# Development of a Tour–Guide Robot Using Dialogue Models and a Cognitive Architecture

Héctor Avilés, Montserrat Alvarado-González, Esther Venegas, Caleb Rascón, Ivan V. Meza, and Luis Pineda

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS)
Universidad Nacional Autónoma de México (UNAM)
{haviles,mon}@turing.iimas.unam.mx, esthervenegas@hotmail.com,
{caleb,ivanvladimir}@turing.iimas.unam.mx, luis@leibniz.iimas.unam.mx

**Abstract.** In this paper, we present the development of a tour–guide robot that conducts a poster session through spoken Spanish. The robot is able to navigate around its environment, visually identify informational posters, and explain sections of the posters that users request via pointing gestures. We specify the task by means of dialogue models. A dialogue model defines conversational situations, expectations and robot actions. Dialogue models are integrated into a novel cognitive architecture that allow us to coordinate both human–robot interaction and robot capabilities in a flexible and simple manner. Our robot also incorporates a confidence score on visual outcomes, the history of the conversation and error prevention strategies. Our initial evaluation of the dialogue structure shows the reliability of the overall approach, and the suitability of our dialogue model and architecture to represent complex human–robot interactions, with promising results.

**Keywords:** Dialogue systems, service robots, human–robot interaction, pointing gestures.

## 1 Introduction

The development of autonomous service robots has received considerable attention in the past decade. These robots should be capable to engage in meaningful, coherent, task–oriented multimodal dialogues with humans to carry out daily–life activities. In addition, these robots must also include a variety of perceptual and behavioural abilities to effectively achieve their goals. When designing such robots, it is desirable to consider conceptual frameworks for the description of the tasks in which domain knowledge, interaction structures, and robot capabilities are integrated in a simple, comprehensible, and coordinated form.

For this purpose, we present the development of a tour–guide robot using dialogue models (*DMs*) [1]. In our context, a DM is a highly abstract interaction protocol, stated declaratively at an intentional level, that allow us to easily specify robot tasks, independently of the input and output capabilities of the robot. DMs are integrated into a novel cognitive architecture that fully coordinates HRI and robot capabilities [2]. Along with DMs, this architecture offers a

A. Kuri-Morales and G.R. Simari (Eds.): IBERAMIA 2010, LNAI 6433, pp. 512–521, 2010.

flexible development framework to design complex tasks for service robots [3]. Our robot is able to conduct a poster session in which speech and pointing gestures can be used by human users to further along the interaction. As part of the tour, the robot visits different posters by request of the user through the scenario. The dynamics of the multimodal interaction is enriched by the use of confidence values and the history of the conversation. Moreover, the flow of the conversation is aided by error prevention and recovery strategies. To evaluate this proposal, we performed an initial evaluation of the overall functionality of the system, the interaction modalities, and the user satisfaction, using a real robot. Our results show that the robot accomplished the task the majority of the time and that the users were satisfied with the experience.

The rest of the paper is organized as follows. Section 2 introduces previous proposals on service robots closely related to our work. A brief description of the cognitive architecture and dialogue models is presented in Section 3. Section 4 is devoted to the specification of the tour–guide task. Section 5 describes the algorithms that support language, vision and motor behaviors of our robot. Section 6 presents the evaluation procedure and results. Finally, conclusions and future work are discussed in Section 7.

## 2 Related Work

One of the first service robots with multimodal communication capabilities is ROBITA [4]. This robot answers questions in spoken Japanese about the location of a person in an office environment using speech and pointing gestures. ALBERT [5] grasps objects following speech and hand posture commands of the type "Hello Albert, take this". Jido [6] tracks head and hands of a user, and answer requests such as "Come here". Markovito [7] recognizes spoken commands in Spanish, delivers objects and spoken messages, and also identifies nine different types of gestures. Notwithstanding the advanced capabilities provided by these robots, none of these examples provides an integration framework focused on the interaction between the robot and the user.

Other approaches have applied dialogue models to describe the task. An example is BIRON [8], a robot that uses pointing gestures and visual information of the environment to resolve spoken references to objects. In [9] a robot controlled by a dialogue system based on simple finite state machines is presented. The dialogue system of the Karlsruhe humanoid robot [10] coordinates object manipulation, localization, recognition and tracking of a user, recognition of gaze and pointing gestures, and interactive learning of dialogues. However, the majority of these examples were build focused on a specific task which complicates the generalization of their frameworks. The overall approach presented in this document is a contribution to solve these problems.

