A Comparison between Dialog Corpora Acquired with Real and Simulated Users

David Griol
Departamento de Informática
Universidad Carlos III de Madrid
dgriol@inf.uc3m.es

Zoraida Callejas, Ramón López-Cózar
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Granada
{zoraida, rlopezc}@ugr.es

Abstract
In this paper, we test the applicability of a stochastic user simulation technique to generate dialogs which are similar to real human-machine spoken interactions. To do so, we present a comparison between two corpora employing a comprehensive set of evaluation measures. The first corpus was acquired from real interactions of users with a spoken dialog system, whereas the second was generated by means of the simulation technique, which decides the next user answer taking into account the previous user turns, the last system answer and the objective of the dialog.

1 Introduction
During the last decade, there has been a growing interest in learning corpus-based approaches for the different components of spoken dialog systems (Minker, 1999), (Young, 2002), (Esteve et al., 2003), (He and Young, 2003), (Torres et al., 2005), (Georgila et al., 2006), (Williams and Young, 2007). One of the most relevant areas of study has been the automatic generation of dialogs between the dialog manager and an additional module, called the user simulator, which generates automatic interactions with the dialog system.

A considerable effort is necessary to acquire and label a corpus with the data necessary to train good models. User simulators make it possible to generate a large number of dialogs in a very simple way, reducing the time and effort needed for the evaluation of a dialog system each time the system is modified.

The construction of user models based on statistical methods has provided interesting and well-founded results in recent years and is currently a growing research area. A probabilistic user model can be trained from a corpus of human-computer dialogs to simulate user answers. Therefore, it can be used to learn a dialog strategy by means of its interaction with the dialog manager. In the literature, there are several corpus-based approaches for developing user simulators, learning optimal management strategies, and evaluating the dialog system (Scheffler and Young, 2001) (Pietquin and Dutoit, 2005) (Georgila et al., 2006) (Cuayáhuitl et al., 2006) (López-Cózar et al., 2006). A summary of user simulation techniques for reinforcement learning of the dialog strategy can be found in (Schatzmann et al., 2006). In this paper, we propose a statistical approach to acquire a labeled dialog corpus from the interaction of a user simulator and a dialog manager. In our methodology, the new user turn is selected using the probability distribution provided by a neural network. By means of the interaction of the dialog manager and the user simulator, an initial dialog corpus can be extended by increasing its variability and detecting dialog situations in which the dialog manager does not provide an appropriate answer. We propose the use of this corpus for evaluating both our user simulation technique and our dialog system performance.

Different studies have been carried out to compare corpora acquired by means of different techniques and to define the most suitable measures to carry out this evaluation (Schatzmann et al., 2005), (Turunen et al., 2006), (Ai et al., 2007b), (Ai and Litman, 2006), (Ai and Litman, 2007), (Ai et al., 2007a). In this work, we have applied our dialog simulation technique to acquire a corpus in the academic domain, and compared it with a corpus recorded from real users interactions with a spo-
ken dialog system

The results of this comparison show that the simulated corpus obtained is very similar to the corpus recorded from real user interactions in terms of number of turns, confirmations and dialog acts among other evaluation measures.

The rest of the paper is organized as follows. Section 2 summarizes the main characteristics of the UAH system. Section 3 describes our statistical methodology for user simulation. Section 4 describes the set of measures used to compare the corpus acquired with real users and the simulated corpus. Section 5 presents the results of this evaluation, and finally, the conclusions are presented in Section 6.

2 The UAH System

Universidad Al Habla (UAH - University on the Line) is a spoken dialog system developed to provide spoken access to academic information about our department (Callejas and López-Cózar, 2005) (Callejas and López-Cózar, 2008). The system is comprised of the five typical modules of current spoken dialog systems, concerned with automatic speech recognition (ASR), dialog management (DM), database access (DB Access), data storage (DB) and oral response generation (RG). In addition, we implemented a module called GAG to automatically create ASR grammars (Callejas and López-Cózar, 2007).

