Grounding Abstract Action Words through the Hierarchical Organization of Motor Primitives

Francesca Stramandinoli*, Davide Marocco, Angelo Cangelosi
Centre for Robotics and Neural Systems, Plymouth University
Drake Circus, PL4 8AA, Plymouth, United Kingdom
Email: {francesca.stramandinoli, davide.marocco, A.Cangelosi}@plymouth.ac.uk
*Telephone: +44 (0) 17525 84908

***Abstract*—Cognitive developmental robotics is facing the challenge of building robots capable of working independently and/or with other agents in non-structured scenarios, which can autonomously react to dynamic changes that occur in the environment. Providing robots with the capability to comprehend and produce language in a "human-like" manner represents a powerful tool for flexible and intelligent interaction between robots and human beings. Robots endowed with linguistic capabilities, in fact, could better understand situations and exchange information; through language robots could cooperate and negotiate with human beings in order to accomplish shared plans.**
This work describes a neuro-robotics model for the acquisition of abstract action words in the iCub humanoid robot. We claim that the acquisition of concepts that refer to such abstract words (e.g. verbs like "*use*", "*make*") can be driven by action's organization. In the presented model the integration of low-level capabilities (e.g. perceptual and sensorimotor skills) enables the hierarchical organization of concepts that leads to the grounding of more general words.

I. The embodiment of abstract language

Can robots use sensorimotor categories to indirectly ground higher-order concepts and more abstract language? This is the question we would like to address through this work. Studies conducted on children's early vocabulary acquisition have shown that, when children learn to speak, they first learn concrete nouns (e.g. object's names) and then abstract ones (e.g. verbs) [1], [2]. While concrete terms refer to tangible entities characterized from an evident mapping to the perceptual world, more general and abstract words cannot be directly linked to sensorimotor knowledge because they are distant from immediate perception and sensorimotor experience [1]. This is why the problem of the acquisition of abstract concepts cannot be simply resolved by directly linking words to the physical entities to which they refer. In contrast to other forms of communication, language is a combinatorial system that permits the conveyance of new messages by combining individual words together [3]. Furthermore, recent evidence has suggested that the human motor system is also hierarchically organized [4]; that is, low level motor primitives can be integrated and recombined in different sequences in order to perform novel tasks. In addition to this, recent studies presented in the neuroscience [5] and behavioural communities [6] have revealed that language is embodied in perceptual and sensorimotor knowledge. Furthermore, different theories proposed in psychology have claimed that embodiment plays an important role even in representing abstract concepts (theories based on "metaphors" [7], "simulations" [8] and "actions" [9]).

II. Neuro-robotics Model

The model we present in this work aims to account for the acquisition of abstract action words in the iCub humanoid robot [10]. In carrying out our experiments we assume that language is *embodied* in perceptual and sensorimotor experience and *situated* in the context in which it occurs. By exploiting the combinatorial organization of language and of the motor system, we propose an architecture that integrates simple motor primitives and words in order to create the semantic reference of terms that do not have a direct mapping to the perceptual world [11]. The semantic referents of these words are formed by recalling and reusing the sensorimotor and perceptual knowledge previously grounded. It is our intent to create a "grounding kernel" from which new concepts can be obtained through linguistic definition alone [12]. In this way, more abstract concepts can be formed through language: they involve a form of higher-order concepts that are based upon the combination of simpler word representations [8].
The architecture presented in this work is based on a 3-layer Jordan partially recurrent neural network [13]. Considering that conceptualization requires the activation of multimodal information [8], our architecture has been conceived to receive different modality inputs. The visual and sensorimotor inputs have been recorded from the iCub sensors while the linguistic inputs are binary vectors for which "one-hot" encoding has been adopted. Vision, motor actions and language are integrated in order to ground abstract action words (e.g. "*use*", "*make*") in perceptual and sensorimotor knowledge.

A. Training Procedure

The implemented training strategy takes inspiration from developmental learning. Studies conducted in developmental psychology and neurophysiology have revealed that perception and sensorimotor learning are pre-linguistic [14]. That is, children acquire some motor behavior and the capability to perceive objects before they learn to name them. Taking inspiration from these studies, we have organized the training of our architecture in three different incremental stages:
(i) First, the model is trained to recognize a set of tools and learn object-related actions.
(ii) Subsequently, the model is trained to name objects and actions. These two stages of the training enable the direct grounding of words into perceptual and sensorimotor inputs.
(iii) In the last stage of the training new words (which refer to more abstract concepts) are grounded, integrating and recalling the visual and sensorimotor knowledge that has previously been directly linked to concrete terms. For these preliminary simulations of the model, the dataset consisted of 18 sequences

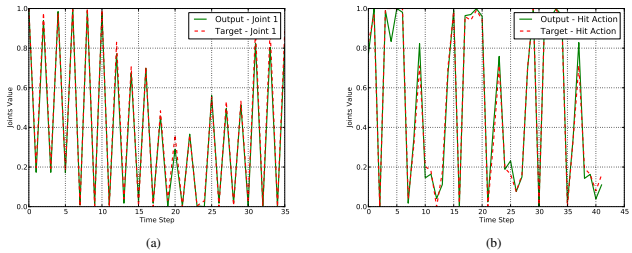
of 6 elements each, which during the three incremental stages of the training permitted learning respectively: the mapping between perceptual and sensorimotor inputs, the name of objects and actions and to acquire higher-order concepts. The network was trained through batch back-propagation for 4000 iterations (incorporating all the three training stages) and for 10 random seeds.

