

Towards a Personal Robot with Language Interface

*L. Seabra Lopes, A. Teixeira, M. Rodrigues, D. Gomes, C. Teixeira,
L. Ferreira, P. Soares, J. Girão, N. Sénica*

Instituto de Engenharia Electrónica e Telemática de Aveiro (IEETA)
Universidade de Aveiro, 3810-193 Aveiro, Portugal {ajst, lsl}@det.ua.pt

Abstract

The development of robots capable of accepting instructions in terms of familiar concepts to the user is still a challenge. For these robots to emerge it's essential the development of natural language interfaces, since this is regarded as the only interface acceptable for a machine which expected to have a high level of interactivity with Man. Our group has been involved for several years in the development of a mobile intelligent robot, named Carl, designed having in mind such tasks as serving food in a reception or acting as a host in an organization. The approach that has been followed in the design of Carl is based on an explicit concern with the integration of the major dimensions of intelligence, namely Communication, Action, Reasoning and Learning. This paper focuses on the multi-modal human-robot language communication capabilities of Carl, since these have been significantly improved during the last year.

1. Introduction

In recent years, robotics-related technologies have reached such a level of maturity that, now, researchers are feeling that the next step is the development of intelligent service robots capable of working in close cooperation/interaction with humans.

It will be necessary for robots of this new generation to comply with three criteria. First, these robots must be animate, meaning that they should respond to changing conditions in their environment. Second, personal robots should be adaptable to different users and different physical environments. Finally, robots should be accessible, meaning that they should be able to explain their beliefs, motivations and intentions, being easy to command and instruct.

In order to meet the animate, adaptable and accessible criteria for intelligent service robots, it is necessary to include in their design such basic capabilities as linguistic communication, reasoning, reactivity and learning. "Integrated Intelligence" is an emerging keyword that identifies an approach to building intelligent artificial agents in which the integration of all those aspects of intelligence is considered. Recent research works explore a variety of alternative paths, leading to architectures in which functionalities are combined in different ways [1].

Given the progress obtained in sub-domains of AI and the maturity of the produced technologies, the "integrated intelligence" challenge seems to be the real challenge to face next. This is the focus of a national-funded project, CARL (Communication, Action, Reasoning and Learning in robotics), started in 1999, aiming to develop a robot capable of understand, using a friendly interface, instructions expressed in a way familiar to the human user. Human-robot communication is one of the main research topics in this project.

This paper describes the current state of evolution of Carl, our prototype of an intelligent service robot, which participated

in the AAI Mobile Robot Competition and Exhibition in 2001 and on 1st International Cleaning Robots Contest in 2002. The main advances with respect to previously published versions of Carl [2] are concerned with language processing, touch interaction and emotional display. Section 2 describes the robot. Section 3 the graphical and touch interaction capabilities. Section 4 describes the language interface.

2. Carl the Robot

Carl is based on a Pioneer 2-DX platform from ActivMedia Robotics. It includes wheel encoders, front and rear bumper rings, front and rear sonar rings, a micro-controller based on the Siemens C166 processor and an on-board computer. The operating system is Linux. A Sony EVI D31 pan-tilt-zoom camera was added to enable such capabilities as object recognition and advanced navigation.

On top of the mobile platform, a fiber glass structure was added, which carries a Fujitsu-Siemens Lifebook laptop computer, also running Linux. The laptop includes a touch screen to enable a touch interaction modality. Additionally, the fiber glass structure carries a VoiceTracker directional microphone array from Acoustic Magic, a speaker and a webcam. Currently, Carl is 1.10 m tall. The microphone array is in a suitable position for speech recognition, since it is at a distance around 1 meter from the mouth of the average adult speaker. The fiber glass structure also includes a recipient for transporting small objects, equipped with an IR sensor for detecting the presence of objects. The base computer and the laptop computer are connected by Ethernet cross-over cable and the robot has the possibility to be controlled and/or monitored by a 3rd computer via Wireless 802.11b WiFi card set on the laptop.

