

Text Generation for Explaining the Behavior of 2D Charts: With an Example of Stock Price Trends

Ichiro KOBAYASHI
Ochanomizu University

Graduate School of Humanities and Sciences, Advanced Sciences
2-1-1 Ootsuka Bunkyo-ku Tokyo, 112-8610 Japan
koba@is.ocha.ac.jp

Naoko OKUMURA
NTT Data Corporation Ltd.

Abstract

We propose a method to generate a verbal report on the trends of stock price. The trends of the stock price is observed by the behavior of numerical data expressed in a 2D chart. Since its behavior reflects the shape of a chart, in order to recognize the behavior in qualitative and quantitative ways, we use the least squares to mathematically recognize the shape and expressed it with words that often appear in news articles reporting the trends of stock prices. Our proposed method can change non-verbal information into verbal information — this provide us with high accessibility and usability for various kinds of information.

1. Introduction

This paper discusses a method to generate a verbal report on the behavior of 2D charts. We often come across many documents with multimodal information like texts, graphs, tables, etc, therefore, technologies to deal with multimodal information have been increasingly required. In particular, if non-verbal information can be expressed with natural language, the information will be easily retrieved with the current text-based information retrieval system. Thus, changing non-verbal information into verbal information will provide us with high accessibility to various kinds of information, and also usability of non-verbal information can also be expanded. In this context, we propose a method to change modality through expressing non-verbal information with verbal information and to expand communication by that. The proposed method is verified by applying it to a problem to generate a stock price report from numerical data.

2 Overview of system

First of all, we show the architecture of a system that generates verbal reports on stock prices trends in Figure 1.

As the data for stock prices dealt with by the system, we use Nikkei stock average.

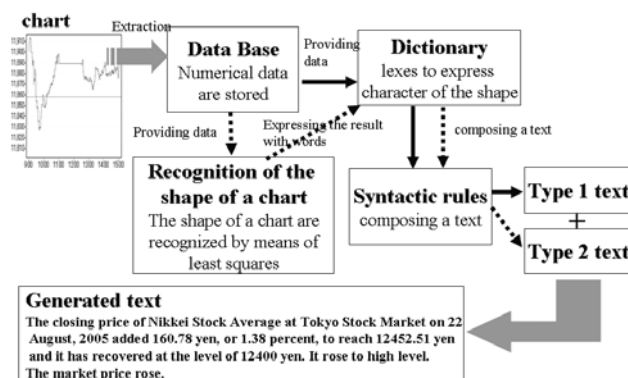


Figure 1. System architecture

The system consists of several modules: a data base to manage numerical data of Nikkei stock average, a module to recognize the shape of 2D charts, a dictionary to express the shape of 2D line chart with words, and a module to provide syntactic rules for generating a text.

The texts generated by the system are categorized into two types as follows:

type 1: The text which does not refer to the shape of 2D charts. The texts of this type are mainly confined to explaining the particular states of numerical data.

type 2: The text which refers to the shape of 2D charts. The texts of this type explain the behavior of the 2D charts trends by recognizing the shape of the line chart.

2.1 Generation process

The algorithm of generating a type 2 text is as follows: (The algorithm of generating a text of type 1 merely omits **step 2** in the following steps.)

step 1. Extracting numerical data from a data base

Necessary numerical data to generate a text are extracted from a data base.

step 2. Recognition of the shape of a chart

The shape of chart is approximately recognized by means of linear least squares.

step 3. Lexical selection for the partial shapes of a chart

To generate a type 1 text, proper lexes are selected to verbally express the numerical data obtained at step 1. And to generate a type 2 text, proper lexes are selected by recognizing the shape of a chart.

step 4. Providing templates and syntactic rules

As for type 1 text generation, the selected lexes are put into the slots of a prepared text template and short sentences to explain the trend of the price are added to output text. As for type 2 text generation, the short sentences to explain the behavior of a line chart are composed with syntactic rules, and then a text explaining the behavior of a line chart is generated.

