GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Communication with Speech and Gestures: Applications of Recurrent Neural Networks to Robot Language Learning

Alexandre Antunes[1*], *Gabriella Pizzuto*[1*], *Angelo Cangelosi* [1]

[1] Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, United Kingdom
[*] These authors contributed equally to this work.

{alexandre.antunes, gabriella.pizzuto, angelo.cangelosi}@plymouth.ac.uk

Abstract

Recurrent neural networks have recently shown significant potential in different language applications, ranging from natural language processing to language modelling. This paper introduces a research effort to use such networks to develop and evaluate natural language acquisition on a humanoid robot. Here, the problem is twofold. First, the focus will be put on using the gesture-word combination stage observed in infants to transition from single to multi-word utterances. Secondly, research will be carried out in the domain of connecting action learning with language learning. In the former, the long-short term memory architecture will be implemented, whilst in the latter multiple time-scale recurrent neural networks will be used. This will allow for comparison between the two architectures, whilst highlighting the strengths and shortcomings of both with respect to the language learning problem. Here, the main research efforts, challenges and expected outcomes are described.

Index Terms: embodied language acquisition, grounded communication, deep learning models, cognitive robotics

1. Introduction

Humans have the inherent ability of learning to communicate via natural language. The benefits of using language is self-evident: through communication, humans exchange ideas, experiences and emotions. However, the understanding of the underlying mechanism in the human brain for acquiring language is still in its inception. Thus, this provides a non-trivial challenge to model on a machine.

Natural language acquisition in robots is still a major challenge, with multiple unresolved issues. Improving this skill in robots has the potential to improve current and future human-robot interaction. As roboticists, we can conceivably achieve this by taking inspiration from research in neuroscience and psychology and implementing it on anthropomorphic machines.

Research in child language acquisition has shown that gestures are the harbinger of language [1]. First, children start using deictic gestures, followed by representational gestures and single-element utterances. Additionally, gestures play an important role in the transition from single words to multi-words [2]. From developmental psychology research, three modalities of gestures are known: (1) equivalent (for example, using the representational gesture and saying ”bye”), (2) complementary (such as pointing to a cup and saying ”cup”) and (3) supplementary (by pointing to a flower and saying ”beautiful”). However, since developmental psychology experiments have illustrated correlations with respect to the infant's language measures in two modalities of gestures (in the complementary mode with vocabulary size and in the supplementary mode with the appearance of the two-word combination), only these two modes will be considered in this first part of the project.

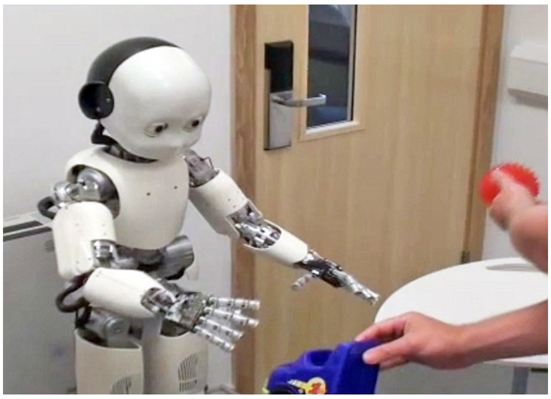

Figure 1: *The iCub humanoid robot used in this project. It is built similar to a 3.5 year old infant and encompasses numerous sensors for multimodal interaction. For the experiments that will be carried out, the robot interacts with the human through gestures and vocal utterances. The human contributes to the robot's language acquisition process by presenting the robot with objects to teach the iCub new vocabulary.*

As the vocabulary starts growing, children start using more complex sentences with an increased usage of verbs. Constructivist theories for language learning have dominated research in that field. Tomasello [3] defends the idea that young children first form verb-islands, learning each verb independently, then generalising these verbs to form grammar rules. A different theory, from Ninio [4], states that children start generalising as soon as they learn different verbs. From the research carried out by Pulvermueller & Fadiga [5], they suggest that language learning is intimately connected to action learning. These theories strongly encourage the idea that children learn not only the vocabulary, but also the structure of language, along with its meaning. In particular, a strong connection has been found between language and motor actions, suggesting a link between language learning and action learning.

The first part of the project focuses on using the Long-Short Term Memory (LSTM) [6] recurrent neural network architecture for implementing, on an iCub humanoid robot [7] (depicted in Figure 1), the gesture-word combination stage in infants before and when transitioning to multi-word utterances. This method will be compared to that of using Multiple Time-scale Recurrent Neural Networks (MTRNNs) [8], which will be developed during the second part of the project and focuses on the implementation and verification of constructivist theories for language learning.

