Controlling Contents in Data-to-Document Generation with Human-Designed Topic Labels

Kasumi Aoki[♭♣] **Akira Miyazawa**[♢♡♣] **Tatsuya Ishigaki**[♠♣] **Tatsuya Aoki**[♠♣] **Hiroshi Noji**[♣]
Keiichi Goshima[♯♣] **Ichiro Kobayashi**[♭♣] **Hiroya Takamura**[♠♣] **Yusuke Miyao**[♮♣]

[♭]Ochanomizu University [♣]National Institute of Advanced Industrial Science and Technology
[♢]The Graduate University for Advanced Studies [♡]National Institute of Informatics
[♠]Tokyo Institute of Technology [♯]Waseda University [♮]The University of Tokyo

{g1120501, koba}@is.ocha.ac.jp miyazawa-a@nii.ac.jp {aoki, ishigaki}@lr.pi.titech.ac.jp
hiroshi.noji@aist.go.jp keiichi.goshima@aoni.waseda.jp takamura@pi.titech.ac.jp yusuke@is.s.u-tokyo.ac.jp

Abstract

We propose a data-to-document generator that can easily control the contents of output texts based on a neural language model. Conventional data-to-text model is useful when a reader seeks a global summary of data because it has only to describe an important part that has been extracted beforehand. However, since it differs from users to users what they are interested in, it is necessary to develop a method to generate various summaries according to users' requests. We develop a model to generate various summaries and to control their contents by providing the explicit targets for a reference to the model as controllable factors. In the experiments, we used five-minute or one-hour charts of 9 indicators (e.g., Nikkei 225), as time-series data, and daily summaries of Nikkei Quick News as textual data. We conducted comparative experiments using two pieces of information: human-designed topic labels indicating the contents of a sentence and automatically extracted keywords as the referential information for generation. Experiments show both models using additional information of target document achieved higher performance in terms of BLEU and human evaluation. We found that human-designed topic labels are superior to extracted keywords in terms of controllability.

1 Introduction

Data-to-text is one of the challenging tasks in natural language generation, which aims to generate summaries of input data such as statistics from sports games (Robin, 1995; Barzilay and Lapata, 2005; Wiseman et al., 2017), financial data (Murakami et al., 2017; Aoki et al., 2018), and database records (Reiter and Dale, 1997; Liang et al., 2009; Mei et al., 2016; Lebret et al., 2016; Novikova et al., 2017; Liu et al., 2018; Wiseman et al., 2017).

Over the past several years, end-to-end neural language generation models have successfully been applied to versatile data-to-text tasks, because they can generate fluent texts without task-specific knowledge and resources.

However, it has also been pointed out that texts generated by neural models suffer from low diversity in expressions (Yang et al., 2019). Especially on the data-to-text tasks, since they are developed under the assumption that the important contents could be uniquely determined, previous methods did not focus on controlling the contents in terms of user's interests.

However, each user may expect different contents in a summary depending on what they are interested in, and thus it is appealing to develop a method to generate various summaries which reflect user's interests.

This paper investigates a method for guiding data-to-document generation in the finance domain, by referring to a sequence of additional information for input financial data. Generating documents consisting of multiple sentences involves an inherent challenge in content selection and ordering (Reiter and Dale, 1997), because one can produce a large variety of documents for specific input data, depending on a focus, intent, readers' interest, etc. Therefore, it is essential for document generation systems to have an additional mechanism to select and order the contents to be represented.

We introduce and empirically compare two types of topic labels, both of which are intended to denote clause-level contents and their orders. One is topical keywords automatically extracted from domain texts (Rose et al., 2010), which was applied for the story generation by using as the contents of the story (Yao et al., 2018).

The other is manually defined topic labels. As our target domain is finance, major topics mentioned in documents are restricted to market indices such as Dow Jones Industrial Average (DJI), Nikkei 225, or foreign exchange rates, etc. We devised a

Proceedings of The 12th International Conference on Natural Language Generation, pages 323–332,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

closed set of domain-specific labels by investigating financial news articles. In the experiments on generating daily summaries of financial markets, we will empirically show the effectiveness of topic labels and potential advantages/disadvantages of this approach.

