株式市場の出来事の長期的視野での理解を支援するニュース記事抽出によるストーリー可視化

木下聖 [1]　西村太一 [1]　亀甲博貴 [2]　森信介 [2]
[1] 京都大学大学院 情報学研究科　[2] 京都大学 学術情報メディアセンター
kinoshita.sho.v85@kyoto-u.jp　nishimura.taichi.43x@st.kyoto-u.ac.jp
{kameko,forest}@i.kyoto-u.ac.jp

概要

近年，政府が投資家の裾野の拡大を目指すうえで，個人投資家の金融リテラシー不足を課題にあげている．本研究では，証券市場特有の株価および市況概況記事を使用し，株式投資家による過去の証券市場の出来事の長期的視野での理解の支援に特化したニュース記事抽出によるストーリー可視化問題を提案する．さらに提案問題を市況概況記事を利用した重要株価材料抽出，類似株価材料抽出および関連記事検索の３つの部分問題に分割して解く手法を提案する．実験により提案手法を用いたストーリー可視化が可能である一方で，検索手法に改善の余地が残されていることも示された．

1　緒論

近年，政府は投資家の裾野を拡大し，貯蓄から投資へのシフトの実現を目指しており，国民の金融リテラシー不足を課題の１つにあげている [1]．世界的に有名な投資家らは，投資で成功するためには歴史から学ぶ重要性，および，多くの投資家は歴史から学ばずに失敗してきたことを指摘している [2, 3, 4]．人々が証券市場の歴史を学ぶには，専門家による解説文献を読むことが最も一般的であるが，以下の２つの問題が含まれる．第一に，証券市場の歴史の解説文献は未来の結末を知る専門家が過去を解説するもので，バイアスを含む可能性が高い．第二に，個人投資家が現在の情勢を知る情報源としてニュースが一般的であり，歴史の解説文献のような未来を含む長期的視野からの情報は存在せず，現在と過去を比較するために使用することが難しい．一方で，過去のニュース記事は当時の専門家が執筆した情報であり，現在のニュース記事と比較することも可能である．そこで，投資家による証券市場の歴史の学習に過去のニュース記事を利用できれば良いと考えられるが，ニュース記事の数は膨大で，個人投資家がすべてのニュース記事を確認することは非現実的である．

出来事の長期的視野での理解を支援するために，長期間の大量のニュース記事をトピック分類し時系列表示してニュースのストーリーを可視化する技術の研究が行われている [5, 6, 7]．証券市場でもあるトピックに関するニュースが長期間報道され株価に大きく影響することは珍しくなく，これらの技術を活用することで過去のニュース記事を用いた証券市場の歴史の学習を支援できる．一方で，これらの技術を証券市場の歴史学習の支援に使用する上で，以下の 3 つがさらに求められる．

1. ニュース記事を投資家が認識する粒度のトピックで分類すること．
2. 株価指数に影響するような全投資家に重要なトピックを抽出すること．
3. 各トピックの大量の関連記事から重要な記事を抽出すること．

そこで本研究では，株式投資家による過去の証券市場の出来事の長期的視野での理解の支援に特化したニュース記事抽出によるストーリー可視化問題を提案する．提案問題の出力例を表 1 に示す．提案問題では，指定された期間の大量のニュース記事から重要な株価材料を抽出し，各株価材料の関連記事で重要なものを時系列表示する．そして，証券市場特有の株価および市況概況記事を利用し，重要株価材料抽出，類似株価材料抽出，および，関連記事検索の 3 つの部分問題に分割して解く手法を提案する．市況概況記事とは，証券市場の毎営業日後に報道される記事で，その日の株価変動および投資家が意識する株価材料を説明する．さらに，提案手法で使用する系列ラベリングの検証のために，我々が作成し

表 1　株価材料「福島原発事故の深刻化に対する懸念」(2011 年 3 月 15 日) の関連記事出力例（一部抜粋）

date	title
2011/3/11	福島第一原発障害で原子力緊急事態宣言を発令，放射能漏れなく被害出る状況ではない＝枝野官房長官
2011/3/15	福島原発でさらなる漏えいの可能性，30 キロ以内の住民に屋内退避求める＝菅首相
2011/3/18	福島第 1 原発からの放射能漏れは局所的，差し迫るリスクない＝WHO
2011/3/25	福島第 1 原発 3 号機の原子炉損傷の可能性，20-30 キロは自主的避難望ましい
2011/4/12	福島第 1 原発事故を「レベル 7」に引き上げ=原子力安全・保安院
2011/4/14	結果的に心配与え遺憾=原発周囲 10 年住めないとの首相発言報道で枝野官房長官

