ニューラルネットワークを用いた株価変動に関するコメントの生成

関野伊吹　佐々木稔
茨城大学院　理工学部　情報工学専攻
21nm734a@vc.ibaraki.ac.jp　minoru.sasaki.01@vc.ibaraki.ac.jp

概要

近年，株価から市況コメントを生成する技術に注目が集まっている. しかし, 現状はアナリストが手作業でコメントを生成している. 本稿では, アナリストの作業負担を軽減するために，市況コメントに含まれる「株価変動に関するコメント」を生成する手法を提案する. 提案手法は､株価の数値変動とそれに対応する記事を学習してコメントを生成し，用意された定型文に付与することで, 市況コメントを完成させるものである. 実験の結果､生成に用いた特徴量が有効であり，提案手法は市況コメントを生成できることを確認した.

1 はじめに

近年，気象・スポーツ・医療・金融など，さまざまな分野でデータの活用が行われている。しかし，データが大規模であったり, 複雑であったりすると、専門知識のない人が理解するのは難しく，専門家であってもデータを理解し，重要な要素を抽出するのに時間がかかるという問題がある．このようなデータを有効活用する方法のひとつとして，Data to Text 技術がある．これは，データの概要を人間が解釈しやすいようにテキストで表現する技術で，近年こういった需要が高まり注目されている．

今回の研究対象である、株価データから市況コメントを生成するタスクも，Data to Text 技術の一種である．現在，市況コメントの生成は，社会情勢などを調査・分析する専門家であるアナリストが行っている。彼らは，株価発表後に株価を分析し，市況コメントを生成している．しかし，アナリストが株価から市況コメント全文を生成するには，多くの時間と労力が必要である．そこで，本稿では，アナリストの市況コメントを生成する労力軽減のために，市況コメントの一部を自動生成する手法を提案する．具体的には，株価数値の値動きとその変動幅に関する表現を手書きの記事から抽出し，機械学習により株価の値動きと表現を学習することで，コメントを生成する．生成されたコメントをあらかじめ用意されたフォーマットに当てはめることで，市況コメントにおける定量的な分析結果を自動生成することが可能となる．その結果、アナリストは本業である要因分析などに集中できるようになる．

本稿では，日経平均株価の市況コメント生成というタスクに基づき，時系列データから様々な特徴を抽出し，テキスト化を行う．まず，株価データの変化を捉えるために，一定期間の日経 225 データを用意する。次に、市況コメント内の表現を生成できるように，NQN(日経クイックニュース)から重要な 12 フレーズを抽出する．これらのフレーズは市況コメントの第一文に頻出するものであり，主なフレーズは「続落」，「続伸」，「反落」，「反落」の 4 つで、これらに「大幅」、「小幅」という変動幅を表す表現を加えた計 12 個を使用する．これらのフレーズを株価の値動きに対応させることで，1 つの学習データを作成した．

実験では、学習済みデータを使用して生成されたフレーズと実際に手書きで書かれた記事から抽出したフレーズの比較のために F 値を用い，提案手法がベースライン手法や先行研究のものと比べて性能が向上していると確認できた．またフレーズ生成の比較として，米株価の影響がどれほどあるかも同時に確認した．米株価は市況コメント内でも触れられる機会が多く，そのほとんどが日経平均株価の変動に関与しているため精度向上を目指して採用したが，精度はほぼ変わらない，あるいは低下という結果になった。

2 関連研究・関連手法

データの要約を人間が解釈しやすいようにテキストで自動生成する Data to Text 技術については，さまざまな研究が行われている．例えば、時系列の気象情報から天気予報のテキストを自動生成する研究[1]，医師や看護師の意思決定を支援する臨床データからのテキスト生成[2]，一定期間内の学習状況を記

録した時系列データから学生へのフィードバックテキストを生成する研究[3]などが行われている．これまで，Data to Text の研究では，人手で作成したルールを用いてテキストを生成することが主流であった．しかし近年，情報通信技術の発展に伴い，大規模かつ複雑なデータを容易に入手できるようになり，データとテキストの大規模な対応関係に基づいてテキストを生成する機械学習型の手法への関心が高まっている。例えば、画像データから説明文を生成する画像キャプション生成[4]や，成形された気象データからの天気予報テキスト生成[5]など，様々なデータからテキストへの機械学習活用の研究が行われている．

市場コメントを生成する技術には，様々な観点からアプローチがなされている．例えば，日経平均株価の値動きに影響を与えたとされる出来事や他の銘柄の情報などの変化要因を生成する技術[6]，日経平均株価のデータに加えて生成する相場コメントの内容を表す話題を入力することで生成する文章を制御する技術[7]，株価の履歴や時間に依存した表現などの特徴を生成する技術[8]などがある．今回の論文で比較する先行研究は，生成するテキスト全文を機械学習によって生成する手法であり，本稿では，テキスト全文ではなく株価の値動きや変動幅を表す単語を適切に選択し、一部のテキストを生成する技術に取り組んでいる．

