AN AUTOMATIC METHOD FOR BUILDING A DICTIONARY FOR CLASS-BASED SPEECH RECOGNITION SYSTEMS

Kohichi Takai[1,2], Gen Hattori[1], Keiji Yasuda[1,2], Panikos Heracleous[1], and Akio Ishikawa[1]

[1]KDDI Research, Inc.
Garden Air Tower, 3-10-10, Iidabashi
Chiyoda-ku, Tokyo, 102-8460, JAPAN
{ko-takai,ge-hattori,ke-yasuda,
pa-heracleous,ao-ishikawa }@kddi-research.jp
[2]Nara Institute of Science and Technology, Japan
ke-yasuda@dsc.naist.jp

Abstract. The mis-recognition of proper nouns reduces the quality of conversation performance of dialog systems or human beings via speech-to-speech translation systems. In this paper, we propose a method for building a proper noun dictionary for automatic speech recognition (ASR) systems that uses class-based language models. The method consists of two parts: a training data building part and a word classifier training part. The first part uses a text sentence corpus containing proper nouns. For each proper noun, the first part determines the class that gives the highest sentence-level automatic evaluation score from the ASR result. The second part trains a Convolutional Neural Network (CNN)-based word class classifier by using the training data yielded by the first step. The training data consists of sentences that contain a proper noun and the proper nouns' classes with the highest scores. The CNN is usually trained to predict the proper noun class, but this proposed method requires no manually annotated training data. By using the proposed method, the experimental results on a speech recognition system show that the dictionary created by the proposed method achieves performance comparable to that of a manually annotated dictionary.

Keywords: Automatic Speech Recognition, Dictionary, Convolutional Neural Network

1 INTRODUCTION

As a result of significant advances in technical innovations of speech processing and natural language processing, spoken dialog systems [1] and speech-to-speech translation systems [2] are becoming realistic tools for travelers. Especially for the travel domain, the coverage of proper nouns for tourist spots, landmarks, restaurants, and accommodation highly influences the system performance.

The ASR system has a role in the input interface for both dialog systems and speech translation systems. Many studies in the field of ASR have considered the problem of proper noun coverage. These studies can be classified into two approaches. The first approach of sub-word modeling is used to handle low occurrence words in the training corpus [3,4]. The second approach deals with out-of-vocabulary (OOV) words in the training corpus by using the class-based language model and other language resources such as a hand-crafted dictionary [5,6]. This paper uses the latter approach.

To handle proper nouns newly produced on a daily basis, the research shown in this paper takes the latter approach, which uses ASR with the class-based language model and dictionary. The class-based language model requires additional cost to manually annotate the word class for each new word. In this paper, we propose a novel method for automatically estimating the word classes. To train a word class classifier, the proposed method only uses surface and phoneme expressions of the new words, and sentences including the words. Since the proposed method does not require manual class annotation nor any kind of speech data, the method has significant merits for practical applications.

Section 2 describes related work. Section 3 explains the proposed method. Section 4 shows the experimental results using the ASR system. Finally, Section 5 concludes the paper and presents directions for future work.

2 RELATED WORK

OOV caused by data sparseness is one of the biggest problems in ASR. To solve the problem, the class-based language model was proposed in the ASR research field. There are several design concepts to introduce the class-based language model to the ASR system.

One of them is a framework that allows posterior proper noun registration. Even after training of the language model, this framework enables the user to register the demanded proper nouns by using the class-based language model. In practical usage, the word registration function is very helpful for users. However, users are expected to understand word class specifications and annotate the demanded new words by themselves.

The other concept of class-based language model usage is a hierarchical language model [7]. Since this method handles OOV by using sub-word or character level modeling, manual word registration is not required. However, there is no way to guarantee the output of demanded new words. Even technology such as end-to-end ASR [8] has the same problem.

As described above, each concept has advantages and disadvantages. In this research, we emphasize practical usage and assume the posterior proper noun registration framework. The proposed method enables the annotation cost for the framework to be reduced.

Some researches of the named entity recognition (NER) [9-10] have been conducted in the Natural Language Processing (NLP) field, but most NER methods require manually annotated data for training. Compared to the conventional NER approach, the proposed method has large merit in practical usage. As a related work, automatic method for building the machine translation (MT) dictionary has been proposed [11]. However, the method only has been applied to an MT's dictionary.

