In: Proceedings of the 23rd International Conference on Artificial Neural Networks (ICANN 2013),
LNCS 8131, pp. 216-223, Springer Heidelberg. Sofia, BG, September 10-13, 2013.

Embodied Language Understanding with a Multiple Timescale Recurrent Neural Network

Stefan Heinrich, Cornelius Weber, and Stefan Wermter

University of Hamburg, Department of Informatics, Knowledge Technology
Vogt-Kölln-Straße 30, D - 22527 Hamburg, Germany
{heinrich,weber,wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM/

Abstract. How the human brain understands natural language and what we can learn for intelligent systems is open research. Recently, researchers claimed that language is embodied in most – if not all – sensory and sensorimotor modalities and that the brain's architecture favours the emergence of language. In this paper we investigate the characteristics of such an architecture and propose a model based on the Multiple Timescale Recurrent Neural Network, extended by embodied visual perception. We show that such an architecture can learn the meaning of utterances with respect to visual perception and that it can produce verbal utterances that correctly describe previously unknown scenes.

Keywords: Embodied Language, MTRNN, Language Acquisition

1 Introduction

Natural language is the cognitive capability that clearly distinguishes humans from other living beings and often is called the key to intelligence. In the past researchers have contributed valuable models to explain the binding of language to experience, but also to ground language in embodied perception and action based on recent neuroscientific data and hypotheses [3, 6]. In addition early models captured the fusion of language and multi-modal perceptions or aimed at bridging the gap between formal linguistics and bio-inspired systems [15, 16].

However, due to the vast complexity of language, some models rely on well-understood Chomskyan formal theories, which are difficult to maintain in the light of recent neuroscientific findings, e.g. of non-infinite-recursive mechanisms and the evident involvement of various – if not all – functional areas in the human brain in language [1, 2, 14]. Other integrating or constructive models are constrained to single words, neglecting the temporal aspect of language [10].

In a recent study Hinoshita et al. claimed that for human language acquisition just an "appropriate" architecture is sufficient and provided a model based on a *Multiple Timescale Recurrent Neural Network* (MTRNN) [11]. They found that such a system composes language hierarchically in a self-organised way, if the architecture includes dynamic interaction of components with different characteristics, e.g. information processing on different timescales. Although the model was reproducing learned symbolic sentences quite well, generalisation was not possible, because the generation of sentences was initiated by the internal state of some neurons, which had to be trained individually for every sentence.

In this paper we incorporate embodied perception based on real world data in an MTRNN model and show that such a novel system is able to generalise to completely new situations by recomposing learned elements, and also self-organises toward the meaning of the learned verbal utterances.

2 Extended MTRNN Model

For our proposed model we employ the MTRNN to process verbal utterances over time [19], extended by several feed-forward layers to integrate embodied perceptions during the processing of utterances. The MTRNN part is composed of an *Input- and Output* layer (IO) and two context layers called *Context fast* (Cf) and *Context slow* (Cs). In general, the MTRNN is an extended *Elman Recurrent Neural Network* (ERNN) on the one hand and a special case of the *Plausibility Recurrent Neural Network* (PRNN) on the other hand [5, 18]. In contrast to the ERNN, the MTRNN allows for full connectivity of neurons to all neurons of the same and of adjacent layers, and introduces a mechanism forcing neurons in the context layers to process information with different timescales. Compared with the PRNN it restricts this concept of *hysteresis* to an increasing slowness from the first to the last layer and also restricts the architecture to one horizontal set of layers. Our extension part consists of an *Embodied Input* layer (EI), an *Embodied Fusion* layer (EF), and an *Embodied Controlling* layer (EC). Fig. 1 provides an overview of our architecture.

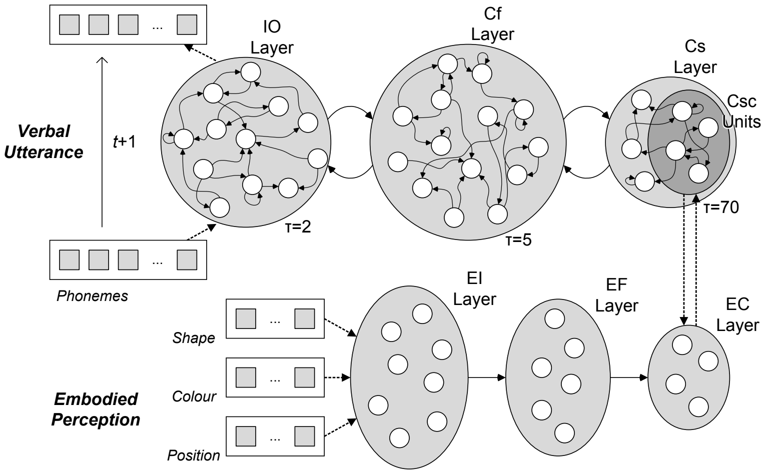


Fig. 1. Architecture of a Multiple Timescale Recurrent Neural Network extended by embodied perception from the scene. A sequence of phonemes (utterance) is processed over time, while the perceived situated information is constantly present.

