Unpredictable Attributes in Market Comment Generation

Yumi Hamazono[1,2] **Tatsuya Ishigaki**[2]
Yusuke Miyao[3,2] **Hiroya Takamura**[2] **Ichiro Kobayashi**[1,2]
[1]Ochanomizu University
[2]National Institute of Advanced Industrial Science and Technology
[3]The University of Tokyo
{hamazono.yumi, koba}@is.ocha.ac.jp
{ishigaki.tatsuya, takamura.hiroya}@aist.go.jp
yusuke@is.s.u-tokyo.ac.jp

Abstract

There are two types of datasets for data-to-text: one uses raw data obtained in the real world, and the other is constructed artificially for a controlled task. A straightforwardly output text is generated from its paired input data for a manually constructed dataset because the dataset is well constructed without any excess or deficiencies. However, it may not be possible to generate a correct output text from the input data for a dataset constructed with real-world data and text. In such cases, we have to provide additional data, for example, data or text attribute labels, in order to generate the expected output text from the paired input. This paper discusses the importance of additional input labels in data-to-text for real-world data. The content and style of a market comment change depending on its medium, the market situation, and the time of day. However, as the stock price, which is the input data, does not contain any such aforementioned information, it cannot generate comments appropriately from the data alone. Therefore, we analyse the dataset and provide additional labels which are unpredictable with input data for the appropriate parts in the model. Thus, the accuracy of sentence generation is greatly improved compared to the case without the labels.The result suggests unpredictable attributes should be given as a part of the input in the training of the text generating model.

1 Introduction

Data-to-text generation is the task of generating textual descriptions from numerical time series or structured data, including sports data (Wiseman et al., 2017) and market data (Murakami et al., 2017). The majority of the recently proposed models for this task are neural network-based language generators conditioned to the structured or numerical data. We usually rely on large training datasets for training neural networks.

However, many datasets used for data-to-text generation are *artificial* because they are constructed as benchmarks for language generation research. Such artificial datasets are usually constructed by crowd workers working under constraints; for example, the target texts should only mention the information in the input data. For instance, the target text in the E2E dataset (Novikova et al., 2017) is constrained to mention all parts of input Meaning Representations (MR) and ToTTo (Parikh et al., 2020) is the dataset for the task of generating a one-sentence description that mentions the highlighted cells of a Wikipedia table. The highlighted cells were manually selected to provide sufficient and not excessive information to generate the sentence description. All the existing datasets above focus on texts that contain only the information in the inputs.

However, some datasets are constructed by extracting real-world data that do not impose these constraints. For example, RotoWire (Wiseman et al., 2017) is a dataset consisting of basketball game summaries aligned with the box scores and the line scores; both input data and target texts are professionally written and extracted from an actual website. The target texts in such a dataset often contain information that is not included in the input box score or line score. Filippova (2020) also highlight that some target texts in the WikiBio dataset are not properly aligned with the input data. These existing studies suggest that non-artificial datasets used for a data-to-text often contain attributes that cannot be estimated only from the input (e.g., who writes the text, and when the text is to be published). Therefore, we investigate such unpredictable attributes in a dataset for the data-to-text task in this study.

As an example of such non-artificial datasets, we focus on the dataset for the task of market comment generation (Murakami et al., 2017; Aoki et al., 2019), where a model generates a market comment given time-series numerical data about indices on the stock market. The market comments are collected from the past real-news data. Additionally, the time-series numerical data are past real indices of the stock market. The market comments of this dataset are in different styles. For example, some market comments are written as prompt reports, while the others are written as regular reports to be published at the fixed times. Additionally, some market comments have a certain writing style in their lexical choices. Such attributes, including writing styles and prompt or regular, might be unpredictable, but can still affect the text to be generated.

To investigate unpredictable attributes in the dataset, we first design the labels of *writing style* and *prompt* or *regular* that cannot be captured from the input data. We then integrate the embedding of these labels into a baseline encoder-decoder model. We compare several models with the label embeddings incorporated into different modules in the baseline model. We then evaluate the models on the task of generating Japanese market comments on the Nikkei Stock Average (Nikkei 225) regarding various metrics, including BLEU (Papineni et al., 2002), accuracy, specificity and F-measure of movement representation generation.

