Recognition Technique of ConfidentialWords Using Neural Networksin Cognitive Infocommunications

Yuya Kiryu[1], Atsushi Ito[1], Masakazu Kanazawa[2]

[1]Utsunomiya University, Graduate School of Engineering
7-1-2 Yoto Utsunomiya Tochigi 321-8505 Japan
mt166519@cc.utsunomiya-u.ac.jp, at.ito@is.utsunomiya-u.ac.jp

[2]TohokenInc.
2041-7 Koshina-cho, Sano-shi, Tochigi, 327-0822, Japan
kanazawa@tohoken.co.jp

Abstract: Cognitive Infocommunications (CogInfoCom) involves communication, especially the combination of informatics and telecommunications. In the future, infocommunication is expected to become more intelligent and supportive of life. Privacy is one of the most critical concernsin infocommunications. A well-recognized technology that ensures privacy is encryption; however, it is not easy to hide personal information completely. One technique to protect privacy is to find confidential words in a file or a website and change them into meaningless words. In this paper, we use a judicial precedent dataset from Japan to discuss a recognition technique for confidential words using neural networks. The disclosure of judicial precedents is essential, but only some selected precedents are available for public viewing in Japan. One reason for this is the concern for privacy. Japanese values do not allow the disclosure of the individual's name and address present in the judicial precedents dataset. However, confidential words, such as personal names, corporate names, and place names, in the judicial precedents dataset are converted into other words. This conversion is done manually because the meanings and contexts of sentences need to be considered, which cannot be done automatically. Also, it is not easy to construct a comprehensive dictionary for detecting confidential words. Therefore, we need to realize an automatic technology that would not depend on a dictionary of proper nouns to ensure that the confidentiality requirements of the judicial precedents are not compromised. In this paper, we propose two models that predict confidential words by using neural networks. We use long short-term memory (LSTM) and continuous bag-of-words (CBOW) as our language models. Firstly, we explain the possibility of detecting the words surrounding an confidential word by using CBOW. Then, we propose two models to predict the confidential words from the neighboring words by applying LSTM. The first model imitates the anonymization work by a human being, and the second model is based on CBOW. The results show that the first model is more effective for predicting confidential words than the simple LSTM model. We expected the second model to have paraphrasing ability to increase the possibility of finding other paraphraseable words; however, the

*score was not good. These results show that it is possible to predict confidential words;
however, it is still challenging to predict paraphraseable words.*

1   Introduction

Cognitive Infocommunications (CogInfoCom) [1] [2] describes communications,
especially the combination of informatics and communications. Cognitive
infocommunications systems extend people's cognitive capabilities by providing
fast infocommunications links to huge repositories of information produced by the
shared cognitive activities of social communities [3]. Future infocommunication is
expected tobe more intelligent and would even have the ability to support life.
Figure 1 shows the idea of CogInfoCom. Clearly, privacy is one of the most
critical concerns in infocommunication. Encryption is a well-recognized
technology used for ensuring privacy; however, encryption does not effectively
hide personal information completely. One technique to protect privacy is to find
confidential words in a file or a website and convert them into meaningless words.
It will be good if a network becomes intelligent and automatically changes private
words into meaningless words. We think that this is one benefit of introducing
cognitive infocommunication into our life.

Based on a Japanese judicial precedents dataset, we discuss a recognition
technique of confidential words using neural networks. The disclosure of judicial
precedents is indispensable to ensure the rights to access legal information from a
citizen. If we use big data analysis technique in artificial intelligence (AI), we can
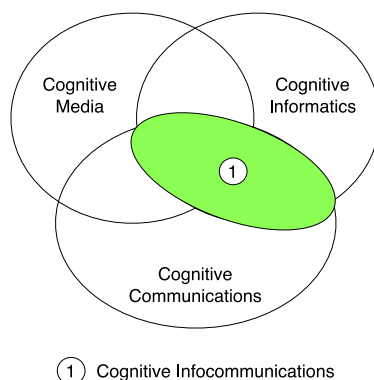

Figure 1
Infocommunication model

receive substantial benefits in the judicial service. However, most precedents are not available publicly on the Web pages of Japanese courts since specifying an individual's name, or other personal details violates the Japanese understanding of privacy. Confidential words, such as personal names, corporate names, and place names, in the precedents available for public viewing, are converted into other words to protect privacy. In Japan, this procedure takes time and effort because this procedure is performed manually. Also, globalization has led to the participation of people from various countries in these trials; therefore, a dictionary of proper nouns would take additional time to create.

