EVALUATION OF A USER-ADAPTED SPOKEN LANGUAGE DIALOGUE SYSTEM

Measuring the Relevance of the Contextual Information Sources

Juan Manuel Lucas-Cuesta, Fernando Fernández-Martínez, G. Dragos Rada
Syaheerah L. Lutfi and Javier Ferreiros
Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain
{juanmak, ffm, georgedragos, syaheerah, jfl}@die.upm.es

Keywords: Spoken language dialogue systems, User interfaces, Contextual information, User profiles, Natural language processing, Real-user evaluation.

Abstract: We present an evaluation of a spoken language dialogue system with a module for the management of user-related information, stored as user preferences and privileges. The flexibility of our dialogue management approach, based on Bayesian Networks (BN), together with a contextual information module, which performs different strategies for handling such information, allows us to include user information as a new level into the Context Manager hierarchy. We propose a set of objective and subjective metrics to measure the relevance of the different contextual information sources. The analysis of our evaluation scenarios shows that the relevance of the short-term information (i.e. the system status) remains pretty stable throughout the dialogue, whereas the dialogue history and the user profile (i.e. the middle-term and the long-term information, respectively) play a complementary role, evolving their usefulness as the dialogue evolves.

1 INTRODUCTION

The design of spoken dialogue systems that can adapt to their users is today a common practice. The goal is not only to modify the behaviour of a system to better react to a particular speaker, but also to play a more proactive role in the dialogue, anticipating the users' desires, and proposing them specific actions that the system foresees. Therefore it is important to accurately model the speakers' characteristics that are relevant for the dialogue (Zukerman and Litman, 2001).

Several evaluation methodologies have been proposed for measuring the relevance of a user model in different research fields (Chin, 2001; Gena, 2005). However, it is difficult to find performance figures from real-world applications that can be extrapolated to other systems or be worldwide accepted, as all of them are directly related to an specific dialogue system. Nonetheless, there is a general agreement on "usability" as the most important performance figure (Dybkjaer et al., 2004), even more than others widely used like "naturalness" or "flexibility".

Different evaluation frameworks that can predict user satisfaction from the analysis of objective metrics (Walker et al., 1997; Möller et al., 2007), as well as several dialogue systems with the ability to change the dialogue initiative or the confirmation mechanisms (Litman and Pan, 2002), have been currently developed. However there are not any standard for assessing spoken dialogue systems, despite different de-facto standards are commonly used (Callejas and López-Cózar, 2008). Furthermore, the definition of metrics to compare the relevance of different contextual information sources in a user-based speech system is still under development.

We have developed a spoken dialogue system with a user-related information manager as part of its contextual information knowledge. We have defined several metrics to assess the usefulness of the information sources that the system takes into account for solving dialogues, and the relevance of the user models when they are used to suggest hypotheses related to the users' preferences and privileges.

The rest of the paper is organized as follows. Section 2 describes the baseline dialogue system and the new user-related information manager. The initial assessment of the system is presented in Section 3. Finally, Section 4 shows the main results of our work.

2 PROTOTYPE DESCRIPTION

We have developed a user-adapted spoken dialogue system (SDS) for controlling different devices. We have evaluated the performance of the prototype when controlling a Hi-Fi device.

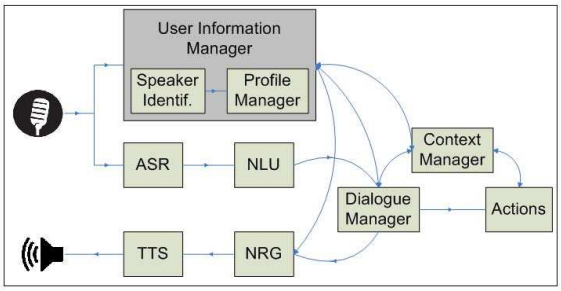

Figure 1: Block diagram of the spoken dialogue system.

Figure 1 shows a block diagram of our conversational interface. The system consists of an automatic speech recognition module (ASR), which translates the audio signal into a text hypothesis of what the user has said; a language understanding module (NLU), that extracts the semantics of the user's utterance; the dialogue manager (DM) (detailed in Section 2.1), which makes use of the semantic information, together with the information gathered during previous dialogues, to determine the actions over the system that the user wants to fulfill, and to provide the user with feedback regarding the current dialogue turn; the context manager (CM), which holds the information of the previous interactions; an execution module, that translates the actions to perform into IR commands; the response generator module (NRG), which makes use of the semantic information provided by the dialogue manager to generate a text output, and a text-to-speech module (TTS), that synthesizes the message to the user. To include user-related information, we also need a speaker identification module and a manager of user information, stored as user profiles. We group both modules into a User Information Manager (UIM), that will be detailed in Section 2.2.

