Generating Market Comments on Stock Price Fluctuations Using Neural Networks

Ibuki Sekino
Major in Computer and Information Sciences
Graduate School of Science and Engineering,
Ibaraki University
email: 21nm734a@vc.ibaraki.ac.jp
4-12-1, Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

Minoru Sasaki
Dept. of Computer and Information Sciences
Faculty of Engineering, Ibaraki University
email: minoru.sasaki.01@vc.ibaraki.ac.jp

Abstract- In recent years, there has been a lot of interest in techniques for generating market comments from stock prices automatically. However, even today, it is still manually generated by analyst. In this paper, we propose a method for generating "comments on stock price fluctuations" included in market comments in order to reduce the workload of analysts. The proposed method learns stock price fluctuations and the corresponding expressions, generates comments, and completes market conditions comments by assigning them to prepared canned sentences. As a result of our experiments, we found that the features used to generate them are effective and the proposed method can accurately generate market comments.

Keywords; Generate market comments; Stock price fluctuation;Nikkei Stock Average

I. INTRODUCTION

In recent years, there has been an increase in the use of data in various fields, such as weather, sports, medicine and finance. However, when the data is large or complex, it is difficult for a person without expert knowledge to understand it, and even if he or she is an expert, it takes time to understand the data and extract the important elements. One way to make effective use of such data is the data-to-text technology. This is a technology that expresses the outline of data in text to make it easier for humans to interpret, and it has been gaining attention due to its increased demand in recent years.

The task of generating market comments from stock price data, which is the subject of this research, is also a type of data-to-text technology. Currently, market commentary is generated by analysts, who are experts in researching and analyzing social conditions, etc. They analyze stock prices after they are released and generate market comment. However, it takes a lot of time and effort for analysts to generate full-text market comments from stock prices. Therefore, in this paper, we propose a method for generating a part of the market comment in order to reduce the effort required for analysts to generate market comment. Specifically, we extract expressions related to the price movements of stock prices and their fluctuation ranges, and then generate comments by learning the price movements of stock prices and expressions through machine learning. By applying the generated comments to the pre-prepared format, the system automatically generates the quantitative analysis results in the market comment, and as a result, analysts can concentrate on their core business, such as factor analysis.

In this paper, we extract various features from the time series data and convert them into text based on the task of generating market comments on the Nikkei Stock Average. First, we form long-term and short-term time series data to capture the changes in the time series stock price data. Next, we extract 12 important phrases from NQN(Nikkei Quick News) so that we can generate a expression in NQN. These phrases are frequent occurrences in the first sentence of the market comment, and the four main expressions are "続落(continued to decline)", "続伸(continued to rise)", "反発(rebound)", and "反落(reactionary fall)", with "大幅(large)" and "小幅(small)" added for a total of 12. By mapping these expressions to the price movements of stock prices, we create a single data set for learning.

In order to compare our results with those of a previous study by Murakami et al [1], we unified the similar expressions among the generated ones. In addition, we omit sentences that do not need to be generated automatically using the neural network. We will verify if there is any change in the experimental results.

II. RELATED WORK/METHODS

In this section, we present a related work and related methods that this paper referred to.

A. *Related works*

Various studies have been conducted on data-to-text technology, which automatically generates a summary of time-series data in text that is easy for humans to interpret. For example, research has been conducted to automatically generate text about weather forecasts from time-series weather information [2], to generate text from clinical data to assist doctors and nurses in decision making [3], and to generate feedback text for students from time-series data that records their learning status within a certain period [4]. In the past, the mainstream of data-to-text research has been the generation of text using manually created rules. However, in recent years, with the development of information and communication technology, large-scale and complex data has become readily available, and interest in machine-learning type methods that generate

ISBN: 978-1-61208-874-7

text based on large-scale correspondence between data and text has been increasing. For example, research has been conducted on the use of machine learning in various data-to-text techniques, such as image caption generation [5], which generates descriptions from image data, and weather forecast text generation from molded weather data [6].

