Out-of-Task Utterance Detection Based on Bag-of-Words Using Automatic Speech Recognition Results

Yoko Fujita*, Shota Takeuchi*, Hiromichi Kawanami*, Tomoko Matsui†, Hiroshi Saruwatari* and Kiyohiro Shikano*
* Graduate School of Information Science, Nara Institute of Science and Technology, Japan
E-mail: {shota-t,kawanami,sawatari,shikano}@is.naist.jp
† Department of Statistical Modeling, The Institute of Statistical Mathematics, Japan
E-mail: tmatsui@ism.ac.jp

***Abstract*—Example-based question answering (QA) is an effective approach for real-world spoken dialogue systems. A limitation of an example-based QA is that a system cannot appropriately respond to a user's question, if a similar question-answer pair does not exist in the question and answer database (QADB). For a robust spoken dialogue system, it is important to classify if a user's utterance is in the task or out of the task. In this paper, we describe our approach for out-of-task utterance (OOT) detection. Using the Support Vector Machines (SVM), the detection model is trained with the bag of words from the 10-best automatic speech recognition (ASR) results. The number of words in a question, the number of unknown words, and the maximum similarity score against QADB are also used as features for the OOT detection. We apply our detection model to the *Takemaru-kun* dialogue system. We evaluate our detection model using adult's utterances of two years and child's utterances of one year spoken to *Takemaru-kun*. Our proposed method decreases the Equal Error Rate (EER) using speech recognition results by 4.4% (from 21.3% to 16.9%) in adult's speech and by 3.6% (from 31.8% to 28.2%) in child's speech, compared with the baseline method.**

I. Introduction

Automatic speech recognition (ASR) has been widely applied to dictation, Voice Search, and car navigation, to name a few.

In this paper, we describe the speech-oriented information guidance system *Takemaru-kun*, which aims to realize a natural speech interface using ASR [1].

Takemaru-kun is a real-environment speech-oriented information guidance system whose task domain is not given beforehand. It is an example-based question answering (QA) system that is flexible to respond to user's questions on demand. An answer to a user's question is selected by referring to the question and answer database (QADB), which can be easily maintained without paying particular attention to the scope of the system.

A serious problem in *Takemaru-kun* is that the system cannot respond to unexpected user's utterances, if a similar example of a question-answer pair does not exist in the QADB. From the analysis of user's utterances, about 5% of valid questions are not included in the QADB. As these questions cannot be answered fundamentaly by the system, they are treated as out-of-task (OOT) utterances. If OOT can detected automatically, the system can employ a strategy to send the question to the Internet search Engine and show the result as the next best way of response. Researches on out-of-task (OOT) utterances detection have been conducted, and training a model to classify ”in” or ”out-of” the task using the Support Vector Machines (SVM) has shown to be effective [2][3].

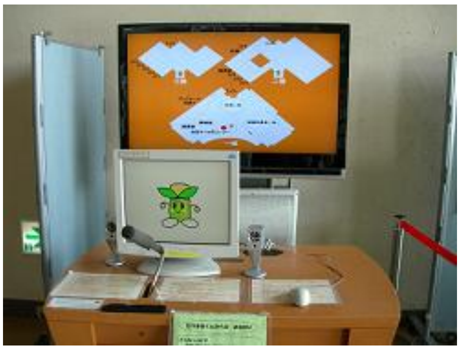

Fig. 1. Speech-oriented guidance system *Takemaru-kun*.

In this paper, we describe our out-of-task detection model, which is trained on the SVM using the bag-of-words from the 10-best ASR results. The outline of the paper is as follows. First, we describe the *Takemaru-kun* system. Then, our proposal, including effective features in the training of the model using SVM, and experimental results, using the real users' utterances, are described. Finally, we conclude our proposal.

II. Speech-oriented guidance system *Takemaru-kun*

A. System overview

Takemaru-kun is a speech-oriented information guidance system that has been in operation since Nov. 2002 at the entrance hall of *Ikoma City North Community Center* (Fig. 1) [1]. The system answers users' questions about the center facilities, services, neighboring sightseeing, agent profile and so on.

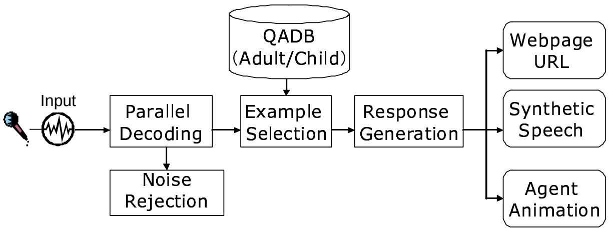

Fig. 2. Processing flow of *Takemaru-kun*.

The system employs a one-question-one-answer strategy. This approach is simple but achieves robust response generation. System response is provided by synthetic speech, Web browser and CG agent animation.

