Linguistic Communication and Grounding in Sensorimotor Cognitive Robots

Abstract

In this paper we present a brief overview of research on language learning in cognitive robotics and describe preliminary work on language grounding using the iCub robotic platform. The study investigates how the language acquired by robotic agents can be directly grounded in perceptual and action representations. Interactive intelligent systems research has focused on the investigation of the relationship between language and action. Studies are characterised by the hypothesis that symbols are directly grounded into the agents' own categorical representations, whilst at the same time having logical (e.g. syntactic) relationships with other symbols (Harnad 1990). This is achieved when cognitive agents and robots learn to name entities, individuals and states in the world whilst they interact with their environment and build sensorimotor representations of it. Such an approach has important technological implications for the design of autonomous systems capable to assist humans in a variety of situations including everyday tasks (e.g. service/household robotics), and highly-specialized situations such as with autonomous systems for defence (e.g. collaborative and multi-agent distributed tasks for exploration and navigation in unknown terrains). Cognitive systems are essential for integrated multi-platform systems capable of sensing and communicating.

This report is organized as follows: Section I provides a general overview of the problem. It also provides a brief review of existing work on language grounding in cognitive robots; Section II gives a summary of the work accomplished up to date during the exchange visit to the IIT (Italian Institute of Technology) at Genoa for a new model of language learning in the iCub robotic platform.

I – Background

In the near future, we expect participation of intelligent robots to grow rapidly in human society. Robots will be able to learn language and world understanding from direct interaction with humans. Therefore, since effective interaction between robots and people will be essential, robots will need to be able to identify a speaker among a group of people and recognise speech signals in a real environment.

Cognitive systems research focuses on the development of natural and artificial information processing systems (e.g. internet agents, adaptive agents, robots) capable of perception, learning, decision-making, communication and action. They are designed to assist humans in a variety of

situations including everyday tasks, such as service/household robotics, and highly-specialized situations, such as in autonomous systems for defence.

Interactive intelligent systems and cognitive robotics research is increasingly focusing on the close integration of language and other cognitive capabilities (Barsalou 1999; Cangelosi et al. 2005; Pecher & Zwaan 2005). One of the most important aspects in language and cognition integration is the grounding of language in perception and action. Grounding can also be considered as the process whereby internal representations are connected to external percepts (Harnad 1990). This is based on the principle that cognitive agents and robots learn to name entities, individuals and states in the external (and internal) world whilst they interact with their environment and build sensorimotor representations of it. For example, the strict relationship between language and action has been demonstrated in various empirical and theoretical studies, such as psycholinguistic experiments (Glenberg & Kaschak 2002), neuroscientific studies (Pulvermuller 2003) and language evolution theories (Rizzolatti & Arbib 1998). This link has also been demonstrated in computational models of language (Cangelosi & Parisi 2004; Wermter et al 2003). The use of this grounded approach to the design of linguistic cognitive systems is vital for overcoming the known difficulties in intelligent agents whose linguistic abilities are purely based on abstract symbolic representations. Equally important, is the reverse: learning abstract categories and situations, which are not directly observed in the world, can only be grounded in language and communications among agents (Barsalou 1999).

Much research has been recently dedicated to modelling the acquisition of categorical representation for the grounding of symbols and language in cognitive agents and robots. Here we focus on the approaches and techniques based on the cognitive grounding principle, i.e. when the language acquired by robotic agents can be directly grounded in perceptual and action representations autonomously developed by the agents. These studies are characterised by the hypothesis that symbols are directly grounded into the agents' own categorical representations, whilst at the same time having logical (e.g. syntactic) relationships with other symbols. First, each symbol is directly grounded into internal categorical representations. These representations include perceptual categories (e.g. the concept of blue colour, square shape, and male face), sensorimotor categories (e.g. the action concept of grasping, pushing, and carrying), social representations (e.g. individuals, groups and relationships) and other categorizations of the agent's own internal states (e.g. emotions and motivations). These categories are connected to the external world through our perceptual, motor and cognitive interactions with the environment. Second, symbols also have syntactic relationships with the other symbols of the lexicons used for communication. This allows symbols to be combined, using compositional rules such as grammar, to form new meanings. For example, the combination of the two symbols "stripes" and "horse", which are directly grounded into the agent's own sensorimotor experience of striped objects and horses in its environment, produces the new concept (and word) "zebra". This new symbol becomes indirectly grounded in the agents' experience of the world through the process of "symbol grounding transfer".

