Study on Visit Sales Detection using Similarity of Paragraph Vector

Rin Hirakawa^{a,*}, Yoshihisa Nakatoh^a

^aKyushu Institute of Technology

1-1, Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, 804-0015, Japan

*nakatoh@ecs.kyutech.ac.jp

Abstract

In Japan, with the increase of elderly people living alone, it is a problem that elderly people get involved in consumer troubles in recent years. Troubles with which elderly people are most likely to get involved are bank transfer fraud and malicious visit sales. The former problem is being solved by research to detect fraud from telephone voice. However, there is still no technical solution for malicious visit sales. In this paper, we proposed a system to detect visit sales from conversation voice of the user and visitor. In our system, cosine similarity of paragraph vector was adopted as criterion of visiting sales. The paragraph vector is acquired by inputting the conversation text output from speech recognition into Doc2Vec. In order to validate the proposed method, we conducted a classification experiment of conversation text. As a result, the area under the ROC curve was 0.92, and it was found that visited sales can be effectively detected by the proposed method.

Keywords: Doc2Vec, Sales Visit, Watching System.

1. Introduction

In recent years, there are various social problems caused by an increase in the elderly population in Japan. Consumer troubles of the elderly are one of them, and it is reported that the number of consultation to the consumer center is increasing year by year. The number of troubles is about 200 thousand, and among them the main ones are bank transfer fraud and malicious visit sales. Regarding bank transfer fraud, police, universities and companies cooperate to study techniques to prevent fraud. In that study⁽¹⁾, they detect bank transfer fraud by judging from conversation voice on telephone whether the user is feeling stress and keywords related to fraud exist in conversation. On the other hand, there are few reference data on visit sales, and no

precedent has been found in the research on detection methods using voice and language information. In this paper, we propose a system to judge whether the content of conversation between user and visitor is related to visiting sales.

2. Outline of the Proposed Method

In our proposed method (Fig.1), after converting the conversation between the user and the visitor into a document by speech recognition, we convert the document into a vector representation. Next, we calculate the similarity between the input document vector and the document vector of the visited sales conversation and judge whether it is a document related to visit sales. In this paper, we will focus on the part that judges whether the text you entered relates to visit sales. Speech recognition used for documenting conversation speech is what you like and it does not matter, but the output format must be a text file. Document vectorization is done using Doc2Vec, which is described in the next section.

Fig .1. Flowchart of the proposed method

2.1 Overview of Doc2Vec

Doc2Vec⁽²⁾ is a Python implementation of Paragraph Vector which is the algorithm proposed by Quoc Le et al. Doc2Vec has extended the algorithm Word2Vec⁽³⁾ which convert words into distributed expressions so that it can be applied to documents.

Word2Vec is beginning to be widely spread as an algorithm to overcome the defect of bag-of-words which has been conventionally used as input of text classification. In bag - of - words, there is a problem that information on the order of words appearing in sentences and the semantic relation between words cannot be effectively expressed. Word2Vec makes it possible to convert it into a distributed expression that takes into consideration the relation between words by letting the neural network learn so that word vectors appearing around can be predicted from a specific word vector. By calculating cosine similarity of the word vector obtained from Word2Vec, it is possible to evaluate the distance between words. When evaluating the distance between sentences using Word2Vec, there is a method of using a weighted average of vectors of all words included in sentences as a vector of the sentences. However, in this method, word order information is lost like bag - of - words. Of course it is possible to combine it with parsing, but this method can only be applied to single sentences, so it cannot cope with multiple sentences.

2.2 Method for Learning Document Vector

Doc2Vec learns word vectors and document vectors in parallel(Fig 2). Since a combination of a word vector and a document vector is used for the input of the hidden layer, it is possible to obtain a distributed expression considering the context. The model obtained by this learning method is called a distributed memory model (PV - DM). There is also a model called Continuous Bag - of - Words (PV - DBOW) that enables faster learning, but does not use document vectors for learning. We adopt the PV-DM model because we want to make use of the context information for judgment of visit sales.

2.3 Calculation of Similarity

After finishing learning the document vector, the Doc2Vec model can infer the document vector for an unknown document. In detecting visit sales, cosine similarity is calculated between the estimated document vector and the document vector of the visited sales conversation learned by the Doc2Vec model(Fig 3). When the average value of the

similarity between the input document vector and all the learned visited sales conversation documents is equal to or greater than a certain threshold value, it is determined as visiting sales. In addition, the document to be learned contains an article of Wikipedia in order to reduce the possibility of unknown words appearing when estimating a vector from the input document.

