The Golem Team, RoboCup@Home 2012

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://golem.iimas.unam.mx

Abstract. In this paper, the Golem team and its robot Golem-II+ are described. 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.
Dr. Iván V. Meza. Language, Dialogue Manager and Cognitive Architecture.
Dr. Gibrán Fuentes. Vision and Object Manipulation.
Dr. Caleb Rascón. Audio, Sound Localization, and Navigation.
Dr. Mario Peña. Electronics and Instrumentation.
Dr. Carlos Gershenson. Cognitive architecture.
M.Sc. Iván Sánchez. Navigation and Spatial Reasoning.
M.Sc. Mauricio Reyes Castillo. Robot's and Team's Image.
M.Sc. Arturo Rodríguez-García. Person Tracking.
M.Sc. Hernando Ortega. Electro-mechanical Devices.
Ms. Lisset Salinas. Dialogue Models.
Mr. Joel Durán Ortega. Electronics and Instrumentation.

Students:

Ms. Varinia Estrada. Lingüistics.
M.Sc. Miguel Salazar. Face Recognition.

2 Group Background

The Golem Group was created within the context of the project “Diálogos 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. 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, and it was able to sustain a simple spoken conversation and to follow simple commands for movement. The group then focused 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 [13].

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. 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 [10,11].

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 model theory and program interpreter, constitutes the theoretical core of our approach [15,14].

At this point, the Golem Team participated in Robocup@Home Istanbul, Turkey 2011, which provided important feedback for the robot’s development. To this point, we have incorporated the capability of spatial reasoning to the navigation subsystem, and added the ability of facing the interlocutor during human-robot interactions [17]. We are also exploring the possibility for the robot to interact with more than one agent at a time, being humans or robots, and are continuing to develop 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) [15,14]. A diagram of IOCA can be seen in Figure 1.

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, but does not assign any meaning to it. These different encodings are the "Modalities" of the system, which present what is being recognized in the form of what we call "Modal Images": codified information without any meaning.

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, building 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 [15,14] 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) [9] 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 then delivered to different *Rendering* devices, which produce changes in the state of the world.

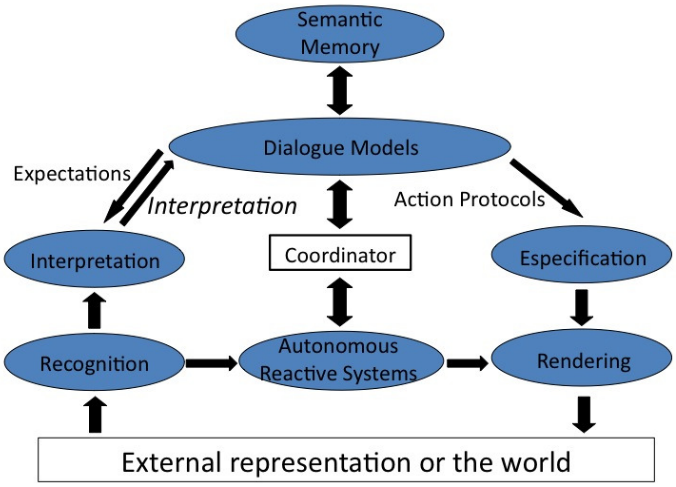

Fig. 1. Interaction Oriented Cognitive Architecture (IOCA).

However, it is not sufficient to model applications based solely in expectations. In order to cope with dynamic environments (unexpected obstacles, the user taking the initiative), IOCA was extended with a set of reactive modules, referred to as *Autonomous Reactive Systems* (ARSs), that relate the input information collected by the recognition devices with the rendering devices directly. Two ARSs are currently implemented: the Autonomous Navigation System (ANS) and the Autonomous Position and Orientation Source of Sound Detection System (APOS) to allow the robot to face its interlocutor reactively. In addition, *The Coordinator* was included as a control structure for coordinating the Dialogue Models with the ARSs.

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: an abstract command such as "go to the kitchen", provided by the Dialogue Manager, will be realized by the appropriate Renderer 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, Tracking and Recognition. OpenCV functions are used to perform face detection and recognition. Face detection is carried out by using the Viola-Jones Method [21], and tracking is carried out via a technique based on Kalman Filtering [7]. Face recognition is based on Eigenfaces technique [20]. We take advantage of the tilt capability of the Kinect camera to enhance the recognition process.