It is not surprising that adding the labels captures the writing style and increases BLEU scores. However, we show that the labels also improve the accuracy of the stock price movements expressed in the generated text. Our model, enhanced with labels, actually outperformed the baseline with a large margin of approximately 15% in terms of the BLEU score. Furthermore, it scored more than 10% higher than the model that uses only the information about the time of publication, which is also used in a study (Murakami et al., 2017). These large improvements imply the unpredictable attributes in the target texts in the dataset and the need for external information in the input data.

2 Related Work

Data-to-text, the task of generating a textual description about input data, has been studied in various domains, such as weather forecasts (Belz, 2007; Angeli et al., 2010), healthcare (Portet et al., 2009; Banaee et al., 2013), and sports (Liang et al., 2009). The task is traditionally divided into two sub-problems (Kukich, 1983; Goldberg et al., 1994): *content selection*, which determines *"what to say"*, and *surface realization*, which determines *"how to say"*, and these are approached using templates (van Deemter et al., 2005) or statistically learned models with hand-crafted features (Belz, 2008; Konstas and Lapata, 2012). Recent models use a neural network-based end-to-end approach for large amounts of data provided from various industries such as finance, pharmaceuticals, and telecommunications. Such large datasets enable us to train, for example, encoder-decoder models (Sutskever et al., 2014), which can capture the relation between input and target texts and produce fluent texts (Mei et al., 2016; Lebret et al., 2016a). However, most studies assume that the input and target texts align correctly, and all necessary information to generate target text is included in the input.

Only a few studies argue that target texts in the dataset used in data-to-text tasks cannot be completely predictable only from the input data. For example, Filippova (2020) mentions that the target texts in well-known datasets such as WikiBio (Lebret et al., 2016b) are not completely predictable only from the input because they include noises that input data and target texts are not aligned correctly. In contrast, in our setting, the numerical input data and target comments are correctly aligned with each other. However, the target texts are not predictable only from the input data because there are multiple target comments for the same input.

The use of external information is known to improve the performance of data-to-text models. For example, Saleh et al. (2019), Iso et al. (2019), Gong et al. (2019) use information about the previous or next match or the author information, which are not in the input box or line scores. While their purpose is to improve the performance of the models, we use external information to investigate unpredictable attributes in the market comment generation task.

3 Previous Study: Market Comment Generation

This section describes the method for generating market comments from stock prices proposed by Murakami et al. (2017), used as the base method in our study.

Murakami et al. (2017) use both long- term and short-term vectors for stock price movements. To represent the long-term price movement, they use the closing price data for M days and represent as $\boldsymbol{x}_{\text{long}} = (x_{\text{long},1} x_{\text{long},2}, \dots, x_{\text{long},M})$. Similarly, they use the N latest values of the 5-minute time frame to represent the short term price movement and set $\boldsymbol{x}_{\text{short}} = (x_{\text{short},1} x_{\text{short},2}, \dots, x_{\text{short},N})$. Through preprocessing, including standardization and moving reference, these vectors are transformed into vectors $\boldsymbol{x}'_{\text{long}}$ and $\boldsymbol{x}'_{\text{short}}$.[1]

In the encoding step, the vectors $\boldsymbol{x}'_{\text{long}}$ and $\boldsymbol{x}'_{\text{short}}$ are fed to multi-layer perceptrons (MLP) to obtain the intermediate representations $\boldsymbol{h}_{\text{long}}$ and $\boldsymbol{h}_{\text{short}}$:

$$\boldsymbol{h}_{\text{long}} = \text{MLP}[\boldsymbol{x}'_{\text{long}}], \quad (1)$$

$$\boldsymbol{h}_{\text{short}} = \text{MLP}[\boldsymbol{x}'_{\text{short}}]. \quad (2)$$

These vectors are combined to obtain the hidden state $\boldsymbol{m}$ of the encoder:

$$\boldsymbol{m} = \boldsymbol{W}_m[\boldsymbol{h}_{\text{long}}; \boldsymbol{h}_{\text{short}}] + \boldsymbol{b}_m. \quad (3)$$

In the decoding step, Murakami et al. (2017) set the initial hidden state $\boldsymbol{s}_0$ of the decoder as $\boldsymbol{m}$ shown in Equation 3, and use LSTM (Hochreiter and Schmidhuber, 1997) as the decoder:

$$\boldsymbol{s}_i = \text{LSTM}([\boldsymbol{w}_{i-1}; \boldsymbol{l}_{\text{time}}], \boldsymbol{s}_{i-1}) \quad (4)$$

where $\boldsymbol{l}_{\text{time}}$ is the embedding of the *time label* obtained from the delivery time of the comment. The *time* label embedding is the hour in which the delivery time of the comment falls into; 9:10 am and 9:30 am are associated with the same *time* label embedding. It is used as an additional input in each step of LSTM in order to generate the word depending on time. This is inspired by the *speaker embedding* introduced by Li et al. (2016). As in the standard LSTM decoder, $\boldsymbol{s}_i$ is fed into a linear layer and a softmax layer to calculate the probability of the next word.