Neural networks are also being increasingly used in natural language processing in recent years. There is ongoing research to predict words and to derive a vector based on the meanings of words [4]. Therefore, in this paper, we discuss ways of applying a neural network to the task of detecting confidential words.

In Chapter 2, we explain how the converted words in the dataset of Japanese judicial precedents are processed. In Chapter 3, we refer to some neural network models. Then, we explain the concept of predicting confidential words and result of preliminary experiment to detect feature around the confidential words in Chapter 4, and we propose models to predict confidential words and result of an experiment in Chapter 5. Finally, we provide our conclusions in Chapter 6.

2   Converting Words

2.1   Confidential Words in Japanese Judicial Precedents Dataset

Some judicial precedents datasets are available for free on the websites of the Japanese courts [5]. Confidential words in these precedents that are available for public viewing on the website are converted into uppercase letters. (In paid magazines and websites, Japanese letters are sometimes used.) Figure 2 shows an example of such changed words.

In the example shown in Figure 2, the personal name "Kiryu" is substituted by the letter "A." If there are some more words that need to be kept confidential, other letters of the alphabets are used (e.g., B, C, and D).
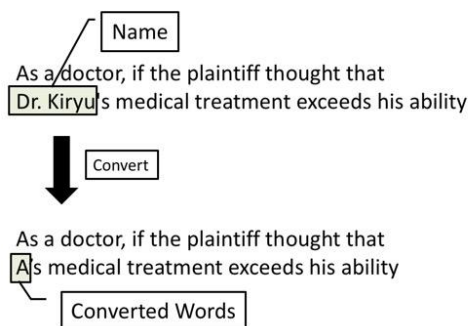
Figure 2

Example of Converted Words

2.2 Problem of Detecting Confidential Words using a Dictionary

Studies are also being conducted in other fields on various methods of hiding confidential proper nouns. The primary method is to create proper noun dictionaries, such as personal name dictionaries or place name dictionaries, and match them with the target documents.

The merit of this method is that the more the dictionary is enhanced, the more the accuracy improves. However, in Japanese, it is sometimes uncertain whether a word is a name or not unless it is read in the context of the entire sentence. Also, as globalization progresses, foreigners often join in trials, and it is difficult to create a dictionary that includes names of people from different continents. Also, certain spelling patterns need to be followed when translating foreign names into the Japanese language.

In this paper, we consider a method for using a neural network to solve the technical problems involved in using dictionaries.

3 Language Models with Neural Networks

3.1 Neural Probabilistic Language Model

The Neural Probabilistic Language Model was published by Bengio in 2003; this model makes predictions from the words that are already present [6]. This method maximizes the probability of the target word with the maximum likelihood

principle in the score of the softmax function. When the word h that has already appeared is given, the probability that w_t appears is

$$P(w_t|h) = \frac{\exp(\text{score}(w_t, h))}{\sum_{all\ w'\ in\ dictionary} \exp(\text{score}(w', h))}. \tag{1}$$

This equation is maximized by using the maximum likelihood method; it is the same as maximizing (2), which is log-likelihood function of (1). The problem with this method is that the number of calculations increases with the increase in the size of the dictionary of members $\sum_{all\ w'\ in\ dictionary} \exp\big(\text{score}(w', h)\big)$. In other words, J_{ML} can be written as follows:

$$J_{ML} = \log P(w_t|h) =$$
$$= \text{score}(w_t, h) - \log\left(\sum_{all\ w'\ in\ dictionary} \exp\big(\text{score}(w', h)\big)\right). \tag{2}$$

3.2   Continuous Bag-of-Words

Mikolov proposed the continuous bag-of-words (CBOW) method to speed up the train of the Neural Probabilistic Language Model and derive embedding vectors to improve the meaning [7].