2.1 Dialogue Management and Dialogue Context

The Dialogue Manager (DM) controls the dialogue flow using the information provided by the speaker, parsed into *dialogue concepts*. We have developed our DM following a probabilistic approach, based on Bayesian Networks (BN), for inferring which actions the user wishes to perform, and which concepts are needed to fulfill those actions (Fernández et al., 2005). We apply two inference mechanisms. The *forward inference* makes use of the dialogue concepts referred by the user, for inferring the *dialogue goals*, that is, the actions the user wants to fulfill.

Using both concepts and goals, the DM applies a *backward inference* to determine whether the given concepts are enough to fulfill these goals. In this case, the system will send the corresponding IR commands to the Hi-Fi device. Otherwise, the system tries to recover those *missing concepts* required to solve the dialogue, using the *Context Manager*.

We have defined 58 concepts, divided into *parameters* (16) to set up (e.g. the volume of the Hi-Fi device), *values* (20) that the different parameters can take, and *actions* (22) to be performed (e.g. modify the volume). We have also defined 15 goals, according to the available functionality of the Hi-Fi device.

The mission of the Context Manager (CM) consists of solving any ambiguity that may arise in the dialogue, using different contextual information handling strategies throughout the dialogue. It consists of three structures, which can be classified according to the recentness of their contents. The *system status* (the short-term one) stores the current values of the Hi-Fi functionalities (CD track, volume, and so on). The *dialogue history* (a middle-term memory) contains the concepts referred by the user since the beginning of the current interaction. A mechanism is applied in such a way that its information is permanently updated coherently to the current state of the dialogue, while discarding information that becomes too old. Finally, the *user profile* (the long-term one), detailed in Section 2.2, stores information of each user since his or her first interaction.

The CM works as follows. When the DM has to recover a missing concept, it first checks the system status. If it contains such a concept, the system recovers it and executes the appropriate action, finishing the current dialogue. Otherwise, the DM checks the dialogue history. If the system is unable to retrieve any concept from it, the DM further checks the user profile, that may suggest one or several concepts based on the user's preferences. Finally, if the system is still unable to retrieve a concept using any of the above strategies, it will request the user to provide the missing concepts, initiating a new turn.

2.2 Profile Management: Modeling User Preferences and Privileges

We have developed a new module, the *User Information Manager*, which adds user-related information to the dialogue flow. As can be seen in Figure 1, this module consists of a speaker identification module, and a *Profile Manager* (PM), that updates the

Table 1: Excerpt of dialogue with the contents stored in, and retrieved from (in boldface), dialogue history and user profile.

Dialogue turn		Dialogue History	User Profile
...	...	$\emptyset$	CD3: preference
User	*Switch the Hi-Fi on.*		**CD3**
System	Turning Hi-Fi on. Playing track 1 of CD 3.	Hi-Fi ON, CD 3, Track 1	
U.	*Play next track.*	**CD3, Track1**	
S.	Playing track 2 of CD 3.	Hi-Fi ON, CD 3, Track 2	
U.	*Volume.*		
S.	What do you want to do with the volume?		
U.	*Raise it to five.*		
S.	Volume 5 selected.	Hi-Fi ON, CD 3, Track 2, Volume 5	

information of the different speakers as the dialogue evolves (Lucas-Cuesta et al., 2009).

A user profile consists of two types of information fields: *static*, used for information such as the speaker's name, gender, age, language, and so on, and *dynamic*, used for information that changes during the dialogue. The dynamic component is composed of two different entities:

- *Usage permissions*, which allow to add restrictions to each user. For instance, a child may not be allowed to listen to an adult-content radio tune.
- *User preferences*, representing the contents that each user prefers to play (i.e. a given CD or radio tune).

User preferences are measured as the frequency of references made by each user of each functionality since the creation of his or her profile. A given functionality becomes a preference for the user when the quotient between its counts and the counts of each of the rest of functionalities exceeds a threshold. We make this comparison among items that share the same type of functionality (i.e. we compare the CD, the tape, and the radio, at a level 'preferred source', each CD unit at a level 'preferred CD', and so on.