2 Related study

Controllability of text generation has been an intensive research focus recently. Examples include suggestive content control such as tense, sentiment, gender, or automatically learned hidden states (Hu et al., 2017; Zhao et al., 2018; Juraska and Walker, 2018; Bau et al., 2019). Another series of work is focused on controlling surface textual features such as length, descriptiveness and politeness (Li et al., 2016; Sennrich et al., 2016; Kikuchi et al., 2016; Ficler and Goldberg, 2017; Shen et al., 2017; Prabhumoye et al., 2018). The target of these previous methods is on controlling generic content-independent features of texts. That is, they aim at varying surface strings while preserving main information content. Wiseman et al. (2018) proposed a neural model that generates diverse texts by learning templates. They control diversity through templates rather than contents or the order of them. The present work is more closely related to methods for controlling topical content by using automatically extracted or human-designed keywords (Wang et al., 2016; Yao et al., 2017, 2018; Miao et al., 2018). Our method resembles the idea of using keywords to control topics of sentences and their orders, but it primarily focuses on describing given data and uses topic labels as auxiliary information. We will empirically attest added effects of introducing topic labels in the data-to-document scenario.

Besides, Gkatzia et al. (2017) and Portet et al. (2009) proposed non-neural language generation models for the data-to-text task with higher controllability on the output. They assumed that the important contents and their descriptions are determined primarily by experts, and their models do not allow users to select the contents directly.

To the best of our knowledge, no previous research tackled a problem with controllability of the content in the data-to-document task.

We believe that the contribution of this paper is the followings: First, we propose explicitly content-controllable data-to-document generator that uses additional clause information. Experiments show the fluency and fidelity of the generated document in terms of BLEU and human-evaluation. Secondly, compared the generated documents between with human-designed labels and automatically extracted keywords, human-designed labels are more useful as the ease of understanding.

3 Generation of Market Comments

Our task is to generate summaries of financial markets. The input is a set of financial time-series data, such as DJI, Nikkei 225, and JPY/USD exchange rate. The output is a sequence of sentences describing movements of the financial data and their relationships.

The overview of our model is illustrated in Figure 1. In the following, we first describe our design of topic labels, then describe our data-to-text model with topic labels.

3.1 Topic Labels

Topic labels are defined as clause-level topics for aiming to guide a sequence of contents to be output. We empirically compare two methods to obtain topic labels: automatically extracted keywords and human-designed labels.

Automatically extracted keywords

This is a straightforward strategy to obtain the labels as the topic of sentences. We use RAKE (Rose et al., 2010) algorithm, which builds document graphs and weights the importance of each word combining several word-level and graph-level criteria to extract the keywords. Using such an automatic keywords extraction system has an advantage on the cost of human annotation while the extracted keywords sometimes do not express writers' intent. For example, RAKE often outputs the word "market" or "observation" as keywords, but they are not appropriate as the topic labels because of the lack of precise information—a system would be unable to understand which market, e.g., Nikkei or DJI, or what kinds of observations, e.g., the growth rate of stock prices or the trends of investments, when generating a text considering these labels.

Human-designed topic labels

We devised a set of topic labels by observing target sentences in the training data and what they often refer to, especially for *Nikkei Quick News* (NQN).

A topic label denotes the objects mentioned in documents and is defined as a triple, each element