The rest of the paper is organised as follows. In Section 2, the motivation behind and the framework for using LSTMs

10.21437/GLU.2017-1

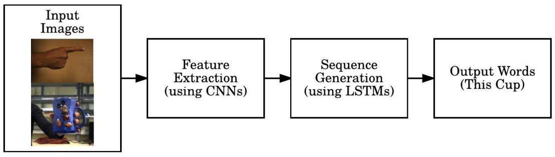

Figure 2: *The block diagram shows a high-level illustration of the architecture for the gesture-language model. The inputs to the network are the images captured by the iCub's cameras, in addition to the images used from the training database. For training, objects such as those used in the iCubworld dataset [12] and gestures similar to those used in the Pointing'04 ICPR Workshop [13] will be used. CNNs are used to realise the visual recognition problem, followed by LSTMs to produce variable-length output words or sentences. In this manner, the output sequence can scale as learning progresses and the vocabulary increases.*

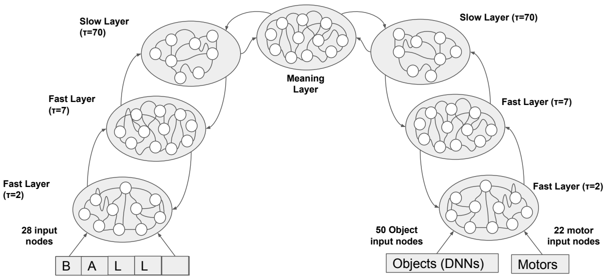

Figure 3: *The MTRNN architecture envisioned for language and motor control learning. The left branch will be dedicated to language, while the right branch will be dedicated to perception/actions. In the centre, the "meaning layer" is responsible for the unification of actions and sentences.*

and MTRNNs for natural language acquisition on a robot will be presented. The experiments that will be carried out to compare the two networks are explained in Section 3. Section 4 presents a discussion on the expected outcomes of this project, together with measurements that will be taken to counteract the challenges that might arise. Section 5 concludes this paper with some final remarks.

2. Recurrent Neural Networks for Language Learning

A vast array of methods can be used to realise the language acquisition task in humanoid robots. These range from Bayesian networks such as hidden Markov models [9] to artificial neural networks [10]. Due to the sequential nature of the data for the language task at hand, Recurrent Neural Networks (RNNs) are effective for fulfilling the task at hand. One of the earlier proposed models was by Elman [11]. However, his model suffers from an inability to model longer-distance dependencies. A deep learning architecture which addresses this problem is the LSTM, where a memory cell is used to encompass information which might have been omitted during the prediction stage. Moreover, with advancements in hardware acceleration technology, the feasibility of using these deep learning methods has become more attainable. Another variation to the RNN architecture is the MTRNN, based on a continuous time RNN that processes information on different timescales. The proposed models for using the LSTM and the MTRNN for this project are presented in Section 2.1 and Section 2.2.

2.1. LSTM

The LSTM network introduces a memory cell to the vanilla RNN with gradient-based learning methods to mitigate the issue of the vanishing gradient, where as a result of the numerous multiplications used during computation, the error cannot be propagated far before it vanishes or explodes. As a result of this modification, the LSTM has shown state-of-the-art in several applications such as humanoid robot control and perception [14] and image and video descriptions [15]. To the best of our knowledge, this architecture has not been previously used for language learning in robots. Thus, our contribution in this project will be to use this network for early language learning, through the use of gestures, on the iCub robot.

The design of the proposed system comprises a LSTM model for the language learning task and Convolutional Neural Networks (CNNs) for extracting features from the images obtained by the robot, as illustrated in Figure 2. The inputs to the network will be the words or phrases uttered by a human acting as the mother or caregiver, the visual input for the objects and gestures performed by a human considered for enhancing the language database of the robot. For the language learning task, the activation of the input layer will be from the words or phrases uttered by the human.

2.2. MTRNN

The MTRNN architecture has been previously used for language learning in the works of Jun Tani [16], [17], [18], where it has been shown that they are very successful in tackling the problem of language compositionality. In other works using the same network [19], [20] it was proved that it could also be very successful at learning the structure for motor actions, where robots would have to learn and execute complex actions. Many different works have recently used MTRNNs both for language and action learning, to very satisfactory levels of success.

The proposed neural network comprises two MTRNNs, with an extra layer connecting them at the control level. These MTRNNs will be dedicated to speech and actions respectively, and the union layer will be the meaning layer, as illustrated in Figure 3. A similar approach has been considered previously by Ogata [21], where the connection between language and behaviour MTRNNs was done through parametric bias neurons. For this project it is proposed instead to use an extra RNN layer for the merging.

