Challenges in Verbal Instruction of Domestic Robots

Guido Bugmann

Robotic Intelligence Laboratory, School of Computing
University of Plymouth, Plymouth PL4 8AA, UK

gbugmann@soc.plym.ac.uk

Abstract:
Domestic robots will need to adapt to the needs and environment of their users. This paper explores the possibility for robots to be programmed by their users through spoken instructions using unconstrained natural language. Issues discussed here are based on intermediate results of an experiment where a miniature remote-brained robot in a model town is given route instructions by naïve subjects. The main issues on the natural language side are: dealing with the ungrammatical nature of many spoken utterances, the detection of errors in speech recognition and interpretation, and the design of intelligent clarification dialogues. On the robot side, the main finding is that complex pre-programmed high-level primitive functions are required to provide a grounding for actions specified in spoken task descriptions. Actions referred to in natural language are adapted to the execution capabilities of human listeners, but are strongly underspecified for robots. Therefore, high-level perception and planning capabilities are required from robots to be able to understand verbal instructions. Other issues include the modular design of domain-specific dialogue systems and robot primitives, and their merging for expanding the robot’s learning capabilities.

1. Introduction

Any robot needs to be programmed by its user to become a functional device. For domestic robots this poses a special problem, as their typical users are naïve in programming and unaware of mechanical and control issues. As for human servants, domestic robots need to learn the particular needs of their employers / owners and the particularities of their environment. Various learning methods have been investigated, such as learning from example (see e.g. Billard et al. 1998), learning by reinforcement, etc. But none of these methods has the power that language has for communicating logical rules and procedural knowledge. Therefore, learning from verbal instructions will be an essential capability in future domestic robots.

Comparatively little research has been devoted to Instruction-based learning (IBL). Huffman and Laird (1995) used textual input into a simulation of a manipulator with a discrete state and action space. Crangle and Suppes (1994) used voice input to teach displacements within a room and mathematical operations, but with no reusability of the learnt procedures in later instructions. In (Torrance, 1994) textual input was used to build a graph representation of spatial knowledge. This system was brittle due to place recognition from odometric data and use of IR sensors for reactive motion control. Knowledge acquisition was concurrent with navigation, not prior to it. Similarly, in (Matsui et al.,1999), the system could learn new actions through natural language dialogues but only while the robot was performing them (i.e. it could only learn a new route from A to B while it was actually moving from A to B and dialoguing with the user).

2. The Plymouth–Edinburgh IBL project

A recent joint project between the Universities of Plymouth and Edinburgh, has re-examined IBL using the latest tools in NL processing and robotics. In this project, learning takes place prior to execution. The robot (fig. 1) learns new navigation procedures from route instructions sampled from a pool of subjects. These subjects used unconstrained natural language, as if they were speaking to another human. These instructions were recorded, transcribed and analyzed for their grammatical structure and lexical content, in order to build a domain-specific restricted speech-recognition grammar. A further functional analysis was performed to produce a list of primitive actions as referred to naturally in human-to-human speech.

The corpus comprising recordings of the subjects and transcripts can be downloaded from the project’s website[1]. The corpus contains instructions for 144 routes given by 24 subjects. 72 of these instructions are used for the development of the system, and the 72 remaining are to be used for evaluating the final system.

The lexicon in this corpus comprises only 330 distinct words and the grammar is a restricted unification grammar that contains only the rules that correspond to utterances in the development set in the corpus. This grammar is compiled into the Grammar Specification Language (GSL) of the speech recognition software NUANCE[2] (Bos, 2002). The semantic analysis of the speech produces results in the form of Discourse Representation Structures (DRS). These are searched for sub-structures corresponding to robot primitives by using rules defined in a Procedure Specification Language (PSL). (Lauria et al., 2002b). The list of primitives found in the instruction is used to construct a piece of program code in the script language Python[3] that stores the learnt procedure.