Object Recognition. This capability is performed using the MOPED framework proposed by Collet et. al. [4]. During the training stage, several images are acquired from each object that is to be recognized, and SIFT features [8] are extracted from each image. Structure from Motion is applied to arrange all the features from all the images and obtain a 3D model of each object. In the recognition phase, the SIFT features are obtained from the visual scene and, with 3D model at hand, hypothesis are made in an iterative manner using Iterative Clustering Estimation [4]. Finally, Projective Clustering [4] is used to group the similar hypothesis and provide one for each of the objects being observed.

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 canceled 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.

Person Tracking. For person identification, the Person Tracker that is part of the OpenNI driver was employed, where several blobs in the visual scene are hypothesized as being person. A person that performs a certain known action is then labeled in the visual scene as the one to be tracked. Then, color features are extracted from the blob section of the visual scene to be used for identification at a later time, specifically when the user walks out of the visual scene or when there are several users in it.

4.3 Arm Manipulation

The Actin-SE software is used to control the ROBAI robotic arm installed in Golem-II+. Because there may be unlimited ways to move the joints to achieve the desired hand motion, it is important to choose an appropriate strategy for coordinated motion and path planning, based on the requirements at hand. This software provides an optimized control system that simultaneously avoids joint limits, singularities, and collisions all while minimizing kinetic energy. This is a IOCA Rendering module.

4.4 Speech Recognition and Synthesis

We use a robust live continuous speech recognizer based on the PocketSphinx software, coupled with the Walt Street Journal (WSJ) acoustic models. For the language models, we hand-crafted a corpus for each of the tasks, and made the ASR be able to switch from one to the other, depending on the context of the dialogue (A yes/no language model for confirmation, a name language model for when the user is being asked their name, etc.). 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.

4.5 Language Interpretation

In this version of the system there are two strategies for the language interpretation. The first strategy is a shallow semantic strategy which uses word and expression spotting. For this strategy, regular expressions and their meanings are stored in the Perceptual Memory. The natural language interpreter tries to match the regular expressions to the orthographic transcriptions that are similar to the expectations of the system. The second strategy is a deep semantic parser based on the GF formalism [16]: several grammars were defined specifically for fine grained semantics tasks, such as General Purpose Service Robot. This is an Interpretation module within the IOCA framework.

4.6 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 as Recognition Modules in IOCA.

Noise Filtering. A preprocessing module was created that stands between the Jack input and the ASR, processing each audio data window, negating the effects of the ambient via a Quantile-based noise estimator [18] that is able to provide real-time estimates of the ambient noise spectral signature even while the user is talking.

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 [17]. This module, in conjunction with the Reactive Navigation Module (described later), compose the Autonomous Position and Orientation Source of Sound Detection System (APOS).

4.7 Navigation

An important issue in modern robotic navigation is the one of auto-localization, primarily the positioning errors that increase the more the robot moves around [19]. Golem-II+ generates hypotheses of its location from environmental ‘clues’ (doors, hallways, or the presence of another robot), basically a Cognitive Map, and can build a spatial model to be used for the correction of its current global coordinates through Spatial Reasoning.

The Spatial Reasoner subsumes lower navigational capabilities which are themselves handled by the Reactive Navigation Module (RNM). In addition to simple moving capabilities (turn θ degrees, move Z meters to the front, etc.), both the Dialogue Manager and the Spatial Reasoner can interact directly with the RNM, by only providing a pre-specified label of a location of where it is desired to move. The label is realized into a set of (x, y) Cartesian coordinates and a route that the robot will partake is deduced via the Dijkstra Algorithm [5], defined by a series of intermediate points. These are obtained from a topological map created from a weighted adjacency matrix overlaid over the map of the area. Between each intermediate point, obstacle evasion is carried out by using, in conjunction, Nearness Diagram [12] (by default) and Smooth Nearness Diagram [6] (as fall-back).

In conjunction, the Spatial Reasoner and the Reactive Navigation Module compose the Autonomous Navigation System (ANS) inside IOCA.

