A SPOKEN DIALOG SYSTEM WITH AUTOMATIC RECOVERY MECHANISM FROM MISRECOGNITION

Norihide Kitaoka  Hirotoshi Yano  Seiichi Nakagawa

Toyohashi University of Technology
1-1 Hbarigaoka Tempaku-cho Toyohashi-shi Aichi 441-8580, Japan

ABSTRACT

We proposed a novel dialog strategy which can recover from misrecognition through a spoken dialog. To recover from the misrecognition without confirmation, our system kept multiple understanding hypotheses at each turn and ‘searched’ for a globally optimal hypothesis across user’s utterances in a whole dialog. As for a dialog strategy, we introduced a new criterion based on ‘efficiency for convergence’ and ‘consistency with understanding hypotheses’ to select an appropriate system response. Using such criterion, the system removes the ambiguity without making the user feel unnatural in relation to the response conflicting with actual user intent. We also proposed to adopt the repetition utterance detection to update the understanding hypotheses. We developed a spoken dialog system using these techniques and showed some dialog examples in which misrecognition was naturally corrected. We also showed that our strategy was efficient in terms of the number of turns.

Index Terms—Speech communication, artificial intelligence

1. INTRODUCTION

When we communicate with computers through a speech interface, misrecognition is inevitable. To address this problem, most dialog systems adopt the turn-by-turn confirmation strategy, which often needs many conversational turns.

In this paper, we propose a novel dialog strategy which reduces confirmation utterances. Without the confirmation, the dialog may proceed with misunderstanding. To solve this problem, our system keeps multiple understanding hypotheses at each turn and finally removes the ambiguity and selects the most probable hypothesis through the dialog with the user. In such a dialog, it is very important for the system to generate appropriate responses, which controls the whole dialog.

Itoh et al. proposed a dialog system which kept multiple understanding hypotheses and rescored them using confidences levels of speech recognition results and dialog histories [3]. The system achieved about 10% relative improvement of understanding rate from the strategy only using the best candidates of speech recognition results. Higashinaka et al. proposed to incorporate discourse features into confidence scoring of understanding (in their case, intention) hypotheses [4]. Other dialog management techniques using confidence measures of speech recognition were also proposed, in which the confidence was used for rejection of words or switching of dialog strategies [5, 6].

Dohsaka et al. proposed the dual-cost method for efficient spoken dialog control [7, 8]. This method tried to minimize the summation of ‘the confirmation cost’ and ‘the information transfer cost,’ and could avoid unnecessary confirmation dialog.

In this paper, we propose a new response generation criterion to remove the ambiguity without making the user feel unnatural in relation to the response conflicting with actual user intent.

This paper is organized as follows: In Section 2, the task of our dialog system was described. We introduce the understanding method to keep multiple understanding hypotheses in Section 3, and a criterion for system response selection based on ‘efficiency for convergence’ and ‘consistency with understanding hypotheses’ in Section 4. We evaluate our system in Section 5 and finally conclude the paper in Section 6.

2. TASK

The task of our system is the destination setting in a car navigation system. The system needs the information on the prefecture, the city, the town and the facility to set the destination, so it asks questions to obtain this information. Some information is implied by the other already obtained information. For example, if ‘Toyohashi’, a city name, is known, then ‘Aichi’, a prefecture name, is also known. In this case, the system does not need to ask the prefecture name. (Of course, the system can select the question regarding the prefecture name according to the criteria described in Section 4, if ‘Toyohashi’ is unreliable.)

Examples of allowed user utterances are as below:

• I will go to Aichi (prefecture).
• Hbarigaoka Tempaku-cho (town).
• I want to go to a gas station (facility) in Shizuoka (prefecture).

Sentences can contain more than one keyword with needed information. The keyword set consisted of 5 prefectures, 22 cities, 139 towns and 29 facilities and the perplexity of the grammar was 91.7. Some facilities are in the hierarchical structure explained in Section 4.1 and some general facilities (for example, convenience stores) are not.