4 Dataset Analysis

This section analyses market comments in the dataset used in the existing studies (Murakami et al., 2017; Aoki et al., 2018) to investigate unpredictable attributes, which makes the task partially ill-posed. Additionally, we show some market comments aligned with exactly the same input, because such examples illustrate the partial ill-posedness of the task.

In the following section, we describe the steps to construct the dataset to explain why there are multiple comments for the same input. We then analyze the factors that cause different characteristics of other comments from the same input data.

4.1 Procedure for Creating the Dataset

We describe the construction of the dataset used in the existing studies (Murakami et al., 2017; Aoki et al., 2018). As the time series in the original data consisted of many data points with very short time intervals (i.e., 15 sec), they needed to be binned into rather larger bins (i.e., periods of 5 min), resulting in a new time series consisting of the representative values (i.e., the last data points) of the bins. The market comments delivered at a time in each bin are all aligned with its representative value. We note that this type of binning process is not uncommon when input data are continuous, and that the issue is not entirely specific to this dataset.

Specifically, the input data are the same when multiple market comments are delivered during a certain 5-minute period. Additionally, the market comments in the dataset have different writing styles; comments in one style tend to start with *Nikkei 225*, and comments in the other style tend to start with *The Tokyo Stock Exchange*. Our validation dataset contains 1,176 unique inputs, with a total size of 1,751 owing to this one-to-many alignment.

4.2 Comparison of Multiple Comments Aligned with the Same Input Data

Table 1 shows examples of multiple comments that are aligned with the same input. We selected one

[1] Refer to Murakami et al. (2017) for details.

Type of Difference	Market Comments
(Pivot comment)	*Nikkei 225 opened with a continual rise. The price is 17024 yen, which is 9 yen higher.* 日経平均、続伸で始まる。9円高の17024円
Content order	*Nikkei 225 started with 17024 yen, 9 yen higher, with a continual rise.* 日経平均、9円高の17024円で続伸して始まる。
Lexical choice	*Nikkei 225 opened with a continual rise. The opening price is 17024 yen, which is 9 yen higher.* 日経平均、続伸で始まる。始値は9円高の17024円
Informativeness	*Nikkei 225 opened with a continual rise.* 日経平均、続伸で始まる。
Others	*Nikkei 225 is up over 100 yen.* 日経平均、上げ幅100円 超える。

Table 1: Example of different market comments with the same input data.

comment as a pivot comment. We then compared this comment with the others for understanding the differences between them. We observed that the other comments were different from the pivot comment in at least four different aspects:

Content order: the orders of the contents expressed in the comments are different.

Lexical choice: the same content is expressed with different words.

Informativeness: the amount of information is different.

Others: other types of differences.

Although Table 1 is an example, the differences listed in the table are common and are discussed further in the next subsection.

4.3 Factors that Cause Differences

We realized that some of the market comments in our data reflect market conditions up to some fixed times, such as 9:00 am, 10:00 am, 11:30 am, 12:30 pm, 2:00 pm, 3:00 pm; they are *regular reports*. The other comments are *prompt reports* delivered at times when some events worth reporting occur. We also observed two different writing styles in our dataset: 62.70% of the market comments in our validation dataset start with *Nikkei 225*, while the others start with *The Tokyo Stock Exchange*. We also realized that some comments are supposed to be important, marked with a special token ⧫ at the beginning of a comment. These three attributes are *regular/prompt*, *writing style*, and *supposed-importance*. In the following section, we examine how these attributes affect the market comments.

Analysis of information amount

We first analyzed the average length of comments, assuming that the comment length was correlated with the amount of information contained in the comment. The average number of tokens in *regular* comments was 8.04, while that of *prompt* comments was 6.14. This implies that regular comments contain more detailed information. Similarly, the average number of tokens in the supposedly-important comments was 8.21, while that of the other comments was 7.24. This implies that the *supposedly-important* comments are more informative than the others. By *writing style*, the average number of tokens for comments starting with *Nikkei 225* was 7.96, while the average number of tokens for comments starting with *The Tokyo Stock Exchange* was 6.99. This implies that the comments starting with *Nikkei 225* are more informative than those starting with *The Tokyo Stock Exchange*.