Fig .2. Distributed memory model

Fig .3. Calculation of similarity

3. Experiment

In order to verify the effectiveness of the proposed method, we conduct classification experiments on conversational documents. The classification performance is evaluated by using the area under the ROC curve of the average similarity. Furthermore, in order to determine the optimum threshold value, a cutoff point is obtained from the drawing result of the ROC curve. The learning of Doc2Vec is carried out using the parameters shown in Table 1 and the classification experiments are performed using the data shown in Table 2. Conversation documents of visiting sales are labeled with three levels of risk according to the maliciousness of contents. The criterion of the risk level is decided by whether or not it violated Article 246 of the Specified Commercial Transaction Law and Criminal Code. What violate the former are considered middle risk level, and what violate both are considered high risk. Also, conversational documents without illegality are labeled low level.

Experiments are conducted on two themes. The first is binary classification as to whether the content of the conversation is visit sales or not, and the second is judgment as to which of the three levels of danger degree visit sales belongs. In the binary classification experiment, all conversation documents labeled with risk are treated as positive data and conversation documents⁽⁴⁾ in daily conversation are treated as negative data. Regarding the determination of the degree of danger, it is assumed that only documents judged to be visit sale by binary classification are handled. In the evaluation experiment, assuming that the binary classification was performed correctly, we conduct experiments for determining the degree of risk using only visited sales conversation documents. Similar to binary classification, average similarity is used as a criterion for determination, but similarity is calculated only for documents with labels of each class.

3.1 Result of Binary Classification

The ROC curve drawn using the average similarity with the visiting sales conversation is shown in Fig. 4, and the cutoff point obtained from the ROC curve is shown in Table 3. The area under the ROC curve was 0.92, which proved that effective classification was possible. As a result of investigating negative data that has high similarity, there was a tendency that topics on money and monetary amount were included during conversation.

Table .1. Parameter used for training

parameter	
vector size	300
alpha	0.0015
window	5
sample	10^{-4}
min count	1
epoch	20

Table .2. Document data used for experiments

document	contents	risk	# of Data
training	wikipedia	zero	2653
	sales visit	high	7
		middle	7
		low	7
test	daily conv	zero	129
	sales visit	high	3
		middle	3
		low	3

Table .3. Cutoff points of classification

	threshold	1-fpr	fpr	tf	tpr
binary	0.55	0.84	0.16	0.04	0.89
high	0.6	0.83	0.17	0.17	1
mid	0.59	0.67	0.33	0	0.67
low	0.62	0.67	0.33	0	0.67

Fig .4. ROC curve of binary classification

3.2 Result of Risk Determination

Figures 5 to 7 show the ROC curves when determining the risk for three classes of low, medium and high. The cut-off points of each class are shown in Table 3. As can be seen from the figure, the classification of low and medium level resulted in lower accuracy compared to high classification. If the visitor does not tell the user the purpose of sale, it violates the Specified Commercial Transactions Law, but it is considered difficult to judge from the overall trend of conversation.

4. Conclusions

In this paper, we proposed a method to detect visit sales from conversation contents using document vector. In the binary classification experiment using the conversation document, the area under the ROC curve was 0.92, showing the effectiveness of the proposed method. It was found that it is difficult to make judgments of low and medium level in the experiment of judging the degree of risk.

As future tasks, it is necessary to improve classification performance by performing document clustering such as K-means method with document vector as input. Since it is considered that the local information of the document is necessary for determining the risk level, we plan to devise a method that can evaluate the document vector in time series.

References

- (1) Naoshi Matsuo et al. : “Development of the System for Stress Situation Detection based on the Voice Information”, DICOMO2014, pp131-137, 2014
- (2) Quoc Le, Tomas Mikolov : “Distributed Representations of Sentences and Documents”, Proceedings of The 31st International Conference on Machine Learning, pp. 1188-1196, 2014
- (3) Tomas Mikolov et al. : “Distributed Representations of Words and Phrases and their Compositionality”, Advances in neural information processing systems, pp3111-3119, 2013
- (4) Fujimura Itsuko, Shoji Chiba, Mieko Ohso, 2012, Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora, Proceedings of the VII th GSCP International Conference. Speech and Corpora, 393-398.

Fig .5. ROC curve of high risk determination

Fig .6. ROC curve of middle risk determination

Fig .7. ROC curve of low risk determination