3. DIALOG UNDERSTANDING METHOD KEEPING MULTIPLE UNDERSTANDING HYPOTHESES

Spoken dialog systems often incorrectly recognize user utterances. To behave robustly against the misrecognition, explicit confirmation utterances are often used. If confirmation utterances were not used, many dialog turns could be reduced. The system, however, may continue to mistake some words for others during the dialog, resulting in failure of the dialog.

Such failures would be caused by using only the best recognition hypothesis obtained from each user utterance. N-best hypotheses may contain correct recognition results, and they should be effectively used to reduce such failures [9].

Our system keeps multiple understanding hypotheses which are constructed at each turn during a dialog as shown in Figure 1. Each row represents an understanding hypothesis and each hypothesis has a confidence score derived from the confidence scores of the recognition results. Each word included in an understanding hypothesis has the posterior probability of the utterance recognition result (log likelihood difference between the sentence recognition result and arbitrary syllable sequence recognition result) as the confidence score, and the scores are summed up to make a confidence score of an understanding hypothesis.
User: I wanna go to a convenience store in Toyohashi (city).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Prefecture</th>
<th>City</th>
<th>Town</th>
<th>Facility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Aichi)</td>
<td>Toyohashi</td>
<td>---</td>
<td>Convenience store</td>
</tr>
<tr>
<td>2</td>
<td>Toyama</td>
<td>---</td>
<td>---</td>
<td>Convenience store</td>
</tr>
<tr>
<td>3</td>
<td>(Aichi)</td>
<td>Toyokawa</td>
<td>---</td>
<td>Convenience store</td>
</tr>
</tbody>
</table>

Fig. 1. An example of a set of multiple understanding hypotheses generated from one user utterance.

Parentheses indicate the words which are implicitly known without being drawn from user utterances in the context of a particular understanding hypothesis and the knowledge of the system. The prefecture name ‘Aichi’ is implicitly determined from the city names ‘Toyohashi’ or ‘Toyokawa’ in the example of Figure 1.

The system combines the current understanding hypotheses with recent N-best recognition results of user utterances to generate new set of understanding hypotheses. The number of hypotheses tends to increase almost exponentially at each user turn. We adopt the concept of ‘search’ as in continuous speech recognition techniques. We consider the understanding of the whole sequence of user utterances in a dialog as finding a globally optimal hypothesis by traversing search space sequentially generated by user utterances as shown in Figure 2. Some combinations of understanding hypotheses and recognition results may be apparently wrong because of the conflicts between hypotheses and recognition results combined. In such a case, either the understanding hypothesis or the recognition result must be wrong, and thus the new combined hypotheses will be assigned a low confidence. Finally, the hypotheses which do not have any conflicts through the dialog have high confidence scores. In this procedure, the recognition result without the best recognition rank may be included in the best understanding hypothesis. The pruning method as in the beam search, of course, can be adopted in this procedure.

With this strategy, many unnecessary dialog turns will be reduced. The dialog, however, may not be completed unless the system behaves well to lead the dialog to converge. We introduce a new criterion for system response selection and utilization of repetition utterance detection in the next two sections.

4. CRITERION FOR SYSTEM RESPONSE SELECTION

In this section, we discuss how to select a system response from an ambiguous understanding status with multiple understanding hypotheses described in Section 3.

The goal is to choose a most probable understanding hypothesis at the end of the dialog. To achieve this goal, a selection criterion based on entropy-inspired information gain has been proposed [10].

In addition to this ambiguity resolution viewpoint, we introduce a measure of ‘naturalness.’ In this paper, we propose a new criterion based on combination of ‘efficiency for convergence’ and ‘consistency with understanding hypotheses.’ From the aspect of efficiency for convergence of the dialog, the more understanding hypotheses the expected answers to the system response (in this case, the question) rejects, the ‘better’ the response is. The confirmation utterance may be the best response under this condition because the answer for the confirmation can reject all the hypotheses except the best one in most cases (that is, the cases in which the 1st best hypothesis is correct). The utterance depending on the 1st best hypothesis, however, may conflict with the ‘true’ situation, with the result that the user feels ‘unnatural’ and notices the system’s misunderstanding. To recover from the misunderstanding without making the user aware of the misunderstanding, an utterance consistent with as many understanding hypotheses as possible is preferable.

