Question and Answer Database Optimization Using Speech Recognition Results

Shota Takeuchi1, Tobias Cincarek1, Hiromichi Kawanami1, Hiroshi Saruwatari1, Kiyohiro Shikano1

1 Graduate School of Information Science, Nara Institute of Science and Technology 
{shota-t, cincare-t, kawanami, sawatari, shikano}@is.naist.jp

Abstract
The aim of this research is a human-oriented spoken dialog system which provides replies to a variety of users’ utterances. The example-based response generation method searches a question and answer database (QADB) for the example question most similar to a user utterance. With this method, the system can answer a question difficult for a model to express. A QADB is constructed from question and answer pairs (QA pairs) by employing a large corpus. In order to enhance robustness to recognition errors of inarticulate utterances such as children utterances, we propose to use speech recognition results, instead of manual transcriptions, as example questions. We also introduce an optimization method that removes inappropriate QA pairs from a QADB to maximize response accuracy. We show that our method improves the response accuracy of utterances especially for children utterances in the open test.

Index Terms: spoken dialog system, example-based response generation, QADB optimization, leave-one-out cross-validation

1. Introduction
As automated speech recognition (ASR) technology has improved, real-environmental operation of spoken dialog systems has been realized. Many studies on spoken dialog systems have been investigated in terms of methodology, theory and real-environmental operation [1, 2, 3, 4, 5, 6, 7].

The authors have been operating a speech-oriented guidance system “Takemaru-kun” [8] (Fig. 1) at a community center for over five years and have collected a large amount of spontaneous utterance data. Additionally, a speech database has been constructed from 2-year speech data by manual labeling speech information.

The main task of the system is information guidance in local sightseeing. Response accuracy of the current system is about 74% for adult users and about 55% for children [9]. This performance is calculated automatically by labeling one correct response; the subjective performance is about 5% higher. This performance shows that our system works favorably in a real-environment for adults. However, it is necessary for that system to improve the response accuracy more for children’s speech. Long operational results show that users, especially children, utter sentences beyond our expectations.

Typical spoken dialog systems perform a single task or employ a task classifier to confirm a correct response. These task-oriented systems are adept at accepting utterances which developers’ expect, but they are inadequate for handling utterances beyond developers’ expectations [9, 10].

To deal with such unforeseen utterances, an example-based response generation method has been studied [4, 8]. This method employs a database which consists of pairs of example question and response (answer). The pair is called a QA pair and the database, QADB. When speech is inputted into the spoken dialog system, this method finds the QA pair which obtains the highest similarity with the input speech through large vocabulary continuous speech recognition (LVCSR) and generates the response of that QA pair.

This method can deal with a variety of utterances without employing task classification. Simply by adding a QA pair to the QADB, the system can answer a question difficult for a model to express. A large amount of 2-year utterance data also enables the system to answer a variety of utterances.

On a conventional example-based response generation, however, there are two problems. The first problem is lower response accuracy with children’s speech. This is due to the greater number of errors in children’s speech, since speech recognition process in an example-based approach is not robust enough to the inarticulate utterances of children. Therefore, we propose to use recognition results as example questions so that matching errors of inputs and examples decrease.

The second problem is interleaving of QA pairs which are labeled inappropriately or which mislead a user input. For the second problem, we have proposed a method to remove those QA pairs automatically by QADB optimization [10]. QADB optimization decides whether each QA pair is useful or not for the system by calculating the improvement of response accuracy when that pair is removed. In this paper, we show the effectiveness of our two methods experimentally.

2. Takemaru-kun speech database
2.1. Spoken dialog system “Takemaru-kun”
Spoken dialog system “Takemaru-kun” is installed in a community center in Nara, Japan, to provide information on the com-
munity center and local sightseeing, etc. This system is also aimed at collecting users’ spontaneous utterances and finding seeds of our researches [9, 11].

Major features of this system are using a character agent, simple dialog and age group classification. First, the system is designed for the user to talk with the character agent “Takemaru-kun”. It makes novice users feel familiar with the dialog system. Second, considering the large number of users, the system works on a one-question-to-one-answer dialog strategy for each utterance. This strategy is fit for that system because it mainly answers simple questions about the community center and local sightseeing. It also performs small talks such as greetings and personal information of the agent. Third, the system executes age classification simultaneously with speech recognition to find a suitable answer for both adult and child.