3 提案手法

3 節では，日経平均株価と NQN から，株価の値動きや変動幅を表す語句を抽出する方法を紹介する．

3.1 概要

図 1 に提案手法のモデルを示す．提案手法ではまず，株価データと記事データの対応付けを行うために，データの成形を行った。記事データにはノイズとなるフレーズが多く含まれているため，生成される一文のみを抽出する．株価データも同様に，ノイズが多いため，学習しやすい形に成形する． 成形したフレーズと株価の対応関係を作り、それを使って学習を開始する．機械学習には，エンコーダとして一般的に使われている MLP（Multilayer Perceptron）を使用した．最後に，学習したデータを用いて，日米の株価データを含むテストデータを入力し，フレーズを予測する．そして生成されたフレーズを，用意されたフォーマットに代入し、市況コメントを完成させる．ただし本稿の評価基準として使用するのは生成されたフレーズとし,フォーマットに代入した全文との比較は行わない．

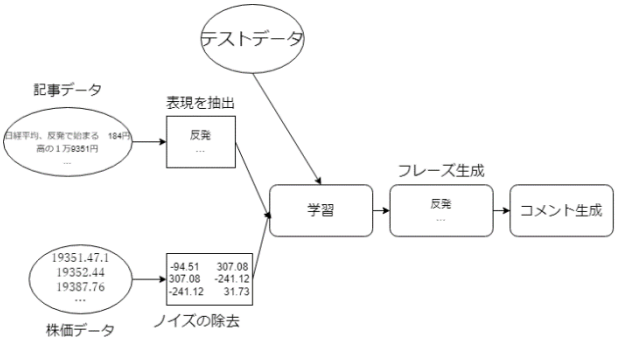

図 1:提案手法の概略

3.2 データの前処理方法

画像処理，自然言語処理など様々な分野で機械学習モデルの汎化やデータからのノイズ除去のために前処理を行うことが一般的である．本論文でも同様に前処理を適用している。NQN からフレーズ抽出は機械的に行い，数値データである日経平均株価終値のデータに対しては標準化と前日比の差分を用いる前処理を施した。記事データから抽出するフレーズに関して，1 日の統計的な市況コメントが生成される 15:00 代の記事からフレーズを抽出する．フレーズが複数個ある場合は，最初に出現したフレーズをその日のフレーズとする．株価データに使用した処理方法の式は以下の二つである。

$$x_{std} = (x_i - \mu)/\theta \quad (1)$$

$$x_{move} = x_i - r_i \quad (2)$$

(1)では，学習に用いたデータ x、平均値 μ，標準偏差 θ を用いて，標準化を行う．

(2)では，前日の終値からの価格変化を捉えるために，各タイムステップの価格 x と前日の終値 r との差を計算する．なお，使用する株価のデータは終値とし，標準偏差・前日比の前処理したデータを同時に使用することがあるので 3 日分を 1 フレーズに対応させる．

3.3 エンコード手法

時系列株価データのエンコード手法の検討として，MLP や CNN，RNN などが考えられる．先行研究の結果から，MLP をエンコーダとしたモデルがベースラインを含めた他すべてのモデルと比較し

てよいスコアがでると判明している．よって本稿でも MLP をエンコーダとして使用するものとする．

4 実験

4.1 データセット

本稿では、株価データとして日経平均株価とダウ平均株価を，記事データとして NQN を使用する．使用するデータは，2014 年から 2017 年までの 4 年間である。表 1 と表 2 に，本稿で使用した日経平均株価とダウ平均株価の例を示す．表 3 に使用した NQN を示す．日経平均株価・ダウ平均株価はともに終値を使用し，NQN は生成する式を抽出できる部分を使用している。

表 1.日経平均株価

日付	価格
2014/1/6	15908.88
2014/1/7	15814.37
2014/1/8	16121.45

表 2.ダウ平均株価

日付	価格
2014/1/6	16425.09
2014/1/7	16530.90
2014/1/8	16462.69

表 3.NQN の一部

No	記事
166	<NQN>◇日経平均先物、夜間取引で下落 60 円安の 1 万 9040 円で終了
392	<NQN>◆日経平均、反発で始まる　184 円高の 1 万 9298 円
411	<NQN>◇日経平均、反発して始まる米株高で市場心理が好転

4.2 評価方法

評価は生成したテキストと手書きで書かれた記事データから抽出したフレーズと比較した F 値で判断を行う．またテキスト生成に使用した素性の有効性を確認するために，2 種類の前処理方法をそれぞれ適用したものと 2 つの前処理を同時に使用したものの計 3 種類の比較を行う. なお，先行研究では，生成されたテキスト全文のうち，フレーズのみの精度も算出しているため，そちらも比較の対象とする．

これらに加えて，市況コメントの 2 文目以降で用いられる米株の影響によっておこる事象のテキスト自動生成の先駆けとして米株が日経平均株価に与える影響についての比較も示す．

