Unknown Example Detection for Example-based Spoken Dialog System

Shota Takeuchi, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology
{shota-t,kawanami,sawatari,shikano}@is.naist.jp

Abstract

In a spoken dialog system, the example-based response generation method generates a response by searching a dialog example database for the example question most similar to an input user utterance. That method has the advantage of ease of system expansion. It requires, however, a number of utterance examples whose correct responses are labeled. In this paper, we propose an approach to reducing the system expansion cost. This approach employs a detection method that screens the unknown examples, the utterances to be added to the database with their correct responses. The experimental results show that the method can reduce the number of utterances required to be labeled while maintaining the system response accuracy improvement as well as full labeling.

1. Introduction

Spoken dialog systems have been investigated in the field of the interfaces on some information search systems [1, 2, 3, 4, 5]. However, few case studies of long-term operation have been reported [6].

In view of a search tool, a dialog system is required to work stably. In view of medium of information or communication, the system is required to be updated the dialog data to deal with the latest information. Especially in long-term operation, reflecting users' interest in the system response (e.g. local information and seasonal information) is important [7]. Thus it is important to choose a system framework that can easily update the contents it provides.

The example-based spoken dialog framework [3, 4, 5] has been investigated for the sake of easy system expansion. This system employs the dialog example database to select the most similar example to the user input speech (recognition results). The framework has two advantages regarding system maintenance. The first is ease of database construction. The second is the ability to treat utterances that are hard to be described by a dialog model. Especially on a one-question-one-answer dialog system [3], this framework only requires the example and response pairs (QA pairs), which are even easy to construct. Nevertheless using this framework, the expansion cost of the dialog system is still expensive since example construction includes utterance labeling costs. Therefore, a method that reduces the expansion cost is important for continuous operation of a dialog system.

In the following sections, we discuss the labeling cost reduction method using a similarity score. We also show the experimental results of reducing the number of newly labeled expansion data.

2. Speech-oriented information guidance system "Takemaru-kun"

We have been operating a spoken dialog system "Takemaru-kun" [3] in the north community center of Ikoma City for over 6 years to investigate a real-environment spoken dialog system. Through the long-term operation, we have also constructed a spontaneous speech database with speech information tags (transcription, age groups, correct response, etc.)

2.1. Dialog strategy and topics

A spoken dialog system "Takemaru-kun" has been installed at the community center.

To make it familiar to a novice user, this system uses the agent character "Takemaru-kun," and the dialog is designed to be simple and friendly. Thus, the dialog strategy is based on the one-answer-one-question dialog strategy, the response message is made simple, and its characteristics are also designed.

Takemaru-kun mainly provides the brief guidance about the facilities of the community center and local sightseeing, and it also provides simple chat-like self-introduction and latest information such as current time, weather or news. Recently, we have improved the web search task by constructing a language model using a web keyword corpus [7].

2.2. Database

All system inputs have been collected from the initial operation. We have constructed a spontaneous speech database from the first two years of collected speech data. In this paper, the database is employed in the experiments, and the features of that database are described in the section 5. More details of the database features are described in the references [7, 8]. We have also been collecting the unlabeled spontaneous speeches for four years and those data are available for an unsupervised training of an acoustic model [6].

3. Example-based spoken dialog system

An example-based spoken dialog system mainly employs an example database that stores pairs of user utterance information and its correct (appropriate) system response. When a user input occurs, the system selects the most similar dialog example to the user input and generates its response [Fig. 1].

Common frameworks of the system [1, 2] have these features; (a) multi-turn dialog is assumed as dialog strategy, (b) the example data includes a user input, (c) the system states (dialog context), the correct response, and the example database are often constructed with RDBMS (relational database management system). In the answering process, a search query is generated by the user input parsing to search dialog examples.

As for the framework in Takemaru-kun, one-question-one-answer dialog is assumed. Thus the framework is much simpler than the common one: the example data does not contain a system state, and the user input is compared with the example question directly so that the RDBMS is not employed. In Takemaru-kun, that example database is called "question and answer database," QADB. This framework can avoids the input parsing error that occurs when the speech recognition fails, especially when using ASR-QADB methodology, which uses the automatic speech recognition (ASR) results as the example question data [9].

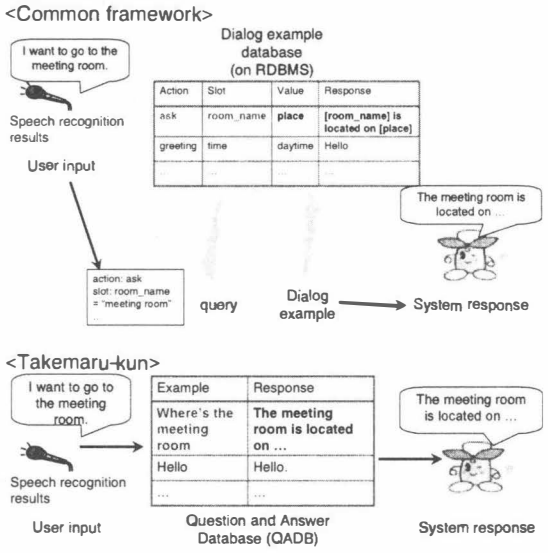

Figure 1. Response generation flow in example-based spoken dialog system.