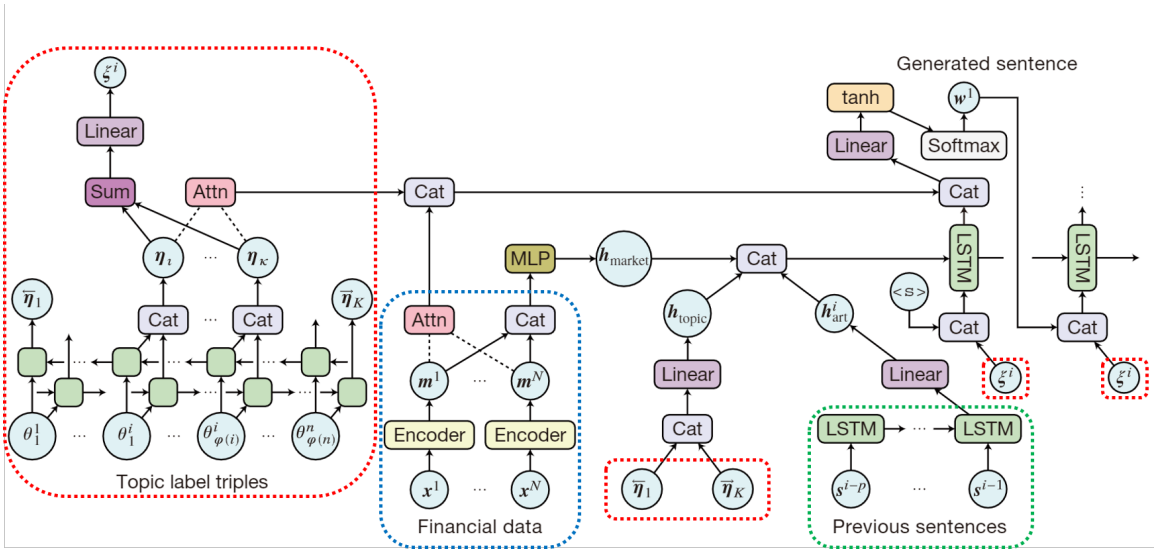

Figure 1: The neural-network architecture of our model with topic labels. The model generates each sentence $\boldsymbol{s}^t = (w_t^1, w_t^2, \ldots)$ separately with document topic labels $\boldsymbol{\Theta} = (\theta_1^1, \ldots, \theta_1^i, \ldots)$, financial data $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^N)$ and also previous sentences $\boldsymbol{s}^{i-p}$ to $\boldsymbol{s}^{i-1}$. Topic labels are encoded to capture both of the document-level and sentence-level information. We denote concatenation as *Cat*, 3-layer MLP as *MLP*. In addition, *Attn* denotes the attention with the hidden states of the decoder.

of which indicates *target*, *actuals or futures*, and *trade or movement*. The latter two are subcategories of *target*; for example, [US Market]-[Actuals]-[Movement] denotes a topic about the movement of actuals of the US market, while [Nikkei 225]-[Futures]-[Trade] means the trading activity of the futures market of Nikkei 225. [Others] is given when these subcategories do not apply.

Table 1 shows all atomic labels, and Table 2 illustrates an example of sentence and its topic labels. Note that a topic label is given to a clause, which means one sentence may have multiple topic labels. On the annotated data, the average number of labels for each sentence is 2.0.

We designed the topic labels aiming to make it easy for a user to control the contents. News articles often refer to trends of each market indices and also their relationships focusing on the general market movements. This means that sometimes the generated summary would not contain the contents expected by the user, especially when the markets attended by the user are not common ones. Therefore, we developed *topic labels* with the concrete names of market indices. The *topic labels* help a user to generate articles which have sufficient content by inputting market names as their interest.

Besides, the granularity of topics is an important factor to design the topic labels. Too rough topics could lead the vagueness of meaning of the topic while too detailed topics would be difficult for some users and take time to annotate. We designed our topic labels in a hierarchical structure to make them adaptive to different levels of granularity.

Human-designed labels are inferior in terms of construction cost, but expected to be superior to automatically extracted keywords in terms of interpretability as discussed later.

3.2 Encoder–Decoder with Topic Labels

Our model is an extended encoder–decoder that conditions on a document topic label sequence and previous sentences, in addition to financial data consisting of multiple numerical sequences. To make learning and generation simpler, our decoder generates each sentence separately, by encoding the sentence that was generated last. The entire neural-network architecture of our encoder–decoder model is shown in Figure 1.

Assume that we have generated $i-1$ sentences in an article and generate the next i-th sentence. Let $s^j = (w_t^j)_{t=1}^T$ $(j = 1, \ldots, n)$ be the j-th sentence in an article, where each w_t^j is a word. We also use $\boldsymbol{w}_t^j$ to denote embedding of w_t^j. In addition, in the following, W_* and b_* are a weight matrix and bias terms in the model parameters, respectively.

3.2.1 Encoders

We employ three encoders for encoding financial indices, previous generated sentences, and topic labels. To generate a next sentence s^i, from these

Topic Labels
Target: [Nikkei 225], [Individual stock (Japan)], [Narrow-based stock indices (Japan)], [TSE1] (First section of Tokyo Stock Exchange), [TSE2] (Second section of Tokyo Stock Exchange), [TOPIX], [US Market], [Individual stock (US)], [Narrow-based stock indices (US)], [DJI] (Dow Jones Industrial Average), [Hong Kong Market], [Market (Other countries)], [Individual stock (Other countries)], [Narrow-based stock indices (Other countries)], [JPY], [JPY/USD], [JPY/EUR], [JPY/AUD], [JPY Others], [USD], [AUD], [EUR], [HKD], [JGB] (Japanese government bond), [JGB (5–10 years)], [JGB (2–3 years)], [US Treasury securities], [US Treasury Notes (5–10 years)], [US Treasury Notes (2–3 years)], [TIBOR], [JPY interest rate], [Economic Index], [Events], [Lack of information for making a decision], [Investors], [Buying operation], [Statement of prominent person], [Others]
Actuals or Futures: [Actuals], [Futures], [N/A]
Trade or Movement: [Trade], [Movement], [N/A]

Table 1: Full set of human-designed atomic topic labels. Each topic is a tuple of values from three categories.