The dialog manager of the UAH system was implemented using VoiceXML documents that are dynamically created using PHP. This way, it can adapt the system responses to the context of the conversation and the dialog state, which improves the naturalness of the interaction. For example, the help messages provided by the system take into account the topic that the user and the system are addressing at a particular moment. The context is used as well to decide the confirmation strategy to use.

The information that the system provides can be classified in four main groups: subjects, professors, doctoral studies and registration, as shown in Table 1. As can be observed, the system must ask the user for different pieces of information before producing a response.

A set of 100 dialogs was acquired with the UAH system by means of its interaction with real users. The acquisition process resulted in a spontaneous Spanish speech dialog corpus with 60 different speakers. The total number of user turns was 422 and the recorded material has a duration of 150 minutes. Figure 1 shows an example of a dialog extracted from this corpus. Turns tagged with an ‘S’ refer to system turns, and turns tagged with ‘U’ refer to user turns.

S1: Welcome to the UAH system. How can I help you?
U1: I want to know information about subjects in Computer Science
S2: Which subject do you want to consult?
U2: Language Processors I
S3: What would you like to know?
U3: Can you inform me about the schedules of Group D?
S4: The schedule of Group D is: Mondays from 9 to 10 and Wednesdays from 9 to 11 and from 11 to 13. Anything else?
U4: Can you tell me the name of the lecturer?
S5: The name of the lecturer of the Group D of Language Processors I is David Griol. Anything else?
U5: No, thank you.
S6: Thank you for using the UAH system.

Figure 1: An example of a dialog acquired with the interaction of the UAH system with real users.

3 The Statistical User Simulation Technique

In the UAH spoken dialogue system, the user simulator developed replaces the functions performed by the ASR and the NLU modules.

The methodology that we have developed for user simulation extends our work for developing a statistical methodology for dialog management (Griol et al., 2008). The user answers are generated taking into account the information provided by the simulator throughout the history of the dialog, the last system turn, and the objective(s) predefined for the dialog. A labeled corpus of dialogs is used to estimate the user model. The formal description of the proposed model is as follows:

Let \( A_i \) be the output of the dialog system (the system answer) at time \( i \), expressed in terms of dialog acts. Let \( U_i \) be the semantic representation of the user turn. We represent a dialog as a sequence of pairs (system-turn, user-turn):
Table 1: Information provided by the UAH system

<table>
<thead>
<tr>
<th>Category</th>
<th>Information provided by the user (including examples)</th>
<th>Information provided by the system</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject</strong></td>
<td>Name</td>
<td>Compilers</td>
</tr>
<tr>
<td></td>
<td>Degree, in case that there are several subjects with the same name</td>
<td>Computer science</td>
</tr>
<tr>
<td></td>
<td>Group name and optionally type, in case he asks for information about a specific group</td>
<td>A Theory A</td>
</tr>
<tr>
<td>Lecturers</td>
<td>Any combination of name and surnames</td>
<td>Zoraida</td>
</tr>
<tr>
<td></td>
<td>Optionally semester, in case he asks for the tutoring hours</td>
<td>First semester</td>
</tr>
<tr>
<td>Doctoral studies</td>
<td>Name of a doctoral program</td>
<td>Software development</td>
</tr>
<tr>
<td></td>
<td>Name of a course if he asks for information about a specific course</td>
<td>Object-oriented programming</td>
</tr>
<tr>
<td>Registration</td>
<td>Name of the deadline</td>
<td>Provisional registration</td>
</tr>
</tbody>
</table>

\[(A_1, U_1), \cdots, (A_i, U_i), \cdots, (A_n, U_n)\]

where \(A_i\) is the greeting turn of the system (the first turn of the dialog), and \(U_n\) is the last user turn. We refer to a pair \((A_i, U_i)\) as \(S_i\), the state of the dialog sequence at time \(i\).