たデータセットを用いて評価する．また，モデル出力例を示し提案モデルを評価する．

2　関連研究

証券市場の毎営業日後に報道される市況分析記事から因果関係を抽出する研究 [8] や，それらを要約し長期間の市況分析を生成する研究 [9, 10] が行われている．市況分析記事における株価材料の記載は非常に簡潔である．本研究では，各株価材料の関連記事を検索することで，これから投資を始める人にもわかりやすく詳細な情報の出力を目指す．

3　市況概況記事を用いたニューストーリー生成手法

3.1　手法概要

本研究では，証券市場特有の株価および市況概況記事を使用し，重要株価材料抽出，類似株価材料抽出，および，関連記事検索の 3 つの部分問題に分割して解く．各部分問題を解く手法の詳細を以下で説明する．

3.2　重要株価材料抽出

指定された期間内の市況概況記事から株式市場における重要度の高い株価材料を抽出する．本研究では日経平均株価の前日比変動率が大きい日の株価材料の重要度が高いと仮定し，日経平均株価の前日比変動率が大きい日に出版された市況概況記事の株価材料を抽出する．日経平均株価の前日比変動率の大きさは以下の式で求める．

$$r_d = \frac{|p_d - p_{d-1}|}{p_{d-1}} \tag{1}$$

r_d：日付 d の前日比株価変動率
p_d：日付 d の株価

本研究では市況概況記事から株価材料を抽出する問題を系列ラベリング問題として定式化して解く．系列ラベリング問題を解くモデルは，Akbik ら [11, 12] が開発する Flair framework を使用して実装し，Flair embeddings の日本語学習済みモデルでベクトル化したものを，LSTM および全結合層に入力する．

3.3　類似株価材料抽出

すべてのニュース記事から 3.2 節で抽出された重要株価材料の関連記事を検索する場合，関連記事検索の正解データ作成コストが大きいこと，および，株式市場に影響しない重要度の低い関連記事も多く抽出されてしまうことが問題となる．そこで，3.2 節で抽出された株価材料に関連する株価材料が市況概況記事に記載されている日付および株価材料を抽出する．これにより，各株価材料の関連記事検索範囲を高々数日に限定でき，正解データ作成コストを大幅に削減できること，株価材料が市況概況記事に記載されるほど重要であった日の記事のみが抽出されることが期待される．市況概況記事から株価材料を抽出するモデルは 3.2 節と共通のものを使用する．株価材料が関連するものか判定する手法として，鈴木ら [13] が日本語 Wikipedia および金融文書を事前学習した BERT 言語モデル”bert-small-japanese-fin”[1] に株価材料を入力し，各 token のベクトルを平均プーリングして得られるベクトル同士の cos 類似度が閾値以上の場合に関連すると判定する．

3.4　関連記事検索

3.2 節で抽出された重要株価材料の関連記事を抽出する．東京証券取引所における内国株式の売買立会時間は 15 時に終了するため，東京証券取引所の前営業日 15 時から 3.3 節で抽出された日付の 15 時までに報道されたニュース記事から，3.3 節で抽出された類似株価材料の関連記事を検索することで実現する．正解データセットの数が不十分で教師あり学習をすることは難しいため，記事検索手法とし

1）”bert-small-japanese-fin”：https://huggingface.co/izumi-lab/bert-small-japanese-fin

表 2　株価材料抽出データサイズ

	2007-2015	2016-2017
総記事数	344	42
総単語数	86,154	9,182
米株 単語/系列数	713 / 179	54 / 15
為替 単語/系列数	1502 / 244	246 / 27
その他 単語/系列数	5,663 / 589	398 / 49

表 3　株価材料抽出の実験結果

	F-score	support
米国株価動向	0.95	15
為替動向	0.96	27
その他株価材料	0.76	49

て，**??**節と同様に株価材料とニュース記事タイトルをベクトル化し cos 類似度が上位のものを抽出することで検索を行った．このとき，株価材料に固有名詞が含まれる場合は，その固有名詞が本文に含まれるニュース記事を優先して検索した．