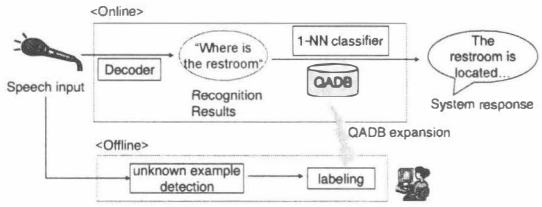

Figure 2. Spoken dialog system operation with QADB expansion.

4. Unknown example detection

For QADB expansion, we collect the example from the user speech for reflecting the users' interests. Temporally for QADB expansion, we collect examples from the user speech to reflect the users' interests. However, it is inefficient to manually attach transcription and a correct answer for all of the speech since some utterances may exist in the QADB. Therefore, we propose a method to detect examples that are "unknown" (not included) in the existing QADB examples [Fig. 2].

In our framework, the similarity of input and example are calculated on the surface expression level (bag-of-words). Thus, for the QADB expansion, it is enough to judge the similarity of collected utterances and existing example questions in the initial QADB.

4.1. QADB expansion with unknown example detection

At the first, noise data are screened out of the collected user speech. Next, valid utterance data are processed by the ASR engine. Then the utterances are screened by unknown example detection using their ASR results.

The detected unknown utterance will be labeled its correct response by a human labeler. ASR-QADB methodology also avoids the transcription process.

4.2. Similarity score

The INMS (intersection normalization by maximum size) method is employed to calculate the similarity score [9]. This method is based on word-matching scoring, and it calculates the ratio of corresponding words over the maximum word count of the two sentences. This score takes the value between 0.0 and 1.0, and the higher value is meant to be more similar between two sentences. If the maximum score of a sentence between QADB examples becomes lower than the pre-defined threshold, the sentence is judged as "unknown."

When the threshold is set lower, the labeling cost (amount of detected unknown utterances) becomes lower, although the true unknown examples will be rarely appended (thus, the system response accuracy will not improve). The proper threshold is on the trade-off point where the labeling cost becomes lower while the response accuracy improves.

5. Experiments

5.1. Procedures

We prepare the three valid utterance datasets, initial QADB construction data, QADB expansion data and the evaluation data. In the experiment, firstly, the unknown example utterances are detected from the QADB expansion data. Next, the expansion QA pairs are constructed by labeling the correct response to detected utterances. Then the updated QADB is constructed by appending the expansion QA pairs to the initial QADB. Finally, the system evaluations are performed on the initial QADB and the updated QADB.

The system evaluation measure is response accuracy and gross detected inputs. Response accuracy is the rate of correctly answered evaluation data by the evaluated system. Gross detection ratio is the ratio of detected utterances in the expansion data. Gross detection ratio is meant to show the utterance labeling cost.

Table 1. Dataset features

Evaluation data	Period	Aug. 2003	
	# of data	Adult	1,053
		Child	6,543
Initial training data	Period	Nov. 2002 – July.2003	
	# of data	Adult	10,588
		Child	28,440
Additional data	Period	Sep. 2003 – Oct.2004	
	# of data	Adult	8,795
		Child	50,906
	Ratio of un-known ex.	Adult	32.2%
		Child	40.2%

Table 2. ASR conditions

ASR engine	Julius[10] Ver 3.5.3, 10-best-sentence output.
Acoustic Model	Retrained JNAS model with 2-year Takemaru speech data 3-state L2R HMM, 2,000 triphone PTM, 64 GMM per state
Language Model	Corpus: Transcription of 2-year Takemaru speech data, morphological analysis with ChaSen[11] Smoothing: Witten-Bell method

In the experiments, we use the valid utterances of "Takemaru-kun" speech data that have already been labeled with the correct responses. See the next section.

5.2. Dataset

We employ the "Takemaru-kun speech database" for the experiments. This database consists of the user speech collected by the dialog system "Takemaru-kun," discussed in the section 2. The numbers of the responses in the whole utterance database are 275 and 285 for adult and child utterances, respectively. In the experiment, we divide the valid utterance data into three parts, 9 months of data for the initial QADB, 1 month for the evaluation data, and 14 months for the QADB expansion data [Table 1]. ASR conditions are shown in Table 2.

