Sentence Embeddings with Sensorimotor Embodiment

*Junpei Zhong (Waseda University and Plymouth University), Angelo Cangelosi (Plymouth University), Tetsuya Ogata (Waseda University)

1. Introduction

Extracting reasonable and efficient features from the input data is one of the essential requirements in machine learning. Recently there have been quite a few research activities about "word embeddings" and "sentence embeddings". For instance, the word2vec model [3] learns conceptual relationships between words when searching the context of the surrounding words. On top of this, a deep recurrent neural network (DRNN) is often utilised to further learn the meaning of a sentence with the vectorized representation of words. However, most of these training methods need a large size of data sets. In this report we use a humanoid robot as a testing platform for our proposed architecture to reduce the need for large training sets, because we argue that the motor data can be used as an association of the text contexts (mostly as verbs), which is similar as the language grounding process [1]. Thus the aim of this report is to examine the proposed method using DRNN based on multiple time-scale recurrent neural network (MTRNN) [4] to ground the language meanings with motor actions executed.

2. The Model

This proposed model is an alternative of the previous experiment [5]; the vocal commands are pre-trained with the word2vec algorithm (in skip-gram) to produce a distributed vector representation of words, so that language modality has a more unified dense structure similar as the vision and motor inputs. On top of that, we used an MTRNN to generalise the embedded verbs and nouns with SOM pre-training (Fig. 1). The inputs to the MTRNN correspond to the language command inputs, the visual inputs as well as the proprioceptive inputs. The neurons in the MTRNN form three layers: an input-output layer and two context layers called "context fast" and "context slow" neurons. With distinct time constants τ, these neurons have different speeds of the adaptation.

The Eq. 1 shows the current membrane potential status of a neuron:

$$\tau_i u_{i,t}^{'} = -u_{i,t} + \sum_j w_{i,j} x_{j,t} \qquad (1)$$

where $u_{i,t}$ is the membrane potential, $x_{j,t}$ is the activity of j-th neuron at t-th time-step, $w_{i,j}$ represents the synaptic weight and τ is the time scale parameter which determines the decay rate of this neuron.

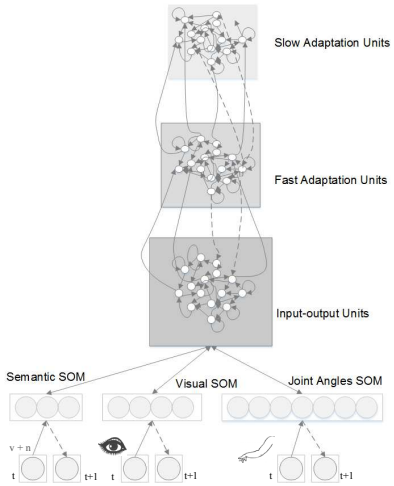


Fig.1: Learning Architecture

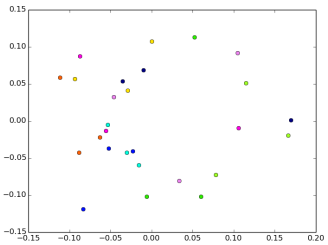


Fig.2: Sentence2vec Training with a Small Corpus The markers with the same colour, depicting that these sentences have the same meaning. were not clustered nearby.

3. Case Studies

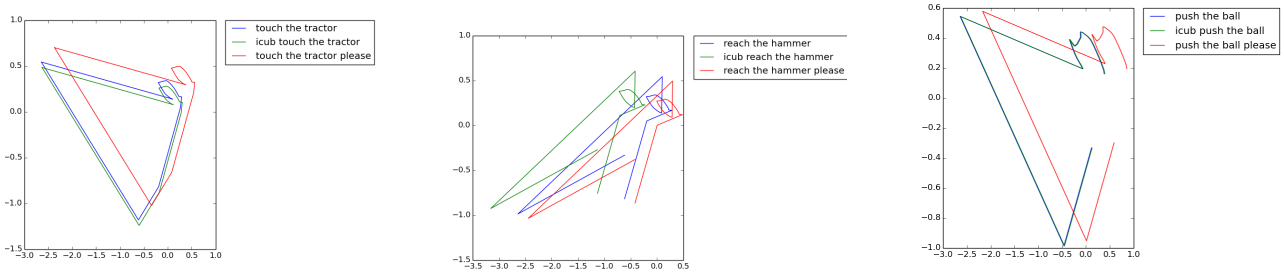
We used an iCub robot [2] to do the object manipulation experiments similar as [5].

