The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

新聞記事からの因果関係を考慮したアナリストレポートの自動要約文生成

Automatic Summarization of Analyst Reports Based on Causal Relationships from News Articles

高嶺 航 *1 Wataru Takamine
和泉 潔 *1 Kiyoshi Izumi
坂地 泰紀 *1 Hiroki Sakaji
松島 裕康 *1 Hiroyasu Matsushima
島田尚 *1 Takashi Shimada
清水 康弘 *2 Yasuhiro Shimizu

*1東京大学大学院工学系研究科 School of Engineering, The University of Tokyo
*2野村證券株式会社 Nomura Securities Co., Ltd.

In this paper, we focused on the causal relationships in both of news articles and analyst reports. We proposed a novel approach for summarizing analyst reports automatically based on the causal relationships extracted from both text data. As a first step toward summarization of analyst reports adequately, we analyzed the validity of the method in extracting causal relationships which can be evaluated from the analyst reports. As a result, the proposed method could extract basis information of analyst's opinions from analyst reports with some accuracy, and we could confirm the styles of analysts in expression of opinions and bases.

1. はじめに

近年，投資家に対する投資判断の支援を行う技術の必要性が高まってきており，投資判断材料の一つであるアナリストレポートの活用に注目が集まってる．アナリストレポートには，証券市場調査・分析の専門家である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た結果がまとめられており，企業の業績や株価に対する証券アナリストの予想と根拠が示されている．記述されている予想の根拠としては，その企業の取り組む事業の近況・財務状況（企業のファンダメンタルズ），事業に影響を与える経済・政治・社会状況（マクロ経済のファンダメンタルズ）などの外部要因についても言及されている．このように，高度な専門知識をもつアナリストによる詳細なレポートは，株価の変動要因にもなりうる [1] ため，彼らの企業の業績や株価に対する予想やその裏付けとなる根拠を投資判断の材料として活用することは有用性が高いと思われる．

しかしながら，アナリストレポートの発行の多くは決算発表の時期に集中し，膨大なレポートの全てを熟読するのは難しく，レポートの内容を十分に把握できない可能性がある．

この問題に対して，近年，自然言語処理やテキストマイニング技術の進展により，膨大な量のアナリストレポートから重要な要点のみを自動で抽出・要約する技術のニーズが高まっており，研究事例も報告されている [2][3]．このように，投資判断材料に必要な情報を要約することができれば，レポートを読む負担が減り，時間の制約がある中でもレポートの内容の要点を把握することができる．しかしながら，これら [2][3] の要約技術は，テキストに記述されている事象の背景にある因果関係を考慮していない．そのため，生成された要約文に，投資判断材料となりうる証券アナリストの予想の根拠が盛り込まれていない場合が想定される．これに対して，[4] では文の因果関係の構造に注目し，原因表現を取り出す手法を提案している．

このように，アナリストレポートの活用として自動的に重要な箇所を要約・抽出，あるいは検索する技術が研究されている．しかしながら，2 つの異なる媒体 (つまり，アナリストレポートとそれ以外のテキスト情報) から一つの要約文を生成する手法はまだ確立されていない．この手法の確立により，アナリストレポート中で業績・株価予想の根拠として言及される情報の特徴を捉えるだけではなく，新聞記事などの媒体からその根拠の背景についての情報を補うことで，より説明できる情報を含んだ要約文の生成が期待できる．さらに，これが可能になれば，要約の過程で抽出する証券アナリストの株価・業績予想につながる根拠や背景としての経済情報を検索できるようになり，証券アナリストのレポート作成支援としても期待できる．

そこで本研究では，因果関係を考慮しながら別の媒体から補填的に情報を抽出し，要約文を自動生成する手法を提案し，その評価を行う．本稿では，その提案手法実現のために，アナリストレポートから根拠情報を抽出する手法の妥当性について実験を行った．

連絡先: 高嶺航, 東京大学大学院工学系研究科技術経営戦略学専攻, 113–8656 東京都文京区本郷 7–3–1 工学部 8 号館 530 室, m2018wtakamine@socsim.org

2. 提案手法

本章では，テキストデータから因果関係表現を抽出し，要約文を自動生成する手法について述べる．2.1 節では本手法の概観，2.2 節ではアナリストレポートおよび新聞記事からの因果関係表現抽出の概要，そして 2.3 節では，要約文における根拠の背景情報の獲得に用いた表現類似度計算手法の概要を述べる．

