Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 319–327
Marseille, 20-25 June 2022

Automating Horizon Scanning in Future Studies

Tatsuya ISHIGAKI[1] **Suzuko NISHINO**[1,3] **Sohei WASHINO**[1] **Hiroki IGARASHI**[1]
Yukari NAGAI[3] **Yuichi WASHIDA**[2] **Akihiko MURAI**[1]
[1]National Institute of Advanced Industrial Science and Technology
[3]Graduate School of Advanced Science, Japan Advanced Institute of Science and Technology
[2]School of Business Administration, Hitotsubashi University
{ishigaki.tatsuya, s.washino, hk-igarashi, a.murai}@aist.go.jp,
{suzuko_k, ynagai}@jaist.ac.jp, b101348r@r.hit-u.ac.jp

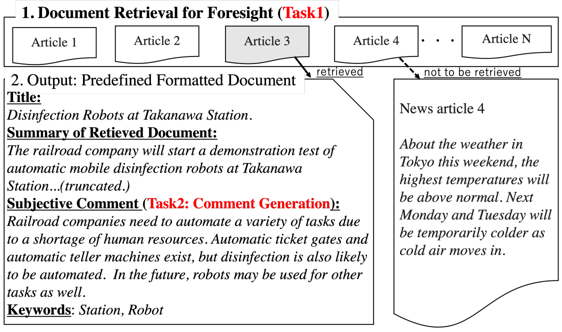
Abstract

We introduce document retrieval and comment generation tasks for automating horizon scanning (Sardar, 2010). This is an important task in the field of futurology that collects sufficient information for predicting drastic changes in the mid- or long-term future. The steps used are: 1) retrieving news articles that imply drastic changes, and 2) writing subjective comments on each article for others' ease of understanding. As a first step in automating these tasks, we create a dataset that contains 2,266 manually collected news articles with comments written by experts. We analyze the collected documents and comments regarding characteristic words, the distance to general articles, and contents in the comments. Furthermore, we compare several methods for automating horizon scanning. Our experiments show that 1) manually collected articles are different from general articles regarding the words used and semantic distances, 2) the contents in the comment can be classified into several categories, and 3) a supervised model trained on our dataset achieves a better performance. The contributions are: 1) we propose document retrieval and comment generation tasks for horizon scanning, 2) create and analyze a new dataset, and 3) report the performance of several models and show that comment generation tasks are challenging.

Keywords: generation, document retrieval

1. Introduction

We introduce document retrieval and comment generation tasks for automating horizon scanning (Sardar, 2010) studied in the field of Futurology study. Futurology studies' methodologies predict possible scenarios of drastic social changes that may occur in the future. Scenarios about the future are particularly important for policymaking by corporations and governments. A common method in this field, the systemic foresight methodology (SFM) (Saritas, 2013), divides scenario planning into two sub-tasks: 1) horizon scanning and 2) planning. The former phase collects much information, for example, articles that contain "future sprouts". These are expressions that imply drastic changes in the future appearing in news articles, blogs, and other existing resources (Palomino et al., 2012). The latter phase aggregates hundreds of collected articles and output scenarios. In this study, we particularly focus on the horizon scanning phase because this step's quality affects the latter steps, and it is currently conducted by costly human experts. The application of language processing technologies can be a solution.



Title:
Disinfection Robots at Takanawa Station.
Summary of Retieved Document:
The railroad company will start a demonstration test of automatic mobile disinfection robots at Takanawa Station...(truncated.)
Subjective Comment (Task2: Comment Generation):
Railroad companies need to automate a variety of tasks due to a shortage of human resources. Automatic ticket gates and automatic teller machines exist, but disinfection is also likely to be automated. In the future, robots may be used for other tasks as well.
Keywords: *Station, Robot*

About the weather in Tokyo this weekend, the highest temperatures will be above normal. Next Monday and Tuesday will be temporarily colder as cold air moves in.

Figure 1: An Example of Horizon scanning.

