A study on suggestion of predicates using the clustering of item names for making RDF of Open Data

 Bo Chen1, Tadanori Hisanaga, Daiki Tomari, Takayasu Fuchida2

Graduate School of Science and Engineering,

Kagoshima University

1chinhaku204@gmail.com,2fuchida@ibe.kagoshima-u.ac.jp

**Abstract:** In recent years, there has been increasing interest in Open Data. The utilization of Open Data has been promoted, and many organizations including the national government, local governments and other organizations are working on publishing and utilizing Open Data. Open data of Kagoshima City has been published in July 2016 as CSV format.CSV format is third star in the 5-star deployment scheme for Open Data and it's not suitable for machine reading. In this research, we focused on the vocabulary that corresponds to the predicate of RDF form for Open Data, and proposed a method to suggest predicates by the clustering of item names for RDF of Open Data. Finally, we present results of our approach, and show that we can learn to predict predicates using the clustering of item names.

**Keywords:** Open Data, RDF, Word2Vec, Namespace

# 1 INTRODUCTION

In recent years, there has been increasing interest in Open Data. Open Data is defined by the Ministry of Internal Affairs and Communications(MIC) as follows: "Open Data is data formats that are suitable for machine reading and data that have been released on the use of rules that allows secondary use, which allows secondary use without requiring a lot of manpower"[1].

The utilization of Open Data has been promoted, and many organizations including the national government, local governments and other organizations are working on publishing and utilizing Open Data. Open data of Kagoshima City has been published in July 2016 as CSV format.CSV format is third star in the 5-star deployment scheme for Open Data and it's not suitable for machine reading. It is a fact that the local governments possess large amounts of public data released as Open Data without cooperation between many ministries and local governments. Namespaces have been proposed in various ways so far, and the classes and properties defined in those namespaces are used as predicates. However, it is very difficult for citizens and private companies to actually understand such namespace and to create RDF using predicates. Therefore, it is necessary to take the measure that can use Open Data as CSV format. If many disclosed data can be easily converted to RDF format, mechanical linkage of multiple Open Data becomes possible and it becomes possible to promote utilization of Open Data.

In this research, we focused on the vocabulary that corresponds to the predicate of RDF form for Open Data, and proposed a method to suggest predicates by the clustering of item names for RDF of Open Data. Finally, we present results of our approach, and show that we can learn to predict predicates using the clustering of item names.

# 2 OPEN DATA IN PRESENT SITUATION

## 2.1 Five stars for Open Data

A 5-star(in Fig 1) deployment scheme for Open Data was suggested by Tim Berners-Lee, the inventor of the Web and Linked Data initiator, as a way of rating the usefulness of openly-released data[2]. It has became easier to read by machine from one to five star.

●One-Star(★): make your stuff available on the Web (whatever format) under an open license

●Two-Star(★★): make it available as structured data (e.g., Excel instead of image scan of a table)

●Three-Star(★★★): make it available in a non-proprietary open format (e.g., CSV instead of Excel)

●Four-Star(★★★★): use URIs to denote things, so that people can point at your stuff

●Five-Star(★★★★★): link your data to other data to provide context

**Fig.** **1.** Five stars deployment scheme for Open Data

**Fig.** **2.** Top page of DATA.GO.JP

## 2.2 Open Data in government

The Open Government Data Strategy was published in 2012 by The IT Strategic Headquarters. The four basic principles is defined as follows: “ (1) The government itself should actively disclose public data (2) Data should be published in machine-readable formats (3) Promote utilization for regardless of commercial and non-commercial use (4) Specific efforts such as publication should be implemented from the possible public data first, to steadily accumulate results”. And “G8 Open Data Charter” [3] was presented in 2013, efforts of Open Data has begun in each country.

In order to promote the utilization of Open Data, Japanese government published the data catalog of possible secondary use, including the data for commercial purpose of the data format suitable for machine reading. Fig.2 shows this site named “DATA.GO.JP”[4].

On this site, 25,737 datasets were disclosed in December,2019. A lot of the data formats which in order of PDF, html, xls, and csv in the same way as local governments, are not suitable for machine reading. This situation is in order to register a lot of data, but on this site, the acquisition of the flow metadata(csv and json format) in RDF format is posted as information for developers.

Open data of Kagoshima City was opened in July,2016 as CSV format[5]. However, it is the fact that local governments possess huge amounts of public data released as Open Data without cooperation of each data.