CBOW predicts w_t from the $2k$ words $w_{t-k}, \dots, w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}, \dots, w_{t+k}$. (We call this number $2k$ as the window size.) There are two aspects of the Neural Probabilistic Language Model. First, $\sum_{all\ w'\ in\ dictionary} \exp\big(\text{score}(w', h)\big)$ is calculated with not all the words in the dictionary but with randomly sampled words in the dictionary. This technique is called negative sampling. Second, each input word vector is compressed into the embedding vector, and all of these are added together. This method reduces the weight matrix to the output layer. An overview of CBOW is shown in Figure 3.
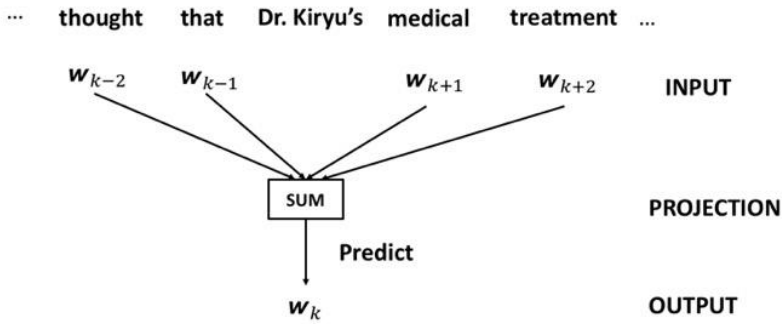
Figure 3
Continuous Bag-of-Words

It is known that the embedding vector derived by CBOW is a vector space based on the word meanings. Even if the spelling of the word is different, if the surrounding words are similar, their embedding vectors will be similar.

3.3   Long Short-Term Memory

Long Short-Term Memory (LSTM) is a kind of recurrent neural network (RNN). Adaptation to the language model was made by Mikolov, et al. (2010) [8]. With RNN as the language model, continuous data (w_i) is input and often handled in the task of predicting the next word. RNN has a feedback structure and calculates the output from the input (w_i) and the feedback (F_{i-1}). Various models have been proposed for this calculation method. However, we have used LSTM in this paper since we would like to start to develop our model from the simple one.
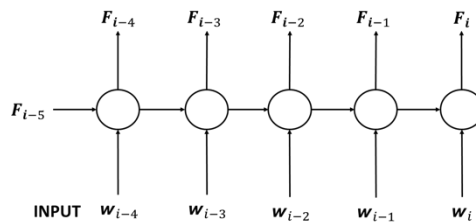

Figure 4
Structure of RNN (LSTM)

4 Prediction Method Using Neural Networks

4.1 Concepts Used

As described in Section 2, we propose a method using a neural network without the use of proper-noun dictionaries. A function is required to recognize the context and determine whether or not the target word needs to be converted to maintain confidentiality. LSTM is one of neural network models and handles continuous data. LSTM is useful in the task of predicting the next word; therefore, we performed our experiment based on this model.

However, the goal of our research was little different from the goal of the Neural Network Language Model (NNLM). In our study, we recognized a common concept that we should change words treated as having different meanings in the corpus (see Figure 5). In the previous tasks, the meanings of the words were used and recognized in the same context. For example, the same predicted confidential word sometimes means "name" and at other times means "place". In this case, it has completely different meaning and return different letter of the alphabet. In other words, the concept of "confidential words" encompasses many words, and it will be difficult to derive this concept as an embedding vector. However, the CBOW model successfully expresses ambiguous meanings that were earlier difficult to express. Therefore, we decided to base this research on the CBOW model.

Previous research

Bob has had two **dogs** for 5 years. ⇨ Predict **dogs**

Alice likes **cats** very much. ⇨ Predict **cats**

Our research

As a doctor, if the plaintiff thought that **Dr. Kiryu's** medical … ⇨ Predict **A** **(Confidential)**
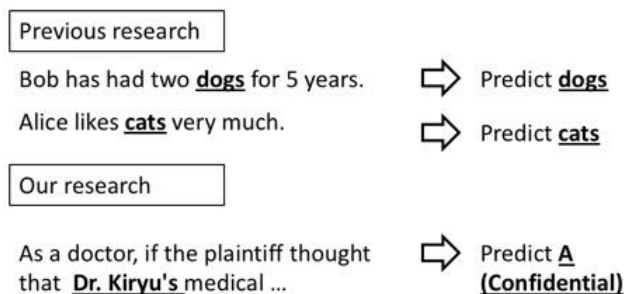
Figure 5

Difference between our research and previous research:

The neural network predicts each word in the sentence from the words that have appeared so far.

