Categorization of Web News Documents Using Word2Vec and Deep Learning

Ryoma Kato/Hosei University
Department of Systems Engineering
Tokyo, Japan
ryoma.kato.ra@stu.hosei.ac.jp

Hiroyuki Goto/Hosei University
Department of Industrial & System Engineering
Tokyo, Japan
goto-h@hosei.ac.jp

***Abstract*— In this research, we examine if Word2Vec can be used as an input for deep learning in categorizing web news. Since each news site has its own categorization policy, we have to search target news at some categories. If we can retrieve target news according to our own categorization policy, it would not be necessary to look for multiple categories that contain target news. We thus categorize web news in this research by machine learning. For the analysis, we use Japanese text data delivered by Japanese web sites. Bag-of-words is a method of vectorial representation of word and it is often used as an input for text classification. Although the method is good for categorization because of higher accuracy, it has a problem in computational complexity in using a neural network. Since it is desirable to reduce the dimension of input layer to resolve its problem, we propose to use Word2Vec for an input to reduce the dimension. Moreover, we examine the accuracy using the same input. Through the experiment, we found that it is practical to express words using Word2Vec as an input of deep learning for categorization document of web news.**

Keywords— text classification, Word2Vec, deep learning, neural network, Web news, unsupervised learning

I. INTRODUCTION

In recent years, the number of Internet users has been increasing by the prevalence of smartphones as well as the improvement of Internet environments, whereby the amount of data on the web has been increased. In particular, the amount of text data is increasing by digitalization of various texts and prevalence of social networking service, and it is supposed to continue for the time being. Therefore, researches analyzing text data automatically is sought. Such researches are called text mining, and its purposes are a finding of unknown features, supporting user that use text data and so on.

In this article, we research by focusing on support in handling the human text data. Specifically, the supports in handling the human text data mean summarizing long text to short text for users to quickly understand the contents of the sentence, giving attributes to text for users search to quickly search objective news and so on. In particular, researches of document summarization are sought since the amount of text data is increasing and we cannot read all contents of text data. However, it is regard as a research of summarizing text is low accuracy yet. We consider that improving the accuracy of the text summary needs to know objective text have what attributes. However, researches of giving attributes to text have some problems yet. Therefore, we aim to solve some problem in giving attributes to text. Moreover, we focus on attribute of the category of the text.

Automatic text classification is researched in various text data and various machine learning techniques. Techniques of general machine learning algorithms of classification such as naive bayes[1] and support vector machine[2] have been successful. However, these techniques are used only supervised learning of the techniques of machine learning, it is not possible to handle only supervised data. Moreover a lot of actual texts in web are not supervised data.

Clustering is known as the classification machine learning algorithm to handle unsupervised data. However the algorithm defines category automatically and we want to classify text to category we decided in this research. Therefore clustering is not considered in this research.

Moreover, there are several problems in handling text data in machine learning. One of them, we need to quantify texts and transform texts into constant dimensions. Currently, there is a bag-of-words to the method that is often used to quantify.

Bag-of-words is a method to consider the text as a set of words. If the word is included in the texts, the value at the word of the texts become 1, otherwise it becomes 0. Therefore, the numbers of dimensions using bag-of-words equal the number of vocabularies in all texts. However it is too large size to learn in machine learning such as neural network using common personal computer if size of texts are large.

In this research, we have aimed at solving the above problems. We resolve problem that we cannot handle unsupervised data in machine learning by using pre-training. We resolve first problem that we cannot handle unsupervised data by using pre-training. Moreover we resolve second problem that bag-of-words is too big to learn in machine learning such as a neural network for quantifying texts by using Word2Vec.

II. RELATED WORK

A. *Text Classification*

Researches of text classification have been done for a long time. It has handled a variety of data and categories, such as emotion classification, categorization of news and so on. However, we explain the only method of text classification, since it does not depend on the data basically.

Almost methods of text classification should prepare large number of supervised text in order to increase the accuracy. However preparing large number of supervised text is too hard to prepare privately. Therefore, we hope to solve the problem that we use large supervised data.

Text classification using small number of supervised data with some contrivances is researched by Lee[3]. In the research, he was measuring the accuracy of the time of performing the learning with a small amount of supervised data.