Each component of the system will be explained in the following.

2.2 Data base

We use numerical data of Nikkei stock average from July 25, 2005 till August 30, 2005 as input information to the system. There are several kinds of input information which are 10 minutes interval data through a day, opening and closing price, and highest and the lowest prices data for past three days.

2.3 Recognition of the shape of a line chart

We apply the linear least squares to make an approximate line from the original line chart, and then verbally recognize the behavior of the original line chart from that of the approximate line chart. Figure 2 shows an example of applying the least squares to the original line chart. As we see from the example, we can recognize the behavior of the original line chart from the approximate line chart as its trend goes down at first and then it rises up.

2.3.1 Least squares

The least squares method assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from a given set of data. Suppose that the data points are, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ where x is the independent variable and y is the dependent variable. The fitting curve $f(x)$ has the deviation (error) d from each data point, i.e., $d_1 = y_1 - f(x_1), d_2 = y_2 - f(x_2), \dots, d_n = y_n - f(x_n)$.

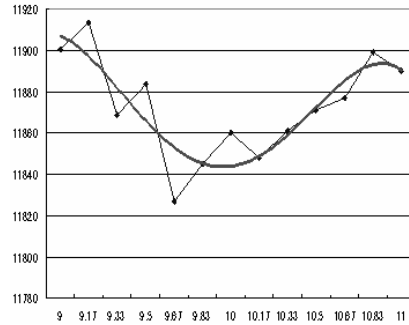


Figure 2. Example of the least squares

According to the least squares method, the best fitting curve has the property that:

$$F(x) = d_1^2 + d_2^2 + \dots + d_n^2 = \sum_{i=1}^n d_i^2$$

$$= \sum_{i=1}^n [y_i - f(x_i)]^2 \rightarrow a \text{ minimum}$$

Through analyzing corpus data — we used 27 news articles during the same period as numerical data — we adopted five dimensional polynomial function to properly approximate the shape of line charts which is because we found that the function is the most suitable polynomial function to be verbalized with the words often used in the news articles explaining the behavior of line charts.

2.3.2 Whole and partial shapes of charts

We have defined 11 types for the shape of a line chart as shown in Figure 3.

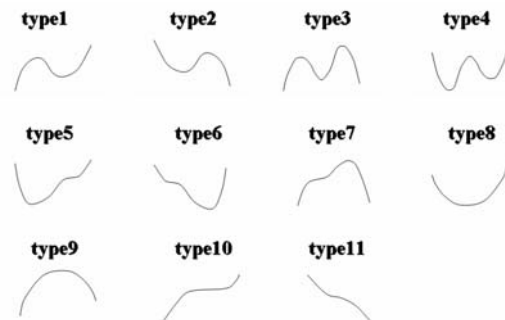


Figure 3. 11 types of whole shape

Furthermore, those 11 types are broken down into sub 13 types of partial shapes as shown in Figure 4.

This categorization for the shapes is decided based on the words often used to explain the behavior of a line chart in the corpus — we focus on the words used in explaining the behavior of a line chart, not on the characteristics of





















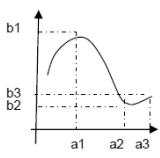
Category	Whole Shape	Partial Shape			
type1					
type2					
type3					
type4					

Figure 4. Example of partial shapes

the shape of a chart, and extracted typical shapes expressed with words.

The algorithm we developed firstly selects a relevant type from the 11 types of whole shape and then the final shape of the line chart is decided through mathematical recognition of the partial shape of the selected whole type (see, Table 1 for example). Afterwards, the proper words to express the final shape are chosen to generate a report on stock price trends.