Table 1 shows an example of how the Dialogue Manager retrieves missing concepts using the contextual information sources, and how the information is stored in the user profile. For example, prior to the first interaction presented in the table (i.e. 'Switch the Hi-Fi on'), the current user had previously interacted with the system. The system thus had a profile available at the beginning of the dialogue. If the profile stores a preference, the system can use it for anticipating on an action over the Hi-Fi before the user explicitly asks for it. In the example, as 'CD 3' is a preference (over CD 2, for instance), the system starts playing the first track of CD 3.

During the next interaction, when the user wants to play the next track (without any explicit reference to neither a CD nor a track), the Dialogue Manager can retrieve the required information for performing the action by checking the dialogue history. This history contains 'CD 3' and 'Track 1' as the last values the user asked for, so it uses them to infer that the user wants to play the second track of the current CD.

Finally, if the user asks for any functionality that is not yet stored in any contextual information source (as is the case of the volume, in the example), the system will start a new dialogue turn, asking the user for the information it needs to fulfill the requested action.

Summarizing, the combined use of the different information sources allows the Dialogue Manager to improve its performance by conducting more efficient dialogues, reducing their number of turns, and reusing any useful information the users provide during both their recent and previous interactions.

3 INITIAL EVALUATION

A total of 9 speakers, 3 female and 6 male, with ages between 24 and 28 years, were recruited. They were classified as 'novice' (5) or 'expert' (4) according to their previous experience in interacting with spoken dialogue systems.

Despite the reduced number of evaluators, which could imply a lack of significance on the results of the evaluation, our main goal was to define a set of metrics regarding the relevance of each information source, and to perform an initial assessment of the performance of the full dialogue system.

We were interested in assessing the relevance of each contextual information source, and the user satisfaction regarding the suggestions our system makes.

The evaluation was divided into two scenarios (Fernández et al., 2008) in which the evaluators controlled the Hi-Fi device by means of speech. In both scenarios the users were allowed to interact with the SDS without any kind of restriction. Before starting each scenario, the dialogue history and the system status were initialized to an empty default state.

Prior to the first scenario, evaluators were in-

formed about the available functionality (i.e. CD, radio, volume, and so on), and the contents of the Hi-Fi media (i.e. rock, news, sports, and so on).

In the *first scenario*, the profile of each evaluator was empty. The system updated it throughout the dialogue with the values referred by the user. The goal of this scenario was to determine the relevance of each information source when the user profile does not exist during the first interactions of a new speaker.

In the *second scenario*, the system made use of the profile created during the first scenario. We measured the contribution of the profile keeping its information when the system status and the dialogue history were empty (i.e. at the beginning of the interaction), and how the contribution of the different information sources varies throughout the scenario. Additionally, we considered the usage privileges. We wanted to measure the relevance of each information source when retrieving missing concepts, as well as the reliability of the permission control.

Table 2: Average number of turns for each scenario (Nov.: Novice, Exp.: Expert).

	Scenario 1		**Scenario 2**	
	Nov.	**Exp.**	**Nov.**	**Exp.**
Turns	80.0	62.5	59.4	53.5

The average number of turns for each scenario an experience level can be seen in Table 2. We have an average of 72.22 turns for the Scenario 1, and 56.78 for the second one, for each evaluator, so despite the reduced number of speakers, the results will show the trends of each of our metrics.

Finally, we asked the evaluators to fill in a subjective survey to gather their opinions about the performance of the system, making a special emphasis on the preference suggestion mechanism and the application of usage privileges.

3.1 Objective Evaluation

We have obtained several automatically-collected metrics (Fernández et al., 2008). As we want to stress the relevance of each information source could be at different times throughout the interaction, we have also defined new metrics to assess the usefulness of each of them. The relevant metrics regarding this usefulness (averaged with respect to the number of evaluators), shown in Table 3, are the following:

- **% Context-dependent Turns** (M1)**:** percentage of dialogue turns in which the Dialogue Manager checks the CM for information.
- **% System Requests** (M2)**:** percentage of dialogue turns in which the system requested information to the user.
- **% Concept Recovery Turns** (M3)**:** percentage of context-dependent turns in which the Dialogue Manager successfully retrieves a concept from the CM.
- **% Recovery Turns from the System Status** (M4)**:** percentage of context-dependent turns in which the system retrieves any concept from the system status.
- **% Recovery Turns from the Dialogue History** (M5)**:** percentage of context-dependent turns in which the system retrieves any concept from the dialogue history.
- **% Recovery Turns from the User Profile** (M6)**:** percentage of context-dependent turns in which the system retrieves any concept from the user profile.

The sum of the three last metrics (M4 to M6) can exceed 100%, because the three information sources can be successfully checked in the same turn (i.e. different concepts can be simultaneously retrieved from the different memories).