With this platform, we are developing an autonomous robot capable, not only of wandering around, but also of taking decisions, executing tasks and learning.

2.1. Software Architecture

Carl has several independent modules coordinated by a central manager module. Each module is a Linux process communicating with the others using TCP/IP.

Human-robot communication is achieved through spoken and written language dialog as well as touch interactions.

Touch screen interaction is controlled through the so called RUBI (Robot User Binding Interface) module. Written language input can be captured through a virtual keyboard displayed on the touch screen. An animated face displays appropriate emotions.

Another module handles general perception and action, including navigation. It is based on Saphira and ARIA API, the software interface for Pioneer robots.

```

state_transition(
interacting,
[heard(tell(Phrase))],
true, % no restrictions
acknowledge_told_fact(Phrase),
[execute_motion(stop),retract_all_times,
memorize_told_fact(Phrase),assert_last_heard_time],
interacting ).

```

Figure 1: Example of state transition.

High-level reasoning (including inductive and deductive inference), natural language parsing and generation are implemented in a different module. Another module provides Carl with learning capabilities.

All computation is done on board. The perception and action process runs on the Pioneer base computer while all other processes run on the laptop computer.

2.2. Execution and Interaction Management

The central manager is an event-driven system. Events originating in the speech interface, in sensors or in navigation activity as well as timeout events lead to state transitions. Such apparently different activities as dialog management and navigation management are integrated in a common unified framework.

The central manager is essentially a state transition function specified as a set of Prolog clauses. Each clause, specifying a transition, has a head of the following form:

```

state_transition(State,Events,
Restrictions,SpeechAct,Actions,NewState)

```

State is the current state; **Events** is a list of events that will cause a transition to **NewState**, provided that the **Restrictions** are satisfied. These events can be speech input events, navigation events, touch screen interface events, timing events, robot body events. **SpeechAct**, if not void, is some verbal message that the robot should emit in this transition. **Actions** are a list of other actions that robot should perform. These can be actions related to navigation, control of RUBI and the animated face, but also internal state update and dynamic grammar adaptation.

The state transition in Fig. 1 is a transition to the same state, in this case the interacting state. The triggering event is the reception of an instance of the tell speech act. The robot immediately stops and acknowledges, then memorizes the told information. The time of this event is recorded, so that the robot may later recognize that the interaction is over, if it didn't finish with an explicit *good bye* from the human interactant.

3. Graphical and Touch Interface

In the previous configurations of Carl, the only available interaction modality was based on spoken language dialog [2]. The touch screen facility, that comes with the laptop computer recently installed in Carl, enables new interaction modalities. For that purpose, a graphical user interface, RUBI, was developed using QT Library. It allows the input of commands and information through touch as well as the display of monitoring and debug information. This way, the usability of the robot could be enhanced. RUBI is organized into several areas. On the top-left corner, an animated face is shown, visible at all times. Below the animated face, a command panel is displayed where some options are offered. Most of the commands can also be issued by voice.

Performative	Description
Register (S,R)	S announces its presence to R
Achieve(S,R,C)	S asks R to perform action C in its physical environment
Tell(S,R,C)	S tells R that sentence C is true
Ask(S,R,C)	S asks R one instantiation of sentence C
Ask_if(S,R,C)	S wants to know if R thinks sentence C is true
Thanks(S,R)	S expresses gratitude to R
Bye(S,R)	S says good-bye to R
Dye(S,R)	S (human master) asks R (robot) to close all execution processes

Table 1: Currently supported performatives (S=sender, R=receiver).

4. Language Interface

The goal of natural language processing is to extract semantics of natural language sentences and, conversely, to generate sentences from specifications of intended semantics.

The human-robot communication process is modelled as the exchange of messages, much like is done in multi-agent systems. The set of performatives or message types in our Human-Robot Communication Language (HRCL) is inspired in KQML. Table 1 shows the currently supported set of performatives.