Analysis of content and order

We analyzed the differences in the content and their orders in the market comments. We denoted temporal expressions as `TIME`, stock prices as `PRICE`, and expressions of stock price movement as `MOVE`, to examine the contents (i.e., `TIME`, `PRICE`, and `MOVE`) mentioned in the comments and their order. For example, "*Nikkei 225 opens with a continual rise. The price is 17024 yen, 9 yen higher.*" is replaced with `[TIME, MOVE, PRICE, PRICE, MOVE]`[2]. In regular comments,

[2] The original dataset is Japanese. We translated the examples for an explanation.

[MOVE, TIME, PRICE, MOVE, PRICE] accounted for the largest proportion (17.23%) and [TIME, MOVE] for the second largest propotion (16.34%) , while the largest proportion of prompt comments was [MOVE, PRICE] (56.54%). This is because the regular comments contain TIME, while the prompt comments do not. Further, 35.65% of the supposedly-important comments were [MOVE, TIME, PRICE, MOVE, PRICE], accounting for the largest proportion , whereas, in the others, this content and their orders were only 1 out of 1,103, while [TIME, MOVE] accounts for the largest proportion (19.95%). The places where TIME entries and also the content amount are different. The contents and their order are different, whether it is regular or prompt reports and the supposedly-important comments. Regarding the writing styles, 21.22% of comments that start with *Nikkei 255* were [MOVE, TIME, PRICE, MOVE, PRICE], and [MOVE, PRICE] accounted for the second largest proportion (20.77%). However, neither of them is in the comments starting with *The Tokyo Stock Exchange*, while [TIME, MOVE] accounted for the highest percentage (28.79%).

Analysis of lexical choice

We examined differences in lexical choice. The third example in Table 1 contains the expression *opening price*, which does not appear in the pivot comment. Such differences in lexical choice cannot often be captured from the input data, and can depend on an unpredictable attribute.

Analysis via attribute prediction

Finally, we examined whether these attributes are unpredictable from the market price data. We used a classifier consisting of the encoder used by Murakami et al. (2017) as the encoder of the market generation model in Section 6 and the softmax layer for classification. In our experiments[3], the accuracy of regular or prompt classification was 78.62%. As 76.87% of the market comments in our validation dataset are regular reports, the result suggests that our regular/prompt classifier is not better than the majority baseline. The prediction accuracy of classification of the supposedly-important comments was 65.87%, which was also not significantly different from 62.99%, the accuracy of the majority baseline. In contrast, 62.70% of the market comments in our validation dataset started with *Nikkei 225*, while the others started with *The Tokyo Stock Exchange*. The accuracy of *writing style* labeling was found to be 73.39% using attribute prediction checking. This is because the writing style that starts with *The Tokyo Stock Exchange* is strongly related to time. The accuracy within a range of 10 minutes after the times at which the market prices are referred to in the regular reports is 65.96%.

These results suggest that these attributes are mostly unpredictable. Therefore, the difference in the market comments derived from these attributes cannot be captured from the input series of stock prices.

5 Label Integration for Market Comment Generation

In Section 6, we show that the unpredictable attributes largely improve the performance of market comment generation, and argue that these unpredictable attributes should be provided as part of the input. In this section, preparing for the experiments, we describe how we integrate the information from the unpredictable attributes discussed in Section 4.2 with market comment generation models. We first design labels representing unpredictable attributes, such as *writing style* and *importance*. We then describe our models that using these labels as the additional input. Note that our objective is not to propose a novel model for market comment generation, but to show the impact of unpredictable attributes and argue that such unpredictable attributes should be given as a part of the input.

5.1 Designing Labels

In Section 4.2, we argued that some differences in comments might be caused by factors such as differences in writing style, regular/prompt, and the supposed-importance. Table 2 shows labels reflecting such factors. In addition to the *time* label for each hour described in Section 3 (Murakami et al., 2017; Aoki et al., 2018), we prepare two new labels: *writing style* and *importance*.

[3] Details of the experiments are the same as the experiments described in Section 6.

