Combining Self-Organizing and Bayesian Models of Concept Formation

Tiina Lindh-Knuutila*, Juha Raitio and Timo Honkela

Adaptive Informatics Research Centre
Helsinki University of Technology P.O. Box 5400
FI-02015 TKK, Espoo, Finland
** E-mail: tiina.lindh-knuutila@tkk.fi*
http://www.cis.hut.fi/research/cog

In this article, we consider contemporary theories of concepts, and Bayesian and self-organizing models of concept formation. After introducing the different models, we present our own experiment. It utilizes a multi-agent simulation framework, in which the emergence of a common vocabulary can be studied. In the experiment, we use jointly the self-organizing maps and probabilistic modeling of concept naming. The results of the experiments show that a common vocabulary to denote prototypical colors emerges in the agent population.

Keywords: concept, concept formation, self-organization, Bayesian inference, multi-agent simulation, language game

1. Introduction

Concept learning is an essential task in any real life machine learning and artificial intelligence application. In this paper, we contrast self-organized with Bayesian approaches in concept formation learning, also considering the different background assumptions. First, we will characterize concepts through a short introduction to the contemporary theories, followed by examination of the process of concept formation as modelled by studies in self-organized and Bayesian fields. After contrasting these approaches, we present our own model that combines probabilistic modeling of concept naming with the self-organization of the underlying conceptual space.

2. Concepts

In our view, a concept is the mediating level between the perception and language (Fig. 1). We require concepts to be grounded. There are several

different views on the grounding problem.[1] We see grounding as the process of transfering sensory perceptions into non-symbolic (connectionist) representations. A concept can also be grounded through its relation to other concepts. Concept as such, does not suppose the existence of a linguistic label, but this label emerges through its function in communication. The *meaning* of a label is established in its association to a concept.

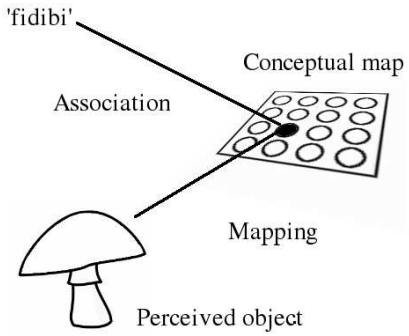

Fig. 1. The three level model of perceptions, concepts and language.

2.1. *Historical theories*

The classical, Aristotelean view sees concepts as a set of necessary and sufficient definitions (e.g bird has wings, beak etc.). This kind of view has a long history in philosophy.[2,3]

The prototype theory emerged in the 1970s as an alternative to the classical theory of concepts. The main idea was to include the experimental findings of typicality effects,[2,4] which means that some instances are better examples of a category than others (e.g. 'robin' is considered a better example of category *bird* than 'penguin'). According to the prototype theory, most lexical concepts are complex representations, whose structure encodes a statistical analysis of the properties their members tend to have: Categories have graded memberships around a certain prototype.

Since 1980s, there has also been a significant amount of research building on artificial neural network models with the aim of connecting the neural and symbolic levels (cf. e.g. the work by P.M. Churchland[5] and P.S. Churchland[6] as early influential accounts).

2.2. *Theory of conceptual spaces*

Theory of conceptual spaces[7] was developed to model conceptual representations in a cognitive framework. A conceptual space is built upon geometrical structures based on a number of quality dimensions. Concepts are not independent of each other but can be structured into domains, e.g., concepts for colors in one domain, spatial concepts in the other domain.

Categories are seen as convex regions in a conceptual space. The concepts are learned from a limited number of examples and by generalizing from them. The similarity of two objects can be defined as a distance between their representation points in the conceptual space, which can then be used, e.g., for categorization: The perceived item belongs to the category whose prototype is the nearest to the mapping of the item in conceptual space. The prototype effects can also be explained in the conceptual spaces. The prototypes are simply those instances of the category that are located in the central parts of these regions.

In general, the theory of conceptual spaces proposes a medium to get from the continuous space of sensory information to a higher conceptual level, where concepts could be associated to discrete symbols. It has been proposed[7] that for example multi-dimensional scaling (MDS) and self-organizing maps (SOM) could be used to model a domain in a conceptual space.