Table 1. Partial shape and verbal expressions

Partial shape	Characteristics	Verbal expressions
	$ b2-b1 / MAX-MIN >0.4$ $ a1-a2 / max-min <0.7$	Sell order was dominant
	$ a1-a2 / max-min >0.7$	Sell order expanded
	$ b2-b1 / MAX-MIN >0.4$ $ b2-b3 / b2-b1 >0.5$ $ a1-a2 / max-min <0.7$	Sell order was once dominant
	$ a3-a2 / max-min <0.7$ $ a3-a2 / max-min >0.5$	Over the middle of the session
	$ a3-a1 / max-min <0.2$ $ a3-a2 / max-min <0.2$	Around the middle of the session
	$ a3-a1 / max-min <0.6$ $ a3-a1 / max-min >0.45$	After the middle of the session

MAX, MIN : the maximum and minimum prices at morning (afternoon) sessions.
 max, min : the starting and closing time of morning (afternoon) sessions.

2.4 Dictionary

We have extracted words often used to explain the behavior of a chart by analyzing 27 Japanese news articles about Nikkei stock average during the same period as the numerical data were obtained, i.e., from July 25, 2005 till August 30, 2005. They are words and short sentences to explain the partial shapes of a chart, words to explain particular conditions recognized from numerical data, words to express time and comparison between the current and past conditions, and conjunctions.

2.5 Syntactic rules for generated texts

Syntactic rules to compose a text differ for each text type. Type 1 texts are generated basically by using templates. Type 2 texts are generated based on verbal expressions of the behavior of a line chart with some syntactic rules and short sentences.

2.5.1 Type 1 text

From the analysis of the 27 Japanese news articles, we have obtained typical three types of template to explain the Nikkei stock average trends. Type 1 text is normally composed based on one of the text templates shown below.

- (a) The closing price of Nikkei stock average at Tokyo market on **DATE** added/lost **PRICE.VALUE**.
- (b) The closing price of Nikkei stock average at Tokyo market on **DATE** added/lost **PRICE.VALUE**, and it got down to **PRICE.LEVEL**.
- (c) The closing price of Nikkei stock average at Tokyo market on **DATE** added/lost **PRICE.VALUE**, and it has recovered at **PRICE.LEVEL**.

DATE, **PRICE.VALUE**, **PRICE.LEVEL** are the variables to be filled with numerical values obtained from the observed data stored in the data base.

And some phrases to explain the conditions of the price are usually added. For example, sentences as follows: ‘*The stock price was moving at high level throughout a day.*’, ‘*The market closed with the high price of the day.*’, etc.

2.5.2 Type 2 text

Type 2 texts are generated by composing words or short sentences to express the behavior of a chart with circumstantial information and conjunctions for the consistency among the sentences. Our system has 2 rules for adding circumstantial information and 4 rules about conjunctions. These rules reflect syntactic patterns used to explain the trends of the stock prices and are obtained through corpus analysis.

3 Generation Example

A screenshot of the developed system is shown in Fig. 5. We see that the system generates type 1 and type 2 texts, and displays the data of Nikkei stock average price on a particular day and also its 2D line chart.

3.1 Generation process of type 1 text

The generation process is explained by following generation steps explained in 2.1.

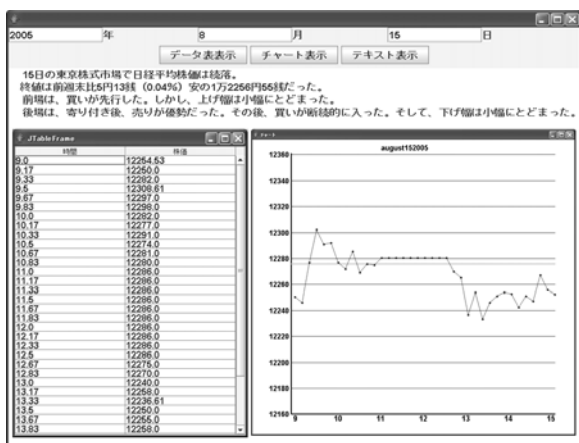


Figure 5. Screenshot of the system

Here, let us assume that we want to get a report of the stock price trends on “August 15, 2005”. This date information is the initial input information to the system.

step 1. Obtaining numerical data from a data base

Numerical data of the stock price on August 15, 2005 and the closing and the opening price, and the highest and the lowest prices data for past three days are obtained from the data base.

step 2. Lexical selection for the shape of a chart

Proper lexes to express the obtained numerical data are selected from the dictionary.