4　データセット

4.1　ニュースデータ

金融機関が契約する大手情報サービス会社として，Bloomberg，ロイターおよび日経 QUICK が有名である．本研究では，2007 年から 2017 年のロイター日本語ニュースを使用する．

4.2　系列ラベリング問題データセット

市況概況記事から株価材料を抽出する系列ラベリング問題の学習および評価に使用するデータセットについて説明する．ロイター日本語ニュースでは，東京証券取引所の売買立会時間終了後に，「東京マーケット・サマリー」をタイトルに含む市況概況記事が掲載される．重要な株価材料が掲載される日は株価変動率が比較的大きくなる傾向があるため，日経平均株価の前日比変動率の大きさが $r_d = 0.02$ 以上の日の最後に報道された市況概況記事を使用する．正解ラベルのアノテーションは，テキストアノテーションツール doccano [14] を使用し人手で行った．東京株式市場の市況概況記事では，株価材料として頻繁に米国株価動向および為替動向について言及されるため，これらの株価材料言及部分には特別なラベルを付与し，それ以外の株価材料には同一のラベルを付与した．ただし，「月末」のように証券市場の前営業日から当日までに関連ニュース記事が報道されないものや，「企業決算」のように関連記事が無数にあるものはラベル付与しない．各ラベルが付与されたフレーズの例と株価材料だがラベルを付与しない例を以下に示す．

例：米国株価動向言及部分

米株が急落，海外株式市場，世界同時株安

例：為替動向言及部分

為替が円高/ドル安，主要通貨，円キャリー取引

例：その他株価材料言及部分

日銀短観，米雇用統計，米サブプライムローン問題，米国の利下げ，FOMC，参院選，原油安

例：株価材料だがラベル付与されないもの

業績不透明感，ポジション調整，月末，企業決算，９月中間期末

さらに，作成したデータセットのサイズを表 2 に示す．

5　評価

5.1　系列ラベリング問題

5.1.1　実験設定

4.2 節のデータセットを使用する．2007 年から 2015 年の記事を 10 分割した交差検証によりモデルの学習を行い，2016 年以降の記事の各単語のラベル分類問題の F-score で評価した．

5.1.2　実験結果

学習および評価を 5 回行った結果の平均は表 3 の通りになった．米国株価動向および為替動向は，表現の種類が少ないため非常に高い精度で抽出できた．一方で，その他株価材料の表現の種類は非常に多く精度が少し下がるが，誤り例を確認すると大多数が抽出範囲の長短によるものであった．以下に誤り例を示す．ただし，下線部分がモデルによる抽出，赤字部分が正解ラベルとする．

抽出範囲の誤り例

... ドル／円が１１８円台を回復したことに加え，輸出・輸入ともにマイナス幅が縮小した１２月中国貿易統計も支えとなった．...

不適切な抽出の誤り例

... 昨年来安値を下回り，量的・質的金融緩和（QQE）第２弾が決定された２０１４年１０月３１日以来，約１年３カ月ぶりの安値となった．時間外で２８ドル割れとなった米原油先物を受けてリスクオフムードが強まり全面安．...

5.2 類似度を用いた関連記事検索

本節では類似度を用いた関連記事検索の出力例を示し考察する．

表 1 は 2011 年 3 月 15 日の重要株価材料「福島原発事故の深刻化に対する懸念」に対する関連記事の出力例で，適切な関連記事を出力している．

一方で，表 4 は 2014 年 3 月 14 日の重要株価材料「ウクライナ情勢に対する警戒感」に対する関連記事の出力例である．こちらでは，タイの政情に関する記事や米国景気に関する記事も出力されている．これは，3.3 節の類似株価材料抽出で，異なる株価材料を同一と判定してしまったことに起因する．3.3 節の類似株価材料抽出および 3.4 節の関連記事検索では，BERT の事前学習モデルによるベクトルの類似度を利用しているが，前後に同様の単語が並ぶ単語の類似度が高くなるために誤った判定をしてしまう場合が多くみられ，特に「米国/中国の利下げ」のような国の違いを区別できない例や，「情勢/景気」のような単語を区別できない例が多かった．

関連記事検索手法の改善が必要であり，教師あり学習をすることが好ましいと考えられるが，ニュース記事は 1 日に数百も掲載され，様々な株価材料で学習するには大量のニュース記事を確認してアノテーションする必要がある．また，キーワード検索を行う方法も考えられるが，市況概況記事での株価材料の表現は市況概況記事特有のものも多いという問題がある．キーワード検索および類似度を用いた検索を併用することで，より優れた検索ができると考えられる．