Moreover, by utilizing the characteristics of such co-occurrence information of words it was aiming also improve accuracy. Supervised data is fully considered, however unsupervised data is not considered at all in the research. We consider that we should not prepare large supervised data if we classify text with using unsupervised text.

Text classification using unsupervised data is researched in some reports introduced by Trinkle[4]. In these methods, we focus on neural network since deep learning produce good results recently in some tasks and deep learning can handle unsupervised data in pre-training.

B. *Word2Vec*

In introduction, we explained bag-of-words is too large dimensions to learn in neural network. Therefore we need to quantify the text in other way and we propose to use Word2Vec in order to quantify the text. Word2Vec is tool invented by Mikolov[5] and it can convert words into distributed vector. Vectors created by Word2Vec are said it can express the words in small dimensions. Moreover it is used for text classification and it gets good results in some researches. For example, Dongwen[6] researched sentiment classification as negative or positive of Chinese comments using Word2Vec and SVM. The accuracy of sentiment classification in the research is over 89%. We consider that the accuracy is very good, despite the large amount of dimensions are deleted. Therefore we decide that Word2Vec is effective method to quantify the text for text classification.

III. PROPOSED METHOD

A. *Overview*

Fig. 1 shows the general framework of proposed method. It is considered that Word2Vec is good tools to quantify the text for text classification. However the research that use deep learning and Word2Vec to handle unsupervised data for text classification do not exist. Therefore, we test and verify whether using deep learning and Word2Vec is applicable to classify text.

We input large unsupervised text into Word2Vec to quantify words. The next, we quantify large unsupervised texts using the vectors of words. Moreover, we use the vectors as input for pre-training in deep learning and cause to learn. We perform the same also and fine-tuning in supervised data.

Also, we need to adjust some parameter in network of deep learning if we hope to improve accuracy. Therefore, we search optimal parameter in this task.

B. *Word2Vec*

Firstly, we should obtain vectors of words to quantify texts. Moreover, we want to decrease number of dimensions of vector since computing time of neural network increase in exponential. Therefore, we use Word2Vec to quantify words with full data set. We should separate texts by a space each words to handle in Word2Vec, moreover Japanese texts are not separated. We use MeCab[7] to separate Japanese texts each words. The next, we input the texts separated by each words to

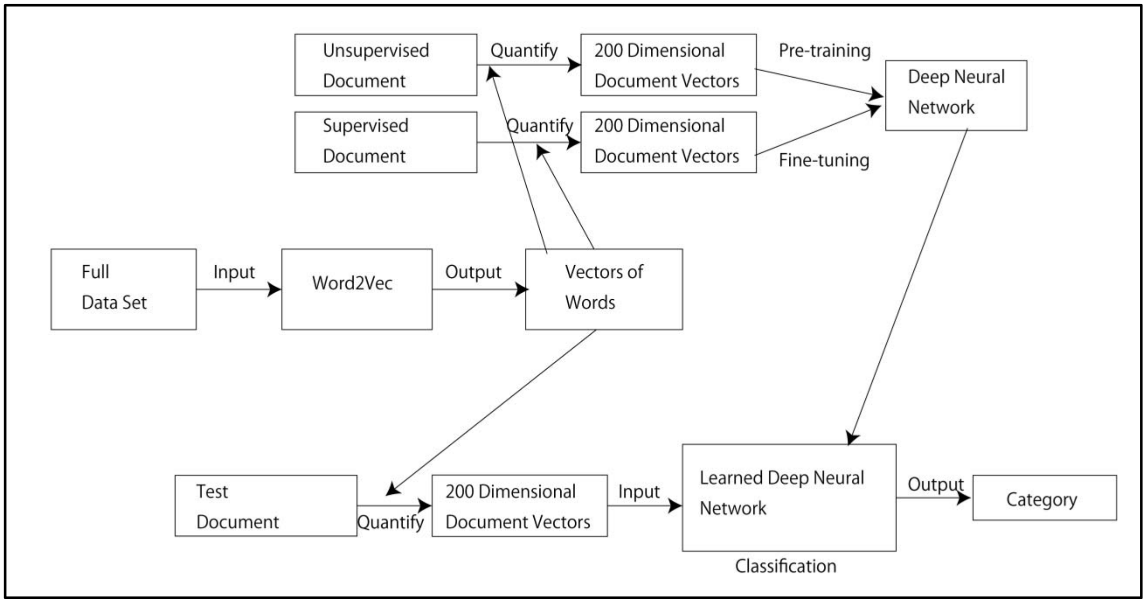

Figure 1 The general framework of proposed method

Word2Vec, and we obtain vectors of words. Also, we define in this way the command of training Word2Vec. "time ./word2vec -train input.txt -output output.bin -size 200 -window 5 -negative 0 -hs 1 -sample 1e-3 -threads 12 -binary 1"

C. Pre-training

We conduct pre-training in neural network to handle unsupervised data. Pre-training is a learning method that conducts learning by using the unsupervised data before supervised learning in deep learning. It is considered that conducting pre-training for neural network can obtain better initial value of weight in network.