3·1 Conceptualisation of Embodied Commands

The objective of this experiment was to examine the embedding multi-modal representations in the MTRNN. Three verbs and three nouns were included for the training sentences(Verbs: Touch, Reach, Push (denoted as *v.*); Nouns: Tractor, Hammer, Ball (denoted as *n.*)). In addition to these 9 combinations, some"noised" words were added in each sentence.

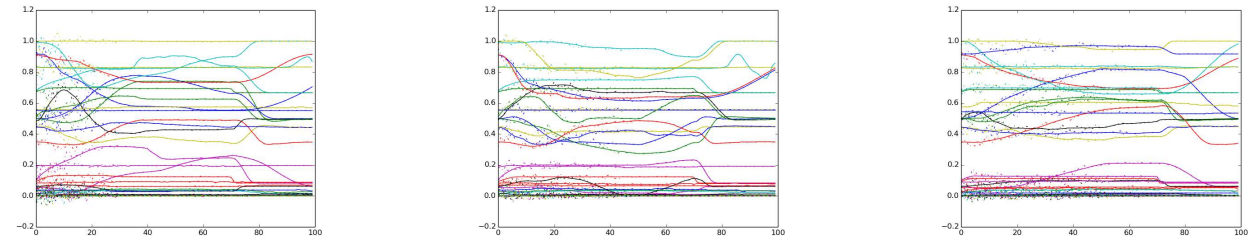
1. icub, *v.* + *n.*
2. *v.* + *n.*
3. *v.* + *n.*, please.

These commands had similar meanings. But



(a) Touch the Tractor (b) Reach the Hammer (c) Push the Ball

Fig.3: Context Fast Neurons with Noised Motor Commands



(a) Generation: touch the Tractor (b) Generation: reach the Hammer (c) Generation: push the Ball

Fig.4: Trajectories Generation based on Distributed Inputs

with sentence2vec they were difficult to be understood when only a small size data-set was available, which can be concluded from the PCA demonstrated vectors by the sentence2vec training in Fig. 2. But after MTRNN training with motor actions, the context slow neuron's activities are shown in Fig. 3. Similar verbs and nouns resulted in a similar neural dynamic on the context fast layers, although the input representation for the MTRNN significantly differed.

3·2 Generation of Motor Actions based on Noised Commands

In this experiment, the output of motor actions given only the language commands were examined. The comparisons between the desired output and the actual output for a few motor action samples are shown in Fig. 4, in which we can see that the actual output of motor action converged to the desired one after a few time-steps.

4. Conclusion

In this report we proposed a deep recurrent neural model based on multiple time-scale recurrent neural network (MTRNN) for learning distributed features of commands for service robots to reduce the big training data-sets for language understanding. During two studies, we discovered that the multi-modal inputs can be utilised for learning language commands. As a result, the noised commands were filtered and its most of the verb information were extracted. Based on this results, we shed light on the language corpus learning for robots should be associated with the action executions at the first stage.

Acknowledgement

This research was partially supported by unit for "Frontier of Embodiment Informatics: ICT and Robotics", Top Global University Project of Waseda University.

References

[1] A. Cangelosi. "Grounding language in action and perception: from cognitive agents to humanoid robots". In: *Phys Life Rev* 7.2 (2010), pp. 139–151.

[2] G. Metta et al. "The iCub humanoid robot: an open platform for research in embodied cognition". In: *Proceedings of the 8th workshop on performance metrics for intelligent systems*. ACM. 2008, pp. 50–56.

[3] T. Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[4] Y. Yamashita and J. Tani. "Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment". In: *PLoS Comput Biol* 4.11 (2008), e1000220.

[5] J. Zhong et al. "Sensorimotor Input as a Language Generalisation Tool: A Neurorobotics Model for Generation and Generalisation of Noun-Verb Combinations with Sensorimotor Inputs". In: *arXiv preprint arXiv:1605.03261* (2016).