The approach is in opposition to other adaptive modelling systems that view language as an independent and autonomous capability of the agent, and are subject to the symbol grounding problems (Harnad, 1990). Language grounding models provide a new route for modelling complex cross-modal phenomena arising in situated, embodied language use. As early language acquisition is overwhelmingly concerned with objects and activities which occur in a child's immediate surrounding environment, these models are of a significant interest for understanding situated language acquisition.

In cognitive robotics literature, there are various language models based on grounded methodologies. Some use real robots interacting in physical environments, while others use simulated adaptive agents. In robotic models, communication results from the dynamical interaction between the robot's physical body, its cognitive system and the external physical and social environment. Some studies stress the grounding in action and sensorimotor processes, such as Marocco's et al. (2003) model of robotic arms and Vogt's (2000) mobile robots. Other robotic models highlight the grounding through social interaction, such as Steels & Kaplan's (1999; 2000) Talking Heads and AIBO robots. For example, Steels and collaborators have investigated the emergence of shared languages in group of autonomous cognitive agents that learn categories of objects. They use discrimination tree techniques to represent the formation of categories of geometric shapes and colours. Cangelosi and collaborators have studied the emergence of language in multi-agent systems performing navigation and foraging tasks (Cangelosi 2001), and object manipulation tasks (Cangelosi & Riga 2006; Marocco et al 2003). They use neural networks that acquire, through evolutionary and epigenetic learning, categorical representations of the objects in the world that they have to recognise and name.

Systems have also been developed which models visually-grounded object descriptors and spatial language to generate whole phrases and sentences in scene description tasks (e.g. Roy & Mukherjee 2005; Herzog & Wazinski 1995; Roy 2002). For example, Roy and Mukherjee use word models which are perceptually grounded in a system capable of scene description understanding whereby speech interpretation is integrated with visual context (Spivey et al. 2001) and modelling visual attention dynamics of situated language comprehension (Tanenhaus et al. 1995; Chambers et al. 2004; Roy & Pentland 2002). Roy & Pentland (2002) have also proposed the cross-channel early lexical learning (CELL) model for language learning in robots. This is capable of learning to break down speech into words and link them to an acquired visual shape and colour categories based on input through video and speech (Robinson 1994). CELL is able to draw distinctions between words by identifying their word boundaries and from there create visual categories and form semantic links between those spoken words and visual categories.

Many robotics projects are looking at various aspect of language emergence such as the development of vocabulary and/or grammar from various forms of experience (Steels 1996, Fitzpatrick 2003, Werker et al 1996, Breazeal 2000; Yu & Ballard 2004). For example, Yu and ballard (2004)

implemented an endpoint detection algorithm for acoustic signals to segment the speech stream into spoken utterances. Each spoken utterance contained one or more spoken words. These utterances were then converted into text for the speech recognition using of the shelf speech recognition software. Varchavskaia et al. (2001) investigated if the speech input has specialised characteristics comparable to those of an infant directed speech. This depends on the nature of the task to which the robot is being applied. Experiments included interactions between the Kismet robot and young children for the purpose of teaching the robot new words as described in Breazel (2000) engaging in proto-conversational turn-taking.

The development and application of sensory-grounded language systems leads the way to a new kind of cognitive model that is able to deal directly with recordings from natural human environments bypassing the need for manual transcription or coding. These systems are able to approach learning from a human perspective, through dealing with natural sensory data.

In terms of what the future holds this type of models, one important aspect is the design of machines which can autonomously learn and question ideas and concepts about the world. They would also be able to subsequently communicate in a natural way about these ideas and beliefs in various problem domains. Automated generation of weather forecasts (Reiter et al. 2005), large-scale image database retrieval by natural language query (Barnard et al. 2003), verbal control of interactive robots, and other human-machine communication systems (Roy 2003; Yu & Ballard 2004; Roy & Reiter 2005; Herzog & Wazinski 1995) are some of the applications which could make use of this kind of emerging technology.