In the system, the tasks are not limited beforehand. As the system offers example-based question answering, the domains of response have been extended on demand of users. The QADB in the system consists of example questions and corresponding answer pairs. As the QADB can be easily updated by adding a question-answer pair, it is easy to expand the response domain or variety of phrases that appear in spontaneous speech.

The system structure is illustrated in Fig. 2. Speech/Noise discrimination using Gaussian Mixture Models (GMM) is executed in parallel with ASR. This process is regarded as the first step of the OOT detection in a broad sense. The GMMs are constructed from five kinds of real input to the system, which are adult's speech, child's speech, laughing, coughing and other noises. If likelihoods in any of the last three are the highest, the input is rejected as a noise.

The N-Best ASR result is used to calculate a similarity score with each example question. The nearest neighbor approach is employed for example selection using (1) [4]. The example question with the highest score is regarded as the user input and the corresponding response is used as the output message.

In the current operating system, users can enter the Voice Search mode by saying "*KeNsaku Kaishi.* (Start Voice Search)" before a Voice Search utterance. The OOT detection process proposed in this paper is planned to be implemented in the system, in which ASR result of OOT utterance is automatically sent to an Internet Search Engine as its query.

Similarity score =

$$\frac{\text{\# of word coincidences in } S_I \text{ and } S_E}{\max\ (\text{\# of words in } S_I, \text{ \# of words in } S_E\ )} \quad (1)$$

$$\text{subject to } S_I \in \ \{ \text{ Input utterances } \}\ ,$$

$$S_E \in \ \{ \text{ Example utterances } \}\ .$$

All system input have been collected from the start of operation. The data of the first two years were manually transcribed with tags concerning noise and labels about age-group and gender. The tags and labels are given by hearing of four trained labelers. These data were used to construct the GMMs and to adapt acoustic models and language models used in the daily operation.

TABLE I
EXAMPLES OF OOT TO *Takemaru-kun* (NOV.2002 TO OCT.2004)

Category	adult	child	Example
Trendy term	78	162	Do you know *Pokemon*?
Person/ organization	109	255	*Beckham*.
Technical term	224	578	Overweight guideline.
Too general term	159	987	Please show me rules of golf.
Ambiguous question	50	482	Six hundred and thirty-one?
Unprepared local info.	91	177	Please show me *Mayumi central park*.
Undefined property of the agent	183	4388	Can you play baseball?
TOTAL	894	7028	

B. Out-of-task utterance

As we mentioned before, in the *Takemaru-kun* system, the task domain is not defined beforehand, because it is preferable that the system to expand the domain to reflect real users' requests.

However, user's questions far from public information should not to be added to the QADB. Table. I illustrates the analysis result of OOT collected by the system in the first two years. We defined five OOT categories, which are *Trendy term*, *Person/organization*, *Technical term*, *Too general term*, *Ambiguous question*. In addition to that, *Unprepared local information* and *Undefined property of the agent* are also treated as OOT in this paper although they are useful for the users. It is because the responses for them are not prepared yet in the current QADB.

The total numbers of speech input are 20,436 and 85,889 from adult and child, respectively. Therefore about 4% of adult speech and 8% of child speech are OOT.

C. Conventional OOT detection in Takemaru-kun

In the future system of *Takemaru-kun*, OOT will be sent to a Web search engine as a query.

However, in the current system, we employ example-based OOT detection. In this approach, OOT examples are included in the QADB. The examples can be seen in Table I. As these OOT are paired with the answer, "I am sorry. I don't know," OOT detection process is included in the process of searching the nearest example question. In this paper, this example-based method is treated as the conventional OOT detection method for comparison.

III. OUT-OF-TASK UTTERANCE DETECTION

Here we describe the SVM-based method in which several kinds of features are combined.

A. SVM

The SVM is a useful machine for data classification [5]. It is basically a supervised leaning binary classifier. When training vectors $\mathbf{x_i} \in R^n, i = 1, ..., l$ (l : number of examples) and their corresponding classes $y_i \in \{1, -1\}$ are given, the SVM estimates a separating hyper-plane with a maximal margin in a higher dimensional space. The soft margin notation, which permits the existence of incorrectly classified data, is also introduced using a slack variable ξ_i and C parameters as

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=-1} \xi_i \qquad (2)$$

$$\text{Subject to } y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0, i = 1, ..., l.$$

C_+ : cost parameter for positive examples

C_- : cost parameter for negative examples.