The inputs to the network will come in two formats: a language input and a physical input. In the language branch a vector with the sentence uttered will be used. This will activate an input layer with 28 nodes (different letters of the alphabet, "space" and "stop"). The subsequent layer will identify words, and sentences will be processed at the third layer. The vector itself will be extracted by language processing methods, independent of the network itself.

For the physical branch of the network, two different types of input will be considered: the objects considered for the action, and the motor inputs from the robot joints. For the object input, a denoising autoencoder will be used to extract the features from the images received through the robot cameras, compressing them into 50 neurons.

For the robot joints, the different joint angles on the humanoid robot platform will be considered. A vector will be considered with 13 cells (3 angles for shoulder, 2 for elbow, 2 for wrist, 2 for thumb, 1 for pointer finger, 1 for middle finger, 1 for other two fingers and 1 neuron to identify the arm, left/right), plus one vector for head and torso movement with 9 cells (3 for head, 3 for torso, 2 for eyes, 1 for emotions). The total number of input nodes will be 72 nodes (50 for the objects, 13 for the arm and 9 for the head, torso and eyes).

3. Methods

The aim of these experiments is to use the architectures illustrated in Figure 2 and Figure 3 to improve the language acquisition mechanism of an iCub humanoid robot. In the experiments that will be carried out, the iCub will be presented with different objects and the human will interact with the robot during the experiment to contribute to the learning process. Experiments will be performed to compare the two architectures in an attempt to find the most adequate one for this task.

The first experiment will validate the applicability of the LSTM for early language learning with the use of gestures to increase the vocabulary, focusing mainly on the transition of single words to multi-words. The human and the robot will interact through the use of familiar and unfamiliar objects. The human will combine mostly deictic gestures, such as pointing (using index finger to point to particular person, object, location or event) at different objects, with vocal utterances.

For training, a dataset made up of the complementary and supplementary modalities of gesture with basic vocabulary, two-word and multi-word combinations will be fed to the network. During the experiment, a human will interact with the robot to teach it new words and phrases. The training process will be similar to that recorded during experiments carried out in developmental psychology experiment [2].

The second experiment focuses on testing the MTRNN. In order to train the network, a dataset will be built that encompasses sentences and their physical counterparts. A table with the different verbs to be used has already been designed, with 31 different verbs. These verbs include action-driven verbs (push, tap, hit), classification verbs (have, be) and emotion verbs (smile, laugh). This table of verbs will have to be converted into actions, where the iCub will execute the different actions with the joint angles and emotion controls being stored and associated with the corresponding verb. Some verbs do not correspond to any actions, but only to attributes of objects (i.e. ”the cube is red”, using the verb ”be”).

During the training phase, this dataset will be used to feed the MTRNN. We will perform several evaluations on the performance of the network in order to detect how it is learning the verbs and testing for generalisation capability in order to verify one of the two constructivist hypothesis.

Finally, an experiment will be designed for a more interactive approach. In this experiment, human participants will be asked to teach the iCub robot the verbs that correspond to the actions it is doing. At a second step, the human will make some requests to the iCub, with verbs that were previously taught to the robot, and it will execute the actions being requested.

4. Discussion

For the first part of the project, where the gesture-language system is implemented, the idea behind using LSTMs is an attempt to facilitate long-term learning as the robot's vocabulary dictionary increases. LSTMs have shown superiority for learning involving sequential inputs and outputs, when compared to other methods; thus, although in practice this method has never been implemented for a language learning task on a robotic platform, their success in image captioning and video description shows great promise for this project. In contrast to this, in robot vision, as a result of different viewpoints and illumination, the images captured by the cameras are not as repeatable as in using image datasets. Thus, in addition to pre-training the model on similar datasets, real images captured by the robot's cameras will also be integrated into training the CNN prior to entering the LSTM stage, to improve the performance of the LSTM model.

The MTRNN proposed could present some difficulties. As was discussed previously in Section 2.2, similar approaches have been tried in order to merge language and action learning. In many cases, these approaches failed due to the complexity of the network, a problem usually solved by keeping the different networks separated and training them individually. Given the scope of the proposed MTRNN, it is likely such issues will present themselves, which could be minimised with a large enough dataset. Possible alternatives involve the training of each branch separately, followed by the training of the meaning layer with input from both branches.