B. *Related methods*

There are various approaches to the technology of generating market comment from various perspectives. For example, there are techniques to generate factors of change, such as events that are said to have affected the price movement of the Nikkei Stock Average and information on other stocks [7], to control the generated text by inputting topics that represent the content of the generated market comment in addition to the Nikkei Stock Average data [8], and to generate characteristics, such as the history of the price of the stock and time-dependent expressions [1].

In this research, we are working on a technique to generate text by appropriately selecting words that represent the direction of price movement and the range of fluctuation of stock prices.

III. PROPOSED METHOD

In this section, we present a method for extracting words and phrases that represent the price movement and fluctuation range of stock prices from the Nikkei Stock Average and NQN, and the data used in this paper.

A. *Overview*

Figure 1 below shows the execution procedure of the proposed method.

First, we molded the data in order to create a correspondence between stock price data and article data. Since the article data contains many noisy expressions, we set the conditions to remove the noise and extracted the original phrases of the expressions generated from the article data. The details will be described later.

Next is the stock price data, which also contains a lot of noise and is inefficient for machine learning, so we molded it into a form that is easy to learn.

We then created a correspondence between three days of stock price data and a single expression and used it to start learning. For machine learning, we used a Multilayer Perceptron (MLP), which is commonly used as an encoder.

Finally, using the trained data, we input test data including stock price data to predict the expression. The predicted expressions are then substituted into the prepared format, and the market condition comments are completed. F value between the predicted expression and the correct data is calculated and used as an evaluation criterion.

B. *Dataset*

In this paper, we use the Nikkei Stock Average as stock price data and NQN as article data. The data used are for the four years from 2014 to 2017. TABLE I and TABLE II show examples of the Nikkei Stock Average and NQN used in this study. The Nikkei Stock Average is the five-minute version, and the NQN is used for the part that can extract the expressions to be generated.

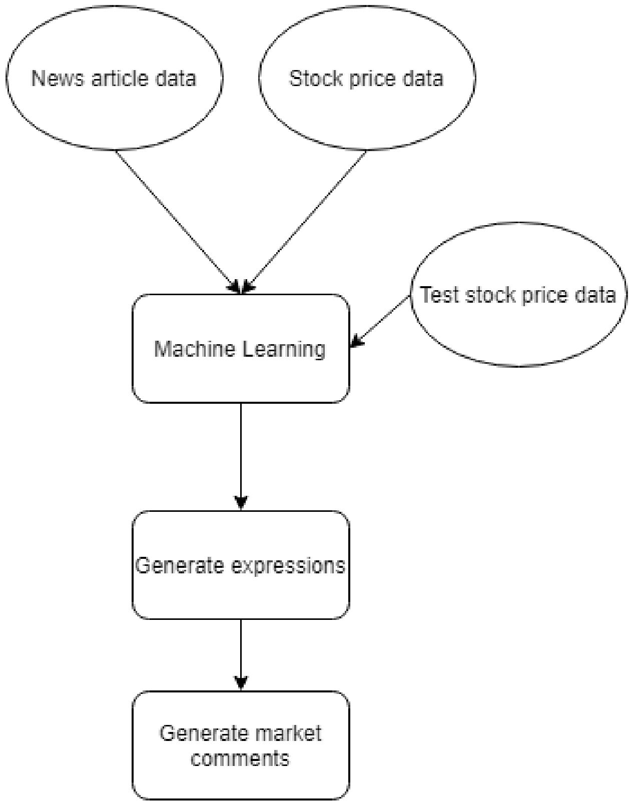

Figure 1. Flow of the proposed method

TABLE I. PART OF THE NIKKEI STOCK AVERAGE 5-MINUTE TTIMEFRAME［2017/01/04］

Time	Opening Price	High price	Low price	Closing price
9:00:00	19298.68	19351.47	19277.93	19351.47.1
9:05:00	19354.42	19362.40	19335.90	19352.44
9:10:00	19358.04	19390.47	19358.04	19387.76

TABLE II. PART OF NQN

No	Headline
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 (Nikkei 225 futures fell in overnight trading, closing 60 yen lower at 19,040 yen.)
392	<NQN>◆日経平均、反発で始まる　184円高の1万9298円 (Nikkei 225 begins to rebound, up 184 yen to 19,298 yen)
411	<NQN>◇日経平均、反発して始まる米株高で市場心理が好転 (Nikkei 225 starts to rebound, market sentiment improves on high US stock prices)

ISBN: 978-1-61208-874-7

C. *Pre-processing*

In various fields, such as image processing and natural language processing, it is common to perform prepr--ocessing in order to generalize machine learning models and to remove noise from data. Also in this paper, preprocessing is applied to the Nikkei Stock Average data, which is numerical data. We used the standardization and the difference from the previous day as the preprocessing methods for the numerical data. The equations of the processing methods used are shown below.

