A connectionist model of acquisition of noun phrases with syntactic bootstrapping

Benjamin Fele
Faculty of Education, University of Ljubljana
Kardeljeva ploščad 16, 1000 Ljubljana
Email: bf9770@student.uni-lj.si

Martin Takáč
Centrum pre kognitívnu vedu FMFI UK
Mlynská dolina, 842 48 Bratislava
Email: takac@ii.fmph.uniba.sk

Abstract

We present a model of acquisition of noun phrases as descriptions of objects in a grounded setting. The model uses cross-situational learning to acquire mappings between words and sensory properties represented in self-organising maps. We also explore syntactic bootstrapping—using the syntactic knowledge to help meaning acquisition and vice versa. The model can reliably reconstruct inputs and is applicable in real-world scenarios.

1 Introduction

The research reported here is a part of a larger project of modeling language acquisition in human-robot interactions in a simple world of geometrical shapes. A prerequisite for understanding sentences is to be able to understand *noun phrases* (NPs)—variable-length descriptions of objects on the scene typically consisting of a noun expressing the object's type and possibly one or more adjectives expressing its properties. In this paper we present a model of acquisition of NPs in a grounded interaction setting, wherein a visual description of an object and its properties (colour, size, position) is paired with a corresponding NP. Visually detected object properties are first internally represented in sensory maps that, during training, gradually learn to consistently represent each property by a particular pattern of activity. The task of grounded language learning is to acquire consistent mappings between activity patterns and words expressing the respective properties. The acquired mapping can then be used for language interpretation (the route from words to sensory map activities to object properties) and language production in object description (the route from object properties to map activities to words).

We model the sensory maps by self-organising maps (SOMs) (Ritter & Kohonen, 1989) and train the associations between words (language modality) and map neurons (visual modality) by Hebbian learning (Hebb, 2005) within cross-situational learning paradigm (Smith & Smith, 2012). The paradigm rests on the assumption that in a long-enough sample of word-meaning pairs spurious correlations will cancel each other out, while the true ones will remain. A combination of SOMs with Hebbian learning was used e.g. in Li et al. (2004); our model differs in using multiple SOMs and exploring syntactic bootstrapping (Fisher et al., 2010), in which the model acquires knowledge about syntactic roles of different words and at the same time using it to guide the word–meaning acquisition and subsequent language production/interpretation.

2 Model

We designed the system to generate word-phrases and visual inputs corresponding to four modalities: *type*: square, sphere, cylinder, cone, *size*: big, small, *colour*: red, green, blue, purple, black, white, and *position*: left, right, top, bottom. The model consists of four 16x16 self-organising maps, a language layer with 16 units (one for each word) and associative links connecting them (Figure 1).

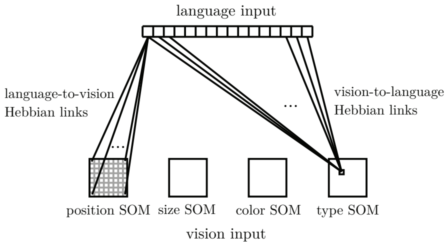

Fig. 1: A simplified schema of our model.

The associative links are directed: one set of them going from language to vision and the other going in the opposite direction. Each of the four self-organising maps takes input from a relevant part of the visual object descriptor, such as RGB values for colour, or centroid of the bounding box for a position. In addition, the model also makes use of syntactic bootstrapping for which it learns two probability tables: the first one stores probabilities of each word corresponding to any of four word categories (corresponding to the modalities). The second table is based on the first one and stores probabilities of a word at certain position within the phrase being from any of the four categories, also depending on the word-phrase length. Information from this table is then used to turn on relevant and turn off irrelevant Hebbian links when training. Information from the first table is used for a similar purpose when reconstructing inputs.

3 Method

The model was trained for 14000 iterations. A standard SOM training was followed by Hebbian training of links between normalised SOM activations and a language vector with active units for each of the words in the NP. The Hebbian rule included an additional factor—a probability of a word at a certain position in the NP having its category matching the modality of the SOM the association links are pointing to. This reflects the use of syntactic bootstrapping (Fisher et al., 2010) in humans.

At the end of each training iteration, all links were normalised depending on their direction. When we reconstruct words from SOM activities, we want the words to compete against each other, that is why weights of all the links from a SOM unit to all the words were normalised to sum to one. When reconstructing a SOM activity pattern for a word, we want the SOM units to compete, that is why all links from a particular word to a SOM were normalised to sum to 1.

To assess the performance of the trained model, we used a number of metrics, but most notably we reconstructed inputs and measured the quality of reconstruction. When reconstructing vision from language, the Euclidean distance between the reconstruction and a matching vision input was calculated. If that distance was the smallest in the set of 20 vision inputs, the reconstruction was counted as correct, from which we obtained a general accuracy of the vision reconstruction. When reconstructing language from vision input, we discretised the output of the network in relation to a certain threshold and then compared it to the original vector. That way, we again computed the accuracy of the reconstruction.

4 Results

Various metrics showed convergence during training of SOMs, Hebbian links and syntactic bootstrapping tables. Many metrics converged to a different degree depending on a category, but results of reconstruction in our tests yielded 91% accuracy for vision and 97% accuracy for language reconstruction. Syntactic bootstrapping, when used during training, only made a difference (sped up learning) during the first 1000 episodes. However, when used during reconstruction, the results were generally better for models trained for up to 8000 episodes. For models trained for longer the difference became negligible (see Figure 2).

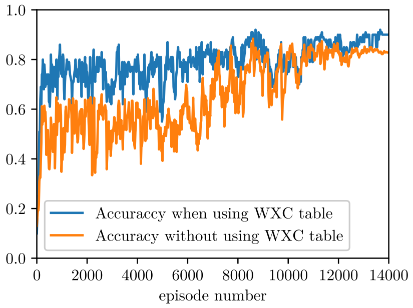

Fig. 2: Accuracy of vision reconstruction with and without using syntactic bootstrapping.

5 Conclusion

Although a convergence measured by prediction entropy differed for different categories, reconstruction accuracy was always above 90%. These differences can be attributed to differences in dimensions and distributions of input data. Since the reconstruction works well regardless of these differences, we conclude that our model works on wide array of inputs having different properties, which gives it a necessary robustness for real-world scenarios. Furthermore, syntactic bootstrapping showed improvement for shorter training, however the advantage disappeared if the training was long enough.

Acknowledgement

This research has been supported by the grant VEGA 1/0796/18 for Martin Takáč and ERASMUS+ scholarship for Benjamin Fele.

References

Fisher, C., Gertner, Y., Scott, R. M. & Yuan, S. (2010). Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149.

Hebb, D. O. (2005). *The Organization of Behavior: A Neuropsychological Theory*. Psychology Press.

Li, P., Farkas, I. & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, 17(8-9):1345–1362.

Ritter, H. & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254.

Smith, A. & Smith, K. (2012). Cross-situational learning. *Encyclopaedia of the Sciences of Learning*, 15:2029–2049.