During learning of the system the MTRNN is trained with verbal utterances and self-organises the neural activity and thereby the internal state values of some of the neurons in the Cs layer (so called *Context Controlling* units (Csc)). These self-organised values are then transferred to the EC layer and associated with the present embodied perception. For training we use an adaptive mechanism based on the *resilient propagation* algorithm [8]. During testing, the system approximates EC values from the visual perception input that are transferred to the Csc units, which in turn initiate the generation of a corresponding verbal utterance.

A full formal description of the MTRNN architecture can be found in the work of Yamashita and Tani [19]. In our model the MTRNN is specified by timescale values of $\tau = 2$, $\tau = 5$, and $\tau = 70$ for the IO, Cf, and Cs layers respectively, based on previous studies [11, 19] and preliminary experiments (not included), which show that these settings work best for the language learning scenario. For the IO layer we employ a soft-max function, while for the neurons in the remaining layers we use the following modified logistic transfer function:

$$f(x) = \frac{1.7159}{1 + \exp(-x \cdot 0.92)} - 0.35795 \quad . \tag{1}$$

The function is modulated in slope and range to capture the characteristics of the synchronic transfer function that has been proposed by LeCun for faster convergence in association tasks [13]. As error function on the IO layer we use the *Kullback–Leibler divergence*:

$$E(W) = \sum_{t} \sum_{i \in I_{\mathrm{IO}}} d_{t,i} \cdot \log\left(\frac{d_{t,i}}{y_{t,i}}\right) \quad , \tag{2}$$

where W represents the weight matrix, y denotes the output of neuron i at time step t, and d identifies the desired activity.

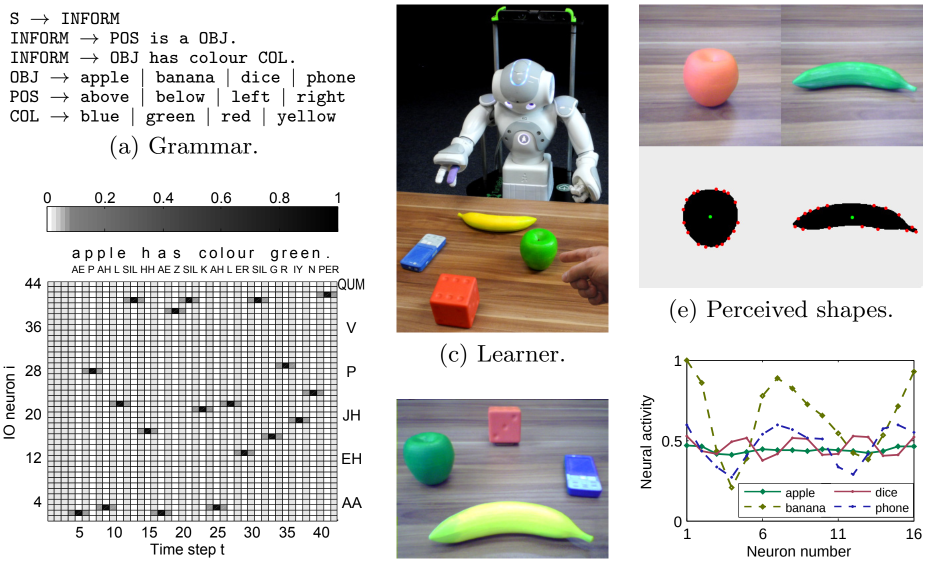
3 Scenario

Our scenario for this model is the interaction between a human teacher and a robotic learner, which is supposed to learn language from scratch by grounding speech acts in its embodied experience, but also is supposed to use its learned language to describe novel situations. The robot is placed in a scene and receives an utterance from the teacher, who describes the scene, e.g. "THE APPLE HAS COLOUR GREEN". The system should learn, in a self-organised way, how to bind the visual scene information with this verbal expression to be able to describe another scene like "THE BANANA HAS COLOUR GREEN" correctly. The focus of this study is on generalisation using possibly learned components.