Figure 1 shows the horizon scanning procedure defined in an existing study (Washida and Yahata, 2021). The input is a large set of articles. In this study, as an example, we focus on news articles written in Japanese as input. The output is a document represented in a predefined format. We aim to automate this procedure by solving two sub-tasks: 1) document retrieval and 2) comment generation. In the first step, documents to be retrieved should contain information indicating future drastic changes. For example, Article 3, shown in Figure 1, is about a robot demonstration implying robots' deployment in diverse industries. Conversely, Article 4 is not to be retrieved because it is about the weather in Tokyo, and thus, does not indicate future drastic changes. Each collected article is represented as a document with a predefined format that contains the title, a summary, and a subjective comment about the article's topic. We can apply existing natural language processing (NLP) technologies, such as headline generation (Rush et al., 2015) or single-document summarization (Nallapati et al., 2016; Cheng and Lapata, 2016) to the content; except for the comment section. Thus, we focus on generating comments as the second sub-task. The comment's section contains various subjective sentences that include a short explanation of the article, opinions, arguments, and/or inferred future expectations.

The document retrieval shown in this paper differs from the existing settings since a model correctly finds expressions that imply future societal changes, which has not been explored in existing document retrieval settings (Mitra et al., 2018). The comment generation task is considered a language generation problem. However, our setting generates comments for future events. Most existing generation tasks target past events or real-time text generation: for example, a summary of sports matches (Puduppully and Lapata, 2021; Iso et al., 2019) from past tracking data or live commentary (Tanaka-Ishii et al., 1998; Taniguchi et al., 2019).

For both tasks, the lack of a dataset is crucial. Thus, we create a novel dataset, comprising 2,266 Japanese articles manually retrieved by experts and students with knowledge of the scanning phase. In this study, we first analyze the characteristics of the retrieved documents regarding the words used and the semantic distance to the general articles. We also analyze the comments' contents by assigning manually designed labels. Furthermore, we apply several automatic methods of document retrieval and language generation.

Our analysis shows that 1) the documents to be retrieved have characteristic words that are different from those of general articles, and the retrieved articles are semantically far from the general articles. Our experiment shows that 2) a BERT-based supervised model can achieve a 70% performance in terms of Precision@100. This is higher than unsupervised or heuristics-based models. 3) A BART-based comment generator outputs fluent text, containing not only a summary of the input article, but also opinions and arguments. However, there is room for improvement in terms of correctness.

The contributions are as follows: 1) we propose document retrieval and comment generation tasks for horizon scanning, 2) create and analyze a new dataset, and 3) report the performance of several models and show that comment generation tasks are challenging [1]

2. Related Work

2.1. Horizon Scanning

Among the futurology methodologies, we particularly focus on the SFM (Saritas, 2013). It involves two phases: the horizon scanning and the planning phases. The former collects much information, for example, news articles. The latter aggregates the collected information by merging (Nagai et al., 2009) and/or grouping (Kunifuji, 2016) the collected articles, and finally generates scenarios from them. Our focus is on the scanning phase as the quality of this step affects the following phase, which can be considered important in automating SFM. The scanning phase is manually conducted by costly human experts, where automation can be beneficial. Furthermore, we are also interested in the retrieved documents' characteristics to understand experts' knowledge. We analyze the characteristics using computational and linguistic methods, and apply several models' retrieval and comment generation tasks.

2.2. Document Retrieval and Language Generation

Document retrievers using neural network-based supervised and unsupervised methods have recently attracted scholarly attention (Mitra et al., 2018; Kwon et al., 2020). Similar to our setting, Trivia Extraction (Kwon et al., 2020) retrieves surprising articles (trivia) from Wikipedia. They use the semantic distance between trivia articles and general articles, calculated using the inverse document frequency (IDF) and neural network-based distributed representations. In this study, we compare the performance of document retrieval using IDF and distances, and a supervised retriever based on a simple BERT-based classifier. The comment generation task can be regarded as one of the language generation problems. In recent studies, there are summary generations for sports matches, including basketball (Puduppully and Lapata, 2021; Iso et al., 2019). However, live commentaries for soccer (Tanaka-Ishii et al., 1998; Taniguchi et al., 2019), baseball (Kim and Choi, 2020), and e-sports (Ishigaki et al., 2021) focus on real-time events. Conversely, only a few studies have focused on generating texts about the future, such as weather forecast generation (Murakami et al., 2021). Our comment generation differs from the above because our focus is on generating subjective comments on medium- to long-term social changes.