II – Language learning in iCub: Ongoing work

The aim of this new model is to extend previous work on language learning and grounding in simulated cognitive agents (Cangelosi & Riga 2006; Cangelosi et al. in press) to the new cognitive robotic platform iCub (Metta, G et al, 2006). In particular, we are interested in the scaling up issues involving current grounded approaches. This scaling up process also involves the productivity of language, i.e. the capacity to generate autonomously new words and concepts from the combination of previously grounded words and actions. The main hypothesis is that such a grounding approach will permit a more efficient development of language capabilities in robots. To achieve this goal, this research tackles the problem from a developmental infant/child point of view, establishing concepts and hypotheses on the psychological development of children including their acquisition of object knowledge and most importantly social skills. The research stages will include modelling the early acquisition of language in robots using artificial neural network controllers.

In the next sections we will briefly overview the ongoing work and the technological choices for this language learning project in the iCub platform.

Speech pre-processing

The human ear can detect and analyse sounds/vibrations frequencies that originate from a sound and distribute it to different nerve cells in the auditory portion of the central nervous system (based on the process of resonance). The human ear is able to take input from the outside world, change the sound waves into a signal, made of nerve impulses, that is sent to the brain.

In order to replicate this process on a computer system for sound processing, we have built a speech analysis software module that uses a Fast Fourier Transform (FFT) of the speech signal. In other words the FFT takes as input sound from the microphone and splits the voice/input into its component frequencies. This allows the system to filter out some of the noises produced by the microphones input.

We have built two microphones with their respective pre-amplifiers in order to be placed on the iCub outer shell head. The microphones are two omni-directional microphones (6x2.7mm) that have a flat frequency response and that are easy to integrate in the iCub platform. The preamplifiers had to be built in order to connect the microphone to the sound card of the computer (using the tip of a 3.5 mm stereo plug).

Initial sound test with speech are promising as voice input is detected adequately, even when the speaker is situated at a distance (2 meters) and noise levels are considerably low. In order to recognize speech with high confidence, the techniques that separate speech signals from various non-speech signals and remove noises from the speech signals have received a great deal of attention.

To further improve the system we plan to place filters directly into the pre-amplifiers in order to filter out excessive unwanted environmental disturbances such as the robots processor cooling fans, computers, and other environmental noises.

Neural network classification of speech

In order for the system to be able to learn from the sound analysis produced by the speech possessing module, we have constructed a self organizing map (SOM; Kohonen 1995), trained using unsupervised learning. The SOM is a single layered feedforward neural network where the output units are arranged in a topological 2D grid. The purpose of the learning in the SOM is to associate various parts of the SOM lattice to respond to different input patterns. This is partially inspired by how the auditory/vision and other sensor information are handled in distinct parts of the cerebral cortex in the human mind (Grossberg 2003). The learning process is competitive and unsupervised, meaning that no teacher is needed to define the correct output (or specify the cell into which the input is mapped) for an input. Only one map node (winner) at a time is activated corresponding to each input. The locations of the responses in the array tend to become ordered in the learning process as if some meaningful nonlinear coordinate system for the different input features were being created over the network (Kohonen, 1995).

Our model has been trained using a vast amount of data collected from various sources. The data comprises of more than 100 single word utterances

from two different speakers (words spoken in isolation), 544 syllable utterances from two different speakers, for determining the ability of the system to distinguish between substantially small differences.

The data was collected using different sources such as: an “off the shelf” microphone, sound files gathered from various users vocalizing words, syllables and utterances (with and without noise) and directly (real time) from the robot’s “ears”. Our system was able to learn and distinguish between the different speech/sounds produced in a substantially small amount of time. Naturally our system took additional time to learn using the “noisy” data but was still able to learn and distinguish between the different or slightly different words. For example words like “ball” and “mall”.

The results of the speech analysis and learning phase still has to be passed to another neural network, as in Cangelosi & Riga’s (2006) language grounding neural system, in order to integrate visual as well as auditory information.

