The Golem Team, RoboCup@Home 2011

Team Leader: Luis A. Pineda
lpineda@unam.mx
http://leibniz.iimas.unam.mx/~luis

Computer Sciences Department
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
Ciudad Universitaria, A.P. 20-726
01000 México D.F. México
http://turing.iimas.unam.mx/~golem

Abstract. In this paper, the Golem team and its robot Golem-II+ are presented. The design of Golem-II+ is based on a conceptual framework that is centered on the notion of dialogue model and the use of an interaction-oriented cognitive architecture (IOCA) with its associated programming environment. The framework provides flexibility and abstraction for task description and implementation, as well as a high level of modularity. This framework is now being evaluated with the tests of the RoboCup@Home competition.

1 Team Members and Responsibilities

Robot: Golem-II+.
Academics:

- **Dr. Luis A. Pineda.** Project's Philosophy, Dialogue Models and Cognitive Architecture. Coordination and Management.
- **Dr. Iván V. Meza.** Language, Dialogue Manager and Cognitive Architecture.
- **Dr. Héctor H. Avilés.** Vision, Gesture Recognition and Navigation.
- **Dr. Caleb Rascón.** Audio, Sound Localization, and Navigation.
- **Dr. Carlos Gershenson.** Cognitive architecture. Implementation of "Enhanced Who is Who" and "Demo Challenge" tests.
- **Ms. Lisset Salinas.** Dialogue Models. Documentation and Website. "Robot Inspection and Poster Session" test.
- **M.Sc. Montserrat Alvarado.** Documentation. Implementation of "Open Challenge" test.
- **M.Sc. Iván Sánchez.** Navigation and implementation of "General Purpose Service Robot" and "Shopping Mall" tests.
- **Ms. Esther Venegas.** Evaluation and Validation.

Students:

- **Mr. Ángel Lee-Reza.** Implementation of "Who is Who" test.
- **Mr. Arturo Rodríguez-García.** Implementation of "Follow Me" test.
- **Mr. Saúl Martínez-Vidals.** Implementation of "Go Get It!" test.

IIMAS' Consultants:

- **Dr. Mario Peña.** Electronics and Instrumentation.

Mr. Joel Durán Ortega. Electronics and Instrumentation.
External Consultants:
Mr. Mauricio Reyes Castillo. Robot's and team's image
Mr. Pablo Rivas Ortiz. Robot's and team's image

2 Group Background

The Golem Group was created within the context of the project "Dilogos Inteligentes Multimodales en Español" (DIME, Intelligent Multimodal Dialogues in Spanish). DIME was founded by Dr. Luis A. Pineda in 1998 at IIMAS, UNAM, with the purpose of developing a theory, a methodology, and a programming environment for the construction of AI systems with spoken Spanish and other input and output modalities (http://leibniz.iimas.unam.mx/~luis/DIME/). The theory and programming environment had to be modular, and also language, domain and application independent; in this regard, Spanish had to be a parameter, so the platform could also be used with other languages. The initial efforts of the group were focused on the analysis of multimodal task-oriented human dialogues, the development of a Spanish grammar, the construction of a flexible platform for speech recognition in Spanish, and the integration of a software platform for the construction of interactive systems with spoken Spanish. Within this context, the Golem project started in 2001 with the purpose of generalizing the theory for the construction of intelligent mobile agents. The group presented the robot Golem in July, 2002. In this initial version, Golem was able to sustain a simple spoken conversation and to follow simple commands for movement. However, the spoken recognition technology was too fragile and, consequently, the group focused on developing a well-founded phonetic alphabet and a large speech corpus for Mexican Spanish [13], on furthering the study on the structure of task-oriented dialogues [14,11] and on developing a pragmatics-oriented interpretation and action theory for interactive applications. From this latter work, a theory for the specification and interpretation of dialogue models emerged, with its associated dialogue manager and programming environment [12]. By the end of 2006, we had a reliable dialogue manager and a more robust Spanish speech recognition platform, and a new version of Golem was presented at UNAM's science museum Universum in June, 2007. This time Golem was able to guide a poster session about the research projects of the Computer Science Department at IIMAS, UNAM. The robot performed well and was widely covered by the Mexican TV, radio, and press. A video about this presentation and Golem's functionality is available at http://leibniz.iimas.unam.mx/~luis/golem/v_proyecto/golem.html. This version of Golem was also widely demonstrated in several academic events in Mexico until 2009. Next, we incorporated computer vision facilities into the platform [1] and in December 2009 we presented the module "Guess the card: Golem in Universum". This application stands in the permanent exhibition of the Universum museum in which the general public can play a game about guessing a card with the system [9,10]. In 2010, we started the development of the Golem-II+ service robot, which is also able to guide a poster session. In addition, the robot is capable of interpreting pointing gestures performed by the human-user during conversation, illustrating the coordination of language, vision and motor behavior [2]. For the development of Golem-II+, we incorporated an innovative explicit cognitive architecture that, in conjunction with the dialogue models theory and program interpreter, constitutes the theoretical core of our approach [16,15]. Currently we are exploring the incorporation of reactive facilities for navigation, and for the detection of the location and orientation of a source of sound, so the robot can face its interlocutor during human-robot interactions [17]. We are now developing the test scenarios of the RoboCup@Home competition to develop further our theory, methodology and programming environment.