Since this proposed method does not require manually annotated data, users can register new words with minimum cost. Additionally, the method has high portability to proper noun specification updates, which can occur in the system development phase.

3 PROPOSED METHOD

The proposed method takes a supervised training approach using automatically built training data. Figure 1 shows the framework for the proposed method. The method consists of three parts: the training data building part, the training part and the classification part.

Firstly, the training data building part is constructed of training data from unsupervised learning (see 3.1). A speech corpus is required to build training data, but only a text corpus is required for actual word class classification in this proposed method. Secondly, by using text sentences including proper nouns and the obtained best class categories, the convolution neural network (CNN) based classifier is trained (see 3.2). Lastly, this CNN based classifier predicts the best class category for proper nouns by processing the sentence including the proper nouns CNN (see 4).

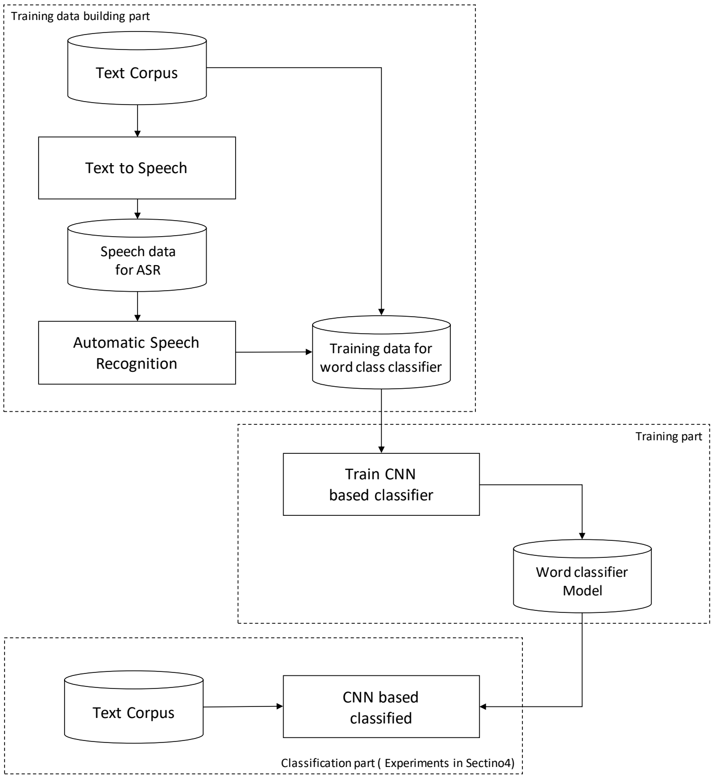

Training data building part
Text Corpus
Text to Speech
Speech data for ASR
Automatic Speech Recognition
Training data for word class classifier
Training part
Train CNN based classifier
Word classifier Model
Text Corpus
CNN based classified
Classification part (Experiments in Sectino4)

Fig. 1. Framework for the proposed method.

3.1 Training data building without human annotation

Figure 2 shows the flow of training data building part. Unsupervised learning is based on determining the word-class category that gives the best ASR performance. By calculating the Word Error Rate (WER) in sentence-units, the best word class is determined for each sentence. In order to evaluate ASR performance by WER, waveform data is necessary. Instead of using human speech, we use the TTS system to yield the synthesized waveform.

