SPOKEN DIALOG STRATEGY BASED ON UNDERSTANDING GRAPH SEARCH

Yuji Kinoshita, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda

Graduate School of Information Science Nagoya University
Furo-cho, Chikusa-ku Nagoya 464-8603, JAPAN

ABSTRACT

We regarded information retrieval as a graph search problem and proposed several novel dialog strategies that can recover from misrecognition through a spoken dialog that traverses the graph. To recover from misrecognition without seeking confirmation, our system kept multiple understanding hypotheses at each turn and searched for a globally optimal hypothesis in the graph whose nodes express understanding states across user utterances in a whole dialog. As for a dialog strategy, we introduced a new criterion based on efficiency in information retrieval and consistency with understanding hypotheses to select an appropriate system response. Using such criterion, the system removes the ambiguity so that users do not feel that a response that conflicts with the actual user intent is unnatural. We developed a spoken dialog system using these techniques and showed 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 a turn-by-turn confirmation strategy that often needs many conversational turns. Without confirmation, the dialog may proceed with misunderstanding. To solve this problem, our system keeps multiple understanding hypotheses as active nodes on the search graph at each turn and finally removes the ambiguity and selects the most probable hypothesis through dialog with a user. In such a dialog, the system must generate appropriate responses that control the whole dialog.

There are many reserches that handles with the system's misunderstandins. Itoh et al. proposed a dialog system that kept multiple understanding hypotheses and rescored them using the confidence level of speech recognition results and dialog histories [3]. The system achieved about a 10% relative improvement of understanding rate from the strategy only using the best candidates of speech recognition results. Higashinaka et al. incorporated discourse features into the confidence scoring of understanding (in their case, intention) hypotheses [4]. Other dialog management techniques using the confidence measures of speech recognition have also been proposed in which confidence was used to reject words or switch dialog strategies [5, 6]. Dohsaka et al. proposed a dual-cost method for efficient spoken dialog control [7]. Their method tried to minimize the summation of the 'confirmation cost' and the 'information transfer cost' to avoid unnecessary confirmation dialogs. Recently, Partially Observable MDPs (POMDPs) has been often used for modeling the uncertainty inherent in spoken dialog systems [8]. However, conventional POMDP has a problem with the number of slots and handles only a few slots and the values. Young et al. proposed a form of POMDP which can be scaled to support practical dialog systems [9]. POMDP has a belief state combined with all the values of slots, on the other hands, our system keeps understanding graph only with the active nodes expanded according to the recognition results obtained from the user utterance. In contrast with POMDP, our system is not affected by the size of state and can also search the graph with heuristics. We also propose a new response generation criterion to remove ambiguity so that users do not feel that responses are unnatural in relation to conflicts with actual user intent.

This paper is organized as follows. In Section 2, the task of our dialog system is described. We introduce the understanding procedure using a graph search in Section 3 and a criterion for system response selection based on efficiency in information retrieval 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 music retrieval from a music database. Users can say the artist name directly, but often does not know the artist or song name. In such situation, the system needs artist-related information, such as genre, the year, and gender to search for a song. The system asks questions about these keywords or confirms them. The following are examples of allowed user utterances:

- "Find a rock song" (genre)
- "Songs from 1990" (year)
- "I want to listen to a song by a female singer" (gender)

The system choices are:

Questions for new information:

- "What genre?"
- "What year?"
- "Male or female"

Questions for confirmation:

- "Is a song by (artist) ok?"
- "Is a (genre) song ok?"
- "Is a song from (year) ok?"
- "Is a song by a (sex) singer ok?"

The keyword set consisted of 76 artists, 13 genres, 70 years/eras, and both sexes.