Given this representation, the objective of the user simulator at time \(i\) is to find an appropriate user answer \(U_i\). This selection, which is a local process for each time \(i\), takes into account the sequence of dialog states that precede time \(i\), the system answer at time \(i\), and the objective of the dialog \(O\). If the most probable user answer \(U_i\) is selected at each time \(i\), the selection is made using the following maximization:

\[
\hat{U}_i = \underset{U_i \in \mathcal{U}}{\operatorname{argmax}} P(U_i | S_1, \cdots, S_{i-1}, A_i, O)
\]

where set \(\mathcal{U}\) contains all the possible user answers.

As the number of possible sequences of states is very large, we establish a partition in this space (i.e., in the history of the dialog preceding time \(i\)).

Let \(UR_i\) be the user register at time \(i\). The user register is defined as a data structure that contains the information provided by the user throughout the previous history of the dialog. The partition that we establish in this space is based on the assumption that two different sequences of states are equivalent if they lead to the same \(UR\). After applying the above considerations and establishing the equivalence relations in the histories of the dialogs, the selection of the best \(U_i\) is given by:

\[
\hat{U}_i = \underset{U_i \in \mathcal{U}}{\operatorname{argmax}} P(U_i | UR_{i-1}, A_i, O)
\]

We propose the use of a multilayer perceptron (MLP) to make the assignment of a user turn. The input layer receives the current situation of the dialog, which is represented by the term \((UR_{i-1}, A_i, O)\) in Equation 1. The values of the output layer can be viewed as the a posteriori probability of selecting the different user answers defined for the simulator given the current situation of the dialog. The choice of the most probable user answer of this probability distribution leads to Equation 1. In this case, the user simulator will always generate the same answer for the same situation of the dialog. Since we want to provide the user simulator with a richer variability of behaviors, we base our choice on the probability distribution supplied by the MLP on all the feasible user answers.

For the UAH task, the variable \(O\) is modeled taking into account the different types of scenarios defined for the acquisition of the original corpus with real users (33).

The corpus acquired with real users includes information about the errors that were introduced by
the ASR and the NLU modules during this acqui-
sition. This information also includes confidence
measures, which are used by the DM to evaluate
the reliability of the concepts and attributes gener-
ated by the NLU module.

An error simulator module has been designed
to perform error generation. The error simulator
modifies the frames generated by the user simula-
tor once the UR is updated. In addition, the error
simulator adds a confidence score to each concept
and attribute in the frames. Experimentally, we
have detected 2.3 errors per dialog in our initial
corpus. This value can be modified to adapt the er-
or simulator module to the operation of any ASR
and NLU modules.

A maximum number of twelve user turns per di-
alog was defined for acquiring a corpus using our
user simulator. A user request for closing the di-
alog is selected once the system has provided the
information defined in the objective(s) of the dia-
log. The dialogs that fulfill this condition before
the maximum number of turns are considered suc-
cessful. The dialog manager considers that the di-
alog is unsuccessful and decides to abort it when
the following conditions take place:

• The dialog exceeds the maximum number of
  user turns.

• The answer selected by the dialog manager
  corresponds with a query not required by the
  user simulator.

• The database query module provides an error
  warning because the user simulator has not
  provided the mandatory information needed
to carry out the query.

• The answer generator provides a error warn-
ing when the selected answer involves the use
of a data not provided by the user simulator.

The semantics selected for the dialog manager
is represented through the 27 possible system an-
wers defined for the UAH task. The selection of
the possible user answers is carried out using the
semantics defined for the user in the NLU module.
A dialog corpus of 3000 successful dialogs was
acquired using the user simulation technique fol-
lowing a set of scenarios covering all the possible
types of queries regarding the information sought.