5.3. Results and Discussions

Figure 3 shows the relations of the response accuracy and the gross detected inputs for each score threshold of the detection method. Each point in the chart shows the result on each step of score threshold $\theta = [0.0, 1.1]$ with 0.1 steps. $\theta = 1.1$ means "over 1.0" in order to treat all the expansion data as unknown. For comparison, manually transcribed data are also evaluated as reference results.

About adult speech data, $\theta = 0.8$ brings a cost reduction to about 50%, with 3.3 points of response accuracy improvement, which is similar to full labeling ($\theta > 1.0$). $\theta = 0.4$ brings a cost reduction to 15% (ratio of about a day per week working period), with 1.3 points of response accuracy improvement. On the other hand, the response accuracy for child data is saturated even if the initial QADB is expanded, since that initial QADB has sufficient examples.

Considering the experimental result, the cost reduction of QADB expansion for the long-term operation is achievable with our unknown example detection method.

6. Conclusions

Labeling cost reduction is desired for the long-term operation of the spoken dialog system. Detecting the unknown example utterances will reduce the labeling cost to expand the QADB while maintaining improvement of the response accuracy.

7. References

[1] A.L.Gorin, G. Riccardi, J.H.Wright, "How may I help you?," Speech Communication, vol.23, pp.113-127, 1997.

[2] Victor Zue, Stephanie Seneff, James R. Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen and Lee Hetherington, "JUPITER: A Telephone-Based Conversational Interface for Weather Information," IEEE Transaction on Speech and Audio Processing, vol.8, no.1, pp.85-96, 2000.

[3] Ryuichi Nisimura, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano, "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability," Proc. of ICASSP2004, vol.1, pp.433-436, 2004.

[4] Hiroya Murao, Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, Kazuya Takeda, Yasuyoshi Inagaki, "Example-based Spoken Dialog System with Online Example Augmentation," Proc. of ICSLP2004, Spec4402p-7, pp.3073-3076, 2004.

[5] Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, Gary Geunbae Lee, "Example-based dialog modeling for practical multi-domain dialog system," Speech Communication, vol.51, is.5, pp.466-484, 2009.

[6] Tobias Cincarek, Izumi Shindo, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Development of Preschool Children Subsystem for ASR and QA in a Real-Environment Speech-oriented Guidance Task", in Proc. of EUROSPEECH 2007, pp. 1469-1472, 2007.

[7] Jumpei Miyake, Shota Takeuchi, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, ``Language Model for the Web Search Task in a Spoken Dialog System for Children," in Proc. of WOCCI, Chania, Greece. October 2008.

[8] Shota Takeuchi Cincarek Tobias, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, " Construction and Optimization of a Question and Answer Database for a Real-environment Speech-oriented Guidance System, " Proc. Of Oriental COCOSDA 2007, pp.149-154, 2007.

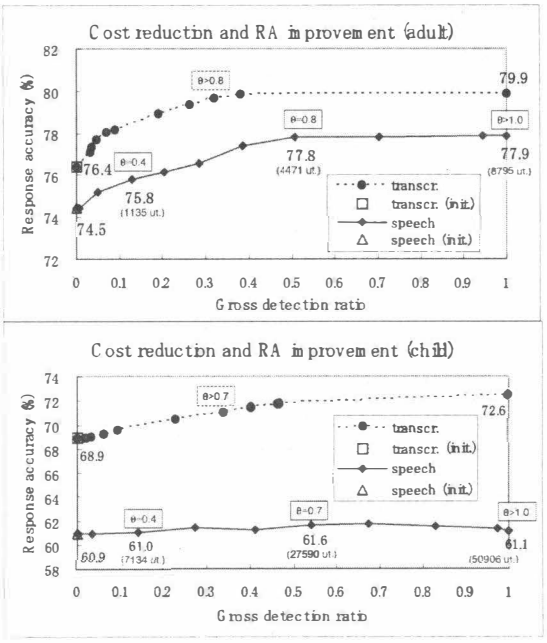

Figure 3. Labeling cost reduction with response accuracy (RA) improvement.
Dotted line with round points shows the reference evaluation (transcription input with transcription QADB, shown by "transcr."). Solid line shows the actual evaluation (ASR result input with ASR-QADB, shown by "speech"). Hollow symbols show the results of initial QADB. Each point shows the result on each step of predefined score threshold $\theta = [0.0, 1.1]$ with 0.1 steps. The number under the point shows the RA on the predefined threshold and the parenthesized number shows the number of detected expansion utterance data.

[9] Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, "Question and Answer Database Optimization Using Speech Recognition Results," INTERSPEECH 2008, pp.451-454, Sep, 2008.

[10] http://sourceforge.jp/projects/chasen-legacy/ .

[11] Akinobu Lee, Tatsuya Kawahara, Kiyohiro Shikano, "Julius --- an Open Source Real-Time Large Vocabulary Recognition Engine," in Proc. EUROSPEECH 2001, pp.1691-1694, 2001.