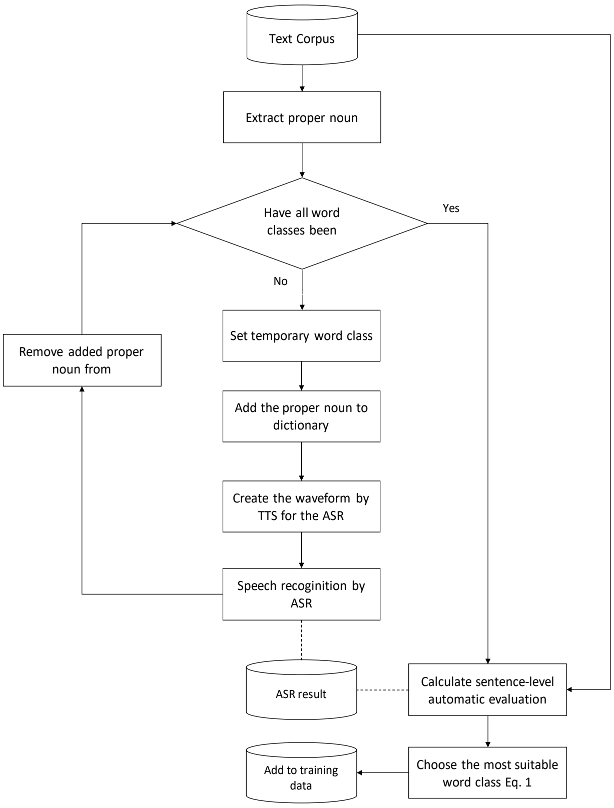

Text Corpus
Extract proper noun
Have all word classes been
Yes
No
Set temporary word class
Remove added proper noun from
Add the proper noun to dictionary
Create the waveform by TTS for the ASR
Speech recoginition by ASR
ASR result
Calculate sentence-level automatic evaluation
Add to training data
Choose the most suitable word class Eq. 1

Fig. 2. Flow of the training data building part.

The text corpus contains proper nouns. For each class category, the proposed method yields multiple ASR results by the following steps.

Step 1 Sentences including proper nouns are extracted from the text corpus.
Step 2 Extracted proper nouns are registered to the ASR and TTS dictionaries as one of the word classes.
Step 3 Speech is synthesized from the sentences by using TTS with the dictionary created in Step 2.
Step 4 The synthesized speech is automatically recognized by using ASR with the dictionary created in Step 2.
Step 5 The registered proper noun entry in Step 2 is removed from the ASR dictionary.
Step 6 Steps 2 to 4 are continued until all kinds of classes are registered to the dictionary.

By using the multiple ASR result and the original sentence in text, the most suitable word class ($\hat{\boldsymbol{c}}$) for the ASR performance is selected by the following formula.

$$\hat{\boldsymbol{c}} = \mathbf{argmin}_{c\in C}\boldsymbol{S}^{\boldsymbol{c}}_{\boldsymbol{WER}}(\boldsymbol{T}_{\boldsymbol{REF}}, \boldsymbol{T}^{\boldsymbol{c}}_{\boldsymbol{ASR}}) \tag{1}$$

Where $\boldsymbol{C}$, $\boldsymbol{T}_{\boldsymbol{REF}}, \boldsymbol{T}^{\boldsymbol{c}}_{\boldsymbol{ASR}}$ are the set of word class, reference, which is the transcription of speech, and an ASR output whose dictionary has entry of the target proper noun as word class *c*, respectively. And $\boldsymbol{S}^{\boldsymbol{c}}_{\boldsymbol{WER}}$ is the sentence-level WER score between $\boldsymbol{T}_{\boldsymbol{REF}}$ and $\boldsymbol{T}^{\boldsymbol{c}}_{\boldsymbol{ASR}}$, calculated by the following formula:

$$\boldsymbol{S}^{\boldsymbol{c}}_{\boldsymbol{WER}}(\boldsymbol{T}_{\boldsymbol{REF}}, \boldsymbol{T}^{\boldsymbol{c}}_{\boldsymbol{ASR}}) = (\boldsymbol{S}^{\boldsymbol{c}}_{\boldsymbol{ASR}} + \boldsymbol{D}^{\boldsymbol{c}}_{\boldsymbol{ASR}} + \boldsymbol{I}^{\boldsymbol{c}}_{\boldsymbol{ASR}})/\boldsymbol{N}_{\boldsymbol{REF}} \tag{2}$$