表 4 2014 年 3 月 14 日「ウクライナ情勢に対する警戒感」関連記事出力例（一部抜粋）

date	title
2014/1/17	タイ・ホットストック＝ AIS など安い，タイ政情を警戒
2014/2/4	[シナリオ] タイの政治的混乱，今後の展開予想
2014/2/13	コラム：米国主導の世界景気回復は本物か＝武田洋子氏
2014/2/27	ウクライナが新閣僚候補発表，ロシア軍は警戒態勢
2014/3/4	アングル：米大統領，ウクライナ情勢で「弱腰」批判再燃も
2014/3/13	不安定なウクライナの状況をロシアは利用＝独首相
2014/3/25	中国国家主席，ウクライナ問題の政治的解決望むと表明＝米ホワイトハウス
2014/5/2	ロシア，ウクライナが「報復的行動」開始と非難
2014/5/4	コラム：対ロシア制裁が効かない理由＝カレツキー氏
2014/8/6	ロシア，ウクライナ国境沿いに軍部隊終結＝ポーランド外相
2014/8/7	ロシア直接介入のリスク高まる＝ウクライナ情勢でポーランド首相

6 結論および今後の課題

本研究では，株式投資家による過去の出来事の長期的視野での理解の支援に特化したニュース記事抽出によるストーリー可視化問題を提案した．証券市場特有の株価および市況概況記事を用いれば，重要株価材料抽出，類似株価材料抽出，および，関連記事検索の３つの部分問題に分割して解くことで実現できることを示した．一方で，事前学習済み言語モデルを利用したベクトルの類似度に基づく関連記事検索手法には問題があり，検索手法に改善が必要であることが示された．

今後の課題として，米国株市場・為替市場の市況概況記事の利用および重要語によるキーワード検索を併用した記事検索手法の開発に取り組むことなどが残されている．

参考文献

[1] 内閣府. 資産所得倍増プラン. 新しい資本主義実現会議（第１３回）, November 2022.

[2] 桑原晃弥. ウォーレン・バフェット 成功の名語録 世界が尊敬する実業家、103 の言葉. PHP 研究所, July 2012.

[3] George Soros. Fallibility, reflexivity, and the human uncertainty principle. **J. Econ. Methodol.**, Vol. 20, No. 4, pp. 309–329, December 2013.

[4] Jim Rogers. 危機の時代 伝説の投資家が語る経済とマネーの未来. 日経 BP, May 2020.

[5] Marieke van Erp, Gleb Satyukov, Piek Vossen, and Marit Nijsen. Discovering and visualising stories in news. In **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 3277–3282, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[6] Philippe Laban and Marti Hearst. newslens: building and visualizing long-ranging news stories. **Proceedings of the Events and Stories in the**, 2017.

[7] Deyu Zhou, Linsen Guo, and Yulan He. Neural storyline extraction model for storyline generation from news articles. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1727–1736, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] 酒井浩之, 坂地泰紀, 和泉潔, 松井藤五郎, 入江圭太郎. 学習データ自動生成による市況分析コメント作成のための要因文と補完情報の抽出. 人工知能学会全国大会論文集, Vol. JSAI2020, pp. 1D3GS1303–1D3GS1303, 2020.

[9] 酒井浩之, 坂地泰紀, 和泉潔, 松井藤五郎, 入江圭太郎. 経済テキストからの市況分析コメントの自動生成. 第 20 回 人工知能学会 金融情報学研究会（SIG-FIN）予稿集, pp. 44–49, March 2018.

[10] 酒井浩之, 坂地泰紀, 和泉潔, 松井藤五郎, 入江圭太郎. 関連記事を用いた市況分析コメントの自動生成. 第 22 回 人工知能学会 金融情報学研究会（SIG-FIN）予稿集, pp. 61–66, March 2019.

[11] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. pp. 1638–1649, August 2018.

[12] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An Easy-to-Use framework for State-of-the-Art NLP. In **Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)**, pp. 54–59, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] 鈴木雅弘, 坂地泰紀, 平野正徳, 和泉潔. 金融文書を用いた事前学習言語モデルの構築と検証. 人工知能学会第二種研究会資料, Vol. 2021, No. FIN-027, p. 05, 2021.

[14] Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. doccano: Text annotation tool for human, 2018.