3 An Interaction-Oriented Cognitive Architecture

The behavior of our robot Golem-II+ is regulated by an Interaction Oriented Cognitive Architecture (IOCA) [16,15]. A diagram of IOCA can be seen in Figure 1.

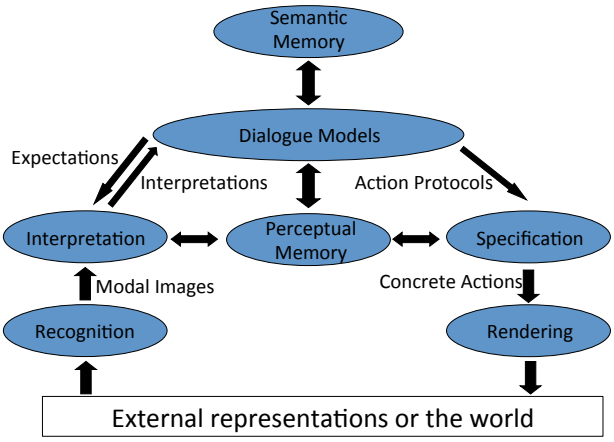

Fig. 1. Interaction Oriented Cognitive Architecture (IOCA).

External stimuli (vision, audio, etc.) are processed by *Recognition* modules. There are different modules for different perceptual modalities. An example of a recognition module is the speech recognition module which encodes "raw" information (audio signal), but does not assign any meaning to it. We called the different encodings the Modalities of the system, while the Modal Image is an encoding of the recognized external acoustic or visual form. The *Interpretation* module assigns meaning to Modal Images. In order to do this, it takes into account the Expectations from the current situation and the history of the interaction. The Interpretation module has access to the Perceptual Memory which relates Modal Images with their meaning. These can be used to build the interpretation. Interpretations are represented in a propositional format, which is modality independent. For example, the same Interpretation can be acquired by speech recognition or by hand gesture recognition.

The *Dialoge Models* are the center of IOCA [16,15] and have been the research focus of our group. Dialogue Models describe the task via a set of situations. A Situation is an information state defined in terms of the Expectations of the agent in a state, so if the Expectations change, the agent moves to a new Situation. Expectations depend on the context and can be determined dynamically in relation to the interaction history. The way that these Situations are linked together is by relating such Expectations with a corresponding Action and a following Situation. Expectations are propositional as well, and in order to be triggered they must be matched with an Interpretation. Once this matching occurs, the corresponding actions are performed and the system reaches its

next Situation. Actions can be external (e.g. say something, move, etc.) or internal (e.g. plan, reason, etc.). Actions are composite roughly along the lines of Rhetorical Structure Theory (RST) [8] and they involve more than one device. Actions are processed by the *Specification* modules—considering the Perceptual Memory—producing a parametric specification of Concrete Actions. These are delivered to different *Rendering* devices, which produce changes in the state of the world.