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

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

x_i represents the stock price.

In (1), standardization is performed using the data x, mean value μ, and standard deviation θ used for learning.

In (2) calculates the difference between the price x_i at each time step from the previous day's closing price r_i in order to capture the change in price from the previous day's closing price.

As in the previous study [1], Prepared “XShort”, a one-day stock price data consisting of 62-time steps and “XLong”, a long-term stock price data using the past closing price as an input, as a short-term time series data to capture short-term and long-term stock price fluctuations. However, it was difficult to extract the expression of the short-term data from the articles, so this paper mainly deals with the long-term stock price data.

As for the molding of the article data, since there is more than one expression extracted from NQN in a day, the first expression of the day is treated as the expression of the date.

IV. EXPERIMENT

TABLE III compares the data set used in the previous study with the data set used in this study. The comparison with the previous paper is made only where the expressions are covered. The reason for the difference in the data used in the training data is that the data of the Nikkei Stock Average for 2013 was in a different format from the data of other years, making it difficult to extract the data. Although there are some differences between the Nikkei 225 data of 2013 and 2017, the differences have been compensated by increasing the number of train data. In addition, the test data are all the same, so the results are expected to be fine.

TABLE III. DATASET FOR PREVIOUS PAPER AND THIS PAPER

	Previous paper	This paper
Training data	Nikkei Stock Average/ Nikkei QUICK News in 2013,2014,2015	Nikkei Stock Average/ Nikkei QUICK News in 2014,2015,2017
test data	Nikkei Stock Average / Nikkei QUICK News in 2016	Nikkei Stock Average / Nikkei QUICK News in 2016
Expressions that describe changes in stock prices	10/ Four expressions were used as references for comparison with this study	12
Others/Differences from this study	the previous study incorporated expressions that depend on the time of day and where the actual stock price values are calculated in the encoder-decoder.	-

V. RESULT

In this experiment, we use a combination of time series data Xlong and Xshort and preprocessing methods std and move, with one time series data as a reference and one or both preprocessing methods applied to it. The number of expressions used in the previous study was four, and they are shown in TABLE IV. TABLE IV includes the experimental results.

TABLE IV. RESULTS

Expression	Xlong_move	Xlong_std	Xlong_move_std	Xshort_move	Previous study
Rebound	0.9	0.85	0.91	0.98	0.803
Reactionary fall	0.94	0.90	0.90	0.98	0.748
Large reactionary fall	0.62	0.38	0.60	-	-
Large rebound	0.55	0.60	0.44	-	-
Large, continued to decline	0.00	0.77	0.00	-	-
Large, continued to rise	0.60	0.69	0.63	-	-
Small, rebound	0.00	0.00	0.00	-	-
Small, reactionary fall	0.00	0.00	0.00	-	-
Small. continued to rise	0.00	0.00	0,46	-	-
Small, continued to decline	0.00	0.00	0.50	-	-
Continued to rise	0.90	0.89	0.88	1.00	0.814

Continued to decline	0.89	0.87	0.90	1.00	0.753

Several methods have been used in previous studies as well as in this paper, and the results here refer to the method that produced the highest F value. The red letters represent the best results within Xlong. The blue letters are the ones with good results, but without the expression for the stock price fluctuation range. This is because when generating comments, NQN does not produce expressions at five-minute version, so I used a rule base to generate expressions without stock price fluctuation ranges. Although it is not directly related to the experimental results, it is described following the execution results of previous studies.

In this table, we can see that the two types of preprocessing give the best results in terms of Xlong alone.

