PREPARING FOR EVALUATION OF A FLIGHT SPOKEN DIALOGUE SYSTEM

Xiaojun WU, Mingxing XU, and Wenhu WU
Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing
[wuxj, xumx]@sp.cs.tsinghua.edu.cn, wuwh@mail.tsinghua.edu.cn

ABSTRACT
Evaluation is needed to test how well the dialogue system works and is helpful for the developer to find some problems and to make the system more satisfactory. We have developed a flight spoken dialogue system and decide to carry out a thorough evaluation. Five dialogue scenarios with brief task descriptions are carefully designed. A questionnaire for user satisfaction is ready, which will be filled out by the subject immediately after the trial. The system will log some important information during the dialogue transactions, so that some objective metrics can be calculated. At last there are discussions about how to arrange the field trial and if there are any defects in the evaluation preparation.

1. INTRODUCTION
There are many spoken dialogue systems in laboratories or open to public. Evaluation of these systems is useful to test how well the systems work and to know which systems work better [1, 2]. It is also helpful to compare different technologies and different strategies [3, 4].

We have developed a Chinese spoken dialogue system named EasyFlight. We aimed to provide flight information inquiry and ticket reservation service to the public with this system. After several test-and-revise cycles, we decide to carry out an evaluation. It will help us to know how well on earth our system works and to find if there are still any bugs. According the practical dialogue results, we will try to modify or enhance the system behavior.

In the next section, the system organization will be described. After that, we will introduce the evaluation preparation in respects of the dialogue scenarios, the questionnaire for user satisfaction, the possible evaluation metrics and the system log contents, and the plan for trail arrangement. Then some discussions will be presented. At last are the conclusion and the future work.

2. SYSTEM ORGANIZATION
Our flight spoken dialogue system is mainly composed of four modules, the speech recognizer, the syntactic and semantic parser, the dialogue manager and the speech synthesizer, as depicted in figure 1.

3. EVALUATION PREPARATION

3.1 Dialogue Scenarios
We have designed five dialogue scenarios as dialogue tasks for evaluation. These tasks are characteristic in the real flight service station.
In the first two scenarios, the subject is supposed to ask for the arrival time of some flight. In scenario one, the information to specify the flight includes the departure city, the arrival city, the flight number and the date. In scenario two, a general time condition, e.g. in the morning, is provided instead of the flight number.

The next three scenarios concern trip planning and ticket reservation. Given several requirements, the subject should query the system to find the proper flights for the trip and book tickets afterwards. In the third scenario, the subject is assumed to fly to Beijing with his/her colleague, so that they can attend a business meeting in time. The next scenario requires the subject to book tickets for a round trip for a tourism according to the date and plane type constraints. In the last scenario, the problem is that the subject has to take a round trip in only one week, but there are not flights available every day between the two cities, as is common for small cities.

We provide the dialogue scenarios in text and after each we give a brief task description. For example, table 1 is the task description for the first scenario, which specifies the cities, the flight date, and the flight number, and indicates that the arrival time is to be queried.

<table>
<thead>
<tr>
<th>Dialogue #1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
</tr>
<tr>
<td>from Beijing to Shenzhen</td>
</tr>
<tr>
<td>Date</td>
</tr>
<tr>
<td>9th March (Saturday)</td>
</tr>
<tr>
<td>Flight Number</td>
</tr>
<tr>
<td>X2 117</td>
</tr>
<tr>
<td>Arrival Time</td>
</tr>
</tbody>
</table>

Table 1: The task description for the first dialogue scenario

Conditions in the dialogue scenarios are carefully designed so that every task has only one correct answer if the dialogue is normally carried out. This will facilitate the corresponding calculation.

### 3.2 Questionnaire for User Satisfaction

In some sense, the user satisfaction is the most important metric of dialogue system evaluation. In order to help the subject to come to a more reasonable conclusion, a questionnaire with questions on several aspects of satisfaction is needed. We prepared a questionnaire too, which will be given to the subject immediately after the trial.