In previous research, learning was performed using the word appearing in the sentence as the correct answer (For example, "dogs" and "cats" in Figure 5). However, in our research, a proper noun must be learned to predict it as a confidential word. In Figure 5, "Dr.Kiryu" is a proper noun; however, it must be predicted as A (confidential word).

Our proposed approach is to predict confidential words from the words surrounding the target words. We assume that there are features in the distribution of words around the confidential words. Therefore, the neural network model can capture that features of the distribution of words around the confidential words. We experimented using CBOW to confirm our assumptions. The details of the experiment are described in Section 5.

4.2   Preliminary Experiment

In this section, we explain the result of the experiment using the method given in Section 4.1. As mentioned in Section 3.2, a feature of the CBOW model is that the embedding vectors will be similar if the surrounding words are similar.

Confidential words in the precedents published by the Japanese courts are usually converted to uppercase letters, such as A, B, and C. The same letters cannot be used for different individuals in the same judicial precedents dataset.

Therefore, if the CBOW model can capture the feature of confidential words, the similarity of each of the converted confidential words (i.e., A, B, C, ... X, Y, Z) would also be high. In this paper, similarity is defined as the cosine similarity, as shown in (3).

$$\cos(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha} \cdot \boldsymbol{\beta}}{|\boldsymbol{\alpha}||\boldsymbol{\beta}|}. \qquad (3)$$

Here, α and β are the embedding vectors of the words to be compared. The closer the cosine similarity is to 1, the higher is the similarity between the words.

The judicial precedents dataset used in this experiment were 20,000 precedents available on the Japanese court website. The various parameters are shown in Table 1. Table 2summarizes the results of calculating the cosine similarities of each confidential word by using the training result.

Table 1

CBOW parameters

Embeddings_Size	200
Batch_Size	200
Window_Size	10
Min_Learning_Rate	0.001
Min_count_vocab	5

Table 2
Top 10 WordsSimilar to Converted Words

Similarity to A		Similarity to B		Similarity to C		Similarity to D		Similarity to E		Similarity to F	
B	0.9352	A	0.9352	D	0.9251	E	0.9483	F	0.9487	G	0.9500
C	0.8833	C	0.9176	B	0.9176	F	0.9345	D	0.9483	E	0.9487
D	0.8071	D	0.8529	E	0.9165	C	0.9251	G	0.9345	H	0.9391
E	0.7612	E	0.8118	F	0.8893	G	0.8743	C	0.9165	D	0.9345
X	0.7363	F	0.7656	A	0.8833	B	0.8529	H	0.8806	I	0.9016
F	0.7094	G	0.7259	G	0.8463	H	0.8316	J	0.8328	C	0.8893
G	0.6772	H	0.7036	H	0.8186	A	0.8071	I	0.8326	J	0.8801
H	0.6517	X	0.6793	I	0.7550	I	0.7835	K	0.8003	K	0.8638
Y	0.6471	Y	0.6690	J	0.7521	J	0.7711	B	0.8118	M	0.8335
K	0.6032	K	0.6498	K	0.7358	K	0.7558	M	0.7743	L	0.8275

Similarity to G		Similarity to H		Similarity to I		Similarity to J		Similarity to K	
F	0.9500	I	0.9548	H	0.9548	I	0.9538	L	0.9589
H	0.9475	G	0.9475	J	0.9538	K	0.9527	M	0.9552
E	0.9345	J	0.9446	K	0.9435	H	0.9446	J	0.9527
I	0.9258	F	0.9391	L	0.9275	L	0.9429	I	0.9435
J	0.9241	K	0.9325	G	0.9258	M	0.9331	H	0.9325
K	0.8845	M	0.9054	M	0.9179	G	0.9241	N	0.9151
D	0.8743	L	0.9015	F	0.9016	N	0.8936	G	0.8845
M	0.8653	E	0.8806	N	0.8751	F	0.8801	O	0.8787
L	0.8569	N	0.8632	O	0.8346	O	0.8717	F	0.8638
C	0.8463	O	0.8356	E	0.8326	R	0.8367	Q	0.8614