In addition, C is divided into parameters C_+ and C_-, as we handle unbalanced data, which have large amount of in-task utterances but small amount of OOT ((3) to (5)). The value of C is set a posteriori.

$$C = C_+ + C_- \qquad (3)$$

$$C_+ = \frac{\text{number of examples in } y_i = -1 \text{ class}}{\text{Total number of examples}} \times C \qquad (4)$$

$$C_- = \frac{\text{number of examples in } y_i = 1 \text{ class}}{\text{Total number of examples}} \times C \qquad (5)$$

B. Features

The following features are used for OOT detection using SVM.
In addition to BOW, number of words in an input utterance and QADB score are picked up as features, which we introduced in our preliminary study to detect a wrong response. Frequency of unknown words is also investigated because it is directly related to OOT.

All features except *QADB score* are modeled and calculated using only ASR results.

Bag of Words vector consists of frequencies of each word in a wordlist. The wordlist consists of words appeared in *in-task* utterances of training data. The dimension of BOW vector is determined by the number of words in the wordlist.

Number of words

Frequency of unknown words is the frequency of words that are not included in the wordlist.

QADB score is the maximum similarity score calculated using (1) against the QADB, which consists of only *in-task* example questions of manual transcription.

The words used for BOW are defined in a *wordlist*. The wordlists used in this paper were constructed with the training data.

QADB score is a similarity score calculated with a QADB whose example questions are transcriptions of training data. OOT examples in the training data are excluded from the QADB. Leave-one-out approach is employed to calculate a QADB score for each sentence in training data.

IV. Experiments

To elucidate the effect of BOW built from ASR, evaluation experiments are conducted using real user's utterances.

TABLE II
Size of the experimental data

		In-task utterances	OOT	Total
Adult	Training Data	18,516	847	19,363
	Test Data	1,026	47	1,073
Child	Training Data	32,788	1,648	34,436
	Test Data	6,305	318	6,623

TABLE III
Experimental condition

		adult	child
ASR	Engine	Julius [6] 3.5.3	Julius 4.0.2
	AM, LM	*Takemaru-model*	
	Output	*10-Best candidates*	
Morphological analyzer		Chasen2.3.3	
Wordlist size (ASR)		5,416	6,294
Wordlist size (transcription)		2,423	4,200
SVM	Tool	LIBSVM [5]	
	Kernel function	Radial Basis Function (RBF)	
	parameter *C*	10, 100, 1000, 10000	

A. Speech database

Table II illustrates the speech database used in the experiments. Experiments with adult and child data are conducted separately. For the experiment with adult's speech, twenty-three months data are used as training data: Nov. 2002 to Oct. 2004, excluding Aug. 2003. The excluded one-month data are used as the test data. On the other hand, as the number of child's speech input to *Takemaru-kun* system is more than three times of adult's speech in average, only one-year data (Nov. 2002 to Oct. 2003) are used as training data and test data. The period of the test data is the same as adult test data (Aug. 2003). Invalid input to the system such as fillers, nonsense utterances are not included in the data.

B. Experimental conditions

The experimental conditions on feature extraction are illustrated in Table III. The acoustic model (AM) and the language model (LM) are separately prepared for adult and child. The AMs are trained using *The Japanese Newspaper Article Sentence database (JNAS)* and adapted by user data spoken to *Takemaru-kun*. The LMs are constructed using the manual transcription of the user's speech to *Takemaru-kun*.

Morphological analyzer *Chasen* is used to split the transcription data into words.

Word accuracy and word correct rate of the speech database are shown in Table IV. It can be seen that word accuracy of OOT is remarkably lower than that of in-task utterances. Especially in child's speech, recognition rates degrade about 20% in OOT from those of the in-task utterances. In addition to that, differences between training data and test data are larger in OOT, especially in adult's speech. These results indicate a possibility that the characteristics of BOW using ASR results are also different between in-task utterance and OOT.

In the following experiments, 10-best ASR candidates are used for constructing wordlists and BOW vectors. The size of the wordlists are also indicated in Table III. The wordlists built by transcription are also indicated. It can be assumed that as

TABLE IV
WORD ACCURACY AND WORD CORRECT RATE OF THE EXPERIMENTAL DATA

		In-task utterances		OOT	
		Acc. (%)	Corr. (%)	Acc. (%)	Corr. (%)
Adult	Training Data	90.1	92.2	79.9	83.5
	Test Data	89.7	92.1	84.2	88.9
Child	Training Data	71.0	75.2	50.5	57.0
	Test Data	69.9	74.8	48.2	55.0

ASR results contain 10-Best recognition results, the wordlists of ASR are larger than that of transcripts. It is expected that characteristics of OOT can be modeled including ASR error tendency by using ASR results.

C. Evaluation measures

We evaluated discrimination accuracy on the basis of the Equal Error Rates (EER) criterion. In binary classification, there are two types of errors. One is false acceptance rate (FAR), which is the percentage of OOT classified into the in-task domain. The other is false rejection rate (FRR), which is the percentage of in-task utterances classified into OOT. EER is the error percentage at which FAR and FRR become equal.