3. INFORMATION RETRIEVAL PROCEDURE AS A GRAPH SEARCH

3.1. Spoken dialog understanding as a graph search

We consider an information retrieval process to be a graph search. Here we assume that a slot-filling type spoken dialog and slot filling through a dialog are regarded as graph searches (Figure 1). Users have a goal (a keyword set that may be ambiguous) to search for a song. By considering a (partial) keyword set as a node of a graph, we can construct a search graph of understanding. In the search process, active nodes are the current understanding status, and the system expands the nodes based on speech recognition results obtained from each dialog step. If an incorrect search advances, backtracking may be required to recover it before it finally reaches a correct understanding: that is, a correct keyword set.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

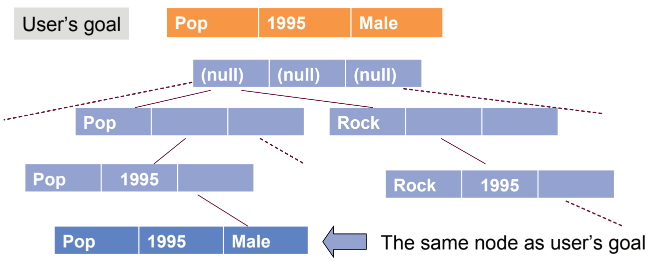
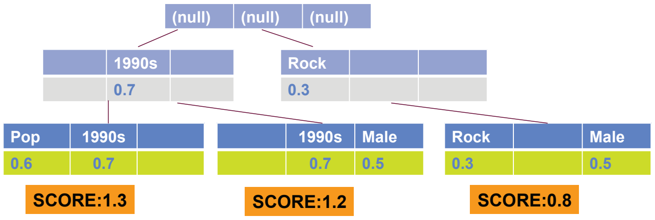


Fig. 1. Example of a graph search



rank	Genre	Year	Sex	SCORE
1	Pop	1990s	(null)	1.3
2	(null)	1990s	Male	1.2
3	Rock	(null)	Male	0.8

Fig. 2. Example of adopting multiple understandings

3.2. Adopting multiple understandings

Spoken dialog systems often incorrectly recognize user utterances. To behave robustly against such misrecognition, explicit confirmation utterances are often used. If confirmation utterances are not used, dialog turns can be reduced. The system, however, may continue to mistake some words for others during the dialog, resulting in dialog failure. Such failures are caused by only using the best recognition hypothesis obtained from each user utterance. N-best hypotheses, which may contain correct recognition results, should be effectively used to reduce such failures [10].

In our former research [1], we proposed a strategy in which the system keeps multiple understanding hypotheses using N-best recognition hypotheses and chooses an appropriate system response so that correct hypotheses can be prioritized. In this paper's method, allowing the existence of multiple active nodes, as shown in Figure 2, is equivalent to keeping multiple understandings. Using N-best recognition results, an active node is expanded to at most N new succeeding nodes.

3.3. Best-first search

There are several search methods in graph search, including breadth-first and depth-first. However, these methods are not suitable for our spoken dialog system because they try to search for every node without using any information obtained from the user utterances. A more practical approach is a best-first search that decides the 'best' node to be expanded with a node scoring strategy. In our system, the confidence score of words from the results of speech recognition is reasonable. Each node has a score, which is the sum of the confidence scores included in the node.

This score is referred to as a search score $g(n)$ in the search process, where n is one of the active nodes. As shown in Figure 3, this method selects the node with the best $g(n)$, generates a question to extract new information, and expands it based on the recognition result of the user response. [1]

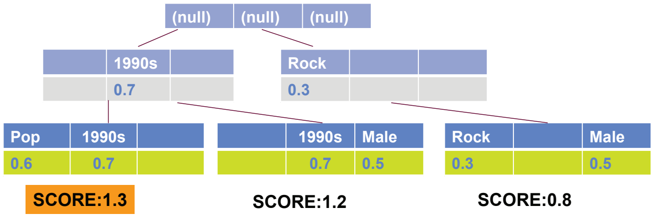


Fig. 3. Best-first search without heuristics

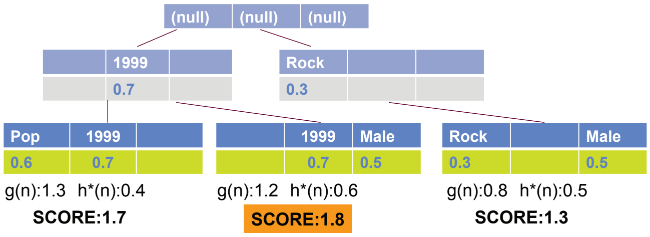


Fig. 4. Best-first search with heuristics

We can use not only $g(n)$ but also heuristic score $\hat{h}(n)$ to evaluate the node. As shown in Figure 4, this method uses both search score $g(n)$ and heuristic $\hat{h}(n)$ and selects a node with the highest score $g(n) + \hat{h}(n)$. Typical $\hat{h}(n)$ is the estimation of future $g(n)$, but here we use the real heuristic score that is used in the system response (question) generation explained in the next section.