In this research, we use the denoising autoencoder in the method of pre-training. Autoencoder is using unsupervised data, it is the learning method to adjust the weights so that the output to reproduce the input in each layer. Denoising autoencoder is a learning method that added the function such as dropout to the function of autoencoder. The reason why we use denoising autoencoder is that we consider adding noise for input can prevent over training. Moreover training algorithm we use is SGD in this research.

D. Fine-tuning

After conducting pre-tuning, we conduct fine-tuning. Fine-tuning is a phase to conduct the learning of the network by using a supervised data after the Pre-training phase. One of the purposes of this research is to reduce the supervised data to use at this stage.

IV. EXPERIMENT

A. Datesets

Full datasets that we use in this research are news data that is published in some web news site such as Yahoo! JAPAN[x], moreover these news data belong to various category. They are unsupervised data that do not have category the news belong. These datasets include the body, the title, publishing site and the date and the time published. We use the body only in this research. The number of news data is 1,728,942 records, and the number of vocabularies in datasets is 218,295. We labeled 800 news data of the full datasets, and 600 of these are used to learn, the other data are used for test. The label is whether the news data belongs to which category. In this research, we define 6 categories that news data belongs to such as "entertainment", "sports", "the economy", "IT, science", "domestic" and "overseas".

B. Tool and environment

Since we measure the time in this experiment, we introduce the spec of experiment environment and the tool we used. Table1 shows our experiment environment. We use Pylearn2[x] to conduct deep learning. Pylearn2 is a tool for carrying out the deep learning in the python. In Pylearn2, we can make deep learning in its own network by setting the shape of a network, such as the input layer. Moreover, denoising-autoencoder and RBM, etc. have also been implemented.

Table 1 Experiment enviroment

Experiment Environment	
OS	OSX(10.10.5)
CPU	Core i5 (I5-4258U)
Memory	8GB
Programming Language	Python(ver.2.7.10)
Tools for Deep Learning	Pylearn2

C. Parameters

In deep learning, we should determine some parameters to increase the accuracy of classification. Moreover, the parameters must be determined by experiment for each task since the optimum value differs for each task. In this experiment, we determine these values of the parameters.

1) The number of hidden layers

The number of hidden layers between the input layer and output layer, we conduct experiments from two layers to five layers.

2) The number of nodes

The input layer and output layer is determined the number of nodes, however the number of nodes in hidden layer is not determined. Therefore, we determine the number of nodes. We conduct experiment with each layer 50, 100, 150, 200.

3) The probability of occurrence of noise

It is a probability of occurrence of noise on the input to be used in the denoising autoencoder. We conduct experiments each 0.1 from 0.1 to 0.9.

4) The number of epochs

This is the maximum number of times of learning if the optimal value of the weight does not fall within a predetermined value. We experiment in the case of 1,50 and 100.

5) The number of batch

In this experiment, we are using a stochastic gradient descent method for minimization of error. Number of batches is the number of computing simultaneously the slope during the optimization. This time we conduct experiment in the case of 10 and 100.

D. Results

We experiment in the case of handling unsupervised data with pre-training, only supervised data and only supervised data in naïve bayes. The results of the experiment, the optimum value, its accuracy and computation time are as shown in Table 2.The accuracy is percentage that calculated the correct result of the whole in this experiment.

1) The number of hidden layers

Even if the number of hidden layer becomes 2 or more, accuracy did not increase and computation time become longer.

2) The number of nodes

If the number of nodes changed 50-1000,the accuracy rise gradually, therefore the accuracy became maximum accuracy at 500.

3) The probability of occurrence of noise

The accuracy was low when the probability of occurrence of noise is 0.1 to 0.3, it was high and similar when the probability of occurrence of noise is 0.4 to 0.9. Moreover, we determined 0,9 as optimal value since the accuracy was best in this experiment.