3. Models of concept formation

We adopt the position that concepts are learned, and they adapt. In the process of concept formation, mental constructs are developed based on sensory experience. Concept formation figures prominently in cognitive development (see e.g. Ref. 8). As a concept emerges, it becomes subject to testing. Consideration of a wide range of possibilities e.g. through playing games, contributes to this process.

3.1. *Self-organization*

The basic principle of the self-organized representations is that internal relations of categories may be derivable from the mutual relations and roles of the data in an unsupervised way.[9]

3.1.1. *The self-organizing map*

The self-organizing map (SOM) is a neural network model developed in the early 1980s.[10,11] It produces a topographic mapping the input space

into an array of nodes. Perhaps the most typical notion of the SOM is to consider it as an artificial neural network model of the brain,[12] especially of the experimentally found ordered cortical "maps".

Each node of the SOM consists of a prototype vector of the same dimension as the input vectors. The SOM is trained according to a competitive learning principle. When an input vector is fed into the system, a prototype vector with the smallest distance to the input (the best-matching unit, BMU) vector is selected. In the adaptation process, the BMU and its neighbors in the topological ordering are moved toward this input in the space. The degree of adaptation depends on the learning function $m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]$, where $h_{ci}(t)$ is the neighborhood function defining how large the neighborhood is, m_i is the ith map unit, $x(t)$ the input vector, and t is the discrete time coordinate.

A detailed description about the selection of the parameters, variants of the map, and many other aspects have been covered in Ref. 11. In this work, we have used the classical SOM architecture due to its reasonably high validity as a neuro-cognitive model. A well motivated alternative in this particular case would also be the Bayesian version of the SOM, i.e. the Generative Topographic Map (GTM).[13]

3.1.2. *Models using self-organization*

Ritter and Kohonen made first experiments of concept learning with the SOM.[9] In the experiment, contextual information for words (preceding and following word of the target word) based on generated three-word sentences was fed to the SOM. Later, Honkela[14] conducted an extended study in the same subject which used the Grimm's tales in English as data. The experiments show that based on the contextual information the target words were indeed organized in a SOM in a way that seems meaningful — nouns in one group, verbs in another. Words with similar usage (e.g., verbs with past tense, nouns describing animate or inanimate objects) could also be found in smaller subgroups.

Schyns[15] demonstrated how simple concepts could be learned with a modular neural network model. The model has two modules, one for categorizing the input in an unsupervised manner and another module for learning the names for the categories in a supervised mode. The input for the SOM, which was used as a categorization module, was pictorial image data varied around 'prototypes' in such a way that the prototypes were never directly shown to the SOM. Instead, the map was fed distortions around these prototypes. In a sense, this image data could then correspond

to certain 'sensory data'. The result of this experiment was that the map learned to represent the prototypes. Schyns sees that the categorization module fills the definitions of the prototype theory. In the second phase, the names for these categories were learned in a supervised manner.

Other examples of SOM applied to symbol processing include Ref. 16, where the representations emerging from color spectrum input and their association to color names were studied. In Ref. 17, a multi-agent simulation of a simple language emergence was conducted. The agents learned very simple concepts in the color domain using SOM as a model for the conceptual memory. In the course of the simulation, symbols emerged, and were associated with areas on the conceptual map, and as a result a simple, shared vocabulary emerged.

The self-organizing models of concept formation often use real-world data, either in textual form or fairly simple data from one conceptual domain. The approach is bottom up, and the learning is self-organizing and unsupervised. The representations obtained are pattern-like, and the inference between concepts is similarity based. It seems, though, that a single SOM cannot be used for representation for a complete set of concepts, but rather, various SOMs are needed for different domains. Additionally, a system to produce the per-concept feature selection for more complex concepts would be needed. See more detailed discussion on this in Ref. 18.

3.2. *Bayesian modeling*

3.2.1. *Bayesian inference*

Bayesian models are very commonly used in modern research in several fields. Bayesian inference utilizes the Bayes' theorem, $P(h|x) = \frac{P(x|h)P(h)}{P(x)}$, where $P(h|x)$ is the posterior probability of a hypothesis h given observation x, $P(x|h)$ the conditional probability (likelihood) of observing x under the hypothesis h, $P(x)$ the probability of the observation, and $P(h)$ the a priori probability of the hypothesis. In Bayesian inference, observations are used to infer the probability that a hypothesis may be true. New observations update this probability.