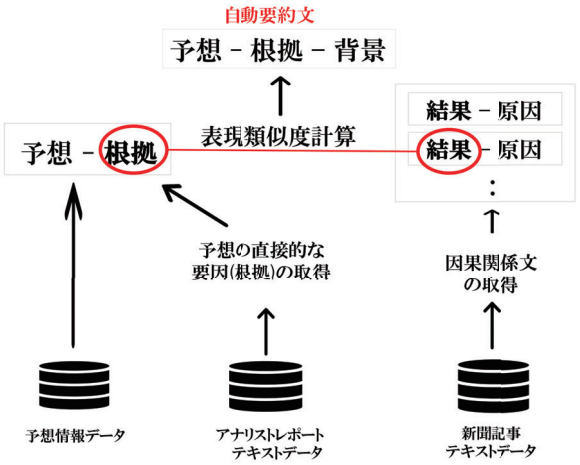

図 1: 提案手法の概説

2.1 要約文生成手法の概説

本節では要約文生成手法の概要を述べる．全体の流れを図 1 に示す．まず，各アナリストレポートにおける証券アナリストの企業業績・株価の予想の情報を獲得する．次に，アナリストレポート本文中に出現する因果関係の構造を抽出し，その中でも結果表現に株価・業績予想が含まれる因果関係を獲得する．獲得した因果関係の原因表現を証券アナリストによる企業業績・株価の予想の根拠情報として獲得する．ただし，この時，証券アナリストの予想情報とその予想の根拠情報は同文章内に出現するものと仮定している．つまり本手法では図 2 のように文章横断的に出現する因果関係表現は抽出しないこととする．また同様に，新聞記事からも因果関係の構造を抽出し，因果関係の結果表現のうち，獲得した証券アナリストの予想の根拠情報と表現が類似する文章を探索する．そして，類似性の高い因果関係における原因表現を根拠情報の背景情報として獲得する．このようにして獲得した，(1) 証券アナリストの企業業績・株価予想，(2) 予想の根拠（直接的な要因），(3) 根拠の背景をまとめ，アナリストレポートの要約文を自動生成する．

本自動要約手法の実装例として，Web サーバー上のシステムとして実装したものの動作画面を図 3 に示す．銘柄，期間を入力すると，すでに生成済みの要約文のうちから入力情報に合致する要約文を出力する．要約文の構成は，一文目に証券アナリストレポートの業績・株価予想，二文目に証券アナリストの予想の根拠，三文目にその根拠の背景を想定している．

文書 1 : 業績予想を下方修正(結果表現)

文書 2 : ○○を織り込んだ。(原因表現)

図 2: 文章横断的に出現する根拠情報の例

図 3: 想定している提案手法を用いたシステムの動作画面

2.2 因果関係抽出手法

アナリストレポート・日経新聞記事から酒井ら [4] の手法を用いて因果関係表現を抽出する．この手法では，因果関係表現を特徴付ける手がかり表現と，手がかり表現に係る節の中で共通して頻繁に出現する共通頻出表現を定義する．最初に少数の手がかり表現と共通頻出表現を与えることで，互いに係り受け関係にある新たな共通頻出表現と手がかり表現が連鎖的に獲得される．この手法を用いる場合，アナリストレポートにおいては特にアナリストの予想を示す文の部分と，その予想の根拠を示す文の部分を分離して抽出する．前者を予想部，後者を根拠部と呼ぶ．

アナリストレポート中から抽出した予想部と根拠部の例を図 4 に示す．この文章の場合，「主力の制御事業の順調な拡大を」を根拠部，「主因に」が手がかり表現，「目標株価を上方修正」が予想部となる．酒井ら [1] は，アナリストレポートからアナリストの予想と根拠情報の抽出を行なっているが，アナリスト予想根拠文の抽出方法として，共通頻出表現の数を用いてアナリストの予想根拠文かどうかを判定している．本手法では，予想の直接的な要因を根拠情報と定義しており，結果表現に業績予想が含まれる因果関係の原因表現を証券アナリストによる企業業績・株価予想の根拠情報として抽出する．

主力の制御事業の順調な受注拡大を

(根拠部)

主因に、目標株価を上方修正

(手がかり表現)　　(予想部)

図 4: アナリストレポートから抽出した予想部と根拠部の例

2.3 表現類似計算手法

本節では，アナリストレポートの根拠の背景情報を新聞記事から獲得するために用いた，二つの文章の表現類似度を計算する手法について述べる．本研究では，表現類似度 s を以下の式のように話題性 t，文の表層 w，極性の一致度 p，文脈の類似性 c の 4 つの構成要素として捉える．

$$s = t \cdot w \cdot p \cdot c \tag{1}$$

- 話題性 : トピックモデル (LDA[5]) による単語の分散表現を用いた文章の話題の類似度を算出
- 文の表層 : Word2vec(Skip-gram[6][7]) による単語の分散表現を用いた文章の表層的な類似度を算出
- 極性の一致度 : 金融極性辞書 [8] を用いた単語の極性を計算し，文章間の極性がどれだけ一致するかを判定
- 文脈の類似性 : アナリストレポートの根拠情報と新聞記事の結果表現の類似度を算出するだけではなく，新聞記事の原因表現との類似度も算出．より根拠情報の文意に沿った文章を抽出する．

LDA と Word2vec を用いて算出された二つのベクトル表現を用いることで，比較する二つの文書の話題性と表層的な類似度を算出する．ベクトル間類似度はコサイン類似度を用いた．A，B はそれぞれ文章 $\vec{A}$，$\vec{B}$ は，それぞれ文書 A，B 内にある名詞・動詞・形容詞の分散表現の相加平均を求めて算出した文書ベクトルである．

$$\text{cosine similarity} = \cos\theta = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\|\vec{B}\|} \tag{2}$$