3. Dataset

In this section, we describe the dataset. It contains 1) gold articles that should be retrieved and 2) subjective comments about the retrieved articles. Creating such a dataset requires expert knowledge of horizon scanning. Thus, we asked experts to manually collect the articles to be retrieved. The experts are professors who regularly teach the scanning phase and MBA students who learned about horizon scanning from professors through a series of classes on scenario planning. Thirty-three experts collected articles to be retrieved and wrote a comment for each article as follows:

1. They searched for news articles on the Web and extracted articles containing information that implied drastic changes in the medium- to long-term future.
2. They wrote a comment, a summary, and a title.
3. They aggregated the comment, summary and title to a predefined format as shown in Figure 1.

The data collection step was conducted over a 17-year period from 2003 to 2020. Thus, they retrieved news articles written in Japanese published between 2003

[1] The data used in the document retrieval experiments and the trained model of the comment generator will be publicly available after approval from our organization [2]

and 2020. They were instructed that the retrieved documents should be maximally diverse, regarding categories and topics. Many students first filtered several news articles by using meta-information, such as categories assigned to the articles. Thereafter they carefully read the headlines and body to determine whether the articles contained expressions about drastic social changes in the future. For each retrieved article, the expert who retrieved it manually wrote a subjective comment with 2.8 sentences on average. Finally, we collected 2,266 articles and documents with the predefined format.

4. Analysis of Dataset

We analyze the collected articles' characteristics and comments. We investigate articles' 1) characteristic words and 2) semantic distance to general articles. We then analyze the comments via manually designed labels as per Table 1.

4.1. Characteristic Words in Collected Articles

The characteristic words may differ from those in general news articles. We analyze this hypothesis by using the gap in Term Frequency-Inverse Document Frequency (TFIDF) between the collected documents and general news articles.

First, let C be the set of retrieved articles, G be the set of general articles, and let TF_C and TF_G of a word x be the values of TF separately calculated for two sets. Let IDF_{C+G} be the value of IDF calculated on the union of C and G. We define the gap between the two TFIDF values as the score of the likelihood of the word x's characteristic:

$$diff(x) = TF_C(x)IDF_{C+G}(x) - TF_G IDF_{C+G}(x) \quad (1)$$

Here, if $diff(x)$ is positive, x characterizes the retrieved articles in C, and if it is negative, x characterizes articles in G. Table 1 shows the top seven words whose values of $diff(x)$ are higher in the positive direction and lower in the negative direction, respectively. We observe that top-scored words often indicate events that rarely occur, for example, "new features of Twitter" or "drought in Taiwan." The terms related to the technology field are also highly ranked. Conversely, in words with lower scores, we observe many words that often appear in articles about daily news, such as articles about criminal cases for example., "interrogation of suspects," or about regular events, such as the governor's press interviews.

4.2. Semantic Distance to General Articles

We assume that the collected articles are semantically far from general articles. To verify this hypothesis, inspired by Kwon et al. (Kwon et al., 2020), we measure the semantic distance between collected articles and general articles by word embeddings. In this analysis, we use articles published in 2020 in Asahi Shimbun [3] as general articles. We obtain word embeddings of nouns in each article by using word2vec (Mikolov et al., 2013). For each article, we calculate a document embedding as the averaged vector over the embeddings of nouns. Finally, we get the centroid of these document embeddings as the representation of general articles. Similarly, we obtain a document embedding of each collected article published in 2020. We obtain the embeddings by averaging word2vec embeddings of nouns. We then compute the distance between the centroid of general articles and each retrieved article X as:

$$\texttt{dist(cent, X)} = 1 - cosine(cent, X). \quad (2)$$

Here, cosine() returns cosine similarity between two vectors. The averaged `dist` over collected articles was 0.34, which is farther than the averaged value over all general articles i.e., 0.30. We use the t-test to confirm that the difference is significant. Therefore, we confirm that the collected articles are different from general articles in the semantic space.

4.3. Contents in Comments

We examine the types of content in the comments. We randomly extracted 30 comments from the dataset and manually assigned at least one of five content labels shown in Table 2. Among the labels in this table, "summary" and "background" represent facts. The "summary" label particularly represents the summarized facts written in the retrieved article. In contrast, the "background" label represents facts that were not in the article, which may be extracted from the writers' knowledge. For example, the comment "Various toys have been recently released." mentions a fact, however, it is not in the retrieved article in Table 2. Thus, we label this comment as "background" not "summary". Conversely, the "argument (present)," "argument (future)," and "expectations" are descriptions of the author's subjective opinions. After assigning labels for each comment, the labels were sorted in their order of appearance in the comment.