Topic labels	Segments of a sentence
[US Market]-[Actuals]-[Movement]	Observing the US stock markets fell during the year-end and New Year holiday, 年末年始の米株式相場の下落を受けて
[Investors]-[N/A]-[N/A]	investors got slightly risk-averse, 投資家のリスク回避姿勢がやや強まり、
[Nikkei 225]-[Actuals]-[Trade]	and selling pressure prevailed in the (Japanese stock) markets. 売りが優勢だった。

Table 2: Example of a sentence and its topic labels. One sentence may have multiple topic labels as in this example.

encoders we first obtain three vectors, $\boldsymbol{h}_{\text{market}}$, $\boldsymbol{h}^i_{\text{art}}$, and $\boldsymbol{h}_{\text{topic}}$, which encode financial data, previous sentences, and all topic labels on the article, respectively. Note that $\boldsymbol{h}_{\text{topic}}$ encodes a sequence of all topics on the article. We also obtain a vector encoding the topic labels of the current sentence, denoted by $\boldsymbol{\xi}^i$, from the topic labels encoder.

Financial data encoder

As the encoder of financial data, we follow Murakami et al. (2017). Given L numerical sequences $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^L$, which are numerical sequences of financial indicators (e.g., Nikkei stock average and the foreign exchange rate of Japanese yen). We concatenate these vectors and feed it to a 3-layer MLP to obtain a single vector $\boldsymbol{h}_{\text{market}}$:

$$\boldsymbol{h}_{\text{market}} = \text{MLP}([\boldsymbol{m}^1 ; \ldots ; \boldsymbol{m}^L]).$$

We use different MLPs to convert numerical sequences into vectors. Note that $\boldsymbol{x}^l$ consists of $\boldsymbol{x}^l_{\text{short}}$ and $\boldsymbol{x}^l_{\text{long}}$, which are short-term and long-term normalized data of $\boldsymbol{x}^l$. $\boldsymbol{x}^l_{\text{short}}$ is composed of the previous prices within one trading day and $\boldsymbol{x}^l_{\text{long}}$ is composed of the closing prices of the seven preceding trading days.

$$m^l = \boldsymbol{W}_l([\boldsymbol{x}^l_{\text{short}} ; \boldsymbol{x}^l_{\text{long}};\ \text{MLP}(\boldsymbol{x}^l_{\text{short}}) ; \text{MLP}(\boldsymbol{x}^l_{\text{long}})]) + b_l.$$

Previous sentences encoder

The hidden state $\boldsymbol{h}^i_{\text{art}}$ representing previous sentences of s^i is obtained from p preceding sentences, $s^{i-p}, \ldots, s^{i-1}$. We input embeddings $\boldsymbol{w}^{i-p}_1, \ldots, \boldsymbol{w}^{i-p}_{|s^{i-p}|}, \ldots, \boldsymbol{w}^{i-1}_1, \ldots, \boldsymbol{w}^{i-1}_{|s^{i-1}|}$ to long short-term memory (LSTM) cells and obtain $\boldsymbol{h}_{\text{art}^i}$ after passing its terminal hidden state to a linear layer, where $|s^j|$ denotes the number of tokens in s^j.

Topic label encoder

The outputs of this encoder are $\boldsymbol{h}_{\text{topic}}$, which encodes the topic sequence of the article, and $\boldsymbol{\xi}^i$, corresponding to the embedding of topics assigned to the target (i-th) sentence s^i. The reason why we encode the document-level topic sequence, rather than sentence topics only, is to make $\boldsymbol{\xi}^i$ context-sensitive, by which we expect the output sentence, conditioned on $\boldsymbol{\xi}^i$, to reflect the position of sentence topics on the entire document topics.

We use a bidirectional LSTM network (see the left part of Figure 1) to encode the sequence of document topics. As an input, we first concatenate all

topic labels of sentences with a token </s>. Then $\boldsymbol{h}_{\text{topic}}$ is obtained from the outputs of this LSTM, by concatenating the outputs at the end tokens of both directions. These are denoted by $\overrightarrow{\boldsymbol{\eta}}_K$ and $\overleftarrow{\boldsymbol{\eta}}_1$ in Figure 1, where $K = \sum_j \varphi(j)$, the length of the document-level topic sequence. $\varphi(j)$ denotes the length of assigned topics for s^j, including the last topic label </s>.

We then extract topic embedding corresponding to the topic labels of s^i as *topic embedding* $\boldsymbol{\xi}^i$. We do this by summing all outputs of LSTMs in a span corresponding to the current sentence *excluding* </s>. Let θ^i_k be the k-th topic in s^i. The bi-LSTM introduced above transforms this input into a vector $\boldsymbol{\eta}_t = [\overleftarrow{\boldsymbol{\eta}}_t ; \overrightarrow{\boldsymbol{\eta}}_t]$, by concatenating the outputs of LSTMs of both directions. Then $\boldsymbol{\xi}^i$ is obtained by summing the outputs in the span, followed by a linear layer:

$$\boldsymbol{\xi}^i = W_\xi \left(\sum_{t=\iota}^{\kappa} \boldsymbol{\eta}_t \right) + b_\xi,$$

in which ι and κ are the indices corresponding to θ^i_1 and $\theta^i_{\varphi(i)-1}$, the start and end topic labels for the i-th sentence. Formally, $\iota = \sum_{j=1}^{i-1} \varphi(j) + 1$ and $\kappa = \iota + \varphi(i) - 1$.