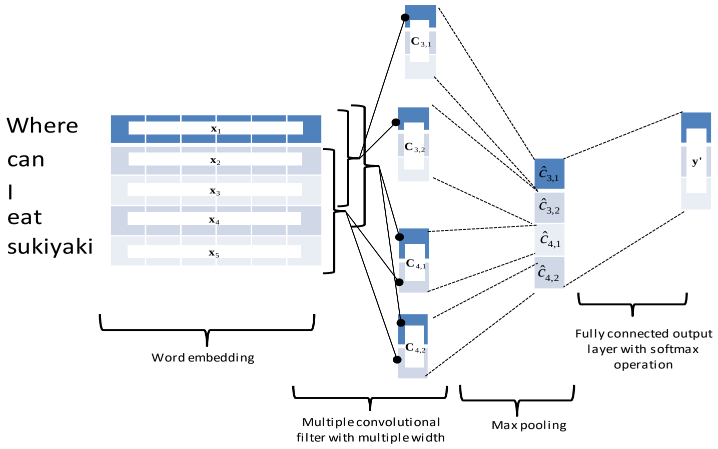
Where $\boldsymbol{S}^{\boldsymbol{c}}_{\boldsymbol{ASR}}, \boldsymbol{D}^{\boldsymbol{c}}_{\boldsymbol{ASR}}$ and $\boldsymbol{I}^{\boldsymbol{c}}_{\boldsymbol{ASR}}$ are the number of substation words, deletion error words, and insertion error words, respectively. And $\boldsymbol{N}_{\boldsymbol{REF}}$ is the total number of words in the reference. In the word class classifier training part, triplets of the proper noun, sentence and $\hat{\boldsymbol{c}}$ which is a teacher signal are used as training data.

3.2 Word class classifier training part

This subsection explains the method for training the word classifier using the training set established by the previous subsection. For proper noun classification, we train classifiers that are configured as the CNN. CNN gave superior performance in the research fields of image processing and speech recognition [12,13]. Currently, CNN has also outperformed the NLP task such as text classification [14,15] by incorporating word-embedding [16] in the input layer.

The network configuration for our proper noun categorization is shown in **Fig. 3.**.

Fig. 3. Convolutional Neural Network for Proper Noun Classification.


Where
can
I
eat
sukiyaki
x₁
x₂
x₃
x₄
x₅
C3,1
C3,2
C4,1
C4,2
ĉ3,1
ĉ3,2
ĉ4,1
ĉ4,2
y'
Word embedding
Multiple convolutional filter with multiple width
Max pooling
Fully connected output layer with softmax operation

Here, $x_i(\in R^k)$ is the word-embedding vector of the *i*-th word in a given sentence. By concatenating word-embedding vectors, a sentence whose length is n is expressed as the following formula:

$$\boldsymbol{x}_{\boldsymbol{1:n}} = \boldsymbol{x}_{\boldsymbol{1}} \cdot \cdot \cdot \boldsymbol{x}_{\boldsymbol{i}} \cdot \cdot \cdot \boldsymbol{x}_{\boldsymbol{n}} \tag{3}$$

The convolutional layer maps the *n*-gram features whose length (or filter window size) is h to the *j*-th feature map by using the following formula:

$$\boldsymbol{c_{h,j,i}} = \tanh(\mathbf{w}_{\boldsymbol{h,j}} \cdot \boldsymbol{x_{i:i+h-1}} + \boldsymbol{b_{h,j}}) \tag{4}$$

where $\mathbf{w}_{\boldsymbol{h,j}}$ and $\boldsymbol{b_{h,j}}$ are the weights for filtering and bias terms, respectively. For each n-gram length and feature map number, concatenate the results of Eq.4 as follows:

$$c_{h,j} = [c_{h,j,1}, c_{h,j,2}, \cdot \cdot \cdot \ c_{h,j,n-h+1}] \tag{5}$$

The max pooling layer chooses the element that has the largest value from all elements in $c_{h,j}$ as follows:

$$\hat{c}_{h,j} = \max_{i=1:n-h+1} c_{h,j} \tag{6}$$

The fully connected output layer uses the softmax operation to yield the probability distribution of the categories ($\hat{y} \in R^{n_c}$) as follows:

$$\hat{y} = \frac{\exp(Z_q)}{\sum_{p=1}^{n_c} \exp(z_p)}, q = 1, \cdot \cdot \cdot, n_c \tag{7}$$

Where n_c is the total number of categories. And, $z(\in R^{n_c})$ are the raw output values from the output layer.

4 EXPERIMENTS

4.1 Evaluation Method

In the experiments, first we build training data by using TTS and baseline ASR. Second, we train the CNN-based classifier using the training data. Third, by using the classifier, we build the Japanese proper noun dictionary for ASR.

As an evaluation of the classifier, we carried out speech recognition experiments using several dictionaries including a handcrafted dictionary and an automatically built dictionary by using the proposed method. In the evaluation experiments, we compare the word accuracy of the ASR systems in several dictionary conditions as follows.