Takemaru-kun employs an example-based response generation (Fig. 2). Each example is assigned to one response message to form a QA pair. This system does not perform task or domain classification explicitly, but it selects a QA pair and presents the response. Responses corresponding to some topics of the day (e.g. current time or weather) are implemented to have slots. The slot is filled during the response presentation process.

2.2. Utterance database features

We have constructed a speech database of “Takemaru-kun” by manual labeling speech information such as estimated age group and gender, transcription, and correct response. The timestamp is available for an identifier of a speech file. We also constructed an utterance database by screening out invalid speeches, which cannot be understood. This paper deals with age group, transcription and correct response tags as the utterance database tag.

3. Response generation using recognition result QADB

3.1. Response generation flow

The example-based response generation method searches a QADB for the example question most similar to user utterances. The system performs three processes to generate a response: automated speech recognition (ASR), example selection, and response generation. First, when a user speech is inputted, the system executes a ASR to generate recognition results of N-best candidates (a set of plain sentences). Second, this system selects a QA pair among the recognition results which is the most similar to the example question. The similarity is based on word matching (discussed in 3.3). At the last step, the system presents the response of the selected QA pair. For responses which concern topics of the day, the system modifies the response before the presentation.

3.2. ASR-QADB

The QADB is constructed from the utterance database by removing duplicate QA pairs (Fig. 3). With the conventional method, the example question of each QA pair employs a manual transcription. To address mismatching between user input and example question, we propose to use the speech recognition results as example questions. This strategy is also expected to reduce system development cost since manual transcription is not required.

In this paper, we use the N-best candidates of the speech recognition results. Hereafter, the QADB whose example questions are recognition results is called ASR-QADB and the QADB whose examples are manual transcriptions is called Transcription QADB for distinction. The ASR processing of the utterances is performed before removing duplicate QA pairs.

3.3. Similarity between recognition results

Our example selection method uses the N-best candidates for a representation of user inputs and example questions. We define similarity of two bag-of-words based on word matching.

Similarity $s(I, E)$ is defined below as:

$$s(I, E) = \frac{\sum_{k \in I \cup E} \min(w_I(k), w_E(k))}{\max(n_I, n_E)}$$

where $I$ is a set of bag-of-words of input and $E$ is the example, $w(k)$ is the average count of word $k$ per sentence, and $n$ is the average number of words per sentence.

The min operator of the numerator expresses the calculation of average count of matched words per sentence. The max operator of the denominator expresses the maximum of composed bag-of-words size by an input and an example (Fig. 4).

3.4. Response accuracy

Response accuracy is the rate of correct answers of a test set given a QADB. In order to calculate the response accuracy automatically, we use the utterance data to which has a corresponding correct response.
Response accuracy $R(T|Q)$ is defined below as:

$$R(T|Q) = \frac{R_e(T|Q)}{N_T}$$  \hspace{1cm} (2)$$

where $T$ is the test data, $N_T$ is the number of test data, $Q$ is the given QADB, and $R_e(T|Q)$ is the number of utterances which the system can answer correctly.

Response accuracy can be calculated by three ways: open test, closed test, and cross-validation test. The cross-validation test is performed in these steps: splitting a data into subsets, calculating response accuracy of one of the subsets when a QADB is constructed from the other subsets, and averaging the results for all subsets.

We define the term leave-one-out cross-validation response accuracy (LOOCV response accuracy), $R^{(cv)}(U|Q)$ in this way: splitting data into one sample, taking a system response of the sample for a QADB from the rest of the data and averaging the number of correct system utterance data (see Algo. 1).

4. QADB optimization

Generally, QADB optimization employs two datasets: QADB construction data and training data. The training data is a data for calculating response accuracy for QADB optimization. The QADB optimization aims at maximizing response accuracy of the training data automatically. The response accuracy for QADB optimization is calculated using speech recognition results. We have proposed QADB optimization using LOOCV response accuracy [10].

This method detects the QA pairs which make the response accuracy lower. The optimizer calculates response accuracy of the training data for a temporal QADB where one QA pair is removed from the original QADB. If the response accuracy increases by removing a QA pair, that QA pair is marked with "useless". We note that the useless QA pair "returns" to the temporal QADB on the next validation step. After all of the QA pairs are validated, the optimal QADB is created by removing all useless QA pairs once. It is supposed that QADB optimization by combinatorial optimization runs along a greedy algorithm about the response error.

