REMOVING 'MIND-READING' FROM THE ITERATED LEARNING MODEL

S. F. WORGAN AND R. I. DAMPER

Information: Signals, Images, Systems (ISIS) Research Group
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, UK.

{sw205r|rid}@ecs.soton.ac.uk

The iterated learning model (ILM), in which a language comes about via communication pressures exerted over successive generations of agents, has attracted much attention in recent years. Its importance lies in the focus on cultural emergence as opposed to biological evolution. The ILM simplifies a compositional language as the compression of an object space, motivated by a poverty of stimulus—as not all objects in the space will be encountered by an individual in its lifetime. However, in the original ILM, every agent 'magically' has a complete understanding of the surrounding object space, which weakens the relevance to natural language evolution. In this paper, we define each agent's meaning space as an internal self-organising map, allowing it to remain personal and potentially unique. This strengthens the parallels to real language as the agent's omniscience and 'mind-reading' abilities that feature in the original ILM are removed. Additionally, this improvement motivates the compression of the language through a poverty of memory as well as a poverty of stimulus. Analysis of our new implementation shows maintenance of a compositional (structured) language. The effect of a (previously-implicit) generalisation parameter is also analysed; when each agent is able to generalise over a larger number of objects, a more stable compositional language emerges.

1. Introduction

Hypothesising that language is a system of compression driven to adjust itself so that it can be learned by the next generation is a relatively new approach in the field of linguistics. Several important simulations (Kirby & Hurford, 1997; Kirby, 2001, 2002; Brighton, 2002; Smith, Kirby, & Brighton, 2003) have illustrated its potential and provide an alternative to established innate accounts of language (Chomsky, 1975; Bever & Montalbetti, 2002; Hauser, Chomsky, & Fitch, 2002). Currently, existing versions of this iterated learning model (ILM) suffer from a number of shortcomings, highlighted by Smith (2005), Vogt (2005), Steels and Wellens (2006). This paper will address some of these while maintaining the positive features of the model.

In the classical ILM, an agent selects an object from its environment and produces a meaning-signal pair that is directly perceived by a listener. The pairing

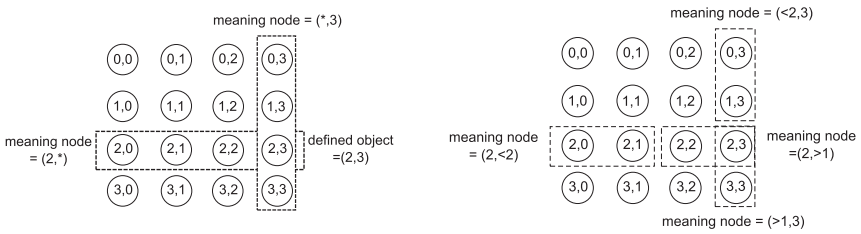
is formed through a weighted connection between a meaning node and a signal node, and is used to adjust the weighted connections between the meaning space and the signal space of the listening agent. In this way, a language evolves across a number of generations. If each agent is only given the associated signal for a small subset of possible objects, it is forced to generalise across the remaining object space, so promoting the formation of a stable compositional language.

2. Shortcomings of the Iterated Learning Approach

In the ILM, the agents' meaning space loosely represents the 'mind' of a language user. In many respects, however, this analogy breaks down, as each agent is created with a perfect knowledge of the surrounding object space, which is never found in reality. We need to consider the nature of the object space and the agents' ability to generalise across it. Also, a learning agent directly observes each meaning-signal pair, and this introduces an element of 'mind-reading', as the learner knows exactly what the adult teacher was thinking when it produced a signal. Obviously, this weakens the ILM's credentials as a simulation of cultural language evolution. Kirby (2002, p. 197) himself acknowledges this criticism, writing "the ready availability of signals with meanings neatly attached to them reduces the credibility of any results derived from these models", whereas Smith et al. (2003, p. 374) write: "This is obviously an oversimplification of the task facing language learners."

We aim to develop a new ILM to address these criticisms. Let the iterated learning approach yield a language, able to describe every object found in the object space, $\mathcal{N}$, through a process of compression, governed by a form of generalisation. This compression is possible by forming a compositional language, which describes common features of objects in the space. Figure 1(a) illustrates how a compositional meaning node is able to define partially a number of objects. In the original ILM, this is automatically determined by the number of values, V, in the object space, e.g., in Fig. 1(a) each compositional meaning node is able partially to define $V = 4$ objects. An implicit generalisation parameter γ then determines the proportion of these V values that each meaning node can generalise over: in Fig. 1(a), $\gamma = 1$. This parameter, ignored in previous work, impacts significantly on the structure of the final compositional language. To understand the role of the environment in the emergence of language, we need to consider what happens when the generalisation parameter γ is not equal to 1. Figure 1(b) shows the compression which results from halving the, now explicit, generalisation parameter. We see that 4 meaning nodes—rather than 2 as previously—are now required to specify the same number of object nodes (i.e., poorer generalisation). In this example, $\gamma = 0.25$ would correspond to a holistic, non-compositional language (i.e., no generalisation).