VI. DISCUSSIONS

In this section, we will discuss the results.

A. *Expressions about stock price fluctuations and Xshort.*

NQN does not produce expressions for every 5 minutes, but only for important time periods (9:00, 12:00, 15:00). Therefore, in the case of short-term data that deals with five-minute data, it is necessary to extract expressions mechanically or by using other data as training data and extracting expressions by predicting them. In this paper, the former method was used. In producing the expressions for short-term data, we used the difference from the previous day in two steps. Specifically, two steps of the previous day's difference are used, with a positive value indicating "続伸(continue to rise)" and a negative value indicating "続落(continue to decline)". However, the thresholds for large or small at this time are not defined, resulting in the results shown in Table3. The NQN shows several instances of large and small falls, but the conditions for their appearance could not be determined because of only two steps of difference from the previous day, so it was not possible to set a threshold. The reason for this is that the analysts who write the market commentary assign "large" and "small" according to their sensitivity.

Therefore, the results show a high F value because there were only four expressions for three years of data. One of the future tasks will be to determine the threshold for mechanically generating expressions related to the range of fluctuation.

B. *Extraction methods considered based on differences with previous studies.*

TABLE V shows the comparable areas in this paper and previous studies. The present study produced high F values for all comparable expressions. This is thought to be due to the fact that similar expressions in the previous studies, such as "反発(rebound)" and "上げに転じる(start to move up)", were treated as the same in this study. In order to improve the accuracy of expression generation, we unified the expressions in this study. It was found that unifying the expressions increased the accuracy by about 10-20%.

TABLE V. COMPARISON OF PREVIOUS STUDIES AND THIS STUDY

Expression	This Paper	Previous paper
Continued to decline	0.91	0.803
Continued to rise	0.90	0.748
Rebound	0.88	0.814
Reactionary fall	0.90	0.753

C. *Number of expressions and number of data References*

The results show, there are some expressions whose occurrence rate is 0, and the problem is that the number of data is too large for the number of expressions prepared.

The following is a table of the number of expressions that exist in the data (TABLE VI.) and the occurrence rate of the expression that represents the fluctuation range of stock prices in the data used (TABLE VII). The red letters in TABLE VI and TABLE VII are the three selected from the lowest values.

Looking at TABLE VII, we can see that there are several expressions that are never generated. As in the case of Short, if the training data is biased, the result will be like this, so it is desirable to have training data where all expressions are generated to some extent. Or it is necessary to review the expressions to be extracted.

TABLE VI. THE NUMBER OF EXPRESSIONS THAT EXIST IN THE DATA

Expression	Xlong_move_std F-value
Continue to rise	184
Rebound	150
Reactionary fall	147
Continue to decline	111
Lage rebound	24
Large, continue to decline	24
Small continue to decline	20
Large Reactionary fall	17
Small continue to rise	17
Small reactionary fall	14
Large continue to rise	13
Small rebound	8

TABLE VII. OCCURRENCE RATE OF THE EXPRESSION THAT REPRESENTS THE FLUCTUATION RANGE

Expression	Xlong_move_std F-value
Large rebound	0.6
Large reactionary fall	0.44

Large continue to rise	0
Large continue to decline	0.63
Small rebound	0
Small reactionary fall	0
Small continue to rise	0.46
Small continue to decline	0.5

VII. CONCLUSION

In this paper, we extracted expressions related to the price movements of stock prices and their fluctuation ranges using the Nikkei Stock Average and NQN, learned the expressions and price movements by machine learning, and generated expressions for given stock prices. We compared the generated expressions with those extracted from the original article and verified which training data was superior in terms of correct answer rate and F value.

In conclusion the results of the training data with two types of preprocessing implemented exceeded those of the previous study. This is thought to be due to the unification of similar expressions in the previous study.

In addition, when generating expressions related to the range of variation of values, such as "大幅(Large)" and "小幅(Small)," it turned out to be difficult to generate them unless the training data contained this expressions with a certain degree of probability.

Future tasks include setting the threshold for generating expressions related to the range of variation mechanically, creating better training data, and reviewing the expressions.

REFERENCE

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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)

[8] 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.