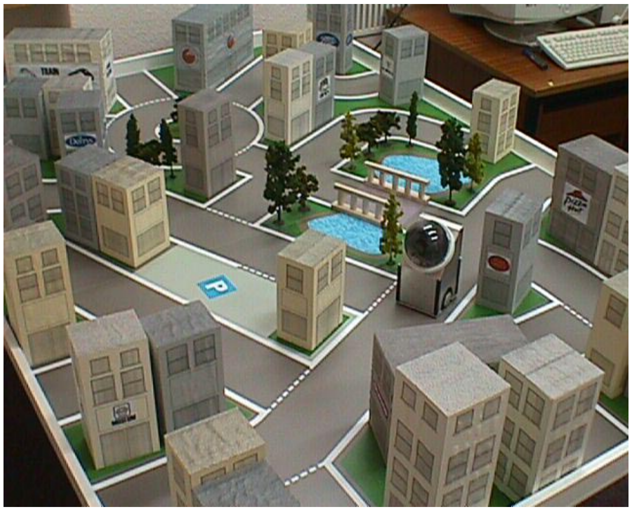

Figure 1. Experimental set-up. A miniature remote-brained robot performs vision-guided navigation in a model town, following a sequence of actions defined in prior instructions given by a user. Video images are sent by wireless from the robot to a PC for processing. The resulting motion commands are sent back by wireless to the robot.

3. Issues in IBL

3.1 Action primitives

The primitives actions found in the corpus (table 1) proved to be rather complex procedures for a robot (Lauria et al., 2002a), requiring the visual localization of landmarks, the identification of navigable areas and the planning of a path to reach a position defined relatively to a landmark.

```
Location_of_landmark_is
enter_roundabout
exit_roundabout
rotate_on_the_spot
move_to_position_relative_to_landmark
enter_road
follow_road
cross_the_road
exit_place
park
take_a_turn
```

Table 1. Primitives found in the corpus.

For instance, subjects used instructions such as “take the second left and you will see it to your left”. These are strongly under-specified commands for a robot and can only be executed if the robot has smart in-built primitives. If the robot had only simple action capabilities, such as “move_forward(*x* cm)” or “turn(*leftwards*, *y* degrees)”, a naive user would first need to learn the set of commands that the robot can understand. Secondly he would probably not be able to construct a sequence of instructions precise enough for the robot to reach its goal. Thirdly, that sequence would be very brittle, depending on a precise starting position and precise execution. Therefore, giving users the freedom to speak in their own words requires quite smart execution functions to be pre-programmed in the robot. This may appear as a constraint, but it has the important benefit of a more robust execution of the instructions, as the robot will gather from the environment the detailed information missing in the instruction. We suggest here that one of the bottlenecks in human-robot communication may not so much be in NL understanding, but in the robot’s ability to execute complex actions.

3.2 Understanding Spoken Language.

Spoken language is different from written language and many utterances that humans understand without problems are deemed ungrammatical if processed by a grammar for written text (Bos et al., 2003

[1] www.tech.plym.ac.uk/soc/staff/guidbugm/ibl/

[2] www.nuance.com

[3] www.python.org

submitted). Unfortunately, there is no accepted general method for extracting the meaning from spoken utterances. This results in a relatively high error rate in speech recognition in this project where no constraint is exerted on the user. In the present stage of development of the system, the grammar covers only 60% of the utterances in the corpus and the word error rate is approximately 40%. In commercial spoken dialogue systems, good performances are achieved by guiding the user through a menu and thereby restricting what he/her is able to say. Hence, progress is needed in the domain of natural language processing, especially spoken language, to enable natural instruction dialogues between man and machine.

3.2 Errors correction.

Speech recognition errors are unavoidable, as they also occur in humans-human communication, and some sentences are inherently ambiguous. Further, users sometime need to use new words not known to the robot (Bugmann et al., 2001). It is important for the safety and the reliability of a service robot that only correctly interpreted instructions are allowed for execution. The standard approach is to build error detection mechanisms into the IBL system. The first stage is to use confidence values produced by the speech recognition engine. This is used to prompt the user to repeat or rephrase an utterance. In the IBL project, a second stage could be implemented, that exploits world knowledge available to the robot. Each primitive and learnt procedure comprises a "prediction" function that specifies the expected result of the action for a given initial state. This enables to simulate "mentally" the execution of a sequence of commands and verify that each command is executable, i.e. is called with a legal set of parameters, and that its preconditions are satisfied by the results of the previous action. We can also check that the sequence of commands lead to the desired goal. In the IBL project, many speech recognition errors were detected through the verification of the executability of the instructions.

Another use of this verification process is the rating of a set of possible interpretations of the users utterances. In this way, the robot can help selecting the more likely interpretation of the user's utterance. This points to potential advantages of coupling speech recognition systems with embedded robotics systems. The latter can provide additional contextual knowledge about actions and their consequences to improve speech understanding. Initial tests of this scheme in the IBL system however show only a minor improvement in speech recognition.