An important part of IOCA is that the Dialogue Models perceive and interact with the world in an abstract manner. The state of the world and the set of possible expectations that can follow are given an abstract meaning, which, in reality, could have been perceived by any number of Recognizers. The same applies when acting upon the world: "explain me the Golem's project poster" is all that the Dialogue Model needs to state for the robot to start moving. This paradigm provides great flexibility in software development, as the Recognizers and Renderers become modular and replaceable, while the task description remains intact. This also secures a framework with which different tasks can be described and that does not require a complete rewrite of the internal software. Moreover, since the Recognizers and Renderers are modular, they can also be reused for different tasks with relative ease.

4 Software

4.1 Dialogue manager

We implement the dialogue manager as an interpreter of specifications of Dialogue Models. The dialogue manager is in charge of managing the execution of a task by defining which is the current situation and its expectations. It also administers the Interpretation and Specification modules of IOCA, by matching the interpretations with the expectations, and for the matched expectation dispatching the specification of the corresponding Actions. Additionally, the dialogue manager keeps tracks of the history of the Interaction.

The dialogue manager is implemented in Prolog and we use the Open Agent Architecture (OAA2 [3]) as a communication channel among the different modules.

4.2 Vision

All of the vision modules belong to the Recognition part of IOCA.

Face detection and recognition OpenCV functions are used to perform face detection and recognition. Face detection is carried out by using Haar-like features [20]. Face recognition is based on Eigenfaces technique [18]. We take advantage of the pan-tilt-zoom capabilities of the Monocular PTZ camera in order to speed up these processes.

Object recognition This capability is performed using the SIFT algorithm [18]. For training, frontal views are used to construct the SIFT feature templates of the objects. For recognition, SIFT features are obtained from the current view and matched against each template. The number of matches for each object is stored in a frequency table R. This table is used to qualify visual outcomes as described in [2]. The object with a maximum number of matches and above a threshold—defined experimentally—corresponds to the classification result. If this criterion is not met, the visual agent informs the dialog manager of this situation, and a simple recovery dialog with the user proceeds to inform the result.

Pointing gestures Currently, our robot understands simple $2D$ pointing gestures and it assumes a static background for a tour–guide robot [2]. To spot the arm of the user, we developed a simple procedure based on the combination of difference, motion, and edge clues. The main assumption behind this idea is that the combination of simple visual clues performs well to contrast between background and foreground objects, and that noise introduced by each clue can be cancelled collectively. Dilation masks are used to fill out foreground regions. This algorithm has been tested extensively under different natural and artificial lightning conditions, and offers a performance suitable for our purposes.

4.3 Speech recognition and synthesis

We use a robust live continuous speech recognizer based on the Sphinx 3 software. For the acoustic models, we use the open US English HUB4 models available with Sphinx. For the language models, we hand-crafted a corpus for each of the tasks. This module is a Recognizer in the IOCA framework.

Similarly, for the speech synthesis we use open tools, in particular the Festival TTS package. From the point of view of the IOCA framework, this is a Rendering module.

4.4 Language interpretation

The interpretation of natural language interaction is achieved with a word and expression spotting strategy. We store in the Perceptual Memory a set of regular expressions and their meanings. The natural language interpreter tries to match the regular expressions to the orthographic transcription but only to the ones which are similar to the expectations of the system. When there is a match, this module returns the meaning associated to the regular expression. This is an Interpretation module within the IOCA framework.

4.5 Audio

The GPL software JACK is used to create an all-encompassing simulated sound card that can be accessed by different audio clients at the same time. Two audio clients were created:

- **Noice Cancellation:** The Sphinx system, used for speech recognition, was modified such that it can access JACK directly. A module inside Sphinx was created to utilize the microphone setup, where two omnidirectional microphones are positioned over the "shoulders" of the robot, and a directional microphone is positioned in front of the robot. This module processes each audio data window before it reaches the rest of the speech recognition system, by canceling out the data of the omnidirectional microphones from the directional, negating the effects of both environmental factors and the robot's voice.
- **Audio-Localization:** This module provides a robust direction-of-arrival estimation in near-real-time manner in mid-level reverberant environments, throughout the 360° range. The signals from the three microphones are set in an equilateral triangle, which provide three measured delay-comparisons. This provides redundancy to the direction-of-arrival estimation, as well as a close-to-linear mapping between delay measurements and direction-of-arrival estimations. This module is based upon our previous work [17]. Audio-Localization is a Recognition module in IOCA.