Similarity to L		Similarity to M		Similarity to N		Similarity to O		Similarity to P	
K	0.9589	L	0.9583	M	0.9508	P	0.9381	O	0.9381
M	0.9583	K	0.9552	L	0.9358	N	0.9225	Q	0.9192
J	0.9429	N	0.9508	O	0.9225	M	0.9221	R	0.9026
N	0.9358	J	0.9331	K	0.9151	Q	0.9128	N	0.8864
I	0.9275	O	0.9221	Q	0.9119	R	0.9096	S	0.8789
H	0.9015	I	0.9179	R	0.9054	L	0.8995	M	0.8683
O	0.8995	H	0.9054	J	0.8936	K	0.8787	L	0.8632
Q	0.8862	Q	0.8973	P	0.8864	J	0.8717	T	0.8449
R	0.8754	R	0.8915	I	0.8751	S	0.8604	K	0.8397
P	0.8632	P	0.8683	S	0.8730	W	0.8419	J	0.8265

Similarity to Q		Similarity to R		Similarity to S		Similarity to T		Similarity to U	
R	0.9212	Q	0.9212	T	0.9296	S	0.9296	W	0.9251
P	0.9192	S	0.9136	R	0.9136	U	0.8977	S	0.9044
O	0.9128	O	0.9096	U	0.9044	R	0.8805	T	0.8977
N	0.9119	N	0.9054	Q	0.9020	W	0.8609	R	0.8947
S	0.9020	P	0.9026	P	0.8789	Q	0.8592	Q	0.8928
M	0.8973	U	0.8947	N	0.8730	N	0.8460	N	0.8624
U	0.8928	M	0.8915	O	0.8604	P	0.8449	M	0.8534
W	0.8883	T	0.8805	M	0.8541	O	0.8418	O	0.8397
L	0.8862	W	0.8774	W	0.8524	M	0.8313	L	0.8390
K	0.8614	L	0.8754	L	0.8424	L	0.8255	V	0.8234

Similarity to V		Similarity to W		Similarity to X		Similarity to Y		Similarity to Z	
W	0.8423	U	0.9251	A	0.7363	Z	0.8119	Y	0.8119
M	0.8271	Q	0.8883	Y	0.7019	X	0.7019	V	0.7378
U	0.8234	R	0.8774	Z	0.6883	P	0.6780	W	0.7361
N	0.8173	N	0.8613	P	0.6807	B	0.6690	Q	0.7321
R	0.8046	T	0.8609	B	0.6793	A	0.6471	O	0.7205
S	0.7919	S	0.8524	C	0.6753	O	0.6453	U	0.7104
Q	0.7891	M	0.8479	V	0.6510	Q	0.6320	P	0.7101
O	0.7887	V	0.8423	O	0.6264	C	0.6315	N	0.7067
T	0.7713	O	0.8419	K	0.6191	W	0.6034	M	0.7065
K	0.7702	P	0.8226	D	0.6180	M	0.6029	K	0.6921

The precedents are written in Japanese; therefore, very few are capitalized. It is more common for English words to have the first letter capitalized than Japanese words. In other words, the judicial precedents were Japanese sentences; therefore,

it was extremely rare that an uppercase letter was used for English words. Table 2 shows the cosine similarity of the top 10 words to the confidential words (appearing as uppercase letters); these are the training results in precedents available for public viewing on the website of the Japanese court. As a result, the top 10 confidential words become uppercase letters.

From the above, we can see that the CBOW model can capture a part of the features of the distribution around the confidential words. Also, a previous study uses the CBOW model as a predictor based on the meanings of words [9]. Therefore, in this paper, we use several neural networks based on the CBOW model to predict confidential words and consider a network model effective for predicting them.