After error detection, an important function is error correction. This requires the design of smart clarification dialogues and poses a number of problems such as that of referencing previous parts of utterances or modifying a speech recognizer and its grammar to recognize a new word. Only a basic clarification dialogue has been implemented in this project, but this functionality will be crucial for the performance of future systems.

3.3 Robot Control Architecture.

When users explain a new procedure to the robot, this is stored in the form a new program that has the same structure as pre-programmed primitives. In principle, it can be re-used to build new, more complex procedures that progressively increase the competence of the robot for executing commands issued by the user. However, experiments have shown that users often refer only to sub-parts of previously taught procedures, for instance, if the robot has been taught previously the route to the post-office, the user may start the explanation of another route with: "take the route as if you were going to the post-office, but turn left after the bridge". This complicates the creation of the program for the new more complex procedure, especially as there may be no reference to the bridge in the previous explanation of the route to the post-office.. In this case, the solution involves interdependent multi-threaded concurrent processing (Lauria et al., 2002b). For instance, the program would go through the sequence of commands leading to the post-office, while at the same time checking if a bridge can be seen in order to switch to the task "turn left after the bridge". This shows that the requirement of understanding unconstrained spoken language has consequences down to the architecture of the robot control system. This issue is far from having been fully explored.

4. Discussion

4.1 Robotic Intelligence, Autonomy and Safety.

All the constraints described in the previous section push in the same direction, that of a more intelligent robot, with the advantage here that "intelligence" is well defined by the analysis of its interaction with the user in an instruction-and-command context. High-level autonomous behaviour such as planning, exploration, etc. is not required, as the knowledge that would result from these processes is directly acquired from the user. However, low-level autonomy is required for the execution of under-specified commands. For instance "take the first left" demands from the robot to look for a left turn in its environment and compute a path to reach it. In general, autonomy poses safety problems and should be avoided. However, some is needed for the understanding of natural language instructions. Where the right balance lies is an open question.

The program generated in the IBL approach comprises not only procedural knowledge but also a function for predicting the consequence of the execution of the learnt procedure. The set of predictive functions of all procedures is equivalent to a graph representing the robots spatial knowledge that could be used for route planning. However, as mentioned above, planning equates to a high level of autonomy and may not be desirable. Previous approaches (e.g. Huffman and Laird, 1995; Mann 1996) have used natural language instructions to generate the graph directly and commands are then executed via planning. However, it has been noted that errors in speech processing result in "noisy" graphs and an unpredictable behaviour of the robot (Mann, 1996). Hence, the direct generation of the execution code under the user's supervision is probably a safer approach.

4.2 The future of Instruction-based Learning

What is the future of IBL? Assuming that all the problems above are solved, our current system would be able to learn reliably the one function `go(start_landmark, goal_landmark)` by adding, for each new pair of landmarks, an execution case built from a set of pre-programmed primitives and previously explained routes. This may be seen as a "route learning package" with domain-specific natural language processing capabilities and the corresponding robot navigation primitives. Domestic robot will accomplish tasks that involve additional functional domains, such as object manipulation, cooking, cleaning, playing games, etc. Each domain can be developed as a package using the current approach. However, it is unclear if a robot would be able to learn from instructions new tasks combining knowledge from distinct packages (e.g. navigation and manipulation). Does the current IBL approach allow the merging of knowledge from different domain packages? The ability to expand IBL systems with new functional packages will be important for users of domestic and service robots. Therefore, this issue needs to be investigated in future work.

Another issue is the handling of deictic reference in situated learning such as "take this glass and moved it over there". So far, the IBL project considered only instructions "from memory", where items to be manipulated by the robot where out of sight of the robot and the instructor at the time of instruction. It is likely that a number of tasks that a user may want to explain to his/her robot are not easily amenable to verbal explanations and may be more comfortable as a combination of in-situ demonstration and explanation. For instance a sorting task: "place objects like this one in this box". In the IBL approach, the reference "like this one" would be mapped to a primitive that locates and records a set of features for later visual recognition of the object, possibly involving a dialogue with the user. Initial work in the IBL project in this direction was aimed at learning labels that define building, for instance the Royal Mail logo on the post-office. Several projects have addressed the question of deictic reference (McGuire, 2002). The recognition of finger pointing movements is usually a key part of such work along with some form of object recognition. A critical issue here is the definition of objects and the storage of reference features for future execution. This requires advanced vision capabilities that are still ineffective for all but the simplest object/background combinations. Here also, limitations of current artificial vision systems are probably more critical than speech recognition problems.