Optimization of QADB $Q$ for training utterance data $U$ is formulated below as (see Algo. 2):

$$Q = \{ \text{Useless QA pair} \}, \quad R(q) = \text{Increase in response accuracy when the QA pair } q \text{ is removed}$$  \hspace{1cm} (3)$$

$$R(U|Q_{\text{opt}}) = R(U|Q) + \sum_{q \in Q} -R(q)$$  \hspace{1cm} (4)$$

where $Q$ is a set of useless QA pair, $R(q)$ is the increase in response accuracy when the QA pair $q$ is removed. $\theta$ is the response accuracy threshold with which a QA pair is judged useless, $Q_{\text{opt}}$ is the optimal QADB, and $R_e(U|Q_{\text{opt}})$ is the response accuracy of the training data for the optimal QADB.

Algorithm 1 Calculating LOOCV response accuracy.

```python
for all $u \in U$ do
    $r_c = 0$
    $R_c(u|Q_{\text{opt}}) = 0$
    for all $q \in Q_{\text{opt}}$ do
        if $R(q) > 0$ then
            $R(q) = 0$
        end if
        $r_c = r_c + 1$
    end for
    $R_e(u|Q_{\text{opt}}) = \frac{r_c}{|U|}$
end for
```

In this paper, $-\theta = 0$, thus the response accuracy of the training data monotonously increases owing to $-R(q) \geq 0$.

5. Experimental evaluations

5.1. Experimental conditions

By introducing LOOCV response accuracy for QADB construction and optimization, the same utterance database was employed. Since it is supposed that the proposed method makes more improvement with each iteration, response accuracy for the optimized QADB is calculated after each iteration step. We have already confirmed the increase of LOOCV response accuracy of the training data using the proposed optimization method [12].

A series of evaluation is also performed to evaluate the effectiveness of the proposed method. The "Takemaru-kun speech database" is employed both for the QADB construction and for the optimization. The "Takemaru-kun speech database" is employed both for the QADB construction and for the optimization. The number of QADB construction (and training) data and evaluation data is 19,383 and 1,053 for adults, 79,346 and 6,543 for children, respectively (see Table 1). The number of response is 275 for adult and 285 for child. Three N-best ASR-QADB are constructed for experiments ($N = 3, 6, 10$).

The ASR engine "Julius Ver. 3.5.3" [13] is employed for the experiments, using different acoustic and language models for adults and children's speech. The speech recognition accuracy is shown in Tab. 2. As our system treats Japanese utterances, we treat Japanese morpheme as the unit of the bag-of-words of user input and QADB example questions. N-best recognition results are separated into morphemes by morphological analyzer "ChaSen."  

$^1$http://julius.sourceforge.jp

$^2$http://chasen-legacy.sourceforge.jp
5.2. Results and discussions

The results of the evaluation show that the response accuracy increased by optimization of the transcription QADB and the proposed ASR-QADB. Figure 5 shows that the response accuracy increases and converges as QADB optimization is repeated. Table 3 shows the response accuracy of the initial QADB and the optimal QADB. The latter gains the best response accuracy. In terms of adult speech input, the maximal response accuracy of the transcription QADB has improved better than the ASR-QADBs. For children's utterances, however, the maximal response accuracy of ASR-QADB has improved better than the transcription QADB. Thus it is suggested that ASR-QADB with QADB optimization is more effective for utterances which may contain more recognition errors such as child utterances.

6. Conclusions

The example-based response generation can deal with a variety of utterances. This method searches a QADB for the example question most similar to a user utterance. We proposed to use the speech recognition results for example questions in order to enhance robustness to recognition errors. The strategy is effective development of advance robotics elemental technologies in Japan.

7. Acknowledgements

This work was partly supported by the NEDO project for strategic development of advance robotics elemental technologies in Japan.

8. References


---

| Table 2: Speech recognition accuracy of utterance data in experiments. |
|------------------------|------------------------|
| **QADB & Training**   | **Evaluation**         |
| adult                  | 90.26%                 | 85.66%                 |
| child                  | 74.44%                 | 66.81%                 |

| Table 3: Response accuracy of the initial QADB and the optimal QADB. |
|------------------------|------------------------|
| **Transcr. QADB**      | **10-best ASR-QADB**   |
| initial                | initial                |
| child                  | child                  |
| 74.2%                  | 75.5%                  |
| 56.9%                  | 59.6%                  |

| Figure 5: Response accuracy improvement for each iteration of QADB optimization. |