Vision

In order to help the system to improve its performance in the learning phase, it is necessary to use a visual processing technology that can support robots to detect and track an object or category of objects situated in its environment. Moreover, using visual processing tools will make the robot not only able to accommodate the sound errors that may occur, but also cast aside unnecessary speech or noise signal/frequencies. The model will therefore be able to augment its performance in the learning phase.

The vision system is still at an early stage of development but is based upon an algorithm that uses approximation techniques for the purpose of detecting round shapes using the OpenCV library. For example, the percentage of roundness will help the system distinguish between objects and categorize them as a being a ball, cube or stick/pen etc. Using OpenCV, the robot receives information concerning the number of objects, the percentage of “roundness” and the coordination of the detected objects.

Simulation Software for iCub

In addition to the above work on language learning, we have developed a prototype of the 3D dynamic simulation software of the iCub robot using Open Dynamics Engine (ODE). ODE is an open source physics and motor dynamic interface. It enables robotic simulators to consider the role of physical constraints within a simulated environment able to compute and resolve forces that emerge through the interaction of objects/entities. ODE includes an interface to OpenGL that facilitates the rendering or objects (boxes, sphere etc).

Our simulation model is a replication, using the exact data and inertia matrices, of the actual iCub platform. This simulator has been developed for the users of the RoboCub project, as an alternative to the physical iCub platform for fast simulating and testing. The idea was to make the simulation

as close as possible to the iCub using the same type of interface. Screenshots of the iCub prototype simulator can be seen in Figures 1 and 2.