B. Robotic Task

The task for the iCub robot consisted of learning to recognize a set of tools (e.g. "*knife*", "*hammer*", "*brush*", etc) and to perform object related actions first (e.g. "*cut*", "*hit*", "*paint*", etc); subsequently the robot learned to name objects and actions. Finally, the robot was trained to learn abstract action words through new linguistic sequences to be interpreted in terms of its own internal motor and language repertoire.

C. Preliminary Results

Simulation results have shown that after only 100 iterations, the training error is smaller than 0.05 and the network successfully learns to ground abstract action words (e.g. "*use*", "*make*") in perceptual and sensorimotor knowledge. From the analysis of the activation values of hidden units, recorded during stage **(ii)** and **(iii)** of the training, we have observed that the hidden units follow similar activation patterns during these different stages. At the end of the training stage **(ii)**, before proceeding with the training stage **(iii)**, we also tested the ability of the network to generalize new linguistic meanings. We wanted to verify if the network for example, after learning the behavior "*cut*" [with] "*knife*", was able to generalize the new linguistic command "*use*" [the] "*knife*" (for which the model was not yet trained). Results of this test have shown that the iCub is able to generalize and therefore to perform the "*cut*" [with] "*knife*" behavior in response to the command "*use*" [the] "*knife*". In (Fig. 1(a)) we show output and target values for one of the 7 joints of the iCub arm, controlled by the network during one of the training stages.



(a) (b)

Fig. 1. Output and target values for one of the 7 joints of the iCub arm controlled by the network (a). Output and target joint values for the "*hit*" action taught to the iCub (b).

From (Fig. 1(a)) we can observe that the network is able to output the appropriate joint values for the iCub arm. In (Fig. 1(b)) we show output and target joint values for one of the actions taught to the iCub. In order to have a quantitative measure of the similarity between the output and target joint values over time, we performed Dynamic Time Warping (DTW) [15] on joint sequences. The result of DTW confirmed that the values of the output joint values over time are very close to their target values. The network that exhibited the best performance in terms of training error and DTW has been used to run tests that permit to better understand the relations between visual, sensorimotor and linguistic knowledge in the iCub robot. Preliminary results have shown that when the visual input is deactivated, the hidden units follow a less structured and more chaotic pattern.

III. Conclusion

The model described in this paper implements a cognitive robotics architecture for the grounding of general words in the iCub robot. Simulation results have shown that the three incremental stages of the training of the model have been successfully accomplished and the network successfully learns to ground abstract action words (e.g. "*use*", "*make*") in perceptual and sensorimotor knowledge. We claim that the hierarchical organization of concepts that the model creates can represent a useful mechanism for the acquisition of more abstract and general concepts in robots.

Acknowledgment

This research has been supported by the EU project RobotDoC number 235065 from the 7th Framework Programme, Marie Curie Action ITN.

References

[1] D. Gentner, *Why nouns are learned before verbs: Linguistic relativity versus natural partitioning*, Champaign, Ill.: University of Illinois at Urbana-Champaign, Center for the Study of Reading, 1982.

[2] B. Mcghee-Bidlack, *The development of noun definitions: a metalinguistic analysis*, Journal of child language. Vol. 18 (2), pp. 417–434, Cambridge University Press, 1991.

[3] K. Frankish, and W. Ramsey, *The cambridge handbook of cognitive science*, Cambridge University Press (to appear).

[4] F. A Mussa–Ivaldi and E. Bizzi, *Motor learning through the combination of primitives*. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 355, 1404, 2000.

[5] F. Pulvermüller, M. Harle, F. Hummel, *Walking or talking? Behavioral and neurophysiological correlates of action verb processing*. Brain and language, Vol. 78, pp. 143–168, 2001.

[6] G. Buccino, L. Riggio, G. Melli, F. Binkofski, V. Gallese, G. Rizzolatti, *Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study*. Cognitive Brain Research, Vol. 24, pp. 355–363, 2005.

[7] G. Lakoff, and M. Johnson, *Metaphors we live by*. Psychological review, Vol. 1111, Chicago London, 1980.

[8] L. W. Barsalou, *Perceptual symbol systems*. Behavioral and Brain Sciences, Vol. 22, pp. 577–660, 1999.

[9] A.M. Glenberg, and M. P. Kaschak, *Grounding language in action*. Psychonomic bulletin & review, Vol. 9 (3), pp. 558–565, Springer, 2002.

[10] G. Metta, G. Sandini, D. Vernon, L. Natale, F. Nori, *The iCub humanoid robot: an open platform for research in embodied cognition*. Proceedings of the 8th workshop on performance metrics for intelligent systems, pp. 50–56, ACM, 2008.

[11] F. Stramandinoli, D. Marocco and A. Cangelosi, *The Grounding of Higher Order Concepts in Action and Language: a Cognitive Robotics Model*. Neural Networks, 32, 165–173, 2012.

[12] S. Harnad, *From Sensorimotor Categories and Pantomime to Grounded Symbols and Propositions*, Oxford University Press, 2010.

[13] M. I. Jordan, *Attractor dynamics and parallelism in a connectionist sequential machine*. Proceedings of the Eighth Annual Conference of the Cognitive Science Society, pp. 531–546, 1986.

[14] M. Jeannerod, *Neural Simulation of Action: A Unifying Mechanism for Motor Cognition*. NeuroImage, Vol. 14 (1), pp. S103–S109, 2001.

[15] H. Sakoe and S. Chiba, *Dynamic programming algorithm optimization for spoken word recognition*. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 1, pp. 43–49, 1978.