Approach of a Word2Vec Based Tourist Spot Collection Method Considering COVID-19

Yuki Nagai\textsuperscript{1(✉)}, Nobuki Saito\textsuperscript{1}, Aoto Hirata\textsuperscript{2}, Tetsuya Oda\textsuperscript{1}, Masaharu Hirota\textsuperscript{3}, and Kengo Katayama\textsuperscript{1}

\textsuperscript{1} Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
\{t18j057ny,t17j033sn\}@ous.jp, \{oda,katayama\}@ice.ous.ac.jp
\textsuperscript{2} Engineering Project Course, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
\textsuperscript{3} Department of Information Science, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700-0005, Japan
hirota@mis.ous.ac.jp

Abstract. The Novel Coronavirus disease 2019 (COVID-19) is raging around the world and is seriously affecting daily lives and economic activities for people. For example, the work and class are now taking place online, and major lifestyle changes are also taking place. It is possible that COVID-19 will continue to change lifestyle habits in the future. To cope with this change, COVID-19 needs to be taken into account in daily life and in economic activities, and measures need to be taken based on a variety of information. A variety of approaches are currently being tried and tested around the world to assist in combating COVID-19. This paper focuses on measures taken in the tourism industry and aims to propose a tourist spot collection method that takes into account COVID-19.

1 Introduction

In this year, the Novel Coronavirus disease (COVID-19) is spreading around the world [1–3]. This has had a serious impact on daily life and economic activities, requiring people to refrain from going out of the house, and has led to major lifestyle changes. The refrain from going out has also had an economic impact on restaurants, hotels and tourist attractions.

On the other hand, since the economy cannot be stopped for long periods of time, daily life and economic activities are being resumed in areas where the number of people infected is declining, taking into account COVID-19. The academic associations also support COVID-19 infection prevention. For example, the Association for the Advancement of Artificial Intelligence (AAAI) provides a dataset on COVID-19 that summarizes the background and symptoms.
of infection in infected individuals to help in the COVID-19 infection prevention. In Japan, the “Go To Campaign” has been running since July 22, 2020 in all regions except Tokyo, where the number of people infected with the disease is severe. This is a project to attract people to restaurants and tourist sites that had been economically affected by the refraining of outdoor activities, with some limited government subsidies for food and travel expenses.

However, the risk of infection COVID-19 has not been completely eliminated. Therefore, tourism must be conducted while taking into account the impact of COVID-19. In this paper, we propose a Word2Vec based tourist spot collection method considering COVID-19 infection prevention.

2 Proposed System

In this section, the processing of the proposed method is described. The process in this paper is divided into four major parts: collection of text data and training of Word2Vec, extraction of similarity from the distributed representation of training words, and extraction of tourist attractions from the word vectors of similarity (Table 1).
### Table 1. Experimental parameters.

| Parameters          | Value |
|---------------------|-------|
| Depth of search     | 3     |
| Waiting time [s]    | 10    |
| Minimum length of string | 1   |
| Maximum length of strings | 10000 |

#### 2.1 Collection of Textual Data

In this section, we describes the data collection system and the data to be collected.

The structure of the COVID-19 related textual data collection system is shown in Fig. 1. The crawler in proposed system load the HyperText Markup Language (HTML) in the website by crawling the website and scraping it to collect the Uniform Resource Locator (URL), strings and the date and time the strings were written. The extracted strings are converted into words using MeCab of a morphological analysis system. Then, the collected data are then classified by the indexer system into URL, strings and words. Also, the string and word data are accompanied by the date and time they were posted.

The indexer in proposed system assign identification numbers to collected URL, words, and strings, and recognizes them to determine where the data is stored and to prevent data duplication. The structure for storing data is based on the Resource Description Framework (RDF) in Linked Open Data (LOD) [6–10], and each URL, word, and string in the same site has the same data structure. This allows for efficient retrieval of interlinked data.

In proposed system, the electronic bulletin board “Open2ch” [12] is the target of the textual data collection. The period of time covered was 6 month period from February to July, when COVID-19 begins to spread, and all COVID-19 related posts posted during this period were considered. The following is an example of RDF in LOD.

```turtle
@prefix bp: <https://corona.go.jp/>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema>.
@prefix geo: <http://www.w3.org/>.
<1> rdfs:label "XXXXXX"@ja;
bp:title "COVID-19 Information and Resources";
bp:url "http://XXXXX.ac.jp";
bp:words "Coronavirus Disease (COVID-19)";
bp:time "1 February, 2020 at 00:00";
```
The indexer in proposed system sends the URL to be searched to the crawler system in the order in which they are found. The proposed system collects data by repeating these processes. The data collected here will be sent to Word2Vec as the training data.

Table 2. Training parameters.

| Parameter       | Value       |
|-----------------|-------------|
| Number of words | 1257401     |
| Number of training | 139270    |
| Window size     | 5           |
| Batch size      | 100         |
| Node of middle class | 100        |
| Initial weight  | Random      |
2.2 Learning Textual Data Based on Word2Vec

In this section, we describe learning textual data based on Word2Vec. The proposed system is shown in Fig. 2. We use Word2Vec to obtain a distributed representation from the collected text data, and the training model is the CBOW model [11]. In addition to data collected from the Open2ch data is used for training. The textual data used in open2ch, writings posted between February and August 9, which is used for the training. The data are combined and trained in Word2Vec as one training data. In order to improve the accuracy of word predictions, if the submitted text contains symbols, the symbols and emojis are removed in advance. The training parameters for Word2Vec are shown in Table 2.