To control our setup, all verbal utterances stem from a small symbolic grammar as presented in Fig. 2a. However, every symbolic sentence is transformed into a phonetic utterance based on phonemes from the ARPAbet and four additional signs to express pauses and intonations in propositions, exclamations, and questions: $\Sigma = \{$'AA', ..., 'ZH'$\} \cup \{$'SIL', 'PER', 'EXM', 'QUM'$\}$, with size $|\Sigma| = 44$.

To encode an utterance $u = (p_1, \ldots, p_l)$ into neural activation over time, we adapted the encoding scheme suggested by Hinoshita et al. [11], but we use a phoneme-based instead of a symbol-based representation: The occurrence of a phoneme p_k is represented by a spike-like neural activity of a specific neuron at relative time step r. In addition, some activity is spread backward in time (rising phase) and some activity is spread forward in time (falling phase) represented as a Gaussian over the interval $[r - \omega/2, \ldots, r - 1, r, r + 1, \ldots, r + \omega/2]$. All activities of spike-like peaks are normalised by the soft-max function for every absolute time step. A detailed description can be found in [11]. For our scenario we set the constants accordingly to $\mu = 4$, $\omega = 4$, $\sigma^2 = 0.3$, and $\upsilon = 2$. The ideal neural activation for an encoded sample utterance is visualised in Fig. 2b.

To encode the visual shape perception into sustained neural activity, we aimed at capturing the salient features of the objects in the field of view, inspired by saccadic eye movements of humans [9]. On an image taken by the NAO robot we employ the mean shift algorithm for segmentation [4], and the Canny edge detection as well as the contour finder for object discrimination. Subsequently, we calculate the centre of mass and 16 distances to salient points around the contour. Finally, we scale the distances by the square root of the object's area and order them clockwise – starting with the largest – to determine the characteristic shape, which is scale and rotation invariant. Fig. 2e provides two example results of this process and Fig. 2f visualises typical characteristics for all employed object shapes (scaled to $[0, 1]$). Encoding of the perceived colour is realised by averaging the three R, G, and B values of the shape, while the perceived position is encoded by the two values of the centroid coordinate in the field of view.



```
S → INFORM
INFORM → POS is a OBJ.
INFORM → OBJ has colour COL.
OBJ → apple | banana | dice | phone
POS → above | below | left | right
COL → blue | green | red | yellow
```

(a) Grammar.

(b) Encoded utterance.

(c) Learner.

(d) Learner's view.

(e) Perceived shapes.

(f) Shape representation.

Fig. 2. Representations and scenario of language learning in human-robot interaction.

4 Evaluation and Analysis

To test and analyse our model, we collected a data set consisting of all possible scenes and their respective verbal description. From the grammar we obtained 32 different combinations, which we set up as scenes and in turn used for collecting different examples. The corresponding verbal utterances were reasonably complex sequences with a length of 32 to 46 time steps (compare Fig. 2b). Subsequently, we ran a series of experiments for which we carefully, but randomly divided the data into a training set and a test set (50:50) – making sure that every scene is included only in one of these sets – and initialised a network. For every setup we repeated this process 50 times with different random seeds. The parameters of the network were mostly chosen based on the experience in [11]: We used $|IO| = 44$ and $|EC| = 21$ constrained by the input representations, but varied the sizes of Cf, Cs and EF to test for robustness. The size of EC depends on and is equal to the size of Csc, which we determined with $|Csc| = \lceil |Cs|/2 \rceil$. In addition, we used a feedback rate $\varphi = 0.1$ and initialised the weights in the interval $[-0.025, 0.025]$ and the initial Csc in the interval $[-0.01, 0.01]$.

4.1 Generalisation

To be able to compare the generalisation capabilities, we use the standard measure F_1-score determined by precision and recall, and defined as follows:

$$p_{\text{precision}} = \frac{tp}{tp + fp}\ ,\ p_{\text{recall}} = \frac{tp}{tp + fn}\ ,\ F_1\text{-score} = 2 \cdot \frac{p_{\text{precision}} \cdot p_{\text{recall}}}{p_{\text{precision}} + p_{\text{recall}}}\ , \quad (3)$$

where we specify all correct and matching sentences as tp (true positives), all correct but not matching sentences as fp (false positives), and strictly all incorrect sentences as fn (false negatives).

The results in Tab. 1 show that the system can be trained perfectly in most cases, and also produces correct utterances for new scenes on a moderate level: For a suitable parameter setting, networks reach an F_1-score of up to 1.0 on the training set and 0.545 on the test set, with an average over all random seeds of 0.999 on the training set and 0.136 on the test set.

Table 1. Comparison of F_1-score for different network dimensions.