5 Predicting Confidential Words

In this section, we describe an experiment to predict confidential words by using neural networks. We propose two models: one model imitates a human being (explained in 5.1.1), and the other model is based on the concept of the CBOW model (explained in 5.1.2).

5.1 Proposed Model

5.1.1 Bi-directional LSTM LR

Bidirectional Long Short-Term Memory Recurrent Neural Network (BLSTM-RNN) has been shown to be very effective for modeling and predicting sequential data, e.g. speech utterances or handwritten documents [10]. From the viewpoints of CogInfoCom, [11] introduces an RNN-based punctuation restoration model using uni- and bidirectional LSTM units as well as word embedding.

Bi-directional LSTM LR (Left to Right) is a model that imitates the anonymization done by humans. When humans perform anonymization, they make a judgment after reading to the left and the right of the target word. Therefore, it becomes a shape as shown in Figure 6 (c). The input order on the back (right side) of the target word is reverse of the sentence order because we assume that the words closer to the target word have higher importance.

5.1.2 SumLSTM

The SumLSTM,which is based on the CBOW model, is a model that validates the effectiveness of the model given in Chapter 4 (Also, see Figure 6 (b)). In addition to the normal LSTM calculation, the total of all the input vectors is calculated and

activated by the softmax function in the output layer. The model combined with the Bi-directional LSTM LR model is shown in Figure 6 (d).

5.2   Corpus and Evaluation Method

We used 50,000 judicial precedents for the training data and 10,000 judicial precedents for the test data. These data included the records of trials from 1993 to 2017. We used the precedent database provided by TKC, a Japanese corporation [12]. We converted all the confidential words into the uppercase letter "A" and separated the Japanese words with spaces by using MeCab, a Japanese morphological analyzer. MeCabw as required because we were using Japanese judicial precedents dataset [13]. Also, word prediction required stop words; therefore, word prediction was not excluded in this experiment.
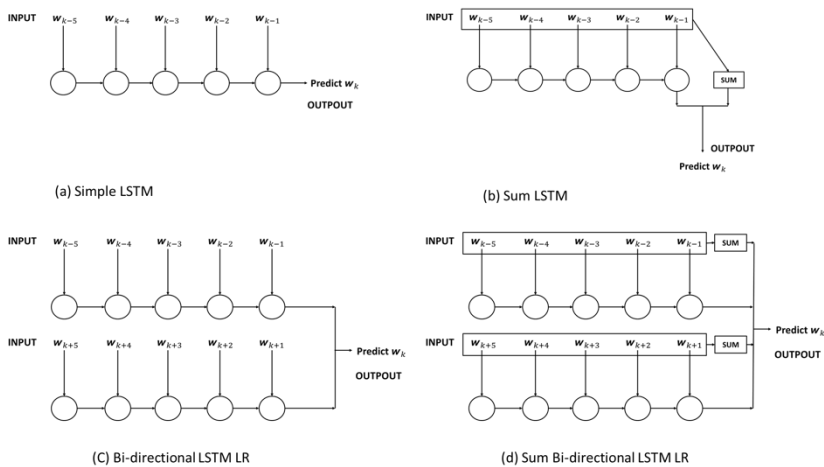

Figure 6

Four neural network models used in experiment. Figure 6(a) is previous simple LSTM model.

Figures 6(b) and (c) are the proposed models.

Figure 6(d) is a combination of the two proposed models.

5.3   Result of the Experiment

In this experiment, we also prepared a simple LSTM model to compare the two models proposed in Section 5.1. This model had a three-layered structure: an input layer, a hidden (LSTM) layer, and an output layer. The size of the hidden units was 200. The input/output layer size was the same as the vocabulary size (approximately 200,000 in our corpus).

The Bi-directional LSTMLR model had two simple LSTM model structures, and SumLSTM also inherited the simple LSTM structure. In addition, we combined the SumLSTM and LSTM LR models and named it SumBi-directional LSTM LR (see Figure 6(d)).

This experiment was conducted using the four models shown in Figure 6. Also, the embedding vector was 200 for all models [12]. For accuracy, we used perplexity (PPL) that was used in a previous research for predicting the next word. PPL was given by the following equation:

$$PPL = \frac{1}{P(correct\ word)}. \qquad (4)$$

In (4), P is the probability, and PPL represents the number of prediction choices that are narrowed down to neural networks. The smaller the value, the better the prediction results.