3.2.2 Decoder

Our decoder is another LSTM conditioned on the outputs of three encoders introduced above. To initialize the decoder, we first concatenate three outputs of encoders and apply a linear layer:

$$\boldsymbol{H}^i_0 = W_H[\boldsymbol{h}_{\text{topic}} ; \boldsymbol{h}_{\text{market}} ; \boldsymbol{h}^i_{\text{art}}] + b_H. \quad (1)$$

Note that $\boldsymbol{h}_{\text{topic}}$ encodes the entire document-level topics, not the topics for the target sentence only. To make the output sentence more relevant to those target topics, we feed $\boldsymbol{\xi}^i$ to the input of the decoder at every step, by concatenating it with the original input vector $\boldsymbol{w}^j_t$.

While this would allow the contents of output sentence to follow the given local topics, the sentence should also reflect the information of global topic sequence, e.g., the relative position of the target sentence in the article. A natural way to encode such context in the decoder is the attention mechanism (Luong et al., 2015), which we apply to the outputs of topic label encoder $\boldsymbol{\eta}_t$, as well as the outputs of financial data encoder $\boldsymbol{m}_l$, to capture the important resources relevant to the current sentence.

We obtain the context vectors $\boldsymbol{c}^{\text{market}}_t$ and $\boldsymbol{c}^{\text{topic}}_t$ for attending the financial data and topic labels from the output of decoder LSTM, and concatenate them before the softmax layer:

$$w_t \sim \text{softmax}(\tanh(W_c[\boldsymbol{c}^{\text{topic}}_t ; \boldsymbol{c}^{\text{market}}_t ; \boldsymbol{H}^i_t] + b_c)).$$

The context vector $\boldsymbol{c}^{\text{market}}_t$ is obtained by a bilinear attention (Luong et al., 2015):

$$\boldsymbol{c}^{\text{market}}_t = \sum_{l=1}^{L} \alpha^{\text{market}}_t(l)\boldsymbol{m}_l,$$

$$\alpha^{\text{market}}_t(l) \propto \exp(\boldsymbol{H}^{i\top}_t W^{\text{market}}_a \boldsymbol{m}_l).$$

$\boldsymbol{c}^{\text{topic}}_t$ is obtained similarly:

$$\boldsymbol{c}^{\text{topic}}_t = \sum_{k=1}^{K} \alpha^{\text{topic}}_t(k)\boldsymbol{\eta}_k,$$

$$\alpha^{\text{topic}}_t(k) \propto \exp(\boldsymbol{H}^{i\top}_t W^{\text{topic}}_a \boldsymbol{\eta}_k).$$

4 Experimental Settings

Each example in our dataset is a pair of aligned time-series data and a corresponding document. We obtained documents by retrieving daily summaries from NQN, which describes market trends in Japanese, as well as aligned time-series data, from Thomson Reuters DataScope Select[1]. Dividing by periods, we obtained 864, 122, and 124 documents (9,337, 1,215, and 1,237 sentences) for train/valid/test sets, respectively. The vocabulary size was 3,025. As topic labels, we used 91 human-designed topic labels and 818 kinds of extracted keywords by the RAKE algorithm. We preprocessed each indicator following Aoki et al. (2018), and used the same parameters for the financial data encoder. Other parameters were tuned by document-level BLEU scores on the validation set.

We compared five different documents; a document written by human writer (Gold), a document generated by our model without topic labels (NoTopicLabel) and three documents respectively generated by our models using topic labels:

- HDTag3: Human-designed topic labels.
- HDTag1: Simplified Human-designed topic labels (only target of Table 1) to see the importance of other factors.
- RAKE: Two keywords extracted by RAKE.

[1] We retrieved five-minute or one-hour charts of 9 indicators (Nikkei 225, TOPIX, DJI, HKHSI, USD/JPY, USD/EUR, JGB (2 years), JGB (10 years), US Treasury Notes (10 years).

Method	BLEU (doc)	BLEU (sent)
NoTopicLabel	21.19±1.16	14.69±0.39
HDTag3	29.21±0.29	22.77±0.40
HDTag1	27.92±0.37	21.25±0.57
RAKE	29.53±0.25	23.35±0.40

Table 3: Result of evaluation in terms of BLEU. Scores were averaged over 5 runs. The values after ± are the standard deviations. We report both the averaged BLEU scores over all the documents (**BLEU (doc)**) and sentences (**BLEU (sent)**).

