拠文を獲得できた．評価の結果，精度 75.13%，再現率 69.60%であり，酒井らの手法[1]と比較しても良好な精度，再現率を達成した．

参考文献

[1] 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.

[2] Hiroyuki Sakai, Shigeru Masuyama, “Cause Information Extraction from Financial Articles Concerning Business Performance”, IEICE Trans. Information and Systems, vol.ED, no.4, pp.959-968, 2008.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.93-98, 2017.