## 3 A Cognitive Architecture Oriented towards HCI

The cognitive architecture that we use is a three–layer structure focused primarily on multimodal human–computer interaction (*HCI*) –See [2] for a graphical

description of this architecture. Perception and action processes are arranged into two main branches that we labeled *recognition and interpretation* and *specification and rendering*.

The recognition and interpretation branch provides an abstract interpretation of the world accordingly to the *expectations* defined by the representation/inference layer –described below. This branch process inputs of different perceptual modalities, and there are as many instances of this branch as required. Each instance specializes in a single perceptual capability of the robot. The specification and rendering branch corresponds to behavioral responses. This branch provides the means in which the robot can influence and act over the world. Similar to the recognition and interpretation branch, there could be as many instances as needed.

The top layer of the architecture coordinates the processes described above, and also keeps track of the interactions between the robot and its environment. This layer is specified by a dialogue model, that is a set of expected multimodal *situations* which defines the interaction protocol. A DM can be described as a directed graph, in which situations are represented as nodes, and edges are labeled with expectation–action pairs. Whenever an expectation is met, the corresponding action is triggered. A simple example is depicted in Fig. 1. This example shows a situation $S_i$, with an edge $\alpha : \beta$ that is directed to the next situation $S_j$. Situations can have one or more input and output expectation–action pairs, and are typed according to the modality of the expected input; the main types are *listening* –or linguistic– and *seeing* –or visual. There is also a special class of situations called *recursive*, in which a full DM is embedded, and allows to modularize the specification. All DMs have one initial and one or more final situations. DMs are based on *recursive transition networks*, augmented with functions standing for expectations, actions and next situations [1].

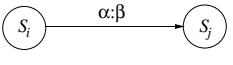


**Fig. 1.** Situations in a dialogue model

This cognitive architecture not only provides a "natural" representation of the flow of different HCI tasks [11], but it also simplifies the integration of perceptual and behavioural algorithms –independently of their implemented modality– through the definition of the expectations. This abstraction could be particularly important whenever the system process inputs perceived by different modalities at the same time.

## 4 Specification of the Task

The dialogue model developed the tour-guide robot application is described as follows. Our main dialogue model $M$ consists of an initial greeting situation $S$ and four recursive situations which direct the execution of the task: a)

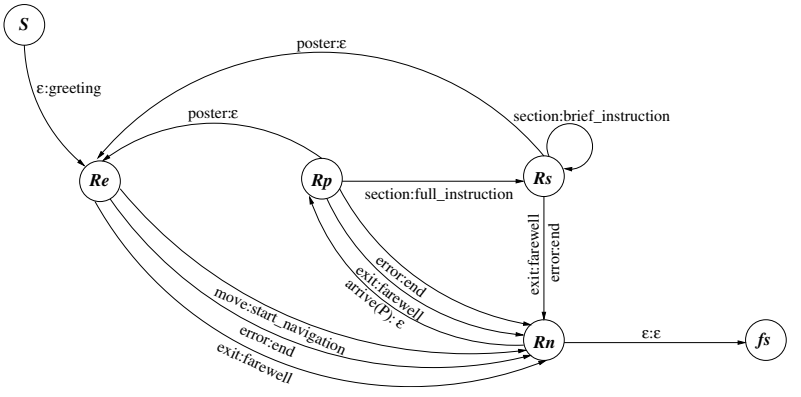


**Fig. 2.** Main dialogue model

the choose–poster sub-dialogue *Re*, b) the see–poster sub-dialogue *Rp*, c) the pointing–section sub-dialogue *Rs*, d) the move–to–poster sub-dialogue *Rn*, and e) the final situation *fs*. The main dialogue model is shown in Fig. 2.

Situation $S$ is triggered by the execution of the program. The robot verbally presents itself and explains the goal of the task. Sub–dialogue *Re* leads the robot to ask the user to choose a poster from a list of available posters –this list is dynamically built depending on the history of the conversation. The robot listens for the poster selected by the user. Then, it navigates around its environment from its initial point to the fixed coordinates of the poster. Once there, it announces its arrival and indicates that it will visually recognize the poster –situation *Rn*. The visual module evaluates the quality of the recognition of the poster and gives the result to the *Rp* sub–dialogue. If the evaluation is high, the dialogue firmly states the poster, if it is medium, the dialogue is rendered as if the robot has some uncertainty on the matter. The robot then offers the user to explain a section of the current poster or to visit another one. If the quality is low or the robot did not see the poster, then the dialogue expresses doubts about the quality of the recognition and only offers the user to visit another poster. The previous design principle, as part of an adequate specification, is an effective error prevention strategy. If the user requires an explanation of a section of the poster, the robot asks him/her to point to the desired text or image, for some seconds over the section, after a beep sound. Then, the visual module evaluates the quality of the identification of the section and gives the result to the *Rs* sub–dialogue. The degree of certainty of the identification is mentioned during the conversation. Then the robot explains the section captured and offers the user to explain another section of the current poster or to visit another one. Our task specification also include recovery strategies for speech interpretation misunderstandings. In these cases, the system provides the appropriate feedback the user in order to proceed with the interaction.