# 3 FLOW OF THE PROPOSED METHOD

In this study, we focused on the vocabulary that corresponded to the predicate of the RDF form for Open Data and we proposed a method to learn word vectors by utilizing the neural network called Keras to suggest predicates.

## 3.1 About Word2Vec

Word2vec[6] is a method for natural language processing that calculates similarity degrees among words in multidimensional vector spaces. Its applications have recently attracted a great deal of attention from the machine learning community. Word2Vec will produce high-dimensional space(vector space) and map many words to the vector space When a large corpus of text is inputted. Word vectors are placed in the vector space in which the words sharing common contexts in the corpus are located very close together.

Fig.3 shows two models of Word2Vec, and Skip-gram model. CBOW predicts the current word based on the context, while Skip-gram predicts surrounding words given the current word.Fig.4 shows composite vector of Words in words space by Word2Vec.

**Fig. 3.** Model of Wor2Vec

**Fig. 4.** Composite vectors of Words in words space

## 3.2. Flow of the method

In order to learn for suggestion of predicates by Keras, we propose a method. This method consists of five steps as

**Fig. 5.** Data flow chart

**Fig. 6.** Process flow chart of algorithm

follows. Fig 5 shows data flow chart and Fig 6 shows flow chart of algorithm.

Step1:Get the csv file from Open Data of local governments

that have already been published.

Step2:Extract item names and item data, then separate words in Japanese with spaces by the word segmentation from the csv file. The word segmentation is called Janome.

Step3:Create vectors of item names and item data with Word2Vec,and then composite vectors of Words in words space as Fig 4 shows.

Step4:Perform hierarchical clustering by vectors of item

names. Fig 7 shows the dendrogram using average's method

**Fig. 7.** Dendrogram using average's method of the cluster

**Table.1** Item name’s cluster by hierarchical clustering

of the hierarchical cluster by cosine distance. Table.1 shows a part of item name’s cluster when the number of cluster is

50.

Step5:Use Keras to learn for suggestion of predicates. Keras is a high-level neural networks API and evaluating deep learning models. Input vectors of item data separated in input layer and the clustering of item names as teacher signal in output layer.

The total number of the pair(dataset) of item name and item data which we extracted in the step2 from the csv file is 1,245,709.

# 4 RESULTS AND DISCUSSION

We used the dataset, which contains 100,000 vectors of item data with 50 different classes. Keras splits it in a training set with 80,000 instances and a testing set with 20,000 instances.

We have learned 100 times by using dataset’s vectors of 50 dimensions. Fig 8 shows accuracy when we learned by using the item name’s cluster and Fig 9 shows loss when we learned. The vertical axis represents accuracy or loss. The horizontal axis represents epochs which is the number of times we train neural network with the dataset.

Table.2 shows a part of results. From these results, predicting predicate depends on the item name’s cluster by using hierarchical clustering. Even if the item data are different ,the results of predicted cluster will be the same. So it comes to the questions that how we can find the best

**Fig. 8.** Accuracy by using the clusters of the item names

**Fig. 9.** Loss by using the clusters of the item names

**Table.2** Results of predict cluster

item name’s cluster by hierarchical clustering and what is the most appropriate number of cluster.

In addition, we have learned under the same condition without using the item name’s cluster as Fig 10 and Fig 11 show, and then compared the accuracy of Fig 8 with that of Fig 10, we know that the accuracy when using the item name’s cluster is higher than that without using it.

# 5 CONCLUSION

This research focused on the vocabulary that

**Fig. 10.** Accuracy by using the item names without cluster

**Fig. 11.** Loss by using the item names without cluster

corresponds to the predicate of the RDF form for Open Data, we proposed a method to learn word vectors by the

neural network mechanism called Deep Learning in order to learn for suggestion of predicates by utilizing existing Open Data, and confirmed the result of the learning. For further study, we will improve the accuracy of the learning method by Keras. we will convert to RDF format by utilizing of the words suggested.

# REFERENCES

[1] Ministry of Internal Affairs and Communications (M

IC) site, http://www.soumu.go.jp

[2] Tim Berners-Lee, the inventor of the web and Linked

Data initiator, suggested a 5-star deployment scheme for Open Data, <http://5stardata.info/en/>

[3]Open Data Charter site,

https://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf

[4]DATA.GO.JP site, http://www.data.go.jp/

[5] Open data of Kagoshima side,

https://www.city.kagoshima.lg.jp/jousys/opendata.html

[6]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013