From an analysis of the order of labels, 83% (25 out of 30) of comments began with facts: a "summary" or "background," and subjective opinions followed. Specifically, 63 % (19 out of 30) of the comments started with a "summary" and 20% (six out of 30) with a "background." Among the 83% of the comments, 96% (24 out of 25) contained subjective opinions such as "expectation," "arguments" or "neutral comments" after the facts. Four out of 30 comments contained only subjective opinions. One exceptional instance contained only a fact.

5. Method

We compare several methods to better understand the task's characteristics.

[3] http://www.asahi.com/information/cd/gakujutsu.html

diff	Words	Example
23.31	Twitter	*A new feature has been implemented that allows you to make and receive donations to your Twitter* account.
22.34	technology	*Femtech is a word coined from the combination of women and technology. This word refers to A business that uses technology to support women's healthcare and lifestyle.*
20.88	humidity	*The IoT monitoring tool provides real-time data on the tea plantation's growth environment, including illumination, temperature, humidity, and soil pH*
20.88	drought	*Taiwan is experiencing the worst drought on record in half a century.*
20.88	number of	*The number of ATMs is declining as the number of bank branches decreases in line with the shift to cashless banking.*
20.88	photosynthesis	*Toshiba aims to commercialize jet fuel by applying the mechanism of artificial photosynthesis...*
20.88	GABA	*Tomatoes with a higher content of GABA, which is effective in reducing elevated blood pressure, were unveiled.*
-0.89	Governor	*Governor Yuriko Koike gave a talk at a task force meeting*
-0.89	next year	*COP26 has been postponed to next year.*
-0.89	night	*An elderly woman in the city who was hospitalized died on the night of the 24th.*
-0.90	test	*PCR test is essential for the early detection and treatment of new coronaviruses.*
-0.90	afternoon	*The polls will be opened at 8:00 p.m. in the third-floor conference room of Yamato Town Hall.*
-0.90	Suspects	*The investigation department has not yet confirmed the identity of the suspects.*
-0.91	City	*The election of mayors will be held in the six cities of Omaezaki, Izu, Fujieda, Shimoda, and Kosai and Yaizu, as well as in Mori Town.*

Table 1: The collected articles characteristic and non-characteristic words. The listed words and examples are translated into Japanese.

Example input news article.
At the end of July, TOMY will release "Evio," a toy that makes the sound of a violin. It can play music like a violinist. The price is 7,000 yen. The company expects demand from a wide range of age groups, especially elementary schoolgirls.

Labels	Definitions	Example Texts
Summary	Summary of the input article.	*Tommy sells a toy that makes the sound of a violin*
Background	Objective information that is not part of the input.	*Various toys have been recently released.*
Argument (Present)	Opinions on summary articles and the background information.	*It is good that there are toys that make use of IT technology.*
Argument (future)	Opinions about possible future events.	*Use of information technology is essential to the evolution of toys.*
Expectation	Description of what may happen in the future.	*Other than toys, karaoke will be replaced.*
Neutral comments	Neutral comments on Input Article and Background	*The combination of toys and violins is interesting.*

Table 2: Definitions and Examples of Labels. An example article and comments were translated into Japanese.

5.1. Document Retrieval

In this task, we are given a large number of articles, and the model returns the top-N scored ones. We use a binary classifier to assign a score for each article. The score represents how likely the article is to be retrieved. We compare two neural network-based models and a heuristics-based model.

5.1.1. Neural Network-based Models

For neural network-based models, we compare supervised and unsupervised models.

For the supervised model, we use a simple BERT-based binary classifier to score the articles. For the training data, the collected articles are used as positive instances. Randomly extracted articles from the Asahi Shimbun corpus are used as negative instances. We first divide the articles into subwords and thereafter assign a special token (CLF) at the beginning. BERT (Devlin et al., 2019)-based encoder converts the subword sequence into a sequence of subword embeddings:

$$\mathbf{e}_{x_0}, \mathbf{e}_{x_i}, ..., \mathbf{e}_{x_n} = \texttt{BERT}(x_0, x_i, ..., x_n). \tag{3}$$

Here, e_{x_i} is the embedded representation of the ith subword; thus, $\mathbf{e}_{x_0}$ is the representation of the [CLF] token. We consider $\mathbf{e}_{x_0}$ to be the representation of the target article. This is used to calculate the probability distribution:

$$\mathbf{y} = \texttt{Softmax}(\mathbf{e}_{x_0}\mathbf{W}_e). \tag{4}$$

Here, $\mathbf{W}_e$ is the size of $\mathbf{e}_{x_0} \times 2$. We obtain the two-dimensional vector $\mathbf{y}$, where each dimension represents a score for the retrieval decision. $\mathbf{W}_e$ is learned to minimize the cross-entropy loss. This model is common for various NLP tasks. However, we verify whether this model can correctly classify articles that imply drastic future changes.