5. Conclusions

With technology moving forward at an unprecedented rate, having biologically-inspired neural networks on artificial machines has become more achievable. In summary, the main goal of this project is to take inspiration from psychology and neuroscience experiments to improve the iCub's language ability through the integration of gestures with the vocal utterances, in a similar manner to how children develop language skills. Additionally, the humanoid robot will also build language compositionality whilst integrating action learning. Our project plans on comparing two of the most popular architectures of RNNs, the LSTM and the MTRNN model, to fulfill the requirements of a language learning task on a humanoid. Although this is a challenging endeavour, the superiority of deep learning methods in pattern recognition and machine learning scenarios shows great promise for the task at hand. This project aims to address some of the open robotics challenges such as improved language learning capabilities, the integration of gestures in human-robot interaction and unite the action and language learning simultaneously, all of which are still in their infancy when it comes to robotic platforms.

6. Acknowledgements

This work is funded by the H2020 Marie Skodowska-Curie Innovative Training Networks through the DCOMM (H2020-MSCA-ITN-2015-676063) and APRIL (H2020-MSCA-ITN-2015-674868) projects.

7. References

[1] C. Butcher and S. Goldin-Meadow, *Gesture and the transition from one- to two-word speech: when hand and mouth come together*. Cambridge University Press, 2000, p. 235258.

[2] O. Capirci, J. M. Iverson, E. Pizzuto, and V. Volterra, “Gestures and words during the transition to two-word speech,” *Journal of Child Language*, vol. 23, no. 3, p. 645673, 1996.

[3] M. Tomasello, *Constructing a language*. Harvard University Press, 2009.

[4] A. Ninio, "No verb is an island: negative evidence on the verb island hypothesis," *Psychology of Language and Communication*, vol. 7, no. 1, 2003.

[5] F. Pulvermüller and L. Fadiga, "Active perception: sensorimotor circuits as a cortical basis for language," *Nature Reviews Neuroscience*, vol. 11, no. 5, pp. 351–360, May 2010. [Online]. Available: http://www.nature.com/nrn/journal/v11/n5/abs/nrn2811.html

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[7] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The iCub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125–1134, 2010.

[8] W. Hinoshita, H. Arie, J. Tani, T. Ogata, and H. Okuno, *Recognition and generation of sentences through self-organizing linguistic hierarchy using MTRNN*, part 3 ed., ser. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6098 LNAI, no. PART 3, pp. 42–51.

[9] O. Al-Dakkak and Y. Harba, "Vocal commands to a robot by an isolated words recognition system using hmm," in *2006 2nd International Conference on Information Communication Technologies*, vol. 1, 2006, pp. 1219–1224.

[10] A. Cangelosi, E. Hourdakis, and V. Tikhanoff, "Language acquisition and symbol grounding transfer with neural networks and cognitive robots," in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, 2006, pp. 1576–1582.

[11] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[12] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale, "Object identification from few examples by improving the invariance of a deep convolutional neural network," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RJS International Conference on (to appear)*, 2016.

[13] M. B. Holte and M. Stoerring, "Pointing and command gestures under mixed illumination conditions: video sequence dataset," 2004.

[14] M. Stollenga, "Advances in humanoid control and perception," Ph.D. dissertation, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), 2016.

[15] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015.

[16] S. Nishide, T. Ogata, J. Tani, T. Takahashi, K. Komatani, and H. G. Okuno, "Motion generation based on reliable predictability using self-organized object features," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 18-22, 2010, Taipei, Taiwan*, 2010, pp. 3453–3458. [Online]. Available: http://dx.doi.org/10.1109/IROS.2010.5652609

[17] W. Hinoshita, H. Arie, J. Tani, H. G. Okuno, and T. Ogata, "Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network," *Neural Networks*, vol. 24, no. 4, pp. 311–320, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2010.12.006

[18] G. Park and J. Tani, "Development of compositional and contextual communication of robots by using the multiple timescales dynamic neural network," in *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL-EpiRob, Providence, RI, USA, August 13-16, 2015*, 2015, pp. 176–181. [Online]. Available: http://dx.doi.org/10.1109/DEVLRN.2015.7346137

[19] S. Nishide, T. Nakagawa, T. Ogata, J. Tani, T. Takahashi, and H. G. Okuno, "Modeling tool-body assimilation using second-order recurrent neural network," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, October 11-15, 2009, St. Louis, MO, USA*, 2009, pp. 5376–5381. [Online]. Available: http://dx.doi.org/10.1109/IROS.2009.5354655

[20] S. Jeong, Y. Park, R. Mallipeddi, J. Tani, and M. Lee, "Goal-oriented behavior sequence generation based on semantic commands using multiple timescales recurrent neural network with initial state correction," *Neurocomputing*, vol. 129, pp. 67–77, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2013.03.050

[21] T. Ogata and H. G. Okuno, "Integration of behaviors and languages with a hierarchal structure self-organized in a neurodynamical model," in *Robotic Intelligence In Informationally Structured Space (RiiSS), 2013 IEEE Workshop on*. IEEE, 2013, pp. 89–95.