4.1. Measure of efficiency for convergence

In this paper, the task is the destination setting for a car navigation system. The user sets a location using natural language. The location names have hierarchical relations (prefecture—city—town—destination facility).

Consider the situation in which the system has both hypotheses including the city name ‘Toyohashi’ in Aichi Prefecture and that including ‘Hamamatsu’ in Shizuoka Prefecture. Using the question to ask the prefecture, the system can reject one of them according to the user’s answer for the question. Intuitively, the hypotheses rejection ability of expected answers is a good measure for a question (or system response). Thus, we define a measure of efficiency as follows:

When the system has a set of understanding hypotheses, \( N \), and asks the question, \( q \), the probability that the system obtains an answer \( a \) from the user can be calculated as:

\[
P(a|N, q) = \sum_{n \in N} P(a|n, q) \cdot P(n|q),
\]

where \( P(a|n, q) \) is the probability of a given a hypothesis \( n \) and a question \( q \) and \( P(n|q) \) is the probability that the hypothesis \( n \) is correct given \( q \). Here, we assumed that the correctness of \( n \) is independent of \( q \), so \( P(n|q) = P(n) \). If the unique \( a \) can be determined by \( n \) and \( q \), \( P(a|n, q) = 1 \), or else, \( P(a|n, q) \) distributes uniformly on possible answers. Strictly speaking, \( P(n) \) has to be estimated a priori depending on the confidence score of \( n \), \( Conf(n) \), but the \( Conf(n) \) has no direct relation with \( P(n) \) and thus the statistics of the relation between \( Conf(n) \) and \( P(n) \) should be estimated from a large amount of training data. In this paper, however, we simply used \( Conf(n) / \sum_{m \in N} Conf(m) \) as \( P(n) \) because of the lack of such data. Finally, we define the efficiency score for a question \( q \) as:

\[
S_e(q) = \sum_{n} (1 - P(n)) \sum_{a} I(a, n) \cdot P(a|N, q),
\]

where \( I(a, n) = 1 \) when the answer \( a \) conflicts with \( n \), and \( I(a, n) = 0 \) otherwise.

4.2. Measure of consistency with understanding hypotheses

Consider the case of understanding status with understanding the hypotheses described in Figure 1. If the system asks a question, “Tell me the city name,” this question will conflict with the first and third hypotheses because the city names were already uttered explicitly by the user in the context of these hypotheses and thus the question is unnatural for the user. The conflict with the ‘true’ situation, with the result that the user feels ‘unnatural’ and notices the system’s misunderstanding. To prevent such utterances, we use the expectation of answers to the system response (in this case, the question) to decide the expected answers.

Finally, the system asks the question, \( q \), with maximum weighted sum of the above two measures:

\[
\hat{q} = \arg\max_{q} \{ w_e \cdot S_e(q) + w_r \cdot S_r(q) \},
\]
The choices of the system are described below:

**Questions for new information:**
- "Tell me the prefecture of destination."
- "Tell me the city of destination."
- "Tell me the town of destination."
- "Tell me the destination facility."

**Questions for confirmation:**
- "Is the destination in - - - ken (prefecture name) ?"
- "Is the destination in - - - shi (city name) ?"
- "Is the destination in - - - cho (town name) ?"
- "Is the destination - - - (destination target facility) ?"

Questions for new information are often consistent with multiple hypotheses. A confirmation utterance is mainly used to confirm whether one or a small number of hypotheses are correct or not, and thus such utterances tend to have low consistency with an understanding status consisting of many understanding hypotheses. When any one of these questions has the score above a certain threshold, the system makes a final confirmation utterance to conclude the dialog.

### 4.4. Dialog management using repetition utterance detection

Our system tries to respond to the user without noticing its own misunderstandings as described in Section 4. However, the system cannot perfectly prevent the user from becoming aware of its misunderstanding. In such cases, it is effective to use user’s correction utterances. Users often correct the misunderstandings/misrecognition using repetition of (a part of) the previously used utterance in the dialog. The repetition of the same content in a dialog is one of the useful cues to recover from the misunderstanding/misrecognition. When a repetition pair is detected, the original and repetition utterances must contain the same words. So the system makes higher confidence scores of the words included in both recognition hypotheses of the utterances (or those of competing words lower. See Section 5.2).

