A spoken dialog system based on automatic grammar generation and template-based weighting for autonomous mobile robots

Takashi KONASHI, Motoyuki SUZUKI, Akinori ITO, Shozo MAKINO

Graduate School of Engineering
Tohoku University, Japan
{konashi,moto,aito,makino}@makino.ecei.tohoku.ac.jp

Abstract

We have been developing a spoken dialog system. Conventional spoken dialog systems need grammar descriptions and scripts of a dialog, that are difficult to develop. The system proposed in this paper is based on semantic frames, and the system generates the recognition grammar from the frames automatically. As the system requires only a frame-based description for a task of dialog, the system can be easily applied to different kinds of tasks. Moreover, the recognition accuracy is improved by sentence weighting based on phrase class template. We evaluated the system by experiments. The system reached the goal with 2.44 user’s utterances in average.

1. Introduction

Autonomous mobile robots are becoming more and more popular in various fields including industry, entertainment and welfare. There are several possibilities for user interfaces for these robots. Conventional user interfaces such as remote controller and touch panel are easy to use. However, these interfaces have some problems. For example, a remote controller is not suitable for robots unless they are used by specific people. Using a touch panel is sometimes dangerous as an interface for a mobile robot. A spoken dialog based interface solves these problems. This interface can be used without contacting the robot, and a user does not need to bring a controller.

Spoken dialog systems developed so far have been designed task by task, that means application of the system to another task requires complete redesign of the whole system. Dialog description languages like VoiceXML help redesign of the system because they split dialog structure from dialog system itself. However, as VoiceXML is a lower-level description language, VoiceXML based redesign will be still painful. Another requirement for this kind of dialog system is accuracy and robustness. A user may talk as sloppily as an utterance to human. Therefore, the dialog system have to be as robust as possible. However, as conventional systems design their recognition grammar by hand, accuracy and robustness of the systems depend on the grammar designers’ skill.

The objective of this system is to develop a spoken dialog system that is robust, accurate and easy to customize. The proposed system exploits frame-based task description, that hides complicated grammatical description from developers. This system can accept more than one kind of tasks. In the following section, overview of the system as well as the result of dialog experiment are described.

2. System overview

The proposed dialog system is designed and developed for form based user interface. As an application of the proposed system, we developed a dialog system for an intelligent care robot. This mobile robot does three tasks, “Bring something to drink” (tea serving task), “open or close curtain” (open and close task) and “throw a garbage away” (bring and take task).

Figure 1 shows the block diagram of the proposed system. The system consists of three subsystems: acoustic analysis subsystem, speech recognition subsystem and dialog control subsystem.

The acoustic analysis subsystem detects the speech and calculates feature vectors from the input speech. The speech recognition subsystem performs matching between feature vectors and the recognition grammar. As the results of recognition, this subsystem calculates the most likely word sequences that conforms the grammar and their likelihoods. The dialog control subsystem receives the recognition results and makes a plan toward the next turn. The dialog control subsystem is divided further into the action controller, the task driver, the information complementor and the answer generator.

In this system, a task is described using semantic frame. Figure 2 shows an example of semantic frame. This example is a task description for tea serving task. This task has three slots: action, drink and quantity. A list of acceptable words are also described for each slot. This task description is used as a frame for speech understanding as well as a source of a recognition grammar.

Figure 3 shows a block diagram of grammar genera-
3. The spoken dialog system

3.1. Task description

A task is expressed by semantic frame. Semantic frame has several slots. Basically, one slot receives a word which is needed to achieve the task. A slot is defined as a tuple of the following specifications:

- The name of the slot
- List of words (word symbols and their pronunciations) that the slot accepts
- Production rules to generate a reply associated with the slot

3.2. Grammar Generation

The grammar should accept many variations of command utterances including the word inflections, inversion of words or phrases, insertion of interjections and filled pauses. The system generates a grammar that covers these phenomena.

Basically a slot accepts a word. For example, “drink” slot in the tea serving task accepts “coffee”, “tea” or “water”. On generating a grammar from the task description, two phenomena should be considered. One is variation of phrase, and the other is phrase order. For “water”, a user may pronounce it as “water”, “some water”, etc. In this system, these expressions are automatically generated from a word “water” using phrase model.
the system generates a grammar that accepts any combination of the generated phrases in any order. One slot corresponds to one state in the generated grammar that accepts all acceptable phrases of the slot. As any order of the slots can be acceptable, all combinations of slot are generated (Figure 5). The transitions that skip any states are generated as well to accept utterance with omission of words. To treat interjections and filled pauses, a filler model is inserted before every states.

3.3. Sentence weighting based on phrase template frequency

In speech recognition using FSA-based grammar, no linguistic likelihood is used. However, in an actual conversation, there are very frequently used sentences while there are sentences hardly used. N-gram based language model can take the frequency into account, but an n-gram requires a large amount of data for training. It is difficult to gather such amount of training data for a specific task.