Label	Values
writing style	Nikkei, TSE
importance	regular ∩ supposedly-important, prompt ∩ supposedly-important, regular ∩ not supposedly-important, prompt ∩ not supposedly important
time	9:00-9:59am, 10:00-10:59am, . . . following Murakami et al.

Table 2: Labels and their values

The *writing style* label comes directly from the writing style attribute discussed in Section 4, and has two values: *Nikkei* and *TSE*[4]. The market comments starting with *Nikkei 225* are annotated with *Nikkei*, while those starting with *The Tokyo Stock Exchange* are annotated with *TSE*.

We define the *importance* label as a Cartesian product of the regular/prompt and the supposed-importance attributes because both these two attributes are related to the importance of comments. Namely, the *importance* label has 4 values: regular ∩ supposedly-important, prompt ∩ supposedly-important, regular ∩ not supposedly-important, and prompt ∩ not supposedly-important.

5.2 Models

Consider the following three different methods for feeding the labels obtained in Section 4.2 (shown in Figure 1); (a) with input data, (b) after the encoder, i.e. just before the decoder, and (c) each step of the decoder. Each value of the label is represented as an embedding, which is obtained through training. We refer to it simply as label embedding, and denote it by $\boldsymbol{l}$. Formally, we change the formula for calculating the states of the network as follows: In the case of (a), Equations (1) and (2) are

$$\boldsymbol{h}_{\text{long}} = \text{MLP}[\boldsymbol{x}'_{\text{long}}; \boldsymbol{l}], \tag{5}$$

$$\boldsymbol{h}_{\text{short}} = \text{MLP}[\boldsymbol{x}'_{\text{short}}; \boldsymbol{l}]. \tag{6}$$

In the case of (b), Equation (3) is

$$\boldsymbol{m} = \boldsymbol{W}_m[\boldsymbol{h}_{\text{long}}; \boldsymbol{h}_{\text{short}}; \boldsymbol{l}] + \boldsymbol{b}_m. \tag{7}$$

In the case of (c), Equation (4) is

$$\boldsymbol{s}_i = \text{LSTM}([\boldsymbol{w}_{i-1}; \boldsymbol{l}], \boldsymbol{s}_{i-1}). \tag{8}$$

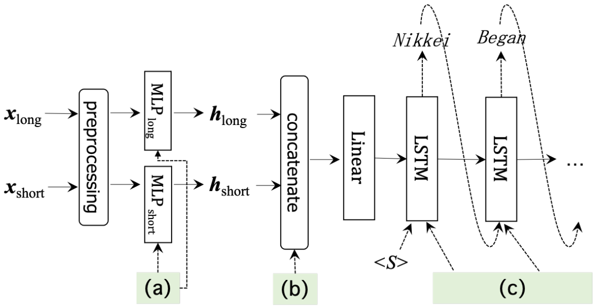
$\boldsymbol{x}_{\text{long}}$
$\boldsymbol{x}_{\text{short}}$
preprocessing
MLP_{long}
$\text{MLP}_{\text{short}}$
$\boldsymbol{h}_{\text{long}}$
$\boldsymbol{h}_{\text{short}}$
concatenate
Linear
LSTM
LSTM
...
Nikkei
Began
<S>
(a)
(b)
(c)

Figure 1: Overview of network

When we use more than one label, we concatenate on the embeddings of the labels. For example, when all three labels we prepare are used at the same time, and $\boldsymbol{l}$ is represented as:

$$\boldsymbol{l} = [\boldsymbol{l}_{\text{time}}; \boldsymbol{l}_{\text{writing style}}; \boldsymbol{l}_{\text{importance}}] \tag{9}$$

where $\boldsymbol{l}_{\text{time}}$ is a *time* label embedding, $\boldsymbol{l}_{\text{writing style}}$ is a *writing style* label embedding, and $\boldsymbol{l}_{\text{importance}}$ is an *importance* label embedding.

6 Experiment

We use the following three labels which are relevant to the output text: (1) *time*, (2) *writing style*, and (3) *importance*. We experiment with the following three ways for injecting these labels: (a) with input data, (b) after the encoder, that is, just before the decoder, and (c) at each step of the decoder.