We conducted both an automatic evaluation with BLEU score in words and a human-evaluation. The human evaluation focused on the fluency and the fidelity and the correctness of each approach. For human evaluation, we sample 15 instances from the test dataset. For each of the 15 instances, evaluators are presented with 5 documents that are respectively generated by a human writer (Gold), NoTopicLabel, HDTag3, HDTag1, and RAKE. Note that NoTopicLabel does not use topic labels, while HDTag3, HDTag1, and RAKE use topic labels. The evaluators are asked to rate the documents on a 1–3 scale with respect to fidelity, correctness and fluency. **Fidelity** measures whether each document reflects the given topic labels. **Correctness** measures whether each document is faithful to the given financial data. **Fluency** measures the fluency of each document without regard to input data. Since the evaluation of **Correctness** is a complicated process which requires the reference to input numerical data, the evaluators are supposed to evaluate only the sentences that satisfy the following two conditions: (i) the sentence starts with "Nikkei stock average" or "The exchange rate of the Japanese yen"[2], (ii) the sentence is labeled with only [Nikkei 225/ Actuals/ Movement] or [JPY//Movement]. Additionally, the evaluators are also asked to conduct sentence-level evaluation of fluency, in which they are presented with 5 sentences generated by the 5 methods including Gold. All the evaluations are conducted by two evaluators, and we compute the average scores for each approach.

[2]The original Japanese phrases are "日経平均株価は (Nikkei stock average)" and "円相場は (The exchange rate of the yen)".

5 Results

Table 3 shows the BLEU scores of different approaches. The models with label information (HDTag3, HDTag1, and RAKE) achieved higher performances in terms of BLEU. RAKE achieved a slightly higher BLEU score than HDTag3, but the difference was not statistically significant. HDTag3 achieved a higher BLEU score than HDTag1. This result suggests that more informative topic labels improved the quality of generated text. In other words, careful design of topic labels helps high-quality generation, although it requires more human cost. It is also encouraging that HDTag3 is comparable to RAKE, in spite of the fact that the labels in the latter are extracted from words in the reference.

The results of human-evaluation are shown in Table 4. There was no statistically-significant difference among the sentence-level fluency scores of all methods. This means that the neural-network based method has the ability to generate a fluent text at least at the sentence level. The methods with topic-label information showed a better document-level fluency than the one without topic labels.

Meanwhile, there was a significant difference between the document-level fluency scores of generated sentences and human-written sentences. We considered it is caused by not considering relationships among topic-labels and also by weak consideration of generated sentences. Specifically, our model possibly generates almost the same content repeatedly as the content of previously generated sentences which are not treated as input resource, and moreover our model could generate different movement descriptions about the same indicator within a document. An example is shown in Table 5, where two sentences describing the same movement of the exchange rate state the contradictory things; *dropped* and *rose*. To solve the above problems, the implementation of additional memories to keep tracking which topics have been mentioned and how topics have been mentioned is interesting avenue for future work.

Besides, we observed the correctness of RAKE is higher than that of the other models. It is not surprising, because topic labels of RAKE are words in the target documents, and the topic labels like *continuously fall* or *rebound* would directly deliver the characteristics of the input data. In comparison between the methods with human-designed labels, HDTag3 is superior to HDTag1. This result is

Method	Fidelity	Correctness	Fluency (doc)	Fluency (sent)
GOLD	–	2.80	2.90	2.93
NOTOPICLABEL	–	1.70	1.13	2.86
HDTAG3	2.70	2.23	1.70	2.90
HDTAG1	2.60	1.93	1.86	2.86
RAKE	1.96	2.53	1.76	2.93

Table 4: Result of human-evaluation. Scores range in [1,3]. **Fidelity** measures whether each document reflects the given topic labels. **Correctness** measures whether each document is faithful to the given financial data. **Fluency** measures the fluency of each document or sentence without regard to input data. **Fluency (doc)** is the document-level fluency, while **Fluency (sent)** is the sentence-level fluency.