**Table 1.** Example of the interaction between the robot and a user; robot and user actions are indicated in italics

| Turn | Actor | English Translation |
|---|---|---|
| 1 | ROBOT | Hi, I am Golem. I can describe information about posters, and also its sections. We have the posters of "Golem", "Mesoamerica" and "Encryption". Which one do you want me to explain? |
| 2 | USER | I want to see the Golem poster. |
| 3 | ROBOT | Ok, let's go to the Golem poster. *The robot moves from its initial position to the poster.* |
| 4 | ROBOT | I've arrived to the Golem poster. Please, hold on a second while I confirm if it's the correct poster. *The visual module indicates that the quality of the recognition of the poster is medium.* |
| 5 | ROBOT | I'm not sure, but I think that it's the Golem poster. Do you want me to explain a section or do you want to see another poster? |
| 6 | USER | No user response. |
| 7 | ROBOT | Sorry, I did not hear you. |
| 8 | USER | Please, explain a section. |
| 9 | ROBOT | Ok, please point to the desired text, image or section, by making a soft movement with your hand for some seconds over it, after the beep sound. *The user do nothing, and the visual module indicates that the quality of the identification is low.* |
| 10 | ROBOT | I could not see you pointing at any region, maybe because you did not point or because I did not see the region you pointed at. Do you want me to explain a section or do you want to see another poster? |
| 11 | USER | Explain a section. |
| 12 | ROBOT | Ok, don't forget to point after the beep. |
| 13 | USER | *The user points to a picture of the robot.* |
| 14 | ROBOT | *The quality of the of the identification is high.* It's me! Do you want me to explain a section or do you want to see another poster? |
| 15 | USER | I want to finish. |
| 16 | ROBOT | Ok, thank you for your visit, goodbye! |

Any time the user wants the robot to finish the task or whenever any of the DMs indicate a major error, the robot returns to its initial position and the dialogue ends –situation *fs*. Table 1 shows an example of an interaction between the robot and a user.

## 5 Robot Capabilities

In this section we describe the visual, speech and navigation capabilities of our robot. The three modules that perform these procedures are herewith explained.

### 5.1 Visual Module

This module is in charge of two main tasks: a) gesture analysis and b) poster recognition. Pointing gesture analysis involves two main steps: i) 2D arm segmentation, and ii) identification of the section pointed by the user. For segmentation, the arm is spotted into a binary image by fusioning motion data, and the absolute difference of the first view of the poster in the actual robot position and the current image in the stream. Data fusion is performed by following a simple logic $AND$ rule.

Simple region growing techniques are used to segment the arm. Least–squares method is used to fit a line to the arm and its tip is selected by comparing the distance from both extreme points of the line to the vertical edges of the poster. Arm segmentation lasts a maximum of 10 seconds. If a section has not been reliably identified at the end of that period, a sub–dialogue with the user starts.

Poster recognition is achieved by the SIFT algorithm. The poster with maximum number of matches and above an empirical threshold corresponds to the classification result. If an appropriate quality in recognition is achieved, the vision agent receives the coordinates of the rectangle that delimits each relevant section of the poster. To adjust these coordinates to the actual view, we calculate a projective transformation matrix using SIFT matches and RANSAC algorithm. Once all visible sections have been calculated, a rectangular window is defined to enclose these sections. An extended explanation of the visual module can be found in [12].

### 5.2 Speech Understanding Module

The Speech Understanding module performs three main functions: i) speech recognition, ii) speech interpretation, and iii) speech synthesis, in Spanish. Speech recognition is based on the DIMEx100 project [13]. This project is aimed to develop general acoustic–phonetic models and pronouncing dictionaries in Spanish for Mexican people of different gender and age. Synthesis of spoken dialogue of the robot is executed with MBROLA synthesizer [14].

Every time a sentence is recognized, speech interpretation takes place. This is done by comparing the sentence with a set of equivalent regular expressions defined for each expectation. If no match is found, the user is warned and requested for another attempt.