Table 3 shows the results. CW_PPL is the average PPL of the test data whose answer reflects the confidential words. However, PPL is the average of all the test data. The PPL scores in Table 3 show that the Bi-directional LSTM LR model decreased by 0.195 as compared with the SimpleLSTM model. Also, the SumLSTM model decreased by 0.247 as compared with the SimpleLSTM model. The combination of the two methods scored the best results, which was 4.462. Therefore, the proposed models are effective for PPL in our corpus.

Let's look at the results of CW_PPL directly related to the task of predicting the confidential words. We will find that the difference of scores is at least 32.492 between PPL and CW_PPL. This result suggests that the task of predicting confidential words is more difficult than the task of predicting other words. Also, each CW_PPL score shows that the Bi-directional LSTM LR model decreased by 18.934 as compared with the SimpleLSTM model. (The score for the Bi-directional LSTM LR model was 37.343, which was the best score). However, the SumLSTM model increased by 16.186 as compared with it. Furthermore, the combination of the two methods recorded the worst score.

In PPL, we found that all the proposed methods were more effective than the simple model. However, for the prediction of confidential words, only the Bi-directional LSTM LR model showed good results. SumLSTM based on CBOW might have produced these results. CBOW is an effective model for paraphrasing words, and SumLSTM also uses this mechanism. Therefore, when SumLSTM predicted a word whose answer is "confidential," the CW_PPL became worse because there was a possibility of paraphrasing words such as "plaintiff,""defendant,""doctor,"and "teacher." Knowing the paraphrased words of the confidential words meant that the embedding vectors of the confidential words could be successfully generated. This meant that the model could recognize the meaning of "confidential." However, the prediction accuracy did not improve; therefore, there was a problem in calculating the probability of the prediction task.

To solve this problem, we could exclude these paraphraseable words from the choices when calculating the probability. It is also important to examine scores other than PPL.

Table 3

Result of prediction with proposed neural networks

	SimpleLSTM	Bi-directional LSTM LR	SumLSTM	SumBi-directional LSTM LR
PPL	4.851	4.656	4.603	**4.462**
CW_PPL	56.277	**37.343**	72.463	77.031

Conclusion

CogInfoCom is a form of communication that involves a combination of informatics and communications. It is expected that in the future, infocommunication will become more intelligent and will even support our life. A brief review of the computationa lintelligence and data mining methods utilized in industrial Internet-of-Things experiments is presented in [14].

Privacy is one of the most important issues in CogInfoCom. Encryption is one of the most well-recognized technologies for providing privacy; however, it is not easy to hide personal information completely. One technique to protect privacy is to find confidential words in a file or a website and convert them into meaningless words. It would be useful to have a network that is intelligent enough to automatically anonymize confidential words. There are several papers relating to privacy and anonimity in CogInfoCom research. [15] is a research on modelling multimodal behavior that often requires the development of corpora of human-human or human-machine interactions. In the paper, ethical considerations related to the privacy of participants and the anonymity of individuals are mentioned. [16] mentions that to realize the social interactions in a virtual workspace, anonymity can influence the process' outputs significantly.

Based on a Japanese judicial precedent, we proposed a recognition technique for confidential words using a neural network. Our proposed model will help solve the privacy problems associated with communication. In the current Japanese judicial precedents dataset, proper nouns that could identify individuals were converted into unrecognizable words, and the Japanese court used these words. Currently, this task is expensive and time-consuming because it is performed manually. Also, using dictionaries is not practical. Globalization has increased the number of trials on foreign subjects, however, it is not practical to include all names of people from all over the world in the dictionary.

Therefore, in this paper, we introduced a technology to predict confidential words using a neural network and without the aid of a dictionary of proper nouns. Firstly, we evaluated CBOW for this purpose. We confirmed that CBOW could capture the features of the words surrounding a confidential word.