Method	Topic label	Text
GOLD		. . ., </s>, "The yen exchange rates rose for four days in a row.", </s>, "There were movements to sell the yen along with the rise of Nikkei 225, but they soon calmed down.", </s>, "In the afternoon, the yen rose and hit a high of around 108.5 yen against the dollar, which was the highest in about three weeks.", </s>, ..., </s>, "円相場は 4 日続伸した。", </s>, "日経平均の上昇と歩調を合わせて円を売る動きがあったが、勢いは続かなかった。", </s>, "午後に 1 ドル=108 円台半ばまで上昇し、約 3 週間ぶりの高値を付ける場面があった。", </s>, . . .
HDTAG3	. . ., </s>, [JPY]-[N/A]-[Movement], </s>, [Nikkei]-[Actual]-[Movement],[JPY]-[N/A]-[Trade], [JPY]-[N/A]-[Movement], </s>, [JPY/USD]-[N/A]-[Movement], </s>, . . .	. . ., </s>, "The yen exchange rates dropped.", </s>, "Observing the steady rise of Nikkei 225, traders started to sell the yen regarding it as a low-risk currency, and it did not rise constantly.", </s>, "The yen exchange rate rose against the dollar.", </s>, ..., </s>, "円相場は反落。", </s>, "日経平均株価が堅調に推移したことで、「低リスク通貨」とされる円を売る動きが出て、円の上値は重かった。", </s>, "円は対ドルで上昇した。", </s>, . . .

Table 5: Sentences generated by HDTAG3.

consistent with the result of BLEU scores.

Moreover, both models with human-designed topic labels show higher fidelity, which means the generated documents reflect the given topic labels. We speculate that the lower fidelity of RAKE is caused by the ambiguity of extracted keywords as discussed through examples in the next paragraph.

We then provide a qualitative comparison of RAKE and human-designed topic labels. Table 6 shows some output examples. As we mentioned, we found that RAKE keywords are often more ambiguous than the human-designed topic labels. This is mainly because the granularity of keywords is not properly defined. Table 6(a) shows an example, where RAKE keywords contain "high", which however does not tell which quantity is high, resulting in the wrong contents in the generated sentence. The human-designed topic labels have a higher interpretability, and the sentences generated with such topic labels are well-controlled.

The quantitative and qualitative evaluations above suggest that human-designed topic labels contribute to a better controllability backed up by high fidelity and interpretability.

Although our approach has the advantage of the controllability in generating sentences, we also found complication in terms of topic design, in particular, the definition of granularity of the topics. We found that the system often generates a wrong description when the topic labels contain a general label, such as `Others` and `Events`. These labels tend to be used as catch-all labels, resulting in diverse contents. Table 6(b) shows an example that `Others` leads to a longer sentence with wrong contents.

To demonstrate that we can control the contents given the same financial data, in Table 6(c), we show how a generated sentence varies by giving topic labels that are different from the actual topic labels (HDTAG3UNSEEN). We can see that a generated sentence properly changes its contents so that it reflects the new topic labels.

(a) Sentences generated by HDTAG3 and HDTAG1, RAKE, for which the topic labels by RAKE are ambiguous.

Method	Topic label	Text
GOLD		. . ., </s>, “Then, Nikkei 225 significantly rose, and traders sold the yen expecting investors to get risk-appetite and regarding yen as a ‘low-risk currency’.”, </s>, ..., </s>, “その後、日経平均株価が大幅高となり、投資家のリスク選好意欲が強まるとみた市場参加者が「低リスク通貨」とされる円の売りを増やした。”, </s>, . . .
HDTAG1	. . ., </s>, [Nikkei 225], [Investors], [JPY], </s>, . . .	. . ., </s>, “A rise of Nikkei 225 improved investor sentiment and prompted traders to buy the yen regarding it as a low-risk currency.”, </s>, ..., </s>, “日経平均株価が上昇したことで投資家心理が改善し、低リスク通貨とされる円を買う動きが優勢だった。”, </s>, . . .
HDTAG3	. . ., </s>, [Nikkei 225]-[Actuals]-[Movement], [Investors]-[N/A]-[N/A], [JPY]-[N/A]-[Trade], </s>, . . .	. . ., </s>, “A significant rise of Nikkei 225 also improved investor sentiment and prompted traders to buy the yen.”, </s>, ..., </s>, “日経平均株価が大幅に上昇したことも投資家心理の改善につながり、円買いを誘った。”, </s>, . . .
RAKE	. . ., </s>, high, preference, </s>, . . .	. . ., </s>, “However, traders bought the yen and sold the dollars to adjust the position in advance of the Fed’s Interest-Rate Announcement.”, </s>, ..., </s>, “ただ、米連邦公開市場委員会 (FOMC) の結果発表を控え、持ち高調整の円買い・ドル売りが入った。”, </s>, . . .

(b) Sentences generated by HDTAG3 with [Others].

Method	Topic label	Text
GOLD		. . ., </s>, “Speculation that the Bank of Japan purchased exchange-traded funds was positive for the markets.” </s>, ..., </s>, “日銀が株価指数連動型上場投資信託 (ETF) 買いを入れたとの思惑も相場を支えた。”, </s>, . . .
HDTAG3	· · ·, </s>, [Others]-[N/A]-[N/A], [Investors]-[N/A]-[N/A], [Nikkei 225]-[Actuals]-[Movement], </s>, · · ·	. . ., </s>, “On alert against the debt problem in Europe, stock prices continued to decline while bonds continued to rise.” </s>, ..., </s>, “もっとも、欧州債務問題への警戒感から株安・債券高が進んだ。”, </s>, . . .