Having acknowledged the role of this (previously-implicit) generalisation parameter, we are now able to remove the 'mind-reading' abstraction from our



(a) $\gamma = 1$ (b) $\gamma = 0.5$

Figure 1. In an ILM, the object space is defined by the number of object values V in each of F dimensions. In this example, $F = 2$ and $V = 4$. In the original ILM in (a), the generalisation parameter γ, representing a proportion of object values, is implicitly set to 1. By varying γ as in (b), where $\gamma = 0.5$, we can vary the level of compression that each compositional meaning node can achieve.

simulations. To do this, we will define the agent's meaning space as a self-organising map (SOM) and γ as a radius around a selected object, removing the two criticisms of IL stated above. An agent no longer has complete and perfect knowledge of the object space, and this knowledge remains private so that each agent develops a different 'understanding' of its linguistic environment.

3. Self-organising maps and iterated learning

Self-organising maps (Kohonen, 1982) have previously been used to good effect to model emergent phonology (e.g., Guenter & Gjaja, 1996; Oudeyer, 2005; Worgan & Damper, 2007). In the present work, SOMs offer a way to model each agent's unique and private understanding of its environment. Our model is based on the neural network model of (Smith et al., 2003, Sect. 4.2.1), but with important differences motivated by the discussion of Section 2 and described explicitly in this section.

In this environment, an object can be defined as, e.g., $x_k = \{1, 2\}$, and in the meaning space as $m_j = \{1, 2\}$. Equivalently, it can be defined as the pair:

$$\begin{aligned} m'_j &= \{1, *\} \\ m'_{j+1} &= \{*, 2\} \end{aligned}$$

where $*$ represents a wildcard. In this example, m_j forms a holistic signal, as this individual meaning node is only capable of defining one object, whereas m'_j and m'_{j+1} together form a compositional signal, as features from the object space are defined by the two meaning nodes and can be combined to define an individual object. These feature definitions can then be used in other combinations to describe other objects. We will maintain this aspect of traditional IL by redefining generalisation as a variable radius around a perceived object.

The weightings on the connections between nodes of the meaning and signal spaces determine the mapping from meaning-to-signal and from signal-

to-meaning. The object space, $\mathcal{N}$, that each agent talks about is represented by a simple coordinate system and a subset of these coordinates is drawn from the object space according to a uniform probability distribution. Each object in turn is mapped directly to the appropriate meaning node in the agent's meaning space. The signals, l_i, are generated by mapping from this meaning space to the signal space, and are represented as characters from an alphabet, Σ as:

$$l_i = \{(s_1, s_2, \ldots, s_i, \ldots, s_l) : s_i \in \Sigma\,,\ 1 \leq l \leq l_{\text{max}}\} \tag{1}$$

from which it is clear that we need a sufficient number of signal nodes to express any of the nodes in the meaning space.

Formally, the object space is:

$$\begin{aligned} \mathcal{N} &= \{x_1, x_2, \ldots, x_k, \ldots, x_N\} \\ \text{with}\quad x_k &= \{(f_1, f_2, \ldots, f_i, \ldots, f_F) : 1 \leq f_i \leq V\} \end{aligned}$$

When required to produce an utterance, an agent will select an object x_k, and each node in the meaning space m_j competes to have the shortest euclidean distance from this point. Formally, if we define the closest node as $m(x_k)$ then:

$$m(x_k) = \arg\min_j \|x - m_j\|\,, \quad j = 1, 2, \cdots, l \tag{2}$$

The winning node is then moved closer to the selected point, better defining the object space as a whole. In addition, neighbouring nodes are moved somewhat closer to the object, allowing the network as a whole to represent the experienced object space. The extent to which these nodes move is determined by a gaussian function, $h_{j,k}$, centred around the selected object (Haykin, 1999, p. 449):

$$h_{j,k} = \exp\left(-\frac{d_{j,k}^2}{2\sigma^2}\right) \quad \text{with } \sigma \equiv \gamma \tag{3}$$

where $d_{j,k}$ is the distance between the winning neuron j and the excited neuron k.