### 5.3 Navigation Module

Our robot navigates in a 2D world. The dialogue manager informs to the agent the target position $(x, y)$ of the poster as well as $\theta$, the final angular orientation.

First, the robot rotates to reach the orientation of the target, and moves along a straight line up to this coordinate. We assume there are no obstacles along this path. Once the robot has arrived, it rotates again up to reach the final $\theta$ orientation. Robot motion is measured using odometry only.

## 6 Experiments and Results

In this section, we present a preliminary round of tests performed to evaluate the overall functionality of our proposal and its results. First we describe our experimental setup.

### 6.1 Experimental Setup

Our robot is an ActivMedia's Peoplebot robot with a Canon VCC5 camera. A 2Gb 1.86GHz Dual–Core laptop computer is attached to the robot through a wired Ethernet connection to distribute processing load. Image frame rate is about 30 FPS. Robot motors and sensors are controlled using Player/Stage libraries. SIFT templates of the posters were taken with the camera at a distance of 1m, and a height of 90cm. The distance threshold to evaluate square Euclidean distances between two SIFT features is fixed to 0.2. The minimum number of SIFT matches to consider a positive classification result is 50. The size of each poster is $1.5 \times .70$m. A cheap headset is plugged to the internal sound card of the laptop and the synthesised speech is played through the robot speakers. We used our lab environment with artificial light. Three informational posters were positioned in a $U$-shaped distribution, located at a distance 2.20m between them.

### 6.2 Evaluation and Results

To evaluate our approach, we defined the tour–guide task as follows. The robot must: i) initiate the interaction with a single user, ii) explain at least one poster and one section selected by the user, and, finally, iii) finish the interaction going back to a predefined initial position. We asked 10 participants to interact with the robot –one at a time– to accomplish this task. Participants are either students or members of the academic staff of our department, each of whom have different levels of expertise in robotics. At the beginning of the exercise, users were advised to follow the instructions provided by the robot, and no special recommendations were given afterwards. Figure 3 shows an example of an interaction. In this case, the user indicates one section of the poster to get further information. At the end of the exercise, each user was requested to fill out a questionnaire that surveys two traditional metrics: the *effectiveness* of the system and the *satisfaction of the user*. Effectiveness measures the accomplishment of the task and the performance of the system modules. Satisfaction ratings assess the perception of the user about the system.

On the one hand, the evaluation of effectiveness shows that 9 out of 10 participants fulfilled the goal. Pointing gestures were successfully recognised 88.9% of

**Fig. 3.** Our tour–guide robot in action. In this example, the user is pointing to a section of the poster. In the background of the scene, our evaluator takes notes of the robot's performance.

**Table 2.** User satisfaction results

| Topic | 4 | 3 | 2 | 1 |
|---|---|---|---|---|
| **TTS Easy to understand** | 8 | 2 | | |
| **ASR System understood** | 5 | 3 | 2 | |
| **Task easy** | 5 | 3 | 2 | |
| **Task easy poster names** | 9 | 1 | | |
| **What to say** | 9 | 1 | | |
| **Expected behaviour** | 4 | 6 | | |
| **Future use** | 10 | | | |
| **Naturalness pointing** | 2 | 6 | 2 | |
| **Naturalness conversation** | 2 | 8 | | |

the time and the posters were always identified correctly; 60% with high confidence and 40% with medium confidence. The robot arrived to the position of the desired poster for all interactions. From the speech recognition side, the request to explain a poster was correctly understood 90% of the time, while the request for a section, 80%. On the other hand, the evaluation of satisfaction included questions from the PARADISE framework regarding *text-to-speech* and *automatic speech recognition* quality, ease of use, user experience, expected behavior and future use [15]. We also considered two additional questions regarding the naturalness of both, the spoken conversation and the pointing gestures. Table 2 shows a summary of the results[1]. In this experiments, all the participants continued interacting with the system beyond the specified goal. We observed that all participants are willing to interact with the system in the future. However, a main issue we are interested in is improving the naturalness the execution of gestures and the spoken conversation.

[1] Users were forced to measure either a positive or a negative response, mapped to a rate from 1 to 4, where 4 is the highest score.

## 7 Conclusions and Future Work

We presented our work on the development of a tour–guide robot using dialogue models. Dialogue models coordinate the human–robot interaction using speech understanding, visual analysis and navigation. Our robot is able to navigate in its environment, and to describe informational posters and sections selected by the user using pointing gestures. The robot also evaluates confidence in its visual results, considers the history of the conversation and modify its dialogues accordingly. Our results showed the effectiveness of the overall approach: all users are willing to interact with the robot in the future and the user satisfaction ratings are, in general, positive.