4.1. Speech Recognition

The recognition block is critical due to the fact of being affected by a relevant set of external conditions, as it is the case of environment noise, parallel conversations and the noise of the robot's own engines. The sudden changes of context by the speaker and the quality of the microphone used are also relevant conditionings for the quality of the recognition.

The speech recognition module of Carl is based on the Nuance 8.0. As language model, Carl has bigrams [3] based on a group of sentences that are susceptible to be said in the context of the CARL project. Thus recognition of correct sentences is improved without restricting what the user can say, allowing it to express itself in a natural way.

Two similar tests were made to the recognizer: one on an investigation laboratory with several computers working, the other tries to simulate the real environment of a robot demonstration through the existence of background noise plus the presence of several people talking at a reasonable sound level.

Table 2 presents the results of the test in the environment with more noise. Results are presented both for the first and second best sentences given by the recognizer, the ones that are further processed by our natural language understanding modules.

4.2. Natural Language Understanding

The current version of this module is based on the Attribute-Logic Engine (ALE), a public domain logic programming and natural language processing system [4].

An implementation in ALE of the generation grammar used by Shieber et al. (1990) [5] to illustrate the semantic-head-driven generation algorithm was used as starting point to develop the grammar for Carl. It contains only 14 words in its lexicon and 5 grammar rules. The top-level types in the signature

Table 2: Nuance Evaluation Results in a noisy environment.

	1st choice		2nd choice	
total sentences	67		63	
% sentences correct	47.76	(32)	7.94	(5)
total words (T)	323		308	
% correct words	79.57		68.18	
% replaced words (R)	13.00	(42)	24.03	(74)
% inserted words (I)	1.55	(5)	2.92	(9)
% deleted words (D)	7.43	(24)	7.79	(24)
WER	20.43		31.82	
Accuracy = (T-I-D-R)/T	78.02		65.26	

```

sem sub [ semrel, sematt, semobj, sem_yes_no, greating].
  semobj sub []
    intro [obj:basicword, rels:sem_list].
  sematt sub []
    intro [attname:basicword, value:basicword].
  semrel sub []
    intro [ relname:basicword,
            obj1:semobj, obj2:semobj,
            prename:basicword,
            vadverb:basicword ].
  sem_yes_no sub [] intro [bool:basicword].
  greating sub [].

```

Figure 2: Type hierarchy for semantics.

section of this grammar are concerned with grammar categories, agreement in sentences, verb forms and semantics. Besides expanding the language coverage of this grammar for Carl, it was necessary to largely restructure its signature section.

The subtree for sentence semantics is presented in Fig. 2. Three main semantic subtypes are considered: objects (**semobj**), attributes of objects (**sematt**) and relations between objects (**semrel**). As relations often correspond to verbs, instances of the **semrel** type have fields for the relation name, objects, a preposition and an adverb. Several ALE macros were created to simplify the specification of lexicon entries. For example, one of the verb macros used in our grammar is:

```

verb(Verb) macro
  tverb,
  vform:nonfinite,
  vsubcat: [ (np,sem:Obj), (np,sem:Subj) ],
  sem: ( relname:Verb, obj1:Subj, obj2:Obj,
         prename:none_, vadverb:none_ ).

```

This macro can be used to define transitive verbs, in non-finite form, subcategorized for a noun phrase as subject and another noun phrase as object. The semantics of the verb is a relation having the verb as name, the subject and object as arguments and having no associated preposition or adverb. The last part of the grammar is mainly constituted by grammar (phrase structure) rules. For example, some verb phrases can be parsed by the following rule:

```

v1 rule(vp, form:Form, subcat:Subcat, sem:Sem)
===> sem\_head> ( verb, vform:Form,
vsubcat: [ (np,sem:Obj) | Subcat ],
sem: Sem ),
cat> (np,sem:Obj).