Table 3: Objective evaluation results (Nov.: Novice, Exp.: Expert).

		Expertise		
Metrics		Nov.	Exp.	**ALL**
M1	% Cont. dep. turns	52.6	47.1	50.19
M2	% Sys. req.	21.5	13.3	17.86
M3	% Concept rec. turns	96.0	95.6	95.85
M4	% Sys. status rec. turns	83.1	80.2	81.79
M5	% Dial. hist. rec. turns	23.1	27.8	25.16
M6	% User prof. rec. turns	12.6	16.9	14.52

About half the total of dialogue turns (50.19%, M1) implies a query in the CM. Most of these turns (95.85%, M3) correspond to retrieval of missing concepts. This value shows the efficiency of our context information handling strategy, which can improve the naturality and flexibility of our dialogue management approach: using the CM, each time the system retrieves a concept, a new request to the user is avoided, thus reducing the number of dialogue turns and allowing speakers to perform actions in less interactions.

The less knowledge of 'novice' users can be seen on the more turns they need to fulfill their goals (69.7 vs. 58 on average) and the higher percentage of system requests (21.5% vs. 13.31%, M2), each of which implies a new turn. The DM also tends to increase context checkings with 'novices' (52.62% vs. 47.16%, M1), which could imply that 'experts' tend to express their goals in a single, more complete turn.

The most relevant information source for the retrieval of missing concepts is the system status (81.79% of turns), as we may expect, since it is the information source firstly checked, consistently with our design of the Context Manager and our checking criterion. The dialogue history and the user profile have an apparently lesser usefulness, due to their update mechanism and their request order (first the dialogue history, then the profile).

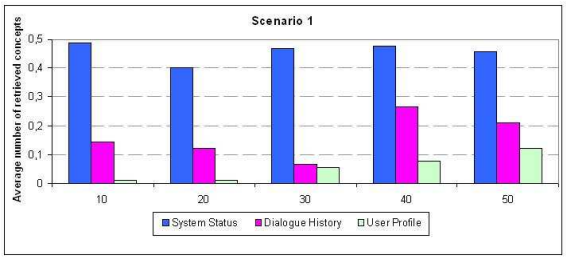

Figure 2: Scenario 1: number of retrieved concepts each 10 turns.

However, these percentages do not present an evolution of the usefulness of each information source as the dialogue evolves. In order to show this, we have considered the number of dialogue concepts retrieved from each information source as a new metric for assessing the usefulness of the CM. Figures 2 and 3 show the evolution of the average number of concepts retrieved from each source during Scenario 1 and Scenario 2, respectively. Since there are too many turns in which the system does not make use of the context information (49.81%), we have averaged the number of retrieved concepts using 10-turn windows.

The relevance of the system status is roughly constant throughout the dialogue in both scenarios, because this is the information source the system first checks, and the one which stores the most recent information.

We want to stress the evolution of the usefulness of the dialogue history and the user profile. At the beginning of Scenario 1 the user profile is empty. Therefore, its relevance is near zero until the interactions allow the system to suggest usage hypotheses (included as retrieval of missing concepts). Figure 2 shows that the system needs an average of 20 dialogue turns to start proposing actions to the users, based upon their preferences. On the other hand, when there are user profiles at the beginning of the interaction (Figure 3), the relevance of the user profile is higher at the initial dialogue turns, when the dialogue history is empty.

The relevance of the dialogue history has a similar behaviour in both scenarios, fitting an increasing trend of their usefulness as the dialogue evolves.

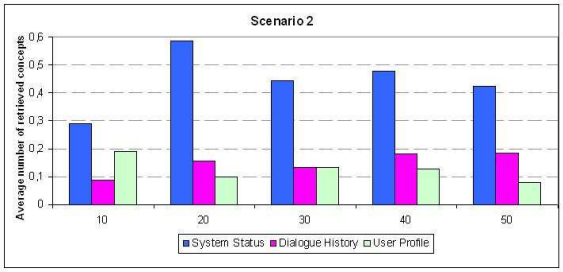

Figure 3: Scenario 2: number of retrieved concepts each 10 turns.

This behaviour relies on the update of both the history and the profiles with the concepts that the users refer to during their interactions.

These results imply that both information sources (dialogue history and user profile) are complementary, and that both are interesting in order to make the interaction with the system more efficient and natural.

3.2 Subjective Evaluation

The results of the subjective evaluation are based on a survey the evaluators were asked to fill in after both scenarios. They had to rate each question with a score between 1 (strongly disagree) and 5 (strongly agree). Table 4 shows the results of the questions related to the User Information Manager.