4.8 Software Libraries

Both the robot internal computer and the external laptop run the Ubuntu 10.04 operating system. Table 1 shows which software libraries are used by the IOCA modules and Golem-II+’s hardware.

Table 1. Software Libraries used by the IOCA Modules and Hardware of Golem-II+

Module	**Hardware**	**Software Libraries**
Dialogue manager	–	Sicstus Prolog
Vision	Kinect, WebCam	SVS, OpenCV, and OpenNI libraries
Voice recognition	Directional Microphone, External Sound Card	JACK, PocketSphinx
Voice synthetizer	Speakers	PulseAudio, Festival TTS
Navigation	IR, Sonars, Bumpers, Breakbeams, Laser	Player/Stage and Gearbox
Object Manipulation	Robotic Arm	Actin-SE Software

5 Description of the Hardware

The "Golem-II+" robot (See Fig. 2) will be used, which is composed by the following hardware:

- PeopleBot^{TM} robot (Mobile Robots Inc.)
  - Three 8-sensor sonar arrays.
  - Two protective IR sensors in the front.
  - Two vertical break beams.
  - Two protective 5-bumper arrays.
  - Speakers.
  - Internal computer VersaLogic EBX-12.
- Dell Precision M4600 laptop computer.
- QuickCam Pro 9000 Webcam
- Microsoft Kinect Camera
- Hokuyo UTM-30LX Laser
- Shure Base Omnidirectional microphones x3
- RODE VideoMic directional microphone
- M-Audio Fast Track Ultra external sound interface
- Infinity 3.5-Inch Two-Way loudspeakers x2
- Robai Cyton Veta 7DOF robotic arm

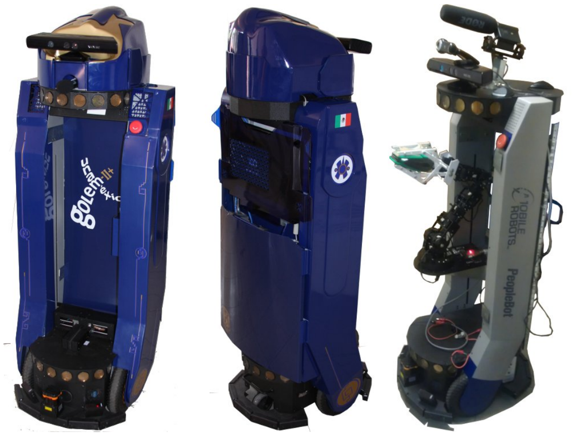

Fig. 2. The Golem-II+ robot.

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. Collet, A., Martinez, M., Srinivasa, S.S.: The moped framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research 30, 1284–1306 (2011)
5. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959), http://dx.doi.org/10.1007/BF01386390, 10.1007/BF01386390
6. Durham, J., Bullo, F.: Smooth nearness-diagram navigation. In: Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on. pp. 690–695 (sept 2008)
7. Kalman, R.E.: A new approach to linear filtering and prediction problems. Transactions of the ASME–Journal of Basic Engineering 82(Series D), 35–45 (1960)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
9. Mann, W.C., Thompson, S.: Rhetorical structure theory: Towards a functional theory of text organization. Text 8(3), 243–281 (1988)
10. 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)
11. 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)
12. Minguez, J., Member, A., Montano, L.: Nearness diagram (nd) navigation: Collision avoidance in troublesome scenarios. IEEE Transactions on Robotics and Automation 20, 2004 (2004)
13. 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)
14. Pineda, L.A., amd Héctor H. Avilés, I.V.M., Gershenson, C., Rascón, C., Alvarado, M., Salinas, L.: IOCA: An interaction-oriented cognitive architecture. Research in Computer Science 54, 273–284 (2011)
15. 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)
16. Ranta, A.: Grammatical framework: Programming with multilingual grammars. CSLI Publications, Stanford p. 340 (2011), iSBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth)
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. Stahl, V., Fischer, A., Bippus, R.: Quantile based noise estimation for spectral subtraction and wiener filtering. In: Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on. vol. 3, pp. 1875–1878 vol.3 (2000)
19. Thrun, S.: Robotic mapping: A survey. Exploring Artificial Intelligence in the New Millenium (2002)
20. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
21. 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)