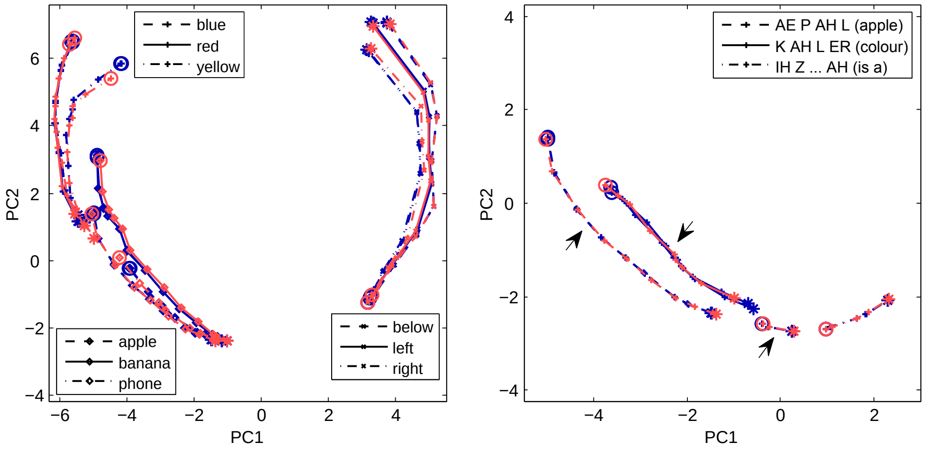
$\|Cf\|/\|Cs\|$ $\|EF\|$	40/11 8	40/11 16	40/11 24	80/23 8	80/23 16	80/23 24	160/47 8	160/47 16	160/47 24
training set best	1.000	1.000	1.000	1.000	1.000	**1.000**	1.000	1.000	1.000
test set best	0.316	0.222	0.316	0.316	0.400	**0.545**	0.316	0.476	0.400
training set average	0.889	0.904	0.908	0.950	0.999	**0.999**	0.979	0.994	0.995
test set average	0.059	0.049	0.043	0.091	0.136	**0.136**	0.096	0.091	0.123

Note that due to the random selection, in several cases the system had to describe a scene, for which it had not seen any aspect (shape, colour, or position) before. This was intended to keep the scenario realistic and observe the effects.

4.2 Network Behaviour

To provide a better understanding of the system, we analysed the neural activity of the Cf layer for the trained networks. We aimed to test, whether this layer had organised itself to represent the words in the utterances (compare [11]). Using *principle component analysis* (PCA) we reduced the dimensionality to visualise trajectories over time for specific words. The starting and the end point of the trajectory were defined as the first highest activity for the first phoneme and the last highest activity for the last phoneme of the word in the IO layer.

The results reveal several characteristics (see Fig. 3 for the trajectories of a typical network): First, the neural activity in the Cf layer is nearly identical for the same words from trained utterances. Second, the same words from untrained utterances have a quite similar activity pattern. Third, words of the same type (shape, colour, or position words) have a very related activity pattern. From the data we can observe, that the networks self-organise patterns for words about shapes, colours, and positions. Forth, words with similar phonetic representation have different activities, if the type of the word is different. Low correlation was found of activity for phonetically similar but semantically different words.



(a) Words of similar type.

(b) Words with similar phonetics.

Fig. 3. Comparison of neural activation in the Cf layer for different words. The dimensionality has been reduced from $|Cf|$ to two dimensions (PC1 and PC2) and the beginning ($*$) as well as the end ($\circ$) of the words have been marked. The dark/blue lines represent words from utterances of the training set and the bright/red lines show words from utterances of the test set. Arrows indicate the same phoneme "AH".

In addition, we found the tendency that the activation of a word primes the activation of other grammatically related words. In terms of trajectories it can be observed that the end point of the word "COLOUR" is close to the starting point of all colour words, and the end point of a position word is close to the starting point of "IS A ..." (compare Fig. 3a and b).

5 Discussion

The combination of visual perception and an architecture that includes different timescales in processing verbal sequences provides a system that self-organises towards the meaning of learned utterances in a real world scenario. Our experiments have shown that such a system apparently is able to understand verbal utterances and describe novel scenes with the correct corresponding verbal utterances. The analysis revealed that novel scenes are described by recomposing the correct words, which have been grounded in the perception of different shapes, colours, or positions.

For some incorrect sentences we observed both cases: Minor substitution errors in terms of a single wrong phoneme or a pause that was too long ("SIL SIL" instead of "SIL"), as well as no meaningful phoneme chains at all. In the first case, listening humans would presumably consider this a normal inaccuracy and automatically correct the recognition. The second case clearly shows that generalisation was sometimes difficult. It is open to clarify, whether this degree of difficulty is inherent, e.g. if the error rate is comparable to certain learning stages in young children during early language learning [12].