4.6 Navigation

Robot navigation is performed on a $2D$ world. To make the robot move, the dialog manager informs to the navigation module the (x, y) Cartesian coordinates of the target position, plus the final angular orientation θ.

The initial position of the robot is obtained automatically by using probabilistic maps [6] and natural landmarks of the environment such as corners, walls and gaps [5]. Local localization is based on particle filters [4]. Path planning and navigation are based on [7,19]. Navigation is a Rendering module of IOCA.

4.7 Description of the software

The robot internal computer has the Debian Etch 4.0 operating system installed, while the external laptop computer runs Ubuntu 10.04.

The robot devices are controlled with the following software:

- Gripper, IR, sonars, bumpers, breakbeams, PTZ camera: Player/Stage libraries.
- Laser: Player/Stage and Gearbox libraries.
- Stereo camera: SVS and OpenCV libraries.
- Webcam: Luvcview code modified for capture and OpenCV library.
- Sound card: M-Audio Fast Track Ultra.

Table 1 shows which software libraries are used by different modules.

Table 1. Software modules of Golem-II+

Module	Libraries
Dialogue manager	Sicstus Prolog V. 3.12.
Vision	OpenCV libraries.
Audio	JACK and PulseAudio.
Voice recognition	Sphinx 3.
Voice synthetizer	Festival TTS
Navigation	Player/Stage libraries. pmap library (map construction).

5 Description of the Hardware

The Golem team will use the "Golem-II+" robot for the competition (See Fig. 2). Golem-II+ is composed by the following hardware.

- PeopleBot^{TM} robot (from Mobile Robots Inc.)
  - Three sonar arrays with eight sensors each.
  - Two protective IR sensors in the front .
  - 2-DOF Gripper with break beam.

  - Two vertical break beams.
  - Two protective bumper arrays with five bumper sensors each.
  - Twin microphones and speakers.
  - Internal computer VersaLogic EBX-12.
- Dell Latitude E6400 laptop computer.
- Color Stereo Cameras Videre STH-MDCS-VAR
- Hokuyo UTM-30LX Laser
- Monocular PTZ Canon VCC5 camera
- QuickCam Pro 9000 webcam
- Shure Base Omnidirectional microphones x2
- M-Audio Condenser Directional microphone
- M-Audio Fast Track external sound interface
- Infinity 3.5-Inch Two-Way loudspeakers x2

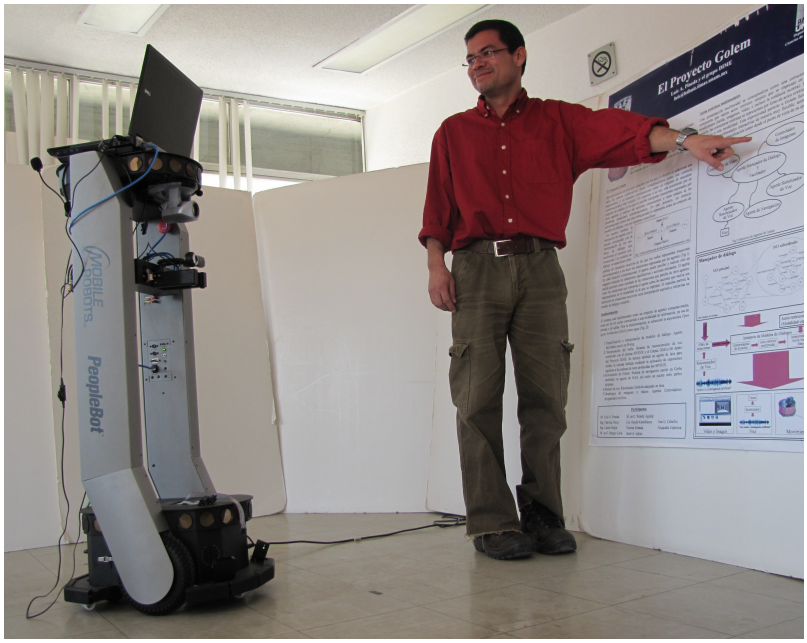

Fig. 2. Golem-II+ robot (at left) recognizing gestures.