Condition 1 Manual word class annotation by a human annotator.
Condition 2 Based on uniform distribution, randomly assign the word class
for each proper noun.
Condition 3 Based on the prior distribution shown in
Table 1. Details of word class in the data set.
.
Condition 4 Automatic word class annotation by the proposed method.

The evaluation experiments are carried out by the 4-fold cross validation manner.

Table 1. Details of word class in the data set.

Category	% in the data set
Accommodation	1.84%
Attraction	3.25%
Building	28.33%
Food	11.75%
Japanese first name	5.09%
Landmark	33.51%
Organization	1.93%
Shop	1.67%
Special name	0.61%
Souvenir	12.02%
Total	100.00%

4.2 Experimental Settings

As a test set of the ASR experiments, we use the Japanese part of speech data from the speech-to-speech translation system field experiments lead by, Global Communication Project funded by the Japanese Ministry of Internal Affairs and Communications. As for the speech data, TTS speech is used in the training data building part. Speech data from the field experiments is used for the ASR evaluation.

From the transcription of the field data, we extract 173 sentences that contain only one proper noun for each sentence. We use the set for our experiments. The unique number of proper nouns is 73. Some proper nouns are observed in multiple sentences in the set. For these kinds of sentences, we arranged for cross validation experiments not to occur in the training data or the validation set.

Table 1 shows the details of the word class specifications and occurrence in the training set. There are 10 word-classes that are related to the travel domain. In the data set, word-class of building names has the highest occurrence. Special names, which is used by animal names, plant names, character names and etc., have the lowest occurrence.

The ASR system used for the experiments has been developed by the National Institute of Information and Communications Technology (NICT). The ASR system is a submodule of the VoiceTra speech-to-speech translation system [2]. In these experiments we use the latest version of ASR. The system is composed of the Deep Neural Network--Hidden Markov Model (DNN-HMM) acoustic model and class-based n-gram language model by using the Weighted Finite State Transducer (WFST). In the language model of the system, around 100 word classes are manually defined for proper nouns. These classes include word classes related to several domains such as medical, travel, disaster and so forth. Except for proper nouns, all words are simply processed in word units by the language model.

Since the speech data set of our experiments is in the travel domain, we manually specified the word class subset related to the travel domain. Table 1 shows the specified subset. For our experiments, we use only the 10 classes shown in the table.

Before training the CNN-based classifier, we trained the word-embedding matrix using Word2Vec[14] on the Wikipedia corpus. The word-embedding matrix is fixed through CNN training. Details of these two corpora are shown in Table 2.

Table 2. Statistics of Training Corpora Used for the Experiments.

Corpus type	# of words	Lexicon size
Experiment corpus	1,572	430
Wikipedia corpus	10,363,151	116,556

Table 3 shows the detailed parameter setting of the CNN-based classifier. As shown in the table, the CNN has 10 output units, and each of them outputs the probabilities of the word classes, which are shown in Table 1. As shown in Table 1, data size varies depending on the word class. To reduce the adverse effect of data imbalances, we sampled training data to be balanced while mini-batch training.

Table 3. CNN Parameter Setting

Parameters	Setting
Maximum length of input sentence	150 words
Mini batch size	64
Dimension of word-embedding vector(k)	100
Filter windows size (ngram length)	3 to 5-gram
Number of filters for each window size	128
Dropout rate for fully connected layer	1
Optimizer	Adam optimizer
# of output units	10

4.3 Experimental results

Figure 4 shows the automatic evaluation results of speech recognition by ASR with several types of class-based dictionaries. In the figure, the vertical axis represents the automatic evaluation score calculated by word accuracy. For all conditions, we used the same class-based ASR system developed by NICT, and only the 10 classes shown in Table 1. The only difference is the way of annotating the word class for the proper noun dictionary. For Conditions 2 and 3, we performed 10 random annotations for each proper noun, then calculated the average score. The error bars in the figure show standard deviations over the 10 random trials. For Condition 4, we carried out 4-cross validations once on the data set.

As the figure shows, the proposed method gives much better performance compared to two random annotations. Additionally, the proposed method gives better performance than the manual annotation. Finally, the dictionary that registered the proper nouns by the proposed method has no adverse effects. The reason may be as follows.