- Date information is obtained from the input information, “August 15, 2005”.
- From the obtained numerical data, the following events are found in the chart: The closing price of the stock is ‘higher’ compared to that of yesterday. The closing price of yesterday is ‘cheaper’ compared to that of the day before yesterday. The closing price of the day before yesterday is ‘higher’ compared to that of three days ago. Therefore, as a proper word to explain the trend of the stock price, ‘rebound’ is selected.
- Compared the closing price to that of yesterday, the price of today is 160.78 yen higher (1.38 % higher), therefore, the actual values, in bold fonts, related to the price conditions were selected. ‘Nikkei stock average added **160.78** yen, or **1.38** percent, to reach **12452.51** yen’.
- By the facts that the closing price of the day was over 12400 yen and the price of yesterday was under 12400 yen, a sentence, ‘The closing prices has recovered at the level of **12400** yen’ is generated.

step 3. Providing templates

- Compared the closing price of the day to that of the day before, the following template was selected — “The closing price of Nikkei stock average at Tokyo

stock market on **DATE** added **PRICE_VALUE** and it has recovered at **PRICE_LEVEL**.”

- The market closed with 25 yen higher of the closing price compared to the price of the day before, therefore, a sentence, ‘It rose to high level’ is added to a generated text.
- The closing price was higher than the opening price at both morning and afternoon sessions, and also, the closing price was higher than the opening price by 100 yen. Therefore, a sentence, ‘The market price rose’ is added to a generated text.
- Compared to the closing price of the day before, the mean value between morning and afternoon sessions was over 25 yen, therefore, a sentence, ‘The stock price was moving at a high level throughout the day’.

The generated text of type 1 is as follows:

“The closing price of Nikkei stock average at Tokyo stock market on August 15, 2005 rebounded. It added 160.78 yen, or 1.38 percent, to reach 12452.51 yen and has recovered at the level of 12400 yen. It rose to high level. The market price rose. The stock price was moving at a high level throughout the day.”

3.2 Generation process for type 2 text

step 1. Obtaining numerical data in a data base

The same process as shown in the generation process for type 1 text happens.

step 2. Recognition of the shape of a line chart

The shapes of a line chart at morning and afternoon sessions are recognized by using least squares with the numerical data obtained at step 1.

In this example, the shapes of the chart at morning and afternoon sessions are recognized as type 1 and type 7 of whole shape category, respectively.

step 3. Lexical selection for the partial shapes of a chart

Proper sentences to express the behavior of a line chart are selected by recognizing the shape of the chart with verbal expressions (see, Figure 6).

- At morning session, ‘Sell order was ahead.’, ‘Trading was steady.’ ‘The width of rising was small.’
- At afternoon session, ‘The prices were continuously rising’, ‘The width of rising expanded’, ‘The prices were decline’.
- ‘At the closing session’ is put ahead of ‘the prices were decline’.

step 4. Providing syntactic rules

Proper conjunctions and temporal information are added to the sentences generated at step 3 by following the syntactic rules for type 2 text generation.

- As for temporal information, ‘At morning session’,

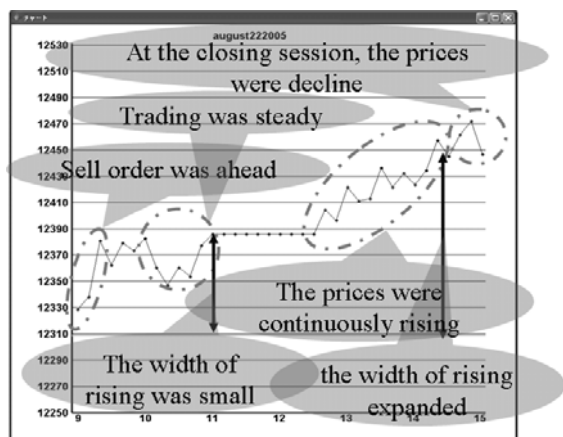


Figure 6. Shapes and their verbal expressions

and ‘At afternoon session’ are added respectively at the head of the first sentence reporting an event of each session.