We used a DTW-based repetition utterance detection method [1] by which the system can detect the positions of repetitions in the utterances and thus the repeated words can be identified. The confidence scores of the hypotheses are adjusted according to the DTW score. The adjustment should be defined based on the “repetition probability,” but we defined it heuristically so far. We obtained a 0.835 recall rate and 0.841 precision of repeated word identification.

### 5. EVALUATION

#### 5.1. Spoken dialog system configurations

We developed a Japanese spoken dialog system. A CFG-based Speech recognizer [2] and repetition utterance detection are adopted. The dialog manager updates the understanding hypotheses using the recognition results, the repetition utterance detection results and the current understanding hypotheses. Based on the updated understanding hypotheses, a system response is selected using the criterion proposed in Section 4 and sent to the speech synthesizer.

The speech recognizer and repetition utterance detection received the 12-dimensional MFCC and their Δ and ΔΔ with Δ power and ΔΔ power derived from 16 kHz sampled speech data windowed by 25 ms Hamming window with 10 ms shift. Acoustic models modeled 114 Japanese syllables using 4 emission states, whose pdfs consisted of 4 mixtures of Gaussians, each with a full covariance matrix.

The system may misunderstand the user’s utterances. If the system displays the current understanding status to the user, then the user may notice the system’s misunderstanding. So the system does not show its understanding status.

**System 1: Please set the destination.**

**User 1:** I will go to a convenience store (facility) in Toyohashi-shi (city)

**Recognition results**
1: Osu Convenience store
2: Toyohashi Convenience store
3: Chofu Convenience store

**Understanding hypotheses**
1: Osu
2: Toyohashi
3: Chofu

**System 2:** Tell me the prefecture of destination.

**User 2:** Aichi-ken (prefecture).

**Recognition results**
1: Aichi-ken
2: Mitsuke

**Understanding hypotheses**
1: Aichi-ken
2: Mitsuke

**System 3:** Tell me the city of destination.

**User 3:** Toyohashi-shi (city).

**Recognition results**
1: Toyohashi
2: Aichi-ken

**Understanding hypotheses**
1: Toyohashi
2: Aichi-ken

---

**Fig. 3. An example of a dialog.**

**5.2. Dialog example**

Examples of the dialogs with the system are shown in Figure 3. The user utterance, recognition results, and understanding hypotheses are shown separately.

In Figure 3, the system asked the prefecture by means of the efficiency criterion (System 2). Then, after obtaining the recognition results of User 2, the system made the confidence score of Chofu, a city in Tokyo Metropolitan Prefecture, which was not in Aichi Prefecture. Then the system asked the city name, and the answer of the user (User 3) was detected as the repetition of User 1. According to this result, the system decreased the confidence score for Osu, which was a competitive hypothesis against Toyohashi.

In this case, the system rectified the misrecognition through the dialogs.

**5.3. Experimental conditions using simulation**

We evaluated the proposed dialog management by the average number of user turns in a dialog. In this paper, we fully evaluated using simulated dialogs on a computer by generating user utterances automatically. We compared our proposed strategy with the conventional turn-by-turn confirmation strategy and likelihood-based confirmation strategy.

The simulation was done as follows:
1. The simulated user first decides the destination which should be set randomly. The simulated user will then try to complete the setting of this destination through the following procedure.
2. The system makes the first utterance, “Please set the destination.”
In this paper, we proposed a novel dialog strategy which could recover naturally from misrecognition through the dialog. Our system kept multiple understanding hypotheses at each turn, resulting in recovery from misrecognition. We adopted a new criterion based on ‘efficiency for convergence’ and ‘consistency with understanding hypotheses’ to select an appropriate system response. We also proposed to reflect the result of the repetition utterance detection to the update of understanding hypotheses to help recovery from misrecognition. We developed a spoken dialog system with proposed dialog management methods and showed some dialog examples in which misrecognitions were naturally corrected. In the future, we refine the understanding hypotheses updating rules to complete a dialog without failure. We have to treat more complex tasks in this framework, for example, the task with mandatory and optional slots. We also evaluate the dialog system subjectively.

7. REFERENCES