6.1 Dataset

We use a five-minute chart of the Nikkei 225 from December 2011 to October 2016 as numerical time-series data, which are collected from Thomson Reuters DataScope Select[5]. As market comments, we use 15,831 headlines from Nikkei Quick News (NQN), which are written in Japanese, describing the behavior of Nikkei 225. We remove descriptions of other economic indices from each headline, such as the U.S. equities and foreign exchange. Additionally, referring to Murakami et al. (2017), we replace numerical values in the headline with generalization tags. The headlines are divided into three parts based on the period of publication: 12,391 for training (December 2011–October 2015), 1,751 for validation

[4] Tokyo Stock Exchange

[5] `https://hosted.datascope.reuters.com/DataScope/`

	(a)	**(b)**	**(c)**
Valid	without labels: 35.82		
time label	37.42	40.78	40.44
writing style label	43.51	43.52	44.28
importance label	48.78	48.50	**50.18**
Test	without labels: 36.79		
time label	40.42	41.40	39.66
writing style label	46.69	46.54	47.36
importance label	**52.77**	51.90	52.15

Table 3: BLEU (%)

In the validation dataset, the highest BLEU score was obtained when an *important* label was injected in (b). The lowest was the *time* label in (a), whereas it was higher than without labels. In the test, the highest score was obtained by injecting an *important* label in (c), the lowest was the setting of previous studies, injecting the *time* label in (c).

(October 2015–April 2016) , and 1,689 for testing (April–October 2016).

6.2 Experimental settings

Consistent with previous studies, we conduct experiments with the following settings: All MLPs in the model have three layers with 256 hidden dimensions. The decoder, LSTM, has a single layer with 256 hidden dimensions. For the length of the short- and long-term vectors, we set $M = 7$ for $\boldsymbol{x}_{long}$ and $N = 62$ for $\boldsymbol{x}_{short}$. Each value of the three labels, *time*, *writing style*, and *importance* is embedded in a 64-dimensional vector. We train the models for 150 epochs with a mini-batch size of 100, using Adam optimizer (Kingma and Ba, 2015) with the initial learning rate 1×10^{-4}, and save the parameters at every epoch, selecting the model with the highest BLEU score on the validation dataset.

We use BLEU (Papineni et al., 2002), which evaluates the quality of text using the similarity between the gold texts and the generated comments. We also assess whether the expressions referring to the movement of the target stock price[6] are generated appropriately in the comments.

6.3 Results

Table 3 presents the results of the BLEU evaluation. BLEU scores increased in all cases with labels, compared to those without labels. Although the *time* label used in the previous studies (Murakami et al., 2017; Aoki et al., 2018) contributed to the improvement of the score, it was not the most effective label. Using *writing style* label and *importance* label, which are found by the dataset analysis, are significantly more effective in raising the scores. The highest BLEU score with one label in the validation case was 50.18% when the *importance* label was used in (c), nearly 15% higher than the score obtained without labels. It is also more than 10% higher than the setting of previous studies with *time* label in (c).

	(a)	**(b)**	**(c)**
Accuracy	without labels: 89.27		
time label	90.26	90.88	90.96
writing style label	90.61	90.64	90.68
importance label	**91.40**	91.27	**91.40**
Specificity	without labels: 93.25		
time label	93.86	94.05	94.08
writing style label	94.06	93.97	94.04
importance label	95.27	95.08	**95.29**
F-measure	without labels: 56.96		
time label	60.91	63.82	64.68
writing style label	63.17	63.36	63.83
importance label	65.01	65.08	**65.19**

Table 4: Evaluation by movement representations with validation dataset (%)

Injecting *importance* label at (c) provides not only the highest accuracy but also the highest specificity and F-measure.

Table 4 shows the evaluation results with movement representation. Notably, the use of labels contributed to improving not only the accuracy of the comment generation but also the selection of the correct movement expressions. As with the BLEU score evaluation, the accuracy is increased in all cases with labels, compared with the case without labels. Furthermore, along with the accuracy, the specificity and the F-measure improved. When using the *importance* label in (c), the F-measure improved more than 8% compared to that without labels. From the above observations, even though the labels themselves do not contain any clue about data movement, labels generate stylistically correct comments and an accurate description of the data.