Next, we describe the unsupervised learning method. As described in Section 4, the articles to be retrieved are semantically distant from the general news articles. Therefore, we use distances. In this model, we use the distance between the centroid of embeddings of the Asahi Shimbun corpus calculated in Section 4 and the target article. We regard the distance as a target article's score to be retrieved:

$$1 - \texttt{cosine}(\mathbf{e}_{x_0}, cent(G)). \tag{5}$$

Here, Cosine returns the two vectors' cosine similarity. $\mathbf{e}_{x_0}$ represents the embedded representation of the article to be scored. $cent(G)$ is obtained by calculating the centroid of the embeddings for the Asahi Shimbun corpus.

5.1.2. Heuristics-based Model

We assume that the articles to be retrieved contain more characteristic words. As a heuristic to capture such a phenomenon, we score each article using the averaged IDF:

$$\frac{\sum_1^n \texttt{IDF}(x_i)}{n}. \tag{6}$$

where $\texttt{IDF}(x_i)$ returns the IDF value of x_i in the target article. The IDF is calculated only for nouns and is computed in advance from the Asahi Shimbun Corpus.

5.2. Comment Generation

Comment generation uses a retrieved article as an input, and the model outputs a comment. As we discussed in Sec.4.3, many comments contained facts e.g., summary of an article, and subjective opinions. Thus, we use BART (Lewis et al., 2020) as a language generator, because it is known useful for summarization. Although the data used for training contains subjective comments, BART itself does not have specific mechanism to output subjective comments. From the view of language generation studies, we acknowledge the importance of taking account such a mechanism and discourse structures of output texts (Ishigaki et al., 2019). In this paper, we focus on reporting the performances obtained from BART as a preliminary results of our settings, and we leave studies that investigate how we can better integrate such mechanisms as one of important future directions. We use a Japanese version of the BART, which was pre-trained on the Japanese Wikipedia. We further fine-tune it on pairs of a retrieved article and its comment in our dataset. Cross-entropy loss is used to finetune this model.

6. Experiments

6.1. Data and Parameters

This section describes the data and the parameter settings used in our experiments. We also describe the compared methods and evaluation metrics for two tasks.

6.1.1. Document Retrieval

To train the supervised model, 320 articles published in 2020 were used as positive examples. For negative examples, we used 8,000 articles published in the same year. These were randomly selected from Asahi Shimbun's corpus. For the training and evaluation of the binary classifier, five-fold cross-validation was conducted. We split the total of 8,320 articles into 3:1:1. The first part was used for training, and the latter two were used for threshold setting and evaluation. We iterated this by changing the combination of the split. As a prepossessing, the articles were tokenized using the Japanese morphological analyzer MeCab (Kudo et al., 2004). The IPA dictionary was used as the dictionary. Furthermore, we applied sentencepiece (Kudo and Richardson, 2018) to separate tokens into subwords. We used a pretrained BERT on the Japanese Wikipedia [4]

6.1.2. Comment Generation

We used the pre-trained BART on the Japanese Wikipedia [5]. The fairseq implementation was used to fine-tune the BART, and we followed the example parameters used for automatic summarization [6].

6.2. Evaluation Metrics

For document retrieval, we used precision, recall, and F-measure as the direct metrics to evaluate the binary classifier. We also used Precision@N, which is frequently used in information retrieval systems' evaluation (Blair and Maron, 1985).