Figure 5 shows the additional experiment. Proper nouns are correct or not in the ASR result. In this figure, the vertical axis represents the automatic evaluation score calculated by the accuracy of proper nouns. In this experiment, the proposed method's result is better than that of Conditions 1,2,3.

A human annotator decided the word category from the proper nouns only. The proposed method automatically annotates word class using a sentence including a proper noun, thus, the proposed method can use context information to annotate multiple proper noun categories. Such context information has one of advantages for the improving the accuracy of proper nouns.

Fig. 4. The word accuracy of sentences including proper nouns

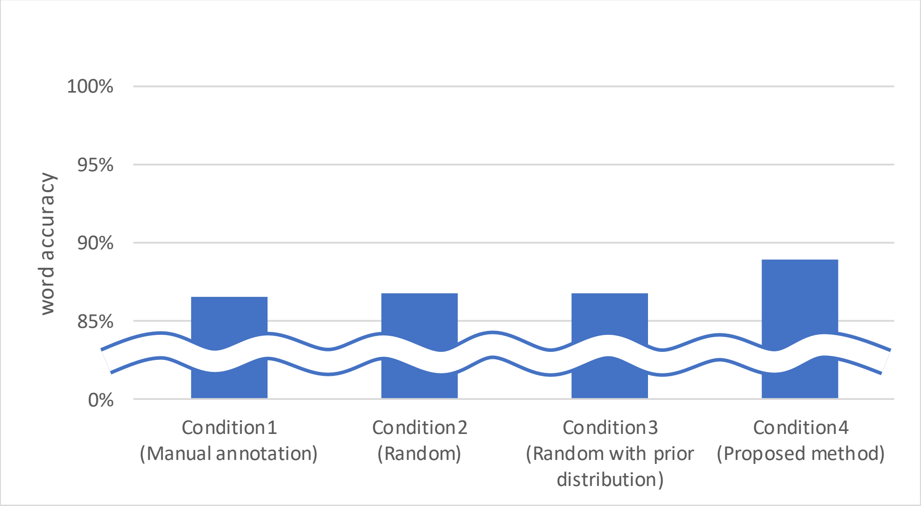
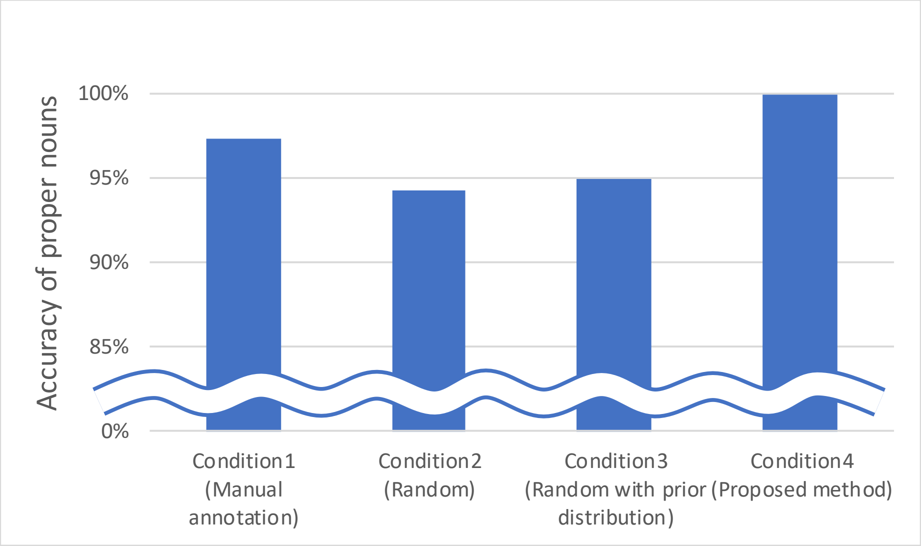


Fig. 5. The accuracy of proper nouns in the ASR result



5 CONCLUSIONS AND FUTURE WORK

According to the method of automatically building for a dictionary, users registered new words better than of Conditions 1. We proposed a method for building a dictionary for a class-based ASR system. We carried out experiments using speech data from the field experiments.