There are seven questions:

- Did the system understand what you said?
- Are the dialogue tasks easy to complete?
- What do you think of the interaction pace?
- Do you think if the system is always able to accept any utterances?
- Did the system responses agree with your expectation?
- Are you willing to use this system to get flight information and to book tickets in future?

Below each question there are five answers to choose, which rank from the most positive to the most negative. Every answer is given a value, so that we can calculate the average user satisfaction.

### 3.3 Objective Metrics and System Logs

There are several common objective metrics for dialogue system evaluation, such as task success rate (transaction completion rate), dialogue/user turns, transaction time (dialogue duration), and so on. For the speech recognition module, word error is universally accepted, and for the syntactic and semantic parser, slot error rate or attribute error rate is often used [7]. Turn error (utterance-level errors not related to recognizer performance) is proposed for dialogue manager [8].

In order to calculate the criterions above, and what’s more, to try to find program bugs if any, we added log functions into the original system. In the speech recognizer, the user’s speech utterance with start time and the recognition result should be saved in every turn. In the syntactic and semantic parser, the parse tree and the semantic frame are put down. The dialogue manager is so complex that we will log the dialogue knowledge before and after the database query, the database query conditions and query results, the dialogue context, the dialogue state, the generated response, and the response time.

When all the subject and object metrics are measured, we can use PARADISE framework to get an overall system performance function [9].

### 3.4 Plan for Trial Arrangement

We will try to test the system with different speech recognizers, i.e. microphone speech input with KWS technology, microphone speech with LVCSR, telephone speech with KWS, and telephone speech with LVCSR. Thus we can find which speech recognition technology is better in microphone and telephone conditions.

The trial subjects are all novice users. They will be divided into two groups. One group will take the first, the third and the fifth dialogue scenarios. The other will take the second, the third and the fourth. There are two reasons for the grouping. First, we want to find the difference between dialogues by novice users that have different requirements. Second, it seems that the third scenario is much more popular than the other ones in our real lives. Therefore the evaluation result will be more reliable for practical use.
4. DISCUSSION

Although we have considered the words and sentences in the dialogue scenarios carefully to make them as natural as possible, we cannot avoid the influence on the subjects' utterances, even if we would only give the brief task description. For example, there are several ways to specify a date, by tomorrow or the day after tomorrow, by the month and the day of month, and by weekly days. We have to express the date constraints in at least one way, which will surely influence the subjects' way of date expression.

The dialogue system is developed to be able to interact with the users as natural as possible. However, because natural language understanding and artificial intelligence are far to perfect, the system can only understand some user utterances. If during the dialogues there are always just several patterns of sentences that can be accepted by the system, the recognizer and parser are expected to achieve better performance. But the users may find the system unsatisfactory. In our present system, there are no constraints on user utterances at any time, i.e. the system never inclines toward any patterns of sentences or any sentences with certain semantic information. During the field trial, the subjects will be aware that they are talking with a machine, which will surely lead to some difference other than talking with a person. We can possibly find how well the users cooperate with the system. Then we can decide under what conditions and how to prove the system performance by narrowing the system acceptance space in every turn.

Easy dialogue tasks often have fewer dialogue turns and higher task success rate than difficult ones. Thus the proportion of easy tasks will affect the overall user satisfaction. If two systems are tested by dialogue tasks with various difficulties, it is hard to tell which system is better only by referring to the task success rate and/or the user satisfaction. We have designed several dialogue scenarios and set the proportion for trial, but it still doesn’t agree with the real situation.

5. CONCLUSION AND FUTURE WORK

We have set up a flight spoken dialogue system and made much preparation for the evaluation. We will try to find at least 100 novice subjects for the field trial, and if possible, from diverse careers. After the trial there are still much work to do with the system logs and subjects’ feedback. The present system is also to be improved.

6. REFERENCES