For comment generation, we used ROUGE-1, ROUGE-2, and ROUGE-N (Lin, 2004), which are common evaluation metrics for automatic summarization. This is because we observed that many comments include summaries, as explained in Section 4. As a comparison model, we used the Lead-3 method, which is a well-known baseline for automatic summarization, to output the first three sentences of the input article. To evaluate whether output comments are correct as horizon scanning outputs, human evaluators judged whether output comments enabled evaluators to imagine future events. We call this novel criterion as *Implication of Future Changes*. Human evaluators also assessed the output comments regarding two commonly used metrics: *Fluency* and *Correctness*. Fluency evaluates whether a comment is grammatically correct. Correctness evaluates whether the generated comment is in contradiction with the fact written in the input article. Evaluators ranked three

[4] https://github.com/cl-tohoku/bert-Japanese.

[5] https://github.com/utanaka2000/fairseq/blob/japanese_bart_pretrained_model/JAPANESE_BART_README.md

[6] https://github.com/pytorch/fairseq/blob/main/examples/bart/README.summarization.md

	Prec.	Rec.	F-measure	P@5	P@10	P@30	P@50	P@100
Neural Network-based								
BERT+finetune (Supervised)	**58.6**	**57.8**	**58.1**	**80.0**	**80.0**	**76.7**	**76.0**	**70.0**
Distance-based (Unsupervised)	5.71	65.6	10.5	40.0	20.0	6.7	4.0	3.0
Heuristics-based								
IDF-based	8.60	33.4	13.7	40.0	20.0	16.7	16.0	15.0

Table 3: Performance of document retrieval models.

	ROUGE-1	ROUGE-2	ROUGE-L
BART	**34.85**	7.52	**21.58**
Lead-3	30.05	**8.15**	19.07

Table 4: Performance of the Comment Generation Model.

	BERT	IDF	Dist.	Articles
1.	0.02	3.75	0.65	*In 1981, I graduated From the University of, Tokyo. The theme of my graduation thesis was "On the ink paintings Hasegawa Tohaku."*
2.	0.82	3.42	0.33	*On April 29, Toshiba announced that developing LSIs, that serve as the brains of electronic devices will be withdrew from their business.*

Table 5: Scores given from different models' articles are translated from Japanese.

	Gold	BART	Lead
Implication of Future Changes			
Gold	-	72	77
BART	29	-	51
Lead	29	41	-
Fluency			
Gold	-	69	37
BART	29	-	33
Lead	46	65	-
Correctness			
Gold	-	49	16
BART	36	-	18
Lead	92	90	-

Table 6: A comparison of Human, BART, and Lead-3 methods by manual evaluation. The numbers indicate the number of times the method presented in a row is better judged.

comments; Gold, BART and Lead-3. They are allowed to judge the tie.

7. Result

In this section, we report the performance of the document retrieval and comment generation tasks.

7.1. Results of Document Retrieval

The results of the document-retrieval task are listed in Table 3. The BERT-based supervised model achieves an F-value of 58.1 for the binary classification, which outperformed other methods. It is inherently difficult to obtain high precision for this classification problem because the balance of positive and negative examples is highly skewed where only 3.8% of the instances are positive. Therefore, the distance-based and IDF-based methods exhibited extremely low performances. Although we show that the semantic distance between the articles to be retrieved and general articles is distant, as described in Section 4, the model that solely uses this information is not optimally functional. This also applied to the model that used only IDF values. The BERT-based model achieves a precision@5 of 0.80, which is more optimal than the 0.40 of the distance-based and IDF-based methods. This shows that the proposed dataset is useful to train supervised models. For distance-based and IDF-based methods, as N in Precision@N increases, Precision@N significantly decreases to 0.03 and 0.09, respectively. Conversely, the BERT-based model achieves a high Precision@100 of 0.7. The supervised models trained on our dataset show a relatively robust retrieval performance when N increases.

7.2. Results of Comment Generation

Table 4 shows the ROUGE scores obtained by BART and Lead-3. The BART method showed higher values for ROUGE-1 and ROUGE-L than Lead-3. ROUGE-2 showed a lower value than Lead-3. This may be because Lead-3 is an extractive summarization method, which copies sentences to the output. When the correct comments are extractive summaries or summaries that are generated by slight paraphrasing. Here, word overlaps between the input and output comments become very large; thus, ROUGE-2 becomes high.