First, the training set for a CNN-based word classifier was automatically built. In order to build the training data without manual annotation, transcription of the field speech and synthesized waveform were used. By using the word classifier, proper nouns in the data set were automatically annotated and added to ASR dictionary. As an evaluation of the proposed method, speech recognition experiments were carried out. Here, ASR system with several dictionary conditions were compared based on word accuracy. According to the evaluation results, proposed method gave the highest speech recognition performance.

Since the proposed method did not require manual annotation and a speech recording, it enabled quick development of the class-based speech recognition system. Then, by using the context information to annotate multiple proper noun categories, the proposed method has advantages for proper nouns that had multiple meanings

The experiment shown in the paper was carried out on only Japanese part of the data from filed experiments of speech-to-speech translation system. As future works, we are conducting experiments on the other languages to show the effectiveness of the proposed method in several languages.

6 ACKNOWLEDGMENTS

This research is supported by the Japanese Ministry of Internal Affairs and Communications as a Global Communication Project.

References

1. Chiori, H., Kiyonori, O., Teruhisa, M., Hideki, K., and Satoshi, N.: Recent Advances in the WFST-Based Dialog System. Proc. of INTERSPEECH, pp. 268-271 (2009).
2. Shigeki, M., Teruaki, H., Yutaka, A., Yoshinori, S., Hidenori, K., Keiji, Y., Hideo, O., Masao, U., Eiichiro, S., Hisashi, K., and Satoshi, N.: Development of the “VoiceTra” Multi-Lingual Speech Translation System. Proc. of IEICE Transactions on Information and Systems, pp. 621-632 (2017).
3. Yamamoto, H., Hiroaki, K., Genichiro, K., Yoshihiko, O., Yoshinori, S.: Out-of-Vocabulary Word Recognition with a Hierarchical Language Model Using Multiple Markov Model.” Proc. of IEICE D-II 87(12), 2104-2111, 2004-12-01 (2004).
4. Lucian, G.: Recognition of Out-of-Vocabulary Words with Sub-Lexical Language Models. Proc. of EUROSPEECH, pp. 249-252 (2003).
5. Yamamoto, H., Shuntaro, I., Yoshinori, S.: Multi-class Composite n-gram Based on Connection.” Proc. of ICASSP, pp. 533–536 (1999).
6. Welly, N., Masatoshi, T., Seiichi, N.: Class Based N-Gram Language Model for New Words Using Out-of-Vocabulary Similarity.” Trans of Proc. of IEICE, pp. 2308–2317 (2012).
7. Yamamoto, H., Genichiro, K., Satoshi, N., Yoshinori, S.: Speech Recognition of Foreign Out-of-Vocabulary Words Using a Hierarchical Language Model., Proc. of INTERSPEECH, pp. 1870-1873 (2006).
8. Dzmitry, B., Jan, C., Dmitriy, S., Philemon, B., Yoshua, B.: End-to-end Attention-based Large Vocabulary Speech Recognition. Proc. of ICASSP, pp. 4945–4949 (2015).

9. Erik, F., Tjong, K, S., Fien, D, M.: Language-Independent Named Entity Recognition. Proc. of CoNLL (2003).
10. Sato, S.: Web-Based Transliteration of Person Names. Proc. of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp.273–278 (2009).
11. Yasuda, K., Panikos, H., Akio, I., Masayuki, H., Kazunori, M., Fumiaki, S.: Building a Location Dependent Dictionary for Speech Translation Systems." Proc. of CICLing (2017).
12. Alex, K., Ilya, S., Geoffrey, E, H.: Imagenet Classification with Deep Convolutional Neural Networks. Proc. of NIPS, pp. 1097–1105 (2012).
13. Ossama, A., Abdel-rahman, Mohamed., Hui, J., Li, D., Gerald, P., Dong, Y.: Convolutional Neural Networks for Speech Recognition. Proc. of IEEE/ACM, pp. 1533–1545 (2014).
14. Kim, Y.: Convolutional Neural Networks for Sentence Classification. Proc. of EMNLP, pp. 1746–1751 (2014).
15. Nal, K., Edward, G., Phil, B.: Convolutional Neural Networks for Modeling Sentences. Proc. of COLING, pp. 655–665 (2014).
16. Tomas, M,, Ilya, S., Kai, C., Greg, C., Jeffrey, D.: Distributed Representations of Words and Phrases and Their Compositionality." Proc. of NIPS, pp. 3111–3119 (2013).