As future work we plan to refine our dialogue model as well as the effectiveness of the speech and visual modules. In addition, we will consider speech and gestures entries in a multimodal manner. We are going to implement 3D pointing gestures using a binocular stereo system. Finally, we will conduct an in–depth study about the relations of speech and gesture to incorporate common grounding, reference resolution and history management into the actual system using these two input modalities.

**Acknowledgments.** This paper has been funded by research grants from CONACyT, Grant 81965, and from PAPIIT–UNAM, Grant 104408.

## References

1. Pineda, L.A.: Specification and Interpretation of Multimodal Dialogue Models for Human-Robot Interaction, In: Sidorov, G. (ed.) Artificial Intelligence for Humans: Service Robots and Social Modeling, pp. 33–50. Sociedad Mexicana de Inteligencia Artificial (2008)
2. Pineda, L.A., Meza, I., Salinas, L.: Dialogue Model Specification and Interpretation for Intelligent Multimodal HCI. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 20–29. Springer, Heidelberg (2010)
3. Rascón, C., Avilés, H., Pineda, L.A.: Robotic Orientation towards Speaker in Human-Robot Interaction. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 10–19. Springer, Heidelberg (2010)
4. Tojo, T., Matsusaka, Y., Ishii, T.: A Conversational Robot Utilizing Facial and Body Expressions. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 2, pp. 858–863 (2000)
5. Rogalla, O., Ehrenmann, O., Zoellner, M., Becher, R., Dillmann, R.: Using gesture and speech control for commanding a robot assistant. In: 11th IEEE Workshop on Robot and Human Interactive Communication, pp. 454–459 (2002)
6. Burger, B., Lerasle, F., Ferrane, I., Clodic, A.: Mutual Assistance between Speech and Vision for Human–Robot Interaction. In: IEEE/RSJ International Conference on Intelligent Robotics and Systems, pp. 4011–4016 (2008)
7. Aviles, H., Sucar, E., Vargas, B., Sanchez, J., Corona, E.: Markovito: A Flexible and General Service Robot. In: Liu, D., Wang, L., Tan, K.C. (eds.) Studies in Computational Intelligence, vol. 177, pp. 401–423. Springer, Heidelberg (2009)

8. Toptsis, I., Haasch, A., Hüwel, S., Fritsch, J., Fink, G.A.: Modality Integration and Dialog Management for a Robotic Assistant. In: European Conference on Speech Communication and Technology, pp. 837–840 (2005)
9. Lee, C., Cha, Y.S., Kuc, T.Y.: Implementation of Dialog System for Intelligent Service Robots. In: International Conference on Control, Automation and Systems, pp. 2038–2042 (2008)
10. Stiefelhagen, R., Ekenel, H.K., Fugen, C., Gieselmann, P., Holzapfel, H., Kraft, F., Nickel, K., Voit, M., Waibel, A.: Enabling Multimodal Human–Robot Interaction for the Karlsruhe Humanoid Robot. Trans. on Robotics: Special Issue on Human–Robot Interaction 23(5), 840–851 (2007)
11. Meza, I., Salinas, S., Venegas, S., Castellanos, H., Chavarria, A., Pineda, L.A.: Specification and Evaluation of a Spanish Conversational System Using Dialogue Models. In: Kuri-Morales, A., Simari, G. (eds.) IBERAMIA 2010. LNCS (LNAI), vol. 6433, pp. 346–355. Springer, Heidelberg (2010)
12. Aviles, H., Meza, I., Aguilar, W., Pineda L.: Integrating Pointing Gestures 2nd International Conference on Agents and Artificial Intelligence, pp. 585–588 (2010)
13. Pineda, L., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., Villaseñor, L.: The corpus dimex100: Transcription and evaluation. Language Resources and Evaluation (2009)
14. Dutoit, T., Pagel, V., Pierret, N., Bataille, F., Van Der Vrecken, O.: The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non–Commercial Purposes. In: 4th International Conference on Spoken Language Processing, vol. 3, pp. 1393–1396 (1996)
15. Walker, M., Litman, D., Kamm, C., Kamm, A.A., Abella, A.: Paradise: A framework for evaluating spoken dialogue agents. In: 35th Annual Meeting of the Association for Computational Linguistics, pp. 271–280 (1997)