2.3 Recommendation Method for Tourist Spot Considering COVID-19 Using Word2Vec

In this section, we will discuss how to extract tourist spot. The process of recommending a tourist destination is shown in Fig. 3. Determine the cosine similarity on the variance representation from the input words. We prepared a database of tourist attractions in advance, and by referring to a database of tourist attractions from a list of words with high cosine similarity, we finally extracted only the tourist attractions. We also obtain the cosine similarity of COVID-19 related words from the variance representation by entering the COVID-19 related words. The extracted word vectors of tourist spots minus the word vectors of related words in COVID-19 are recommended as tourist attractions with low relevance to COVID-19.
3 Case Study

In this section, we present a case study of the proposed system.

3.1 Word2Vec Learning Results

We describe the results of the study of Word2Vec. Firstly, crawling of COVID-19 related pages of Open2ch was performed by the proposed system. As a result, we scraped 615 pages and collected 206051 lines of text. Based on this string, we generated training data for 1257401 words and used this data to train Word2Vec. Next, the training process of Word2Vec is shown in Fig. 4 and the post-training distributed representation is shown in Fig. 5. Figure 4 shows the cross-entropy error for each training session, and it can be seen that the error decreases as the training progresses and Word2Vec is learning. The study was repeated 139270 times in a sequence of forward to reverse propagation and took 348 minutes to complete. The variance representation in Fig. 5 is represented using the weights of the input layer of Word2Vec after training, with each dot representing a word. From the coordinates of the words in this distributed representation, we calculate the cosine similarity between the words.
3.2 Experimental Results

We describe the experimental results of the proposed system. In this paper, “Travel” was used to extract tourist spots and “Corona Virus” was used as
Table 3. Tourist spots and COVID-19 related words.

| Travel       | Corona          |
|--------------|-----------------|
| Gokayama     | Virus           |
| Yakushima    | New model       |
| Tushima      | Pollen          |
| Sado         | Pneumonia       |
| Higashihongannzi |               |
| Kansai       | Bacteria        |
| Kinpu-san    | Each country    |
| Minamiizu    | China           |
| Yunoyamaonnsen | Prefectures    |
| Tochigi      | Infection       |
| Yokohama     | Confusion       |
| Shirakawago  | Mutation        |
| Osaka-Jo     | Gene            |
| Tohoku      | Stock price     |
| Kyushu       | Company         |
| San’in       | Trend           |
| Kusiro       | GoToTravel      |
| Makuhari Messe | Treatment      |
| Shiretoko    | Prevalence      |
| Kawai        | Kanto           |
| Kisarazu     | Hospital        |
| Wakasa       | Proportion      |
| Ichihino     | Vaccine         |
| Hiroshima    | Splash          |
| Takehu       | Inspection      |
| Niseko       | Japan           |
| Kabuki-za    | High fever      |
| Ichihuzi     | Hokkaido        |

a related term of COVID-19. A list of tourist spots extracted from “Travel” and examples of words extracted from “Corona Virus” are shown in Table 3. From “Travel” it was possible to extract 28 tourist spots, which included the names of tourist spots and the place names of tourist spots. Examples of words extracted from “Corona Virus” are randomly selected from the extracted words. The result of subtracting the words extracted from “Corona” from the list of tourist spots is shown in Fig. 6. The graph shows the degree of similarity between the tourist sites and “Corona Virus”. In this experiment, we were able to extract 16 places. The 16 tourist spots extracted, 3 were tourist spots and 16 were place
names of tourist spots. The maximum similarity to “Corona Virus” was 73.15 for “infection” and the average similarity was 22.98, so these 16 tourist sites were extracted as having a low association with “Corona Virus”.

4 Conclusion

In this paper, we proposed a of Word2Vec based tourist spot collection method considering COVID-19. The proposed system allowed us to extract 16 tourist sites in this paper. Future goals are to increase the amount of data we use for training so that we can increase the number of places we can extract, and to learn the coordinates of places so that we can make recommendations that take into account the location of tourist spot.

Acknowledgement. This work was supported by Grant for Promotion of Okayama University of Science (OUS) Research Project (OUS-RP-20-3).

References

1. Mehta, P., et al.: COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet 395(10229), 1033–1034 (2020)
2. Bai, Y., et al.: Presumed asymptomatic carrier transmission of COVID-19. J. Am. Med. Assoc. (JAMA) 323(14), 1406–1407 (2020)
3. Fauci, A., et al.: COVID-19 - navigating the uncharted. New Engl. J. Med. 382(13), 1268–1269 (2020)
4. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS-2013), vol. 2, pp. 3111–3119 (2013)
5. Kudo, T.: MeCab: yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab/. Accessed 20 Aug 2020
6. Fernandez, J., et al.: Compact representation of large RDF data sets for publishing and exchange. In: Proceedings of The 9th International Semantic Web Conference (ISWC-2010), pp. 193–208 (2010)
7. Ermilov, I., et al.: Linked open data statistics: collection and exploitation. In: Proceedings of The 4th International Conference on Knowledge Engineering and the Semantic Web (KESW-2013), pp. 242–249 (2013)
8. Ristoski, P., et al.: Mining the web of linked data with rapidminer. J. Web Semant. 35(3), 142–151 (2015)
9. Kamdar, M., et al.: Enabling web-scale data integration in biomedicine through linked open data. npj Digit. Med. 2(90), 1–14 (2019)
10. Maki, T., et al.: Resource propagation algorithm considering predicates to complement knowledge bases in linked data. Int. J. Space-Based Situat. Comput. 8(2), 115–121 (2018)
11. Goldberg, Y., Levy, O.: Word2Vec explained: deriving Mikolov et. al.’s negative-sampling word-embedding method. arXiv:1402.3722, pp. 1–5 (2014)
12. Open2ch. https://open2ch.net/. Accessed 20 Aug 2020