- A conjunction, ‘afterwards’ is inserted between ‘Sell order was ahead’ and ‘Trading was steady.’
- A conjunction, ‘Therefore’ is put ahead of ‘The width of rising was small.’

The generated text of type 2 is as follows:

“At the morning session, sell order was ahead. Afterwards, trading was steady. therefore, the width of rising was small. At the afternoon session, trading was continuously rising. The width of rising was expanded. At the closing session, the prices were decline.”

4 Related work

There are several studies on text generation for explaining the behavior of a line chart. Boyd [1] has applied Wavelet analysis to recognizing the characteristics of time-series data expressed with a line chart and proposed a method to express the characteristics of the chart with natural language based on the analysis result. Carberry et al. [3] is a work about summarizing graph information as well as textual information — here, summarizing graph information means expressing the graph information with natural language. Kobayashi [8] has investigated the relation between patterns of line chart behaviors and natural language to express the patterns, and has proposed an algorithm to generate a text explaining the behavior of a line chart. Furthermore, as studies to deal with time-series data, Goldberg et al. [6] and Coch [4] have developed systems that generate weather reports from weather data. As the works that extract particular patterns from time-series data, Barndt et al. [5] dealt with economic data, Bakshi et al. [2] dealt with process control data, Himowitz et al. [7] dealt with medical data, etc.

5 Conclusions

As a method to transform non-verbal information into verbal information, we have proposed a method to generate a text by recognizing the shape of a 2D line chart. In particular, we used the linear least squares to make an approximate line chart against the original chart and then expressed the shape of the approximate line with words to generate a text reporting the stock price trends. This approach is, unlike the approach using Wavelet analysis, to verbalize the behavior of charts from the viewpoint of the words often used to verbalize time-series data, especially stock price trends, in the case of this paper. We verified our proposed method by comparing the generated reports with actual news articles. Although we used the same data for generating reports and for evaluation of the system, we have got 89.8 % precision rate between the generated report and news article for the same period of 27 days. This result shows us our system can properly verbalize the behavior of time-series data.

The proposed method is useful for reporting the trends of time-series data, and also for mining knowledge from the data with natural language. Furthermore, since the method changes visual information, i.e., 2D charts, into verbal information, it has possibility to provide blind people with an interface which can report what is happening in time-series data. In this context, the proposed method provides us with high accessibility and usability to various kinds of information. This realizes high interaction among multimodal information.

References

- [1] S. Boyd. Trend: A system for generating intelligent descriptions of time-series data. *Proc. IEEE-ICIPS.*, 1998.
- [2] B.R.Bakshi and G.Stephanopoulos. Reasoning in time. *Advances in Chemical Engineering*, 22:485–548, 1995.
- [3] S. Carberry, S. Elzer, K. Green, N.and McCoy, and D. Chester. Extending document summarization to information graphics. *Proc. ACL Workshop on Text Summarization.*, 2004.
- [4] J. Coch. Interactive generation and knowledge administration in multimeteo. *Proc. The 9th International Workshop on Natural Language Generation (INLG-98)*, pages 300–303., 1998.
- [5] D.J.Berndt and J.Clifford. Finding patterns in time series: A dynamic programming approach. *Advances in Knowledge Discovery 1996*, 1996.
- [6] E.Goldberg and N.Driedger. Using natural language processing to produce weather forecasts. *IEEE Expert.*, 1994.
- [7] I.J.Himowitz, P. P. Le, and I.S.Kohane. Clinical monitoring using regression-based trend templates. *Artificial Intelligence in Medicine*, 7:473–496., 1995.
- [8] I. Kobayashi. A study on text generation from non-verbal information on 2d charts. *In Computational Linguistics and Intelligent Text Processing*, pages 226–238, 2001.