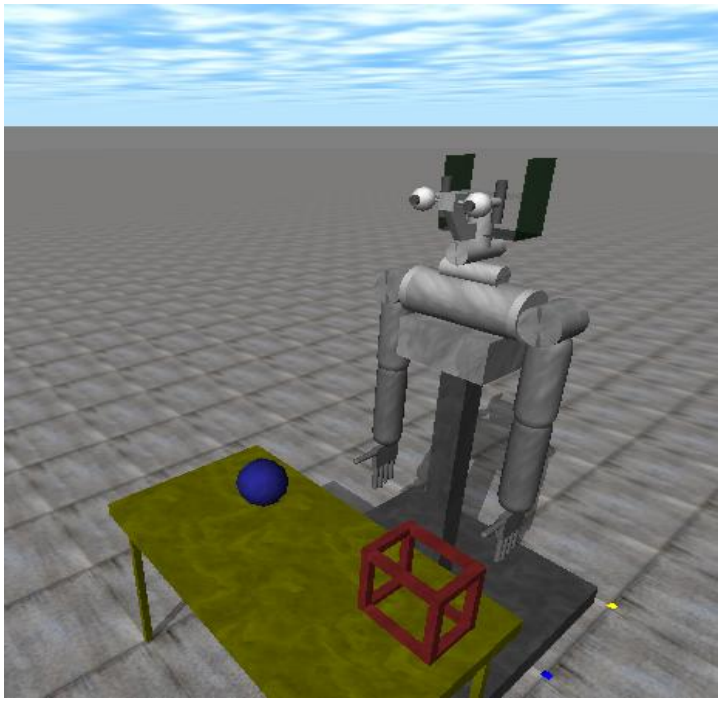

Figure 1: Simulation setup of the iCub (wihtout head cover) and objects

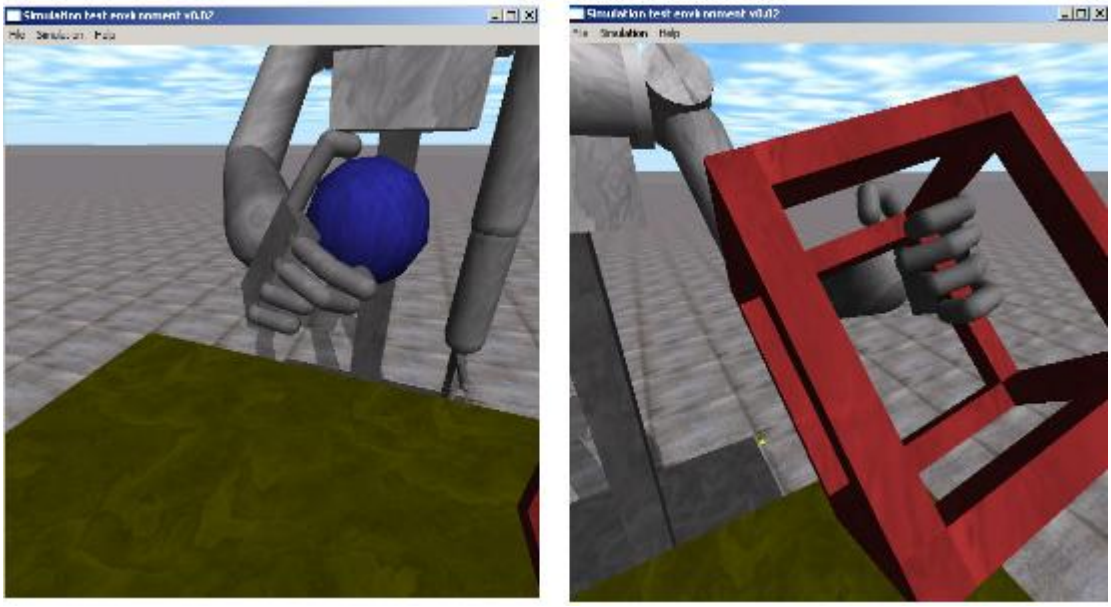

Figure 2: Right arm ball grabbing and left arm cube grabbing

Future work

The plan for the future work to complete this prototype model of language learning in iCub includes:

- Completion of visual categorisation module
- Integration of visual and SOM modules for language grounding as in Cangelosi & Riga (2006) neural architecture for language grounding
- Experiments on grounding transfer and production of new words
- Experiments on motor control and naming of actions
- Experiments on object categorisation and object manipulation using linguistic instructions
- Further work on the iCub simulator

Conclusion

The general research aim of this report is to focus on the use of grounding approaches and developmental robotic methodologies to study language acquisition and human robot interaction and communication. It also provides a brief description of an ongoing language learning study with the iCub robot.

The potential technological and practical implications for this grounding approach to language learning are great in the fields of robotics, artificial intelligence and cognitive systems design. The successful design of linguistic cognitive agents that are able to interact with their environment (including humans) provides an innovative approach to the field of interactive intelligent system design. Such systems are need in much for example in the domains of service robotics, household systems, exploratory autonomous systems applications, and even commonly used software applications such as search engines and natural language interfaces.

Acknowledgement

This work has been supported by the euCognition support action NA97-2.

References

Barnard, K., et al., Matching words and pictures. The Journal of Machine Learning Research 2003. 3(Special issue on Machine learning methods for text and images): p. 1107-1135.

Barsalou, L., Perceptual symbol systems. Behavioral and Brain Sciences, 1999. 22: p. 577-609.

Breazeal, C. (2000). Sociable Machines: Expressive Social Exchange Between Humans and Robots. PhD thesis, MIT Department of Electrical Engineering and Computer Science.

Cangelosi A. (2001). Evolution of communication and language using signals, symbols and words. IEEE Transactions on Evolutionary Computation. 5(2), 93-101

Cangelosi, A., G. Bugmann, and R. Borisyuk, Modeling Language, Cognition and Action:. Proceedings of the 9th Neural Computation and Psychology Workshop. 2005, Singapore: World Scientific.

Cangelosi, A. and D. Parisi, Simulating the Evolution of Language. 2002, London: Springer-Verlag.

Cangelosi A., & Riga T (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots, Cognitive Science. 30(4), 673-689

Chambers, C.G., M.K. Tanenhaus, and J.S. Magnuson, Action and affordances in syntactic ambiguity resolution Journal of Experimental Psychology: Learning, Memory and Cognition, 2004. 30: p. 687-696.

Fitzpatrick P. From First Contact to Close Encounters: A developmentally deep perceptual system for a humanoid robot. PhD thesis at MIT, 2003.

Glenberg, A. and M. Kaschak, Grounding language in action. Psychonomic Bulletin & Review, 2002. 9(3): p. 558-565.

Grossberg S., How Does the Cerebral Cortex Work? Development, Learning, Attention, and 3-D Vision by Laminar Circuits of Visual CortexBehav Cogn Neurosci Rev, March 1,2003; 2(1): 47 – 76

Harnad, S., *The Symbol Grounding Problem.* Physica D, 1990. 42(335-346).