To form a compositional signal, we build valid decomposition sets from the meaning space, governed by the generalisation parameter, γ. We can then define a set, $\mathcal{K}_k$, containing all of those meaning nodes which fall inside the radius around x_k. Formally:

$$\mathcal{K}_k = \{m_j\ :\ \|x_k - m_j\| \leq \gamma\} \tag{4}$$

Considering all possible decompositions in turn, the agent will pick the signal, with the highest combination of corresponding weight values according to:

$$g(\langle l_i \rangle) = \sum_{i=1}^{|\mathcal{S}|} \sum_{j=1}^{|\mathcal{K}_k|} \omega(K(x)_j) \cdot W_{K(x)_j N_{Si}} \tag{5}$$

which is similar to Smith et al.'s equation on p. 380, in that $\omega(K(x)_j)$ "... is a weighting function which gives the non-wildcard proportion of ..." $K(x)_j$, so favouring compositional meaning nodes.

All meaning and signal nodes that correspond to a possible decomposition of the object are activated, with activations a_{s_i} and a_{m_j}, respectively. If two active nodes are connected, the weight on that connection is increased. If there is a connection between an active node and an inactive node the weight is decreased. Weights between two inactive nodes remain unchanged. The learning displayed by this Hebbian network can be formalised as follows:

$$\triangle w_{ij} = \begin{cases} +1 & \text{iff } a_{s_i} = a_{m_j} = 1 \\ -1 & \text{iff } a_{s_i} \neq a_{m_j} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\triangle w_{ij}$ is the weight change at the intersection between s_i and m_j, $s_i \in N_S$ and $m_j \in N_M$.

While listening to each utterance, the weight values of the agent are adjusted—extending its knowledge of the current language. This hypothesis allows it to generalise to objects it has not encountered before, resulting in a meaningful expression. Therefore, a poverty of stimulus causes the language to generalise across an object space. Additionally, by having a limited number of nodes form the meaning space, the agent does not have an infinite memory resource to draw upon, forcing compression through limited memory as well as limited stimuli.

Using this model, we will vary γ in order to assess how this affects the stability, S, of the final compositional language:

$$S = \frac{S_c}{S_c + S_h} \tag{7}$$

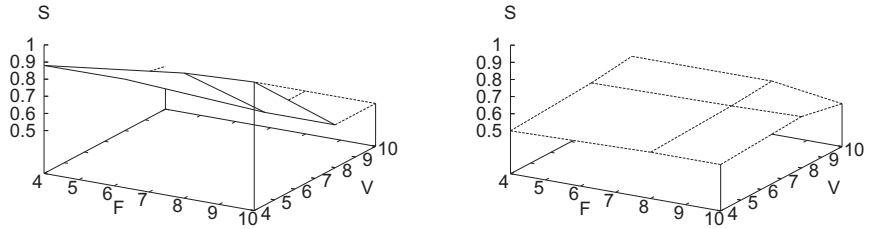
where S_c represents the proportion of compositional languages and S_h defines the proportion of holistic languages, which emerge over cultural time. The higher the value of S, the more likely is a compositional language to emerge—see Smith et al. (2003, p. 377).

In the new model, each agent's meaning space is undefined at birth (randomly initialised) and will need to learn the structure of the object space as each object is encountered. Consequently, the meaning space gradually comprehends the object space but also remains potentially unique to each agent, as a different subset of objects is encountered.

4. Results

We first ran the new SOM iterated learning model under the same conditions as the previous implementation, see Figure 2. As we can see from the results, compositional languages emerge ($S > 0.5$) under a similar set of circumstances

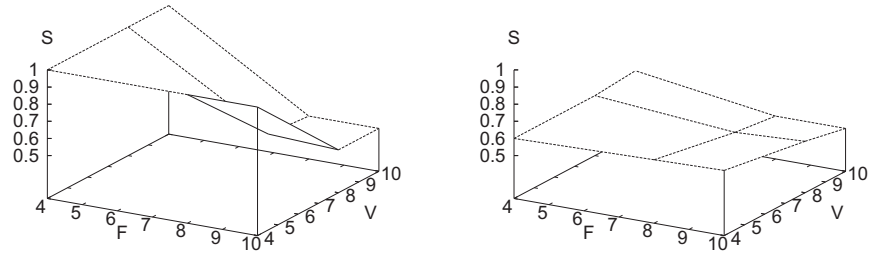
to Smith et al.'s (2003) previous implementation. Therefore, the requirements for a tight bottleneck and a structured meaning space remain in this implementation.



(a) 10% (b) 90%

Figure 2. Stability of the resulting languages, calculated according to equation 7, when each agent is exposed to some percentage of the object space (Smith et al.'s "bottleneck" parameter).