(c) Sentences generated by HDTAG3 and HDTAG3UNSEEN.

Method	Topic label	Text
GOLD		<d>, “Nikkei 225 rebounded greatly.”, </s>, “It was the fourth largest leap in this year.”, </s> . . . . . <d>, </s>, “日経平均株価は大幅に反発した。”, </s>, “上げ幅は今年 4 番目の大きさとなった。”</s> . . .
HDTAG3	<d>, [Nikkei 225]-[Actual]-[Movement], </s>, [Nikkei 225]-[Actual]-[Movement], </s>, . . .	<d>, “Nikkei 225 rebounded greatly.”, </s>, “The closing price was the highest since <month/> <day/>.”, </s> . . . <d>, </s>, “日経平均株価は大幅反発した。”, </s>, “終値は<month/><day/>以来約<months/>ぶりの高値を付けた。”</s> . . .
HDTAG3UNSEEN	<d>, [Nikkei 225]-[Actual]-[Movement], </s>, [US Market]-[Actual]-[Movement], [Nikkei 225]-[Actual]-[Trade], </s>, </d>	<d>, “Nikkei 225 rebounded greatly.”, </s>, “Observing the rise in the US stock markets yesterday, traders tended to buy stocks from this morning.“, </s>, </d> <d>, </s>, “日経平均株価は大幅反発した。”, </s>, “前日の米株式相場の上昇を受けて朝方から買いが先行した。”, </s>, </d>

Table 6: Examples of generated sentences.

6 Conclusion

We proposed a data-to-document generator which can be controlled by a sequence of topic labels. We compared two topic labels, the human-designed topic label and automatically extracted keywords, and conducted experiments with a financial dataset. Our experiments empirically showed that the model using topic label information achieved higher performance in terms of BLEU and human-evaluation. Furthermore, the model using the human-designed topic labels has an advantage on controllability of the output documents without reducing BLEU scores. In addition, experiments showed that the granularity of topic labels influences the generation quality.

As future work, we will employ the network architectures which have additional memories to keep tracking which topics have been mentioned and how topics have been mentioned for high topical coherence in the sentences. In addition, future work should include reducing the inconsistency between a generated text and the actual movement of input financial indicators because even one conflict could be fatal to the reliability of the generated text.

Topic labels should also be easy to handle for human users, who actually use the system to generate a document. We also need to evaluate topic labels in terms of the easiness of use.

Acknowledgements

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

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 *Proc. of INLG 2018*, pages 135–139.

Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 331–338, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104. Association for Computational Linguistics.

Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2017. Data-to-text generation improves decision-making under uncertainty. *IEEE Computational Intelligence Magazine*, 12:10–17.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Juraj Juraska and Marilyn Walker. 2018. Characterizing variation in crowd-sourced data for training neural language generators to produce stylistically varied outputs. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 441–450, Tilburg University, The Netherlands. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Rémi Lebret, David Grangier, and Michael Auli. 2016. 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. 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. 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. Association for Computational Linguistics.

Tianyu Liu, Kexiang Wang, Baobao Chang Lei Sha, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence.*

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

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. Association for Computational Linguistics.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2018. CGMH: constrained sentence generation by metropolis-hastings sampling. *CoRR*, abs/1811.10996.

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 *Proc. of ACL 2017*, pages 1374–1384.

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, Saarbrcken, Germany. Association for Computational Linguistics.

Francois 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(7-8):789–816.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Nat. Lang. Eng.*, 3(1):57–87.

Jacques Pierre Robin. 1995. *Revision-based Generation of Natural Language Summaries Providing Historical Background: Corpus-based Analysis, Design, Implementation and Evaluation*. Ph.D. thesis, New York, NY, USA. UMI Order No. GAX95-33653.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining. Applications and Theory*, pages 1–20.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6830–6841. Curran Associates, Inc.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060. The COLING 2016 Organizing Committee.

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. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2018. Learning neural templates for text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3174–3187, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng Yang, Lei Li, Fuli Luo, Tianyu Liu, and Xu Sun. 2019. Knowledgeable writer: Enhancing topic-to-essay generation with external commonsense knowledge. In *Proc. of ACL 2019*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-and-write: Towards better automatic storytelling.

Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2199, Copenhagen, Denmark. Association for Computational Linguistics.

Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5902–5911, Stockholmsm ä ssan, Stockholm Sweden. PMLR.