Figure 2 shows an example of a dialog cor-
responding to a scenario with two objectives (to
know timetables of a given lecturer and subject,
and information about the registration in its spe-
cific degree). The confidence scores generated by
the error simulator are shown between brackets.
Sentences in natural language, which are equiva-
 lent to the labeled system and user turns, have
been incorporated to clarify the explanation.

| S1: Opening |
| Welcome to the UAH system. How may I help you? |
| U1: Lecturers [0.1] |
| I want to know information about lecturers. |
| S2: (Confirmation:Lecturers) |
| Do you want to know information about lecturers? |
| U2: Affirmation [0.9] |
| Yes. |
| S3: (Question:Lecturer_Name) |
| Tell me the name of the lecturer. |
| U3: Lecturer_Name: Ramón López-Cózar [0.8] |
| Ramón López-Cózar. |
| S4: (Question:Lecturer_Information) |
| What would you like to know about Mr. López-Cózar? |
| U4: (Question:Office) [0.9] |
| The location of his office please. |
| S5: (Answer:Office) (New-Query) |
| {The office of Ramón López-Cózar is located in the third floor of...} Anything else? |
| U5: (Question:Registration) [0.9] |
| Degree: Computer Science[0.9] |
| How to register in Computer Science. |
| S6: (Answer:Registration) (New-Query) |
| {Registration information} Anything else? |
| U6: Negation [0.9] No. |
| S7: (Closing) |
| Thank you for using the UAH system. |

Figure 2: An example of a dialog acquired by
means of the simulation technique

4 Evaluation of the Corpora

We used a set of measures to carry out the evalua-
tion of the acquired corpora based on prior work
in the dialog literature. (Schatzmann et al., 2005)
proposed a comprehensive set of quantitative eval-
uation measures to compare two dialog corpora.
These measures were adapted for our purpose and
can be divided into three types:
High-level dialog features

<table>
<thead>
<tr>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of turns per dialog</td>
</tr>
<tr>
<td>Percentage of different dialogs</td>
</tr>
<tr>
<td>Number of repetitions of the most seen dialog</td>
</tr>
<tr>
<td>Number of turns of the most seen dialog</td>
</tr>
<tr>
<td>Number of turns of the shortest dialog</td>
</tr>
<tr>
<td>Number of turns of the longest dialog</td>
</tr>
</tbody>
</table>

Dialog style/cooperativeness measures

- **System dialog acts**: Confirmation of concepts and attributes, Questions to require information, and Answers generated after a database query.
- **User dialog acts**: Request to the system, Provide information, Confirmation, Yes/No answers, and Other answers.

Figure 3: Evaluation measures used to compare the acquired corpora

- High-level dialog features: These features evaluate the duration of the dialogs, the amount of information transmitted in the individual turns, and how active the dialog participants are.

- Dialog style/cooperativeness measures: These measures analyze the frequency of the different speech acts and study, for example, the proportion of actions which are goal-directed vs. dialog formalities.

- Task success/efficiency measures: These are computations of the goal achievement rates and goal completion times.

We have defined six high-level dialog features for the evaluation of the dialogs: the average number of turns per dialog, the percentage of different dialogs without considering the attribute values, the number of repetitions of the most seen dialog, the number of turns of the most seen dialog, the number of turns of the shortest dialog, and the number of turns of the longest dialog. Using these measures, we tried to evaluate the success of the simulated dialogs as well as their efficiency and variability with regard to the different objectives.

For dialog style features, we have defined a set of system/user dialog acts. On the system side, we have measured the frequency of confirmations, questions that require information, and system answers generated after a database query. We have not taken into account the opening and closing system turns. On the user side, we have measured the percentage of turns in which the user carries out a request to the system, provide information, confirms a concept or attribute, Yes/No answers, and other answers not included in the previous categories.

We have not considered task success/efficiency measures in our evaluation, since only the dialogs that fulfill the objectives predefined in the scenarios have been incorporated into our corpora. We have considered successful dialogs those that fulfill the complete list of objectives defined in the corresponding scenario. Figure 3 summarizes the complete set of measures used in the evaluation.

5 Evaluation Results

To compare the two corpora, we have computed the mean value for each corpus with respect to each of the evaluation measures shown in the previous section. Then two-tailed t-tests have been employed to compare the means across the two corpora as described in (Ai et al., 2007a). All differences reported as statistically significant have p-values less than 0.05 after Bonferroni corrections.