3.4. Node expantion

When obtaining a user response, not only the best active node but also other active nodes can be expanded. To avoid generating the same question when backtracking occurs, these nodes are also expanded simultaneously.

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 the 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 [11].

In this paper, we propose a new criterion based on the combination of consistency with understanding hypotheses and efficiency in information retrieval. A confirmation utterance may be the best response under this condition because the answer to the confirmation can reject all the hypotheses except the best one in most cases: that is, cases in which the first best hypothesis is correct. The utterance depending on the first best hypothesis, however, may conflict with the 'true' situation, resulting in a situation where the user feels the response is '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 from the point of

[1] Question generation strategy is discussed in the next section.

view of consistency. In addition, the system needs to suggest a song as fast as possible. If the system has two choices of questions, one that narrows down the retrieval results is more preferable. Considering these two aspects, the system must choose an appropriate system response at that time.

4.1. Measure of consistency with understanding hypotheses [1]

Consider the case of understanding status by understanding the hypotheses described in Figure 2. If the system asks, “What genre?” this question conflicts with the first and third hypotheses because genre was already uttered explicitly by the user, and thus the question is unnatural. The confirmation, “Is it from the 1980s?,” conflicts with the third hypothesis because the decade is the “1990s” in this context. To prevent such utterances, we adopt a consistency measure:

$$S_c(q) = \sum_{n \in N} (1 - I(q,n)) \cdot P(n), \tag{1}$$

where n is one of the active nodes N, and $I(q,n) = 1$ when question q conflicts with n, and $I(q,n) = 0$ otherwise. $P(n)$ is the probability that hypothesis n is correct and weighs the score to prefer hypotheses thought to be correct. Strictly speaking, $P(n)$ has to be estimated a priori depending on the confidence score of n, $Conf(n)$, but $Conf(n)$ 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 $P(n)$ due to a lack of such data.

4.2. Measure of efficiency in information retrieval

In an information retrieval task, a question that greatly narrows the search space is efficient. A question that does not narrow the search at all is just a waste of time. Thus, we use mutual information as a measure of retrieval efficiency to estimate how much entropy can be decreased. The mutual information of multiple understandings is defined as:

$$S_e(q) = I(X; q|N) = H(X|N) - H(X|q, N), \tag{2}$$

where $H(X|N)$ and $H(X|q,N)$ are defined as:

$$H(X|N) = -\sum_{n \in N} \sum_{x \in X} p(n,x) \log_2 p(x|n), \tag{3}$$

$$H(X|q,N) = -\sum_{n \in N} \sum_{x \in X} \sum_{a \in A_q} p(n,x,a) \log_2 p(x|n,a), \tag{4}$$

where A_q is a set of the possible answers given by the user by asking question q and X is a set of the retrieval results from the database.

4.3. Final system response decision

Finally, we have to balance the above two measures. Here, the system selects question $\hat{q}$ with its maximum weighted sum:

$$\hat{q} = \arg\max\{w_c \cdot S_c(q) + w_e \cdot S_e(q)\}. \tag{5}$$

4.4. Using these measures as heuristics

We used these measures as heuristic $\hat{h}(n)$ described in Section 3.3. With a heuristic, we search considering efficiency to reach the goal faster than by only using search score $g(n)$. When calculating the score considering only best node N in Eqs. (1-4), include only the node. The system can also consider the heuristic more globally. In such a case, N can include other active nodes and then it might decrease the topic jumping caused by backtracking.