During training of the system, we found that the connection weights from the Cf to the Cs layer as well as from the IO to the Cf layer converged towards zero in many cases. This means that the highly dynamic networks organised themselves towards a directed flow of information from the context to the phonetic output instead of a mutual exchange of information. This is plausible in the light of neuroscientific evidence [10], but for future experiments implies that the MTRNN architecture might already be more complex than necessary and should be tested with less initial connectivity. In addition we found that incorporating an adaptive training mechanism and a novel transfer function already allowed to reduce the complexity of the training itself.

Parameter exploration has shown that for this architecture a good balance of the number of neurons and the number of training samples is important. This is in line with experience from associator networks [13], but less desirable. Further investigations should include the consideration of architectures that are dynamic in connectivity as well as in size. In addition the architectures should be tested with more complex scenes and verbal descriptions, including interrelations of multiple objects and embodied experience of a broader set of real world situations.

In conclusion, our study supports that the embodiment of language in perception and a hierarchical structure with different timescales are important aspects of an appropriate architecture for language. For such an architecture a feasible constraint can be our mostly feedforward but compositional structure, also suggested for the (visual) cortex [7]. In the future we need to further refine the architectural characteristics to identify the most important building blocks for natural language processing. The understanding of the brain's architecture for language can explain the humans' most important cognitive capability, but also can inform future software frameworks for service robots that should interact with and understand humans.

Acknowledgments. The authors would like to thank A. Saleh and H. Vöcking for support with the real world data acquisition as well as J. Dávila-Chacón and S. Magg for critical and inspiring discussions.

References

1. Barsalou, L.W.: Grounded cognition. Annu. Rev. Psychol. 59, 617–645 (2008)
2. Borghi, A.M., Gianelli, C., Scorolli, C.: Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics. Frontiers in Neurorobotics 4(3), 8 (2010)
3. Cangelosi, A.: Grounding language in action and perception: From cognitive agents to humanoid robots. Physics of Life Reviews 7(2), 139–151 (2010)
4. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. on Pattern Anal. and Mach. Intell. 24(5), 603–619 (2002)
5. Elman, J.L.: Finding structure in time. Cognitive Science 14(2), 179–211 (1990)
6. Frank, S.L.: Strong systematicity in sentence processing by an echo state network. In: Proc. 16th International Conference on Artificial Neural Networks (ICANN 2006). LNCS 4131, pp. 505–514. Springer Ber. Hdb., Athen, GR (Sep. 2006)
7. Friston, K.: A theory of cortical responses. Philosophical Transactions of the Royal Society B 360, 815–836 (2005)
8. Heinrich, S., Weber, C., Wermter, S.: Adaptive learning of linguistic hierarchy in a multiple timescale recurrent neural network. In: Proc. 22nd International Conference on Artificial Neural Networks (ICANN 2012). LNCS 7552, pp. 555–562. Springer Ber. Hdb., Lausanne, CH (Sep. 2012)
9. Henderson, J.M.: Human gaze control during real-world scene perception. Trends in Cognitive Sciences 7(11), 498–504 (2003)
10. Hickok, G., Poeppel, D.: The cortical organization of speech processing. Nature Reviews Neuroscience 8(5), 393–402 (2007)
11. Hinoshita, W., Arie, H., Tani, J., Okuno, H.G., Ogata, T.: Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. Neural Networks 24(4), 311–320 (2011)
12. Karmiloff, K., Karmiloff-Smith, A.: Pathways to language: From fetus to adolescent. Harvard University Press (2002)
13. LeCun, Y., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. In: Neural Networks – Tricks of the Trade, LNCS 1524, pp. 9–50. Springer Ber. Hdb. (1998)
14. Pulvermüller, F., Fadiga, L.: Active perception: sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11, 351–360 (2010)
15. Rohde, D.L.T., Plaut, D.C.: Connectionist models of language processing. Cognitive Studies 10(1), 10–28 (2003)
16. Roy, D.K., Pentland, A.P.: Learning words from sights and sounds: A computational model. Cognitive Science 26(1), 113–146 (2002)
17. Steels, L., Spranger, M., van Trijp, R., Höfer, S., Hild, M.: Emergent action language on real robots. In: Language Grounding in Robots, chap. 13, pp. 255–276. Springer New York (2012)
18. Wermter, S., Panchev, C., Arevian, G.: Hybrid neural plausibility networks for news agents. In: Proc. National Conference on Artificial Intelligence (AAAI-99). pp. 93–98. Orlando, US (Jul. 1999)
19. Yamashita, Y., Tani, J.: Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. PLoS Computational Biology 4(11), e1000220 (2008)