As we can see, 'novice' users seem to have difficulties to perceive the usefulness of the information management strategies (i.e. Q1), despite the system checks the Context Manager more often than when interacting with 'experts', as the objective metrics proved. Even so, 'novices' perceive that the system can propose their preferences (Q3).

Additionally, despite the good rating of the overall behaviour of the system (4.22, Q4), the 'expert' users are less bound to use speech for controlling a Hi-Fi device. This happens because their higher knowledge of the possibilities and limitations of the system. Nevertheless, their score is still above 3, which suggests their positive opinion about the possibilities of using speech control for these systems.

4 CONCLUSIONS

We have presented an evaluation of a user-adapted spoken dialogue system. We include user models as part of the Context Manager, applying usage privileges and suggesting actions from the user profile when the system searches the Context Manager to retrieve missing concepts, reducing the number of turns

Table 4: Subjective evaluation results (Nov.: Novice, Exp.: Expert).

Survey		Nov.	Exp.	ALL
Q1	Was the system able to act coherently with its context?	3.6	4.5	4
Q2	Was the system easy to use?	4.2	4.25	4.22
Q3	Did the system suggest your preferred styles?	4.2	3.75	4
Q4	Global rate	4.2	4.25	4.22
Q5	Would you use this system instead of a remote control?	4.4	3.5	4

needed to fulfill actions.

Our system is able to propose usage suggestions under certain hypotheses, thus giving an important degree of proactiveness to the system. We are currently working on more ambitious schemes for better exploiting user profiles.

Regarding the initial evaluation, we have defined several metrics to assess the usefulness of each information source (system status, dialogue history, and user profile). The results show that the most relevant source of information is the short-term one (the system status). The dialogue history and the user profile have a complementary behaviour, supporting each other as the dialogues evolve. The subjective evaluation shows that the users perceive the information manager as a useful element of the dialogue system, as it can anticipate the actions required, and can apply usage privileges.

Now we are applying smoothing strategies to improve the decision of the speaker identification module included in the User Information Manager, using dialogue-based information (a dialogue with incomplete actions, the time between successive turns, and so on) in an effort to reduce identification errors.

We are also defining a new scenario in which several speakers alternatively interact with the system. We will measure the accuracy of the identification module, as well as the application of preferences and privileges for different users.

ACKNOWLEDGEMENTS

This work has been partially supported by the Spanish Ministry of Science and Innovation under contracts TIN2008-06856-C05-05 (SD-TEAM UPM) and DPI2007-66846-C02-02 (ROBONAUTA), and the Spanish Ministry of Education under the Program of University Personnel Training (AP2007-00463).

REFERENCES

Callejas, Z. and López-Cózar, R. (2008). Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication*, 50:646–665.

Chin, D. (2001). Empirical Evaluation of User Models and User-Adapted Systems. *User Modeling and User-Adapted Interaction*, 11:181–194.

Dybkjaer, L., Bernsen, N., and Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication*, 43:33–54.

Fernández, F., Blázquez, J., Ferreiros, J., Barra, R., Macías-Guarasa, J., and Lucas-Cuesta, J. (2008). Evaluation of a Spoken Dialogue System for Controlling a Hifi Audio System. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT08)*, pages 137–140.

Fernández, F., Ferreiros, J., Sama, V., Montero, J., San-Segundo, R., and Macías-Guarasa, J. (2005). Speech Interface for Controlling a Hi-Fi Audio System Based on a Bayesian Belief Networks Approach for Dialog Modeling. In *Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH05)*, pages 3421–3424.

Gena, C. (2005). Methods and Techniques for the Evaluation of User-Adaptive Systems. *The Knowledge Engineering Review*, 20(1):1–37.

Litman, D. and Pan, S. (2002). Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction*, 12:111–137.

Lucas-Cuesta, J., Fernández, F., Salazar, J., Ferreiros, J., and San-Segundo, R. (2009). Managing Speaker Identity and User Profiles in a Spoken Dialogue System. *Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN)*, 43:77–84. ISSN: 1135-5948.

Möller, S., Smeele, P., Boland, H., and Krebber, J. (2007). Evaluating Spoken Dialogue Systems According to De-Facto Standards: a Case Study. *Computer, Speech and Language*, 21:26–53.

Walker, M., Litman, D., Kamm, C., and Abella, A. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings ACL/EACL*, pages 271–280.

Zukerman, I. and Litman, D. (2001). Natural Language Processing and User Modeling: Synergies and Limitations. *User Modeling and User-Adapted Interaction*, 11:129–158.