The hyper-parameters C are set a posteriori.

D. Experimental results

In Figs. 3 and 4, the EERs of the conventional method (QADB with OOT question examples), SVM using BOW, and SVM using BOW and other three features (Number of words, Number of unknown words, QADB score) are listed.

It is shown that BOW consisting of ASR results are effective for OOT detection. EERs are decreased from the QADB-based conventional method by 4.4% (from 21.3% to 16.9%) in adult's speech and by 3.8% (from 31.8% to 28.0%) in child's speech, compared with the baseline. The proposed approach overcomes the conventional method under the condition that the word accuracy is even 70% or less. (Table IV).

In addition to that, EERs decrease to 13.0% and 25.8%, when using Number of words, Number of unknown words and QADB score together. As we mentioned in Section III, all features except for QADB store are obtained from speech automatically. In this paper, QADB scores are calculated referring the QADB with manual transcription. This QADB can be replaced with a QADB with ASR results as example questions because it is shown that introducing ASR results to QADB is effective, especially when large scale speech data is obtained [4]. Therefore, all the features used here are assumed to be obtained automatically.

V. CONCLUSION

ASR result is employed to construct BOW, which is introduced to OOT detection based on SVM. The OOT detection rate is evaluated using user's speech received by the spoken dialogue system *Takemaru-kun*.

The experimental result confirms that SVM using BOW consisting of ASR results decreases EERs both in adult and child data.

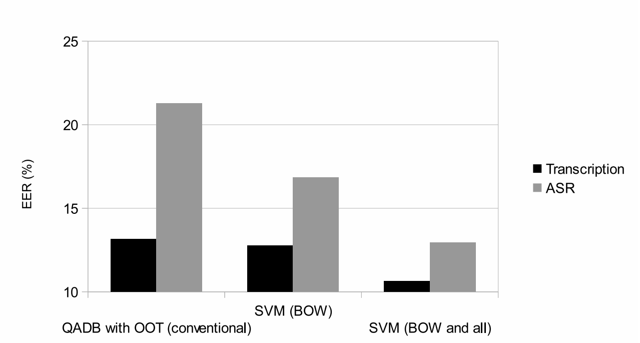

Fig. 3. Results of experiment (adult)

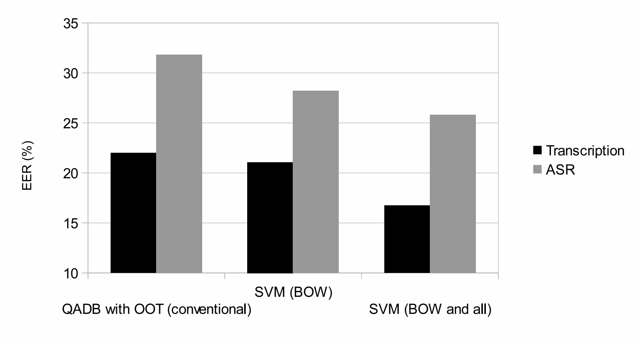

Fig. 4. Results of experiment (child)

In the future work, how to treat several feature vectors with different attribute, will be reported from the viewpoints of vector combining method, multiple kernel method and scaling of feature values.

ACKNOWLEDGMENTS

This work was partially supported by CREST (Core Research for Evolutional Science and Technology), Japan Science and Technology Agency(JST).

REFERENCES

[1] R. Nisimura, A. Lee, H. Saruwatari and K. Shikano, "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," *In Proc. ICASSP2004*, vol. 1, pp. 433-0436, 2004.

[2] I. Lane, T. Kawahara, T. Matsui and S. Nakamura, "Out-of-Domain Utterance Detection Using Classification Confidences of Multiple Topics," *IEEE Trans. Audio, Speech & Language Process.*, Vol. 15, No. 1, pp. 150–161, 2007.

[3] K. Komatani, S. Ikeda, T, Ogata and H. Okuno, "Managing Out-of-Grammar Utterances by Topic Estimation with Domain Extensibility in Multi-Domain Spoken Dialogue Systems," *Speech Communication*, Vol. 50, pp. 863–870, 2008.

[4] S. Takeuchi, H. Kawanami, H. Saruwatari and K. Shikano, "Question and Answer Database Optimization Using Speech Recognition Results," *INTERSPEECH2008*, pp. 451–454, 2008.

[5] C.-C. Chang, C.-J. Lin "LIBSVM : A Library for Support Vector Machines," http://www.csie.ntu.edu.tw/˜cjlin/libsvm, 2001.

[6] A. Lee, T. Kawahara and K. Shikano, "Julius - an Open Source Real-time Large Vocabulary Recognition Engine," *Proc. Eurospeech2001*, pp. 1691–1694, 2001.