```

Table 3: Results, as percentage of correct decisions, of sentence/non-sentence tests. First row indicates percentage of sentences used for training.

%	1	5	10	20	40	60	80
Train	152	762	1.5k	3k	6k	9k	12k
Test							
1	85.4	91.5	93.8	95.0	95.9	96.2	96.4
2	81.7	93.7	92.7	94.9	95.5	96.2	96.1
...							
9	83.8	94.0	94.3	95.0	95.5	96.3	96.4
10	72.1	92.7	92.9	94.8	95.6	95.7	96.2
mean	79.6	92.7	93.9	95.0	95.5	95.9	96.1
std	10.7	1.2	0.7	0.2	0.2	0.3	0.4

The verbs acceptable by this rule are verbs that subcategorize a noun phrase as object, as it happens in the macro presented above. The *v1* phrase structure then fits into other rules until a complete sentence can be parsed. If the sentence is well formed, the parsing process delivers the semantics of the sentence, otherwise it fails. After obtaining the typed feature structure representing the semantics of a given sentence, the last step is to convert it to a list of Prolog terms that can be asserted in the Prolog database (in tell speech acts) or matched with facts already existing in the database (such as in ask and ask_if speech acts). For instance, the semantics of the sentence *peter is in the car of sandy* would be given by the following list:

```

[name_(X, peter), type_(Y, car),
name_(Z, sandy), of(none_, none_, Y, Z)].

```

This has a direct correspondence in first-order logic. The current grammar has approximately 150 entries in the lexicon section and approximately 30 phrase structure rules.

The voice recognition made by the recognizer isn't always interpretable nor, most of the times are the sentences that it returns grammatically correct. Due to the fixed structure analysis, ALE can not retrieve any information from such sentences.

Supposing that the ALE manages to withdraw information from all the grammatically correct sentences, it becomes necessary to find another way of detecting and interpreting incomplete and badly recognized sentences, because each sentence that is not analyzed is information that is not available for Carl.

As a first step towards the new module, a program to filter the sentences from the recognizer was prepared. The objective is to identify the grammatically correct sentences, from the hypothesis given by the recognizer. Just this type of sentences are passed to ALE, being all the others ignored for now. The Memory-Based Learning (MBL) approach was adopted. The option is justified by the possibility of using training examples, created to serve our needs. The classification program TiMBL [6] was used. To increase the information concerning each sentence, a tool that makes the attribution of a part-of-speech (POS) tag to each word [3, p. 298] was used.

The formation of the training file information vectors is the following: the two first features are the first word of the sentence and respective tag; the third and fourth positions are the second word of the sentence and respective tag, and in this way successively. The 19th position is the sentence classification. Results are presented in Table 3. With 762 examples in the training set, correct decisions are above 90 %. 3000 examples

[**responsible(carl_proj)**] professor seabra is the coordinator.
 [**projects**] projects are: carl project and f. c. portugal.
 [**researchers**] researchers are: seabra lopes antonio teixeira.
 [**relation(researcher,carl_proj,X)**] professor antonio teixeira is a researcher of the carl project.
 [**cabinet('antonio teixeira',X)**] professor antonio teixeira is in the second floor in the cabinet 210.

Figure 3: *Some examples of information lists and corresponding sentences generated by the NLG module.*

are enough to obtain 95 % correct decisions. Results are very similar for the 10 different runs, leading to a low standard deviation.

4.3. Natural Language Generation

One of advantages of ALE is its natural language generation capabilities. This means that, provided a description of the intended semantics, ALE can use the grammar to derive the corresponding sentence. Unfortunately, generation is an intrinsically non-deterministic process. What we observed was that, as new rules were added to the grammar, the generation process was becoming increasingly slower. For this reason, instead of using grammar-based generation, Carl continues to use a simpler template-based generation approach, introduced in the previous version of the language module. Also some experiments were made using ASTROGEN [7], developed at the University of Stockholm.