5.1 High-level Dialog Features

As stated in the previous section, the first group of experiments covers the following statistical properties: i) Dialog length in terms of the average number of turns per dialog, number of turns of the shortest dialog, number of turns of the longest dialog, and number of turns of the most seen dialog; ii) Number of different dialogs in each corpus in terms of the percentage of different dialogs and the number of repetitions of the most seen dialog; iii) Turn length in terms of actions per turn; iv) Participant activity as a ratio of system and user actions per dialog.
Table 2: Results of the high-level dialog features defined for the comparison of the three corpora

<table>
<thead>
<tr>
<th></th>
<th>Initial Corpus</th>
<th>Simulated Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average number of user</td>
<td>4.99</td>
<td>3.75</td>
</tr>
<tr>
<td>turns per dialog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage of different</td>
<td>85.71%</td>
<td>77.42%</td>
</tr>
<tr>
<td>dialogs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of repetitions</td>
<td>5</td>
<td>27</td>
</tr>
<tr>
<td>of the most seen dialog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of turns</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>of the most seen dialog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of turns</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>of the shortest dialog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of turns</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>of the longest dialog</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the results of the comparison of the high-level dialog features. It can be observed that all measures have similar values in both corpora. The more significant difference is the average number of user turns. In the four types of scenarios, the dialogs acquired using the simulation technique were shorter than the dialogs acquired with real users. This can be explained by the fact that there was a number of dialogs acquired with real users in which the user asked for additional information not included in the definition of the corresponding scenario once the dialog objectives had been achieved.

5.2 Dialog Style and Cooperativeness

Tables 3 and 4 respectively show the frequency of the most dominant user and system dialog acts. Table 3 shows the results of this comparison for the system dialog acts. It can be observed that there are also only slight differences between the values obtained for both corpora. There is a higher percentage of confirmations and questions in the corpus acquired with real users due to its higher average number of turns per dialog.

Table 4 shows the results of this comparison for the user dialog acts. The most significant difference between both corpora is the percentage of turns in which the user makes a request to the system, which is lower in the corpus acquired with real users. This is possibly because it is less probable that simulated users provide useless information, as it is shown in the lower percentage of the users turns classified as Other answers.

6 Conclusions

In this paper, we have presented a comparison between two corpora acquired using two different techniques. Firstly, we gathered an initial dialog corpus from real user-system interactions. Secondly, we have employed a statistical user simulation technique based on a classification process to automatically obtain a corpus of simulated dialogs. Our results show that it is feasible to acquire a realistic corpus by means of the simulation technique. The experimental results reported indicate that the simulated and real interactions corpora are very similar in terms of number of user turns, user and system dialog style and cooperativeness, and most frequent dialogs statistics. As future work, we plan to employ the simulated dialogs for evaluation purposes and for extracting valuable information to optimize the current dialog strategy.

References


Z. Callejas and R. López-Cózar. 2008. Relations between de-facto criteria in the evaluation of a spoken
Table 3: Percentages of the different types of system dialog acts in both corpora

<table>
<thead>
<tr>
<th></th>
<th>Initial Corpus</th>
<th>Simulated Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmation of concepts and attributes</td>
<td>13.51%</td>
<td>12.23%</td>
</tr>
<tr>
<td>Questions to require information</td>
<td>18.44%</td>
<td>16.57%</td>
</tr>
<tr>
<td>Answers generated after a database query</td>
<td>68.05%</td>
<td>71.20%</td>
</tr>
</tbody>
</table>

Table 4: Percentages of the different types of user dialog acts in both corpora

<table>
<thead>
<tr>
<th></th>
<th>Initial Corpus</th>
<th>Simulated Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Request to the system</td>
<td>31.74%</td>
<td>35.43%</td>
</tr>
<tr>
<td>Provide information</td>
<td>21.72%</td>
<td>20.98%</td>
</tr>
<tr>
<td>Confirmation</td>
<td>10.81%</td>
<td>9.34%</td>
</tr>
<tr>
<td>Yes/No answers</td>
<td>33.47%</td>
<td>32.77%</td>
</tr>
<tr>
<td>Other answers</td>
<td>2.26%</td>
<td>1.48%</td>
</tr>
</tbody>
</table>