Acknowledgments

Golem-II+ was financed by CONACyT project 81965 and by PAPIIT-UNAM projects IN-121206, IN-104408 and IN-115710.

References

1. Aguilar, W., Pineda, L.A.: Integrating graph-based vision perception to spoken conversation in human-robot interaction. In: IWANN 2009, part I, LNCS, vol. 5517, pp. 789–796. Springer (2009)
2. Avilés, H., Alvarado-Gonzalez, M., Venegas, E., Rascón, C., Meza, I.V., Pineda, L.A.: Development of a tour-guide robot using dialogue models and a cognitive architecture. Advances in Artificial Intelligence - IBERAMIA 2010 6433, 512–521 (2010)
3. Cheyer, A., Martin, D.: The open agent architecture. Journal of Autonomous Agents and Multi-Agent Systems 4(1), 143–148 (March 2001), http://www.ai.sri.com/~oaa/
4. Dellaert, F., amd Wolfram Burgard, D.F., Thrun, S.: Monte carlo localization for mobile robots. In: IEEE International Conference on Robotics and Automation (ICRA99) (1999)
5. Hernández-Alamilla, S.F., Morales, E.F.: Global localization of mobile robots for indoor environments using natural landmarks. In: IEEE International Conference on Robotics, Automation and Mechatronics. pp. 651–656 (2006)
6. Howard, A.: Multi-robot simultaneous localization and mapping using particle filters. International Journal of Robotics Research 25(12), 1243–1256 (2006)
7. Kavraki, L.E., Svestka, P., Latombe, J.C., Overmars, M.H.: Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transactions on Robotics and Automation 12(4), 566–580 (1996)
8. Mann, W.C., Thompson, S.: Rhetorical structure theory: Towards a functional theory of text organization. Text 8(3), 243–281 (1988)
9. Meza, I., Pérez-Pavón, P., Salinas, L., Avilés, H., Pineda, L.: A multimodal conversational system for a museum application. Procesamiento del Lenguaje Natural 44, 131–138 (2009)
10. Meza, I.V., Salinas, L., Venegas, E., Castellanos-Vargas, H., Alvarado-González, M., Chavarría-Amezcua, A., Pineda, L.A.: Specification and evaluation of a spanish conversational system using dialogue models. Advances in Artificial Intelligence - IBERAMIA 2010 6433 (2010)
11. Pineda, L., Estrada, V., Coria, S., Allen, J.: The obligations and common ground structure of practical dialogues. Inteligencia Artificial 11(36), 9–17 (2007)
12. 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. SMIA, Mexico (2008)
13. Pineda, L.A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., Villaseñor, L.: The corpus DIMEx100: Transcription and evaluation. Language Resources and Evaluation 44, 347–370 (2010)
14. Pineda, L., Castellanos, H., Coria, S., Estrada, V., López, F., López, I., Meza, I., Moreno, I., Pérez, P., Rodríguez, C.: Balancing transactions in practical dialogues. In: Gelbukh, A. (ed.) CICLing 2006, LNCS, vol. 3878. Springer (2006)
15. Pineda, L.A., amd Héctor H. Avilés, I.V.M., Gershenson, C., Rascón, C., Alvarado, M., Salinas, L.: IOCA: An interaction-oriented cognitive architecture (2011), submitted
16. Pineda, L.A., Meza, I.V., Salinas, L.: Dialogue model specification and interpretation for intelligent multimodal HCI. Advances in Artificial Intelligence - IBERAMIA 2010 6433 (2010)
17. Rascón, C., Avilés, H., Pineda, L.A.: Robotic orientation towards speaker for human-robot interaction. Advances in Artificial Intelligence - IBERAMIA 2010 6433, 10–19 (2010)
18. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
19. Vargas, B., Morales., E.: Solving navigation tasks with learned teleo-reactive programs. In: IEEE International Conference on Intelligent Robots and Systems (IROS 2008) (2008)
20. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 1, pp. I–511 – I–518 (2001)