Herzog, G. and P. Wazinski, VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions, in Integration of Natural Language and Vision Processing: Computational Models and Systems, P. McKevitt, Editor. 1995, Kluwer,: Dordrecht. p. 83-95.

Kohonen T., 1995 Self-Organizing Maps. Springer, Berlin, Heidelberg

Marocco, D., Cangelosi, A., & Nolfi, S. (2003). The emergence of communication inevolutionary robots. Philosophical Transactions of the Royal Society of London – A 361, 2397-2421.

Metta, G., P. Fitzpatrick, and L. Natale, YARP, Yet Another Robotic Platform. International Journal on Advanced Robotics Systems Special Issue on Software Development and Integration in Robotics 2006.

Pecher, D. and R.A. Zwaaan, Grounding cognition: The role of perception and action in memory, language, and thinking. 2005, Cambridge: Cambridge University Press.

Pulvermuller F. (2003). The Neuroscience of Language. On Brain Circuits of Words and Serial Order. Cambridge University Press.

Reiter, E., et al., Choosing Words in Computer-Generated Weather Forecasts. Artificial Intelligence, 2005. 167: p. 137-169.

Rizzolatti G. & Arbib M.A. (1998). Language within our grasp. Trend Neurosciences, 21(5), 188-194.

Robinson, T., An application of recurrent nets to phone probability estimation. IEEE Tansaction on Neural Networks, 1994. 5: p. 298-305.

Roy, D., Learning visually grounded words and syntax of natural spoken language. Evolution of Communication, 2002. 4(1): p. 33-56.
Roy, D., Grounded Spoken Language Acquisition: Experiments in Word Learning. IEEE Transactions on Multimedia, 2003. 5(2): p. 197-209.

Roy, D. and N. Mukherjee, Towards Situated Speech Understanding: Visual Context Priming of Language Models. Computer Speech and Language, 2005. 19(2): p. 227-248.

Roy, D. and A. Pentland, Learning words from sights and sounds: A computational model. Cognitive Science, 2002. 26: p. 113-146.

Roy, D. and E. Reiter, Connecting Language to the World. Artificial Intelligence, 2005. 167(1-2): p. 1-12.

Spivey, M., et al., Linguistically mediated visual search. Psychological Science, 2001. 12: p. 282-286.
Steels, L. (1996). Emergent adaptive lexicons. In Proceedings of the fourth international conference on simulation of adaptive behavior, pages 562–567, Cape Cod, MA.

Steels L. and Kaplan, F. (1999). Situated grounded word semantics. In Dean, T., editor, IJCAI99. Morgan Kaufmann Publishers.

Steels L., & Kaplan, F. (2000). AIBO's first words: The social learning of language and meaning. Evolution of Communication, 4, 3-32.

Tanenhaus, M.K.e.a., Integration of visual and linguistic information during spoken language comprehension. Science, 1995. 268: p. 577-609.

Varchavskaia P, Fitzpatrick P and Breazeal C . Characterizing and processing robot-directed speech. Proceedings of the IEEE/RAS International Conference on Humanoid Robots 2001, Tokyo, Japan, Nov. 22-24, 2001

Vogt P. (2000). Bootstrapping grounded symbols by minimal autonomous robots. Evolution of Communication, 4(1): 89-118.

Werker, J., Lloyd, V., Pegg, J., and Polka, L. (1996). Putting the baby in the bootstraps: Toward a more complete understanding of the role of the input in infant speech processing. In Morgan, J. and Demuth, K., editors, Signal to Syntax: Bootstrapping From Speech to Grammar in Early Acquisition, pages 427–447. Lawrence Erlbaum Associates: Mahwah, NJ.

Wermter, S., Elshaw, M., and Farrand, S., 2003, A modular approach to self-organization of robot control based on language instruction. Connection Science, 15(2-3): 73-94.

Yu, C. and D. Ballard, A multimodal learning interface for grounding spoken language in sensorimotor experience. ACM Transactions Applications perception, 2004. 1: p. 57-80.