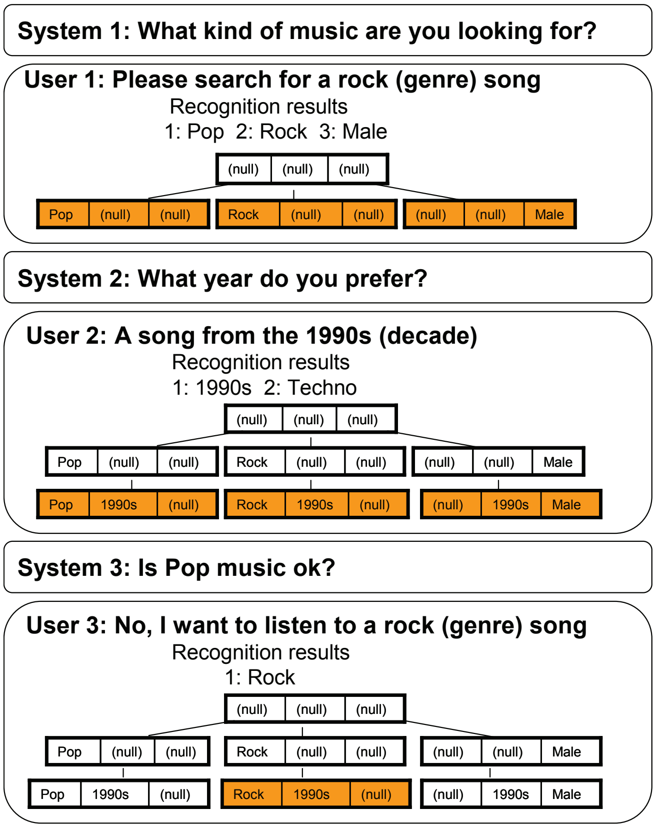


Fig. 5. Example of a dialog

5. EVALUATION

5.1. Spoken dialog system configurations

We developed a Japanese spoken dialog system by adopting a Julius speech recognizer [2]. The dialog manager updates the understanding hypotheses using the recognition results by expanding the active nodes. Based on the active nodes, a system response is selected using the criterion proposed in Section 4 and sent to the speech synthesizer. The system may misunderstand the user utterances. If the system displays the current understanding status, then the user may notice the system’s misunderstanding. So the system does not show its understanding status.

5.2. Dialog example

An example of dialogs with our system is shown in Figure 5. User utterances, recognition results, and search space status are shown separately.

In Figure 5, the system got three active nodes expanded by the recognition results of User 1. Then the system asked about the year by the proposed criterion (System 2). After obtaining the recognition results of User 2 (techno was omitted because it conflicted with System 2), all active nodes were expanded. Then the system confirmed the genre. According to the utterances of User 3, the system narrowed down the active nodes.

In this case, the system reached the correct node 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 automatically generating user utterances. We compared our three proposed strategies with conventional turn-by-turn confirmation and likelihood-based confirmation strategies [5]. The simulation was done as follows:

1. The simulated user first decides the goal (keyword set) that is set randomly. The simulated user then tries to complete this goal ’ s setting through the following procedure.
2. The system makes the first utterance, “What kind of music are you looking for?”
3. The simulated user replies to comply with the system utterance.
4. The simulated recognizer makes pseudo recognition results and their confidence scores from the utterance of the simulated user based on a predefined recognition rate.
5. The system expands the active nodes to update the understanding hypotheses using the pseudo recognition results.
6. If the understanding status does not satisfy the termination conditions, the system generates the next utterance. Go to 3.

In Step 4, we predefined recognition rate R at 60-90%. In our proposed system, if all slots were filled or the retrieval results are less than three, then the system suggested the best song. If the song matches the user goal, the user replies “yes” to the system’s suggestion, terminating the system. Until then, the system continues to ask questions and offer suggestions. In Step 6 in the above simulation procedure, the method explained in Sections 3 and 4 was used in our proposed method, where Eq.(3)-(4) were calcultated by counting the number of results obtained by searching a database and weights in Eq.(5) were set to 1.0 and 0.2 respectively. In turn-by-turn confirmation strategy, the system made a confirmation utterance for the user’s new information input or a question for new information after the user’s “yes.” In all methods, the simulated user corrected the misrecognition with a repetition utterance when noticing it. 1000 simulations were done for each method.