Next, we proposed two models to predict confidential words from neighboring words. The two proposed models are effective for predicting all the words. However, only the Bi-directional LSTM LR model was effective for predicting confidential words. This could have happened because SumLSTM was based on CBOW. CBOW is an effective model for paraphrasing words; therefore, SumLSTM also has that mechanism. Therefore, when SumLSTM predicts a word whose answer is "confidential," the CW_PPL became worse because there was a possibility of paraphrasing words such as "plaintiff," "defendant," "doctor," and "teacher." The CBOW mechanism that was good at paraphrasing showed good performance for the recognition of "confidential" words; however, it was not effective for prediction.

Also, the parameters of these models were based on Mikolov's paper [7]. However, there were other important parameters as well. Therefore, to improve accuracy, we plan to optimize with Bayesian optimization in the future.

Our proposed method aims to compensate for the drawbacks of the method that uses the dictionary, but our method does not work on a single unit. In the future, we need to consider how to use the output of the two methods. Another method is to use a proper noun dictionary as input to a neural network. By using the dictionary information for input to a neural network, we expect the accuracy to increase. If the target word appears in the dictionary, it is easy to provide the correct answer. Even if the target word does not appear in the dictionary, and it is a confidential word, the input vectors would have an easy pattern. Therefore, we expect to achieve high accuracy if we combine a dictionary and the prediction method. However, parameter adjustment would still remain a major concern. Also, we would like to detail the requirements between identity and anonymity in a judicial precedents by referring [17].

Finally, we can conclude that the function to find a confidential word can be realized and will be an important function of future CogInfoCom.

Acknowledgement

References

[1]   Baranyi,P., Csapo, A. (2012). Definition and Synergies of Cognitive Infocommunications, Acta Polytechnica Hungarica, Vol. 9 No. 1, pp. 67-83

[2]   Baranyi, P., Csapo, A., Sallai, G. (2015). Cognitive Infocommunications (CogInfoCom), Springer International, ISBN 978-3-319-19607-7

[3]   Graham Wilcock, Kristiina Jokinen, (2017). Bringing CognitiveInfocommunications to small language communities; Cognitive Infocommunications (CogInfoCom 2017)

[4]   Lai, S., et al. How to generate a good word embedding. IEEE Intelligent Systems, 2016, 31.6: 5-14

[5]   http://www.courts.go.jp/app/hanrei_jp/search1 [Dec, 31, 2017]

[6]   Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. Journal of machine learning research, 3(Feb), 1137-1155

[7]   Mikolov, T., et al. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

[8]   Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Honza Cernocky, Sanjeev Khudanpur. (2010). Recurrent neural network based language model. Proc. INTERSPEECH2010

[9]   Shunya Ariga, Yoshimasa Tsuruola. (2105). Extension of synonym words according to context by vector representation of words. The 21st Annual Conference of Natural Language Pro-ceedings (NLP2015), pp. 752-755

[10]   Peilu Wang, Yao Qian. (2015). A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding. arXiv:1511.00215 [cs.CL]

[11]   Mate Akos Tundik, Balazs Tarjan, and Gyorgy Szaszak. (2017). A Bilingual Comparison of MaxEnt- and RNN-basedPunctuation Restoration in Speech Transcripts. Proceedings of 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017)

[12]   http://www.tkc.jp/law/lawlibrary [Dec, 31, 2017]

[13]   http://taku910.github.io/mecab/ [Mar, 9, 2018]

[14]   Jouni Tervonen, Ville Isoherranen. (2015). A Review of the Cognitive Capabilities and DataAnalysis Issues of the Future Industrial Internet-of-Things. CoginfoCom2015, 6th IEEEInternational Conferenceon Coginfocm

[15]   Maria Koutsombogera and Carl Vogel. (2017). Ethical Responsibilities of Researchers andParticipants in the Development of MultimodalInteraction Corpora. 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2017)

[16]   Laura Kiss, Balázs Péter Hámornik, Dalma Geszten, Károly Hercegfi. (2015). The connection of the style of interactions and the collaboration in a virtual work environment. 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2015)

[17]   Gary T. Marx. (2001). Identity and Anonymity:Some Conceptual Distinctions and Issues for Research. In J. Caplan and J. Torpey, Documenting Individual Identity. Princeton University Press