As mentioned in Section 4, in many cases, the gold comments first mention facts through a "summary" or "background,", and subjective opinions follow. Therefore, BART also learns such discourse structures. We observed many generated comments with such a struc-

1. Input Article:
Japan Railway East announced that it would start a demonstration experiment for an automatic mobile disinfection robot at Takanawa gateway station. The robot will be operated at night to verify its effectiveness. Three robots will be introduced (truncated.).
Gold Comment:
Railroad companies cannot secure human resources for various tasks required to maintain business. They try to automate them. Automatic ticket gates and payment machines are common products of this trend, however, it was finally applied to moving objects. I believe that in the future, it will be possible to apply this technology to various aspects of society
Generated Comment:
An automatic mobile disinfection robot equipped with sprays is introduced. This technology may be applied for disinfection and cleaning in the foodservice industry.
2. Input Article:
Grilas, a start-up company from the University of Tokushima, raised a total of 230 million yen to expand its edible cricket business to set up fully automated production equipment for edible crickets. The company plans on using this financing as an opportunity to develop a structure to go public. In May, Grilas formed a business alliance with JTEKT to jointly develop a fully automated production system
Gold Comment:
Recently, financiers made from 10 and 30 crickets were sold in Japan, and they were sold out in two days. I have never seen any food products made with crickets except for rice crackers, but it seems that there are signs that insect-eating is finally making its way into society
Generated Comment:
Grilas, a start-up company from the University of Tokushima, has raised a total of 230 million yen to expand its edible cricket business. The company has raised a total of 230 million yen from venture capitalists to expand its edible cricket business. The company is working on the joint development of a fully automatic production system
3. Input Article:
Two suspects were arrested by the Metropolitan Police Department for vandalizing vending machines. According to an investigation by the Mitaka Police Department,. they used an electric drill to open the keyhole in an old vending machine. (truncated) .
Gold Comment:
ATMs and vending machines are in danger. In Tokyo and Kanagawa, there were more than 100 incidents per year. With such an increase in [the] damage, there is [is there] no final need for a full scale computerization of vending machines that handle money?
Generated Comment:
Two suspects were arrested during the COVID-19 pandemic. The Japanese food boom has become very popular around the world. Japanese food culture has the potential to spread to the rest of the world.

Table 7: Examples of Generated Comments.

ture.

7.3. Manual Evaluation by Humans

The results of the human evaluation are presented in Table 6. Each number in the table shows the times that a method on the row is judged better than one on the column. We show the numbers in terms of different criteria; 1) Fluency, 2) Correctness and 3) Implication of Future Changes. From the numbers in terms of Implication of Future Changes, we observe that the comments generated by BART were 51 times judged better than Lead, which is 41 times judged better than ones generated by BART. This implies that automatically generated comments have more implications of future changes, which are important for the comment generation task. Gold comments are judged 72 times better than BART, which shows room for improvement still exists on this task. However, in terms of correctness, BART is only 18 times better than Lead, which is 90 times better than BART. This suggests the need for mechanisms to improve correctness.

8. Error Analysis

In this section, we analyze automatically retrieved articles and generated comments.

8.1. Outputs of Document Retrieval

The three sample scores calculated by BERT, Distance-based, and IDF-based methods are shown in Table 5. We show two different input articles and their scores. Note that both articles are negative instances not to be retrieved.

BERT-based model's Errors The BERT-based model correctly provided low scores in the first example.

However, we observed some instances where BERT-based approaches failed to determine the articles to be retrieved, as shown in the second example. For this example, the BERT-based model gave a high score of 0.82, even though the article was not retrieved. This article included characteristic words, such as "electronics," "semiconductor," and "development" development, which also appeared in the articles to be extracted. The BERT-based model may overlook the negative expression "withdrew" in the input. We observed several similar errors where the negative expressions were not correctly recognized.

Distance-based model's Errors We observed many cases where the distance alone could not distinguish between articles to be retrieved and those not retrieved. For example, the first example in Table 5 is rare in the Asahi Shimbun corpus. This is because it is an essay rather than a normal news article. Thus, the distance to the general articles was far, which may be considered an example that the distance alone could not capture the information that implied drastic changes in the future.

8.2. Outputs of Comment Generation

Finally, we discuss the outputs of the comment generation task. We show three examples of an input article, gold, and the generated comments in Table 7. The generated comment for the first input correctly includes a summary and future expectations: "*This technology may be applied to disinfection and cleaning in the food-service industry*". The second and third examples are erroneous.

The generated comment in the second example was generated by extracting sentences from the input article. It lacked subjective opinions and prospects. Even if the generated text was fluent, it could be considered an error in our setting. In the future, mechanisms that enforce the inclusion of subjective opinions or expectations will be necessary to generate better comments.