5.4. Simulation results

The experimental results are shown in Figure 6. It compares the average number of turns among our three proposed strategies (best-first searches without heuristics, with heuristics, and with global heuristics) and the conventional turn-by-turn confirmation and likelihood-based confirmation strategies with a recognition rate at 60-90%. From these results, the average number of turns was reduced using our strategies at any recognition rate. These results showed that adding a heuristic score for node selection to be searched was effective.

6. CONCLUSION

In this paper, we proposed novel dialog strategies that consider an information retrieval process as a graph search problem and choose an appropriate system response in each dialog step. We adopted a new criterion based on search efficiency and consistency with understanding hypotheses not only to select an appropriate system response but also to score the search nodes. We developed a spoken dialog system with our proposed dialog management methods and showed dialog examples in which misrecognitions were naturally corrected. These strategies were compared to conventional strategies and reached the user goal faster. In the future, we will refine the measure for choosing an appropriate question to complete a dialog without failure. We have to treat more complex tasks in this framework, for example, tasks with mandatory and optional slots. We will also subjectively evaluate the dialog system.

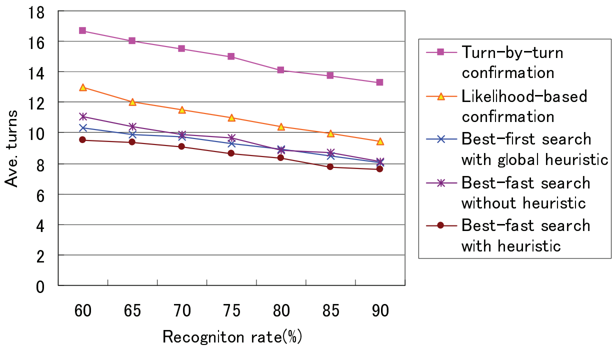


Fig. 6. Evaluation results of number of turns by simulation dialogs

7. REFERENCES

[1] N. Kitaoka, H. Yano, and S. Nakagawa, “A Spoken Dialog System with Automatic Recovery Mechanism from Misrecognition,” IEEE 2006 Workshop on Spoken Language Technology (SLT2006), pp. 202-205, Dec. 2006.

[2] A. Lee, “Large Vocabulary Continuous Speech Recognition Engine Julius ver. 4,” Vol. 2007, No. 129(20071220) pp. 307-312 2007-SLP-69-(53).

[3] T. Itoh, A. Kai, Y. Itoh, and T. Konishi, “An understanding strategy based on plausibility score in recognition history using CSR confidence measure,” Proc. ICSLP2004, pp. 2133-2136, 2004.

[4] R. Higashinaka, K. Sudoh, and M. Nakano, “Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems,” In Proc. ICASSP2005, Vol. I, pp. 25-28, 2005.

[5] K. Komatani and T. Kawahara, “Flexible mixed-initiative dialogue management using concept-level confidence measures of speech recognizer output,” In Proc. of COLING, Vol. 1, pp. 467-473, 2000.

[6] T. J. Haze, S. Seneff, and J. Polifroni, “Recognition confidence scoring and its use in speech understanding systems,” Computer Speech and Language, Vol. 16, pp. 49-67, 2002.

[7] K. Dohsaka, N. Yasuda, and K. Aikawa, “Efficient dialogue control depending on the speech recognition rate and system’s database,” Proc. EUROSPEECH-2003, pp. 657-660, 2003.

[8] J. Williams, P. Poupart, and S. Young, “Partially observable Markov decision processes with continuous observations for dialogue management”, In Proc SIGdial Workshop on Discourse and Dialogue, Lisbon, 2005.

[9] S. Young, J. Williams, J. Schatzmann, M. Stuttle, and K. lhammer, “The hidden information state approach to dialogue management”, Technical Report CUED/FINFENG/ TR.544, Cambridge University Engineering Department, 2006.

[10] B. Souvignier, A. Kellner, B. Rueber, H. Schramm, and F. Seide, “The thoughtful elephant: Strategies for spoken dialog systems,” IEEE Transactions on Speech and Audio Processing, Vol. 8, no. 1, pp. 51-62, 2000.

[11] T. Misu and T. Kawahara, “Speech-based information retrieval system with clarification dialogue strategy,” Proc. HLT/EMNLP, pp. 1003-1010, 2005.