5 実験結果

F 値による評価の実験結果を表 4 に示す．

表 4. 各 F 値

フレーズ	move	std	Move+std	+米	先行研究
反発	0.81	0.88	0.87	0.78	0.803
反落	0.85	0.76	0.80	0.84	0.748
大幅反発	0.67	0.73	0.29	0.18	-
大幅反落	0.00	0.14	0.00	0.67	-
大幅続伸	0.60	0.55	0.50	0.73	-
大幅続落	0.36	0.73	0.00	0.38	-
小幅反発	0.00	0.00	0.00	0.00	-
小幅反落	0.00	0.00	0.57	0.00	-
小幅続伸	0.00	0.00	0.00	0.40	-
小幅続落	0.00	0.00	0.00	0.29	-
続伸	0.86	0.83	0.85	0.89	0.814
続落	0.83	0.91	0.70	0.71	0.753

5.1 先行研究との差異

先行研究では，表 3 に表したような記事テキスト全文を学習し，市況コメントを生成する手法をとっており，その中で生成されたフレーズの精度が表 4 の先行研究の欄に示されている．本稿の結果を見ると，ほぼすべての素性を用いたものでも結果が良くなっている．これは先行研究で使用されているフレーズが多すぎるというのが推察できる．例えば先行研究では，株価の価格がいったん下がって上がるという「反発」というフレーズに対して，「反発」のみではなく「上げに転じる」というフレーズを用いている．これは文章の流暢性の確保のために行っていることだが，文章としては「反発」，「上げに転じる」のどちらでも通じると判断し，フレーズの統一化を行った結果，約 10 パーセント近くの精度向上を確認できた．

5.2 米株の影響

手書きの市況コメント本文には一行目に日経平均株価の変動について書かれ，2 行目以降で米株について触れられることが頻繁にある．その中で記述される中に「前日の米株式市場で主要株価指数がそろって上昇した流れを受け、(中略)日経平均株価は反発した。」というような文章が現れる．この文章からも読み取れるように日経平均株価は米株の影響を色濃く受けていると推察できるため，本稿で有効性の確認を行った．結果として，主なフレーズ 4 種類(続落・続伸・反発・反落)の F 値はほぼ変わらなかったものの，小幅・大幅のつくフレーズの生成率が全体的に増加している．日経平均株価だけでなく米株価の値を入力したことによってよりフレーズに対する株価の細分化ができたからだと推察できる．

終わりに

本稿では，日経平均株価と NQN を用いて株価の値動きとその変動幅に関するフレーズを抽出し，機械学習によりフレーズと値動きを学習させ，与えられた株価に対するフレーズを生成することを試みた。生成した表現と元記事から抽出した表現を比較し、どの学習データが優れているかを F 値で検証した。結論として，2 種類の前処理を実施した学習データが,先行研究の結果を全体的に見ると上回りました．これは，先行研究において類似したフレーズを統一していたためと考えられる．

また、米株が日経平均株価に与える影響を調べるために米株価(ダウ平均株価)を新たな入力として与えることで，生成率に変化があるかを調べた．結果として米株価を入力として与えた時，主なフレーズの生成率にはあまり影響なかったが，株価の変動幅を表すフレーズの生成率は全体的に向上していた．

参考文献

[1] B. Anja, “Probabilistic Generation of Weather Forecast Texts” Association for Computational Linguistics, pp. 164-171, 2007.

[2] F. Portet, E. Reiter, J. Hunter, and S. Sripada “Automatic Generation of Textual Summaries from Neonatal Intensive Care Data” Artificial Intelligence, Volume173, pp. 789-816 , 2009.

[3] D. Gkatzia, H. Hastie, and O. Lemon “Comparing Multi-label Classification with Reinforcement Learning for Summarization of Time-series Data.” Association for Computational Linguistics, pp. 1231-1240, 2014.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan “Show and Tell: A Neural Image Caption Generator” IEEE, Accession Number: 15524253, 2015.

[5] H. Mei, M. Bansal, and M. R. Walter “What to talk about and how? Selective Generation using LSTMs with Coarse-to-Fine Alignment” Association for Computational Linguistics, pp. 720-730, 2016.

[6] T. Aoki, A. Miyazawa, T. Ishigaki, K. Goshima, K. Aoki, I. Kobayashi, H. Takamura, and Y. Miyao “Generating Market Comments Referring to External Resources” Association for Computational Linguistics, pp. 135-139, 2018.
(RELATED WORK/METHODS)

[7] K. Aoki, A. Miyazawa, T. Ishigaki, T. Aoki, H. Noji, K. Goshima, I. Kobayashi, H. Takamura, and Y. Miyao “Controlling Contents in Data-to-Document Generation with Human-Designed Topic Labels", Association for Computational Linguistics, pp. 323-332, 2019.

[8] S. Murakami, A. Watanabe, A. Miyazawa, K. Goshima, T. Yanase, H. Takamura, and Y. Miyao, “Learning to Generate Market Comments from Stock Prices” Proceeding of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1374-1384, 2017.