4) The number of epochs

Table 2 Results of experiment

Proposed Method	Computation Time[sec]					Accuracy[%]
Fine-tuning Only	58.87	57.12	53.55	56.14	59.36	33
Pre-training + Fine-tuning	1353.34	1390.63	1367.92	1364.5	1388.27	78

Naive Bayes	Computation Time[sec]					Accuracy[%]
Test Datasets1	7.76	4.73	4.75	4.89	4.87	67
Test Datasets2	4.97	5.10	4.97	5.00	5.01	68

The accuracy is increased when number of epoch became 10-1000. Since difference of accuracy was not observed when number of epochs became 500-1000, we considered that the number of epochs is the best at 500.

5) The number of batch

Because the number of the batch was raised more than 5% of accuracy on average towards the 100 than 10, we considered that the best number of the batch was 100.

V. CONSIDERETION

A. Parameters

In the deep learning, it is considered that increasing hidden layers can improve the accuracy. However, the accuracy did not improve if we increased hidden layers in this experiment. We think that because using the compressed data by using Word2Vec. We consider that the reason we can improve accuracy with increasing hidden layers is that increasing hidden layers can raise power of expression. Moreover, vectors of words created by Word2Vec have enough power of expression.

B. Using Word2Vec and deep learning

Firstly, we consider about using Word2Vec. The number of vocabulary in the datasets we used is 218,295 and the number of dimensions in vectors of words compressed by Word2Vec is 200. In our environment, since we took about 20 minutes in 200-dimensional experiment, it is impossible to calculate with using the original number of dimensions. Despite compressing the number of dimensions to about 1/1,000, the accuracy was very good. Therefore we consider that we use Word2Vec for text classification is practical.

The next, we consider about using deep learning. We used deep learning to handle unsupervised data in pre-training in this experiment. Table. 2 shows that we can obtain better 45% accuracy with pre-training than without pre-training. Therefore, using unsupervised data in pre-training for text classification is practical.

Compared with the naïve Bayes, the calculation time is increasing the accuracy is increased sufficiently. Moreover, the computation time can be shorten if we use actually since the calculation of the pre-training is performed only once. Except if the shortening of the computation time is sought, we consider that the proposed method a method that is practical and effective.

VI. CONCLUSION AND FUTURE WORK

Different from most of the conventional methods for text classification, our research use unsupervised data with using Word2Vec and deep learning. Moreover, we were compared to experiment with the proposed method with the conventional naive bayes method.

As a result, the proposed method was superior accuracy than the naive bayes method. However, the computation time was inferior to naive bayes. Therefore, our future work is understanding codes of Pylearn2 to shorten computation time. Moreover, since Word2Vec cannot give vectors to unknown words, it is necessary to give some vectors to unknown words in the future work.

REFERENCES

[1] McCallum. A. & Nigam. K, “A comparison of event models for naive Bayes text classification.” AAAI-98 Workshop on Learning for Text Categorization,1998

[2] Joachims. T, “Text categorization with Support Vector Machines: Learning with many relevant features.” Machine Learning: ECML-98, Tenth European Conference on Machine Learning, pp. 137–142, 1998

[3] Lee, K. H, “Text Categorization with a Small Number of Labeled Training Examples.” Unpublished Doctor of Philosophy, University of Sydney, 2003

[4] P. Trinkle, “An Introduction to Unsupervised Document Classification”, unpublished, 2009.

[5] Mikolov. T, Sutskever. I, Chen. K, Corrado. G, and Dean. J, ”Distributed representations of words and phrases and their compositionality.” Advances in Neural Information Processing Systems 26, pp. 3111–3119, 2013.

[6] Zhang. D, Xu. H, Su. Z, and Xu. Y, “Chinese comments sentiment classification based on word2vec and svm”, Expert Systems with Applications, Vol.42, pp.1857–1863, 2015

[7] Kudo, T. “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”. http://mecab.sourceforge.net/

BIOGRAPHY

Ryoma Kato was received his B. E degree from Hosei University in 2014. He is now a master course student of Hosei University. His research interests include machine learning, natural language processing, and data analysis.

Hiroyuki Goto is a professor in the department of Industrial & System Engineering, Hosei University, Japan. His research interests include operations research and high-performance computing.