Advanced sensory capabilities are key to robust robot programming. The IBL project has made an indirect contribution to this problem via the design of high-level primitives. For instance, in order to execute "turn left", the robot must be able to recognize and localize a left turn. In (Kyriacou et al., 2002) this was implemented through a road-surface extraction procedure followed by a template matching procedure. There were templates for different road features and only those required for the task at hand were used. For instance, the image would not be searched with a "right_turn" template if the task was to turn left. However, if all templates were used with each picture, then such as road-landmark recognition capability could theoretically be used to teach a route to a robot simply by passively guiding it along the route. The robot would be able to build a high-level representation of the route in terms of turns, intersections, etc. This has the advantage of being much more robust than for instance odometric recordings of displacements. This representation also allows the robot to later report verbally on its experience in terms understandable by naive users.

5. Conclusions

In conclusion, experiments with a real robot and spoken language input have generated clear pointers to where future research is needed to enable the instruction and command of domestic and service robots using natural language. The possibly unsurprising message is that the robot needs to have intelligent execution capabilities closer to those of humans in order to understand human language. Processing spoken language has its problems that could possibly be eliminated by using a more directed dialogue approach. However the requirement for a high-level of functionality on the robot side remains.

Acknowledgments:

The author is grateful for discussions with Ewan Klein, Johan Bos, Stanislao Lauria and Theocharis Kyriacou.

References:

Billard A., Dautenham K. and Hayes G. (1998) "Experiments on human-robot communication with Robota, an imitative learning and communication doll robot", Contribution to Workshop "Socially Situated Intelligence" at SAB98 conference, Zurich, Technical Report of Centre for Policy Modelling, Manchester Metropolitan University, CPM-98-38.

Bos J. (2002) "Compilation of Unification Grammars with Compositional Semantics to Speech Recognition Packages" COLING 2002. Proceedings of the 19th International Conference on Computational Linguistics. Pages 106-112.

Bugmann G., Lauria S., Kyriacou T., Klein E., Bos J. and Coventry K. (2001) " Using Verbal Instruction for Route Learning: Instruction Analysis ", Proc. TIMR 01 – Towards Intelligent Mobile Robots, Manchester 2001. Technical Report Series, Department of Computer Science, Manchester University, ISSN 1361 – 6161. Report number UMC-01-4-1 .

Crangle C. and Suppes P. (1994) Language and Learning for Robots, CSLI Lecture notes No. 41, Centre for the Study of Language and Communication, Stanford, CA.

Huffman S.B. and Laird J.E. (1995) "Flexibly Instructable Agents", *Journal of Artificial Intelligence Research*, 3, pp. 271-324.

Kyriacou T., Bugmann G. and Lauria S. (2002) "Vision-Based Urban Navigation Procedures for Verbally Instructed Robots" Proceedings of the 2002 IEEE/RSJ Intl. Conference on Intelligent Robots and Systems (IROS'02) EPFL, Lausanne, Switzerland, October 2002. pp.1326-1331.

Lauria S., Bugmann G., Kyriacou T., Klein E. (2002a) "Mobile Robot Programming Using Natural Language" *Robotics and Autonomous Systems*, 38 (3-4): 171-181

Lauria S. Kyriacou T., Bugmann G., Bos J., Klein E. (2002b) "Converting Natural Language Route Instructions into Robot-Executable Procedures" Proceedings of the 2002 IEEE Int. Workshop on Robot and Human Interactive Communication (Roman'02), Berlin, Germany, pp. 223-228.

Mann GA, (1996) "Control of a Navigating, Rational Agent by Natural Language", PhD Thesis, School of Computer Science and Engineering, University of New South Wales, Australia.

Matsui T., Asoh H., Fry J., Motomura Y., Asano F., Kurita T., Hara I. and Otsu N. (1999) Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services, Proc. AAAI/IAAI, pp. 621-627.

Torrance M.C. (1994) Natural Communication with Robots, MSc Thesis submitted to MIT Dept of Electrical Engin. and Comp. Science.

P. McGuire, J. Fritsch, J. J. Steil, F. Röthling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter (2002). "Multi-modal human-machine communication for instructing robot grasping tasks". In Proc. IROS 2002, pages 1082--1089. IEEE, 2002.