To introduce a word-order-based weighting into FSA-based language model, we utilized phrase class template. A phrase class is a class corresponds to a phrase. For abstraction, we used only five phrase classes: a noun phrase, a verb phrase, a modifier phrase, a conjunctive phrase and an interjective phrase. We regard a sequence of phrase classes of a sentence as a phrase class template of the sentence. For example, a phrase class template of a sentence “Give me a cup of coffee” is \(<verb noun noun>=0\). We observed frequencies of phrase templates. Analysis result of higher rank using two corpora[?] is shown in table 1.

Table 1: List of phrase template frequency

<table>
<thead>
<tr>
<th>template</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun</td>
<td>1025</td>
</tr>
<tr>
<td>verb</td>
<td>724</td>
</tr>
<tr>
<td>noun noun</td>
<td>569</td>
</tr>
<tr>
<td>noun verb</td>
<td>376</td>
</tr>
<tr>
<td>noun noun verb</td>
<td>116</td>
</tr>
<tr>
<td>verb noun</td>
<td>90</td>
</tr>
</tbody>
</table>

Using these phrase class templates and their frequencies, probability of a sentence is calculated as follows. First, considering that a slot corresponds one phrase, phrase class templates that contains more classes than number of slots can be excluded from calculation. Second, log-probability \(P_L(c)\) of phrase class template \(c\) is calculated as

\[
P_L(c) = \log \frac{n_c}{\sum_{i \in C_A} n_i}
\]

where \(n_c\) is frequency of \(c\) and \(C_A\) is a set of templates that are not excluded for the task. If a template of possible phrase order does not exist in the training data, the frequency of that template is regarded as 1.

Likelihood of a sentence \(P\) is calculated as

\[
P = P_A + wP_L
\]

where \(P_A\) is acoustic likelihood, \(P_L\) is language likelihood, \(w\) is language weight parameter. We use \(w = 2.0\) that gives best score on evaluation.

3.4. Information complement

The recognition result obtained from the speech recognition subsystem may contain error. Therefore, it is risky to use one most likely recognition result for dialog control. If the result contains misrecognized word, it may increase the cost of the dialog (number of turns to achieve the task). To avoid misunderstanding, a task and a slot are identified not only by the candidate with highest score but also by confidence of the candidate. The confidence is calculated based on scores of multiple candidates.

To judge if the required information is missing or not, this system classified all slot into two classes: essential slot and optional slot. Essential slots are indispensable slots in order to achieve the task. When value of an essential slot is not given in dialog, system must ask a user to fill the slots. On the other hand, optional slots are the slots which do not need to be given.

Not all values for the slots have to be obtained from the user’s utterance. Some of them could be inferred from situation of the dialog or personal preference. To realize this, our system introduces ‘default’ and ‘history’ mechanism.

The following sections describe the detail of the slot classes and history mechanism.

4. Evaluation of the system

An experiment was carried out to evaluate the dialog system. Table 2 shows the experimental conditions. In this experiment, eight users were asked to achieve four kinds of goals as shown in table 3. The vocabulary sizes currently used in each task were 31 for Tea serving, 18 for
Table 2: Experimental conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>4 males and 4 females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dialog</td>
<td>8 dialogs</td>
</tr>
<tr>
<td>Number of task</td>
<td>4 tasks</td>
</tr>
<tr>
<td>Number of words</td>
<td>134 words</td>
</tr>
<tr>
<td>Sampling frequency</td>
<td>16kHz</td>
</tr>
<tr>
<td>Window length</td>
<td>25ms</td>
</tr>
<tr>
<td>Shift period</td>
<td>10ms</td>
</tr>
<tr>
<td>Feature</td>
<td>MFCC(12coeff.)</td>
</tr>
<tr>
<td></td>
<td>+\Delta MFCC(12coeff.)</td>
</tr>
<tr>
<td></td>
<td>+\Delta Pow(Total 25coeff.)</td>
</tr>
<tr>
<td>Acoustic model</td>
<td>triphone HMM</td>
</tr>
<tr>
<td></td>
<td>left-to-right 4states, 3loops</td>
</tr>
<tr>
<td></td>
<td>training set ATR set-C</td>
</tr>
</tbody>
</table>

Table 3: Goals for evaluation

<table>
<thead>
<tr>
<th>Goal</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal 1</td>
<td>Drink:coffee, Quantity:2</td>
</tr>
<tr>
<td>Goal 2</td>
<td>Drink: green tea, Quantity:1</td>
</tr>
<tr>
<td>Goal 3</td>
<td>Drink: tea, Quantity:3</td>
</tr>
<tr>
<td>Goal 4</td>
<td>Drink: oolong tea, Quantity:2</td>
</tr>
</tbody>
</table>

Open and Close, 32 for Bring and Take and 16 for Urgent.
The words used by all tasks were 134 words.

The grammar accepted by the system were lectured to the subject before the experiment. After the lecture, goals were given to a subject and a subject did dialogs to achieve the given goals. Figure 6 shows the number of utterances in each dialog. The average number of user utterances per one dialog is 2.44. Since a dialog always needs confirmation, at least two utterance is required to achieve one dialog. Therefore, this result shows that the dialog was performed smoothly.

5. Conclusion

This paper proposed the spoken dialog system for autonomous mobile robots. The system does frame-based dialog management and it can treat various phenomena in spontaneous speech automatically.

6. References