The objective was to make Carl capable of answering questions about IEETA (the research institute where Carl is being developed) and about himself. The database was built in Prolog to facilitate the interaction with other programs. The information that it holds is stored under the form of predicates, providing information regarding entities attributes (laboratories, projects, professors and Carl), relations between the several entities and IEETA internal map. The type of all entities is also defined. The main predicates are: `type`; `relation`; `attribute`; `cabinet`; `place`. It's using these attributes that the AI module specifies which is the information to transmit. Figure 3 represents some examples of sentences generated by the NLG module.

4.4. Speech Synthesis and Animated Face

Conversion from text to speech continues to be made by using the IBM ViaVoice TTS system. An exploratory experiment with limited domain synthesis using Festival framework was done recently. Due to the limited vocabulary used by Carl to convey information to the user, it was possible with a small amount of recording and processing to develop a new voice with a more natural quality.

The functionality of a robot, in terms of tasks that can be performed, is not all that matters. Many users will prefer robots that interact in a friendly way. We thought that an animated face might contribute to that. For the face animation we use the muscle model approach, developed by K. Waters based on earlier work by Parke [8]. For a more realistic behavior, when Carl is speaking, the mouth movement is synchronized with speech synthesis process and some work was performed in adding some expressions of emotions. Random small movements were also added to give some naturalness to the face. Such movements include blinking of an eye or a lip, or moving up and down from

time to time. We believe this to somehow break the "general robot concept" which is something still and mechanical.

5. Carl's Participation in a Recent Contest

A version of Carl, including the human-robot interaction capabilities described above, has been demonstrated at the welcome reception of IROS'2002 as part of the 1st International Cleaning Robots Contest event (Lausanne, October 2002). Some pictures of this demonstration are available at the conference website (http://iros02.epfl.ch/gallery/view_album.php?set_albumName=Events). This demonstration was a great success, as it attracted a lot of public attention as well as media attention.

6. Conclusion and Current Work

This paper has presented the latest developments of a mobile robot with language interface. We have briefly described the software architecture and the solutions found for each module, with more emphasis in natural language understanding.

The integration of several areas of knowledge (speech processing, machine learning, robot navigation, etc..) is one of the biggest challenges of this project. In pursuing our goal significant work was done in testing, combining, changing and configuring several tools.

The CARL project as a strong pedagogic component. A significant portion of the work was developed by undergraduate and MsC students. The aim is not just to develop a robot with natural language interface but also to build an integrated learning platform.

As an ongoing project, Carl is continuously evolving. We are now working on the application of partial parsing to the natural language understanding module. A new syntactic analyzer is being developed using LCFlex for interpreting incomplete and badly recognized sentences. Planned for the near future are: improvements to the NLG module, improving connection between NLG and speech synthesis, and adaptation of the speech recognition acoustic models.

7. References

- [1] L. Seabra Lopes and J.H. Connell. Semisentient robots: Routes to integrated intelligence. *IEEE Intelligent Systems*, pages 10–14, 2001.
- [2] L. Seabra Lopes. Carl: from situated activity to language level interaction and learning. In *Proc. IEEE IROS*, pages 890–896, 2002.
- [3] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Prentice Hall, 2000.
- [4] B. Carpenter and G. Penn. *ALE: The Attribute Logic Engine. User's Guide. Version 3.2.1*. U. Toronto, 2001.
- [5] S.H. Shieber, F.C.N. Pereira, G. van Noord, and R.C. Moore. *Semantic-Head-Driven Generation*. Computational Linguistics, 1990.
- [6] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosh. *TiMBL: Tilburg Memory Based Learner Reference Guide, v. 4.1*, 2001.
- [7] H. Dalianis. A general validation tool by natural language generation for the STEP/EXPRESS Standard, 1999.
- [8] F.E. Parke and K. Waters. *Computer Facial Animation*. A.K. Peters, 1994.