In the third example, the second sentence in the generated comment mentions "the Japanese food boom." However, there is no relevant information in the source article. This may be considered as an error regarding correctness. It appears difficult to include relevant "background" information on the input article.

9. Conclusion

In this paper, we proposed document retrieval and comment-generation tasks intended for automating horizon scanning in futurology. We created a dataset with 2,266 manually retrieved documents and comments aligned to articles. We analyzed the retrieved articles and applied several models for retrieval and comment-generation tasks to this dataset. We found that 1) retrieved documents had different characteristics from general news articles and 2) a supervised model performs better for the retrieval task regarding Precision@N. This demonstrates the dataset's effectiveness. In the future, we will extend the comment-generation model to include more subjective comments.

Acknowledgments

We thank the reviewers for their careful reading and many suggestions. This work was supported by JST SPRING, Grant Number JPMJSP2102. We thank students and experts at Hitotsubashi University who created scanning material used in our experiments. We also appreciate human evaluators. Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used. Asahi Shimbun corpus was used in our experiments.

10. Bibliographical References

Blair, D. C. and Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3):289–299, mar.

Cheng, J. and Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494, Berlin, Germany, August. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL2019)*, pages 4171–4186, Minnesota, USA. Association for Computational Linguistics.

Ishigaki, T., Kamigaito, H., Takamura, H., and Okumura, M. (2019). Discourse-aware hierarchical attention network for extractive single-document summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 497–506, Varna, Bulgaria, September. INCOMA Ltd.

Ishigaki, T., Topic, G., Hamazono, Y., Noji, H., Kobayashi, I., Miyao, Y., and Takamura, H. (2021). Generating racing game commentary from vision, language, and structured data. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK, August. Association for Computational Linguistics.

Iso, H., Uehara, Y., Ishigaki, T., Noji, H., Aramaki, E., Kobayashi, I., Miyao, Y., Okazaki, N., and Takamura, H. (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.

Kim, B. J. and Choi, Y. S. (2020). Automatic baseball commentary generation using deep learning. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 1056–1065, New York, NY, USA. Association for Computing Machinery.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Kunifuji, S. (2016). A japanese problem-solving approach: The kj ho method. In *Knowledge, information and creativity support systems: Recent trends, advances and solutions*, pages 165–170. Springer.

Kwon, J., Kamigaito, H., Song, Y.-I., and Okumura, M. (2020). Hierarchical trivia fact extraction from Wikipedia articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4825–4834, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 7871–7880, Online. Association for Computational Linguistics.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Mitra, B., Craswell, N., et al. (2018). *An introduction to neural information retrieval*. Now Foundations and Trends.

Murakami, S., Tanaka, S., Hangyo, M., Kamigaito, H., Funakoshi, K., Takamura, H., and Okumura, M. (2021). Generating weather comments from meteorological simulations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1462–1473, Online, April. Association for Computational Linguistics.

Nagai, Y., Taura, T., and Mukai, F. (2009). Concept blending and dissimilarity: factors for creative concept generation process. *Design Studies*, 30(6):648–675.

Nallapati, R., Zhou, B., dos Santos, C., Gulçehre, Ç., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Palomino, M., Bardsley, S., Bown, K., De Lurio, J., Ellwood, P., Holland Smith, D., Huggins, B., Vincenti, A., Woodroof, H., and Owen, R. (2012). Web-based horizon scanning: Concepts and practice. *Foresight*, 14:355–373, 08.

Puduppully, R. and Lapata, M. (2021). Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September. Association for Computational Linguistics.

Sardar, Z. (2010). The namesake: Futures; futures studies; futurology; futuristic; foresight—what's in a name? *Futures*, 42(3):177–184.

Saritas, O. (2013). Systemic foresight methodology. In *Science, technology and innovation policy for the future*, pages 83–117. Springer.

Tanaka-Ishii, K., Hasida, K., and Noda, I. (1998). Reactive content selection in the generation of real-time soccer commentary. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Taniguchi, Y., Feng, Y., Takamura, H., and Okumura, M. (2019). Generating live soccer-match commentary from play data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7096–7103, Jul.

Washida, Y. and Yahata, A. (2021). Predictive value of horizon scanning for future scenarios. *Foresight*, 23(1):17–32.