Next, we considered the effect of varying the generalisation parameter, γ, as shown in Figure 3. The higher the generalisation, the greater the stability, S, of the compositional language and, conversely, the lower the generalisation, the lower the stability. This highlights the importance of the previously implicit generalisation parameter on the final stability of the compositional language. Accordingly, a reasonable level of generalisation is required to enable cultural emergence.

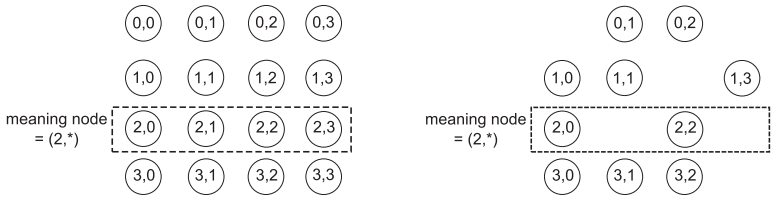


(a) $\gamma = 2$ (b) $\gamma = 0.5$

Figure 3. Stability of the resulting languages when each agent is exposed to 10% of the object space, with different degrees of generalisation: (a) $\gamma = 2$, (b) $\gamma = 0.5$. Here γ has been reformulated as a gaussian width, as shown in equations 3 and 4

Figure 4 shows how structuring the object space allows each meaning node to generalise over a greater number of objects, increasing the stability S. As we can see, the potential generalisation of each meaning node is not as effective as fewer objects are located in each generalisation area, the compositional meaning node can only generalise across two objects in the unstructured object space of

Fig. 4(b). This gives us greater insight into Smith et al. (2003)'s comparison of structured and unstructured meaning spaces. By considering these results in terms of γ we can see how these meaning spaces indirectly affect the level of potential generalisation.



(a) Structured space (b) Unstructured space

Figure 4. In a structured object space, each meaning node generalises over a greater number of objects.

5. Conclusions

In this paper, we have addressed some criticisms of the well-known iterated learning model of cultural language emergence, most notably the 'mind-reading' aspect of earlier ILM implementations. This was achieved using self-organising maps to model each agent's meaning space. The result is a closer analogy to real cognitive spaces. Specifically, the meaning spaces are limited in the amount of memory resource they have available, and are not omniscient. Rather they are private and unique to each agent. The SOM does not have a high enough capacity to completely define the agents' environment—forming a further motivation to generalise. We have made explicit the generalisation parameter that was previously implicit to earlier ILM's and demonstarted its role in promoting emergence of compositionality. As well as being unique to each individual, the learning displayed by the SOM demonstrates another property of real language learners: namely, change over time with each new encountered object.

These enhancements, or improvements, to the classical iterated learning framework are gained without compromising the essential tenets of the paradigm. As with the classical framework, stable, compositional languages emerge through use (i.e., inter-agent communication related to structured object spaces) over cultural time. Further, the poverty of stimulus encountered both in reality and in our simulations remains essential in the evolution of a structured language, rather than a 'problem' as in the Chomskyian tradition.

Although in this work, we have relaxed or removed some of the weakening assumptions in the classical ILM, much remains to be done. There are still many strong simplifications and abstractions concerning the nature of language and communication utilised in our computer simulations. One important direction

for future work is to move towards acoustic ('speech') communication—having agents produce and perceive sounds coupled to meaning, as suggested by Worgan and Damper (2007).

References

Bever, T., & Montalbetti, M. (2002). Noam's ark. *Science*, *298*(22), 1565–1566.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, *8*(1), 25–54.

Chomsky, N. (1975). *Reflections on language.* New York, NY: Pantheon.

Guenter, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(22), 1569–1579.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (Second ed.). Upper Saddle River, NJ: Prentice Hall.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, *8*(2), 185–215.

Kirby, S., & Hurford, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands & I. Harvey (Eds.), *Fourth european conference on artificial life* (pp. 493–503). Cambridge, MA: MIT Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Oudeyer, P.-Y. (2005). The self-organization of speech sounds. *Journal of Theoretical Biology*, *233*(3), 435–449.

Smith, A. D. M. (2005). The inferential transmission of language. *Adaptive Behaviour*, *13*(4), 311–324.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386.

Steels, L., & Wellens, P. (2006). How grammar emerges to dampen combinatorial search in parsing. In *Third international symposium on the emergence and evolution of linguistic communication (eelc 2006).* Published in *Symbol grounding and beyond*, Springer Verlag LNAI Vol. 4211, pp. 76–88.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial Intelligence*, *167*(1–2), 206–242.

Worgan, S. F., & Damper, R. I. (2007). Grounding symbols in the physics of speech communication. *Interaction Studies*, *8*(1), 7–30.