There were no significant differences in the results between the label injection ways (a), (b), and (c),

[6] fall down, continual rise, rebound, fall back, up, down, high, low

Gold	
(I)	*Nikkei 225 continued to fall, closing down 69 yen to 17697 yen.* 日経平均、 続落 大引けは69円安の17697円
(II)	*Nikkei 225 closed with continual fall.* 日経平均大引け、続落
(III)	*TSE closed with continual fall.* 東証大引け、続落
Prediction	
Without labels	
	Nikkei 225 continued to fall, closing down 69 yen to 17697 yen. 日経平均、続落 大引けは69円安の17697円
With all labels	
(I)	*Nikkei 225 continued to fall, closing down 69 yen.* 日経平均、続落 大引けは69円安
(II)	*Nikkei 225 closed with continual fall.* 日経平均大引け、続落
(III)	*TSE closed with continual fall* 東証大引け、続落

Table 5: Generated market comments with/without labels from the same input data of stock price

whereas, in the choice of labels, the use of the *importance* label gave the highest scores. As defined in Section 5.1, the *importance* label is a combination of the regular/prompt and the suggestedly-important attributes, which both had lower classification accuracy than the majority baseline in the results of *Analysis via attribute prediction* in Section 4.3. Therefore, these results suggest that using attributes that are not predictable improves the scores.

Table 5 shows examples of the generated comments with the same input data; Without labels, the generated comments remain the same. In contrast, with labels, they generate different comments, and each comment is similar to its gold output text.

7 Conclusion

Our research focused on generating market comment text with time-series market data, one example of the data-to-text using real-world data and text. By re-validating the existing dataset, we found that there are unpredictable attributes in the dataset, making it impossible to generate the correct market comments from the input. We created two additional labels out of the unpredictable attributes, and fed the labels to the neural network-based model. Although it is not surprising that the use of labels improves the generation performance, we also observed the positive side effect that the labels improved the selection of movement expression. The impact of the unpredictable attributes suggests that the text generation model should not be trained in the presence of unpredictable attributes, but instead such unpredictable attributes should be given as a part of the input. We leave the way of automatically discovering such unpredictable and accuracy-enhancing attributes as future work.

Acknowledgments

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was also supported by JSPS KAKENHI Grant Number JP21J14335. For experiments, computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 502–512, Cambridge, MA, October. Association for Computational Linguistics.

Tatsuya Aoki, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. 2018. Generating market comments referring to external resources. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 135–139, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

Kasumi Aoki, Akira Miyazawa, Tatsuya Ishigaki, Tatsuya Aoki, Hiroshi Noji, Keiichi Goshima, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. 2019. Controlling contents in data-to-document generation with human-designed topic labels. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 323–332, Tokyo, Japan, October–November. Association for Computational Linguistics.

Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. 2013. Towards NLG for physiological data monitoring with body area networks. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 193–197, Sofia, Bulgaria, August. Association for Computational Linguistics.

Anja Belz. 2007. Probabilistic generation of weather forecast texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 164–171, Rochester, New York, April. Association for Computational Linguistics.

Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online, November. Association for Computational Linguistics.

Eli Goldberg, Norbert Driedger, and Richard I Kittredge. 1994. Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Heng Gong, Xiaocheng Feng, Bing Qin, and Ting Liu. 2019. Table-to-text generation with effective hierarchical encoder on three dimensions (row, column and time). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3143–3152, Hong Kong, China, November. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hayate Iso, Yui Uehara, Tatsuya Ishigaki, Hiroshi Noji, Eiji Aramaki, Ichiro Kobayashi, Yusuke Miyao, Naoaki Okazaki, and Hiroya Takamura. 2019. Learning to select, track, and generate for data-to-text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2102–2113, Florence, Italy, July. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ioannis Konstas and Mirella Lapata. 2012. Unsupervised concept-to-text generation with hypergraphs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 752–761, Montréal, Canada, June. Association for Computational Linguistics.

Karen Kukich. 1983. Design of a knowledge-based report generator. In *21st Annual Meeting of the Association for Computational Linguistics*, pages 145–150, Cambridge, Massachusetts, USA, June. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016a. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas, November. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016b. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas, November. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August. Association for Computational Linguistics.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore, August. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, San Diego, California, June. Association for Computational Linguistics.

Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. 2017. Learning to generate market comments from stock prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1384, Vancouver, Canada, July. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany, August. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online, November. Association for Computational Linguistics.

François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173(7):789 – 816.

Fahimeh Saleh, Alexandre Berard, Ioan Calapodescu, and Laurent Besacier. 2019. Naver labs Europe's systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 273–279, Hong Kong, November. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Kees van Deemter, Mariet Theune, and Emiel Krahmer. 2005. Real versus template-based natural language generation: A false opposition? *Computational Linguistics*, 31:15–24, 03.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark, September. Association for Computational Linguistics.