表 1: 実験に用いた手がかり表現の概要

手がかり表現の数	手がかり表現の例
109	織り込んで，見込んで，をきっかけに，背景に，考慮し，踏まえ など

3. 実験

提案手法では，おおよそ同一文章内で因果関係表現が出現し，予想とその根拠情報が獲得できるという仮説に基づいて，酒井ら [4] の因果関係抽出手法を使用している．本実験では，対象としているアナリストレポートにおいて，同一文章内で予想の根拠情報が獲得できた件数の割合を示し，根拠情報抽出をする手法の妥当性を検証する．実験には，表 1 に示す手がかり表現を用いた．

実験データには，2011 年から 2016 年までの間に発行された 7927 件のアナリストレポートのうち文章内で因果関係表現が抽出できた 7716 件を用いた．アナリストレポートから抽出した因果表現を含む文章の中から結果表現の部分に「目標株価」および「業績予想」に関する記述がある場合，その原因表現を予想の根拠情報として抽出している．なお，因果関係を抽出するにあたって，本実験では形態素解析器としては Mecab を用い，係り受け解析器としては Cabocha[9] を用いた．

4. 実験結果と考察

評価方法に関しては，文章内で因果関係抽出ができたレポートの件数に対する証券アナリストの予想の根拠情報の抽出ができたレポートの件数の割合を Precision(精度) とした．目標株価に対する根拠情報，業績予想に対する根拠情報，そしてどちらか一方に対する根拠情報が抽出できた割合の 3 項目を算出した．

実験結果を表 2 に示す．検証した全項目で 5 割を下回る精度となり，因果関係ができた抽出した文のうち，結果表現にて目標株価と業績予想のいずれかに言及しているアナリストレポートの件数は，4 割程度であった．必ずしも予想に対して直接的な表現を使用している訳ではないことが分かる．この要因として次の 3 点が挙げられる．

1. 証券アナリストの予想に表記揺れがある（例：野村予想を上方修正，利益予想を引き上げる）
2. 明確な根拠表現を回避する
3. 文章横断的な根拠表現が抽出できない

このうち，1 の予想情報の表記揺れについて検討する．具体的には結果表現に含まれる記述として「目標株価」,「業績予想」に加え,「利益予想」,「野村予想」,「収益」,「売上高」等，計 13 個のフレーズがある場合，予想の根拠情報として抽出を行なった．目標株価および業績予想の根拠情報の抽出割合における表記揺れの考慮の結果を表 2 に示す．

表記揺れを考慮することによって 8 割程度まで精度を向上することができた．予想情報の言及に関して，証券アナリストは複数の言い回しをしており，その表記揺れを考慮に入れて根拠情報を抽出する必要があることが分かった．本結果より，今回実験を行なったアナリストレポートの 8 割程度が予想の根拠情報を同一文章内にて言及しており、本手法における因果関係表現抽出手法の有用性が示すことができた．

表 2: 各予想に対する精度（Precision）

	Precision
目標株価のみ	0.20
業績予想のみ	0.33
目標株価あるいは業績予想	**0.44**

表 3: 表記揺れを考慮に入れた結果（Precision）

	Precision
表記揺れを考慮しない場合	0.44
表記揺れを考慮した場合	**0.83**

5. まとめ

本研究では，新聞記事からの情報を活用し，文章内に出現する因果関係表現と文章間の表現類似性に着目したアナリストレポートの自動要約手法を提案した．その提案手法実現のために，アナリストレポートから根拠情報を抽出する手法の妥当性について実験を行った．表記揺れを考慮することによって，8 割程度のアナリストレポートで同一文章内に証券アナリストの予想とその根拠情報が出現していることが分かり，本論文において用いた因果関係抽出手法の有用性が示された．また，実験結果を通じて証券アナリストのレポート内の書きぶりに関しても考察を行った．今後の課題として，要約文の評価データセットの作成，2 節で紹介した表現類似度計算の精度向上に関する手法の考案，因果関係抽出の精度向上に寄与する手がかり表現の語義曖昧性解消手法の考案などが考えられる．

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. 人工知能学会第 17 回金融情報学研究会, pp. 25–30, 2016.

[2] Jahna Otterbacher, Güneş Erkan, and Dragomir R Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 915–922. Association for Computational Linguistics, 2005.

[3] Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 246–254. Association for Computational Linguistics, 2009.

[4] Hiroyuki SAKAI and Shigeru MASUYAMA. Cause information extraction from financial articles concerning business performance. *IEICE Transactions on Information and Systems*, Vol. E91.D, No. 4, pp. 959–968, 2008.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[8] 伊藤友貴, 坪内孝太, 山下達雄, 和泉潔. テキスト情報から生成された極性辞書を用いた市場動向分析. 人工知能学会全国大会論文集 2017 年度人工知能学会全国大会 (第 31 回) 論文集, pp. 2D21–2D21. 一般社団法人 人工知能学会, 2017.

[9] 工藤拓, 松本裕治ほか. 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.