3.2.2. *Models using Bayesian inference*

There are several researchers who use Bayesian inference as a tool for modeling some cognitive phenomenon. We will briefly describe some of the research relevant to this article. Bayesian modeling of concept learning has been considered e.g. in Tenenbaum's model[19,20] which is based on the fol-

lowing basic building blocks: (1) A constrained hypothesis space of possible extensions of a concept, (2) a prior distribution over the hypothesis space reflecting the learner's relevant background knowledge, (3) the size principle for scoring the likelihood of hypothesis, favoring smaller consistent hypotheses, and (4) hypotheses averaging: integrating the predictions of multiple consistent hypotheses.

More specifically, the task of learning simple concepts was defined as the task of learning axis-parallel rectangles, based on a small number of positive examples only. The likelihood $p(x|h)$ was computed based on the background assumption of randomly sampled positive examples (strong sampling criterion[20]) $P(x|h) = \begin{cases} \frac{1}{|h|} & \text{if } x \in h \\ 0 & \text{otherwise} \end{cases}$, where h indicates the size of the region. This leads to the *size principle*: smaller hypothesis that just cover the observed samples become more likely than larger ones. They define the purpose of concept learning to help in the decision making: As an example, they give a doctor who needs to learn a doctor to learn which levels of cholesterol can be considered healthy[19] or a baby bird that needs to learn which worms are edible edible based on the worm color shades.

Dowman studies the Bayesian concept learning principle in the color domain,[21,22] using the Bayesian approach described earlier. His work does not consider the pre-linguistic categorization, instead phenomenological color space consisting only of the hue of color is assumed to be the representational level for colors. A further assumption is that all humans are able to produce a representation of the color space in a similar way and that there is a 'correct' denotation of the color term and color association as a norm of the speech community as a whole. It does not make a distinction between color words and color categories which is often made e.g. in Ref. 23, but simply names a range of colors directly without any reference to a prelinguistic category.

3.3. *Combining self-organizing and Bayesian models*

The self-organizing models assume a prototype or conceptual spaces theory of concepts. The main purpose of these models is in transferring sensory percepts in some way to the conceptual level. The Bayesian models studied here are based on classical and prototype theory and these models assume that the conceptual representations are the same for each learner, which makes a crucial difference to the self-organizing approach presented later in this paper. Though the extent of the area each term occupies in the repre-

sentational space can vary as in Refs. 21,22, the underlying representations for the conceptual color domain are the same.

In the following, we employ the self-organizing map for representing color perceptions and combine it with probabilistic modeling of concept naming.

4. Experiment in the color domain

As a basis of the language emergence, we use language games introduced by Wittgenstein.[24] We implemented a version of the naming game[23] using the SOM. The setup follows closely an earlier the multi-agent simulation framework,[17] where a SOM is used as a model of the conceptual memory of an agent. We use data from the color domain which is often used used[16,21,22] since the data is simple and easily obtained. Following the conceptual spaces theory, we are then treating one domain of integrated dimensions, the color space. For practical reasons, we use the RGB color space, even though any other color space should be equally good.

Each agent thus has a conceptual map which is based on a SOM. Prior to simulation, each of these maps is trained with color data. The training data sets for agents share similarities, but are not the same, yielding individual (albeit similar) conceptual memories for each agent. After the initial training, which takes place before the simulated naming games, the SOMs are not changed. Similarly, an additional set of color vectors is created to serve as the topics of the naming games. In the simulation, agents play naming games, which proceed as follows:

1. Two agents are randomly selected from the population. They are assigned the tasks of the speaker, and the hearer, respectively. Similarly, a topic vector is randomly selected.
2. Both agents map the topic to their conceptual memories. In the SOM, this corresponds to finding the BMU in the map.
3. The speaker searches for a word that could best match the given topic. If no word is found, a new word is invented and communicated to the hearer. The decision making process is described in more detail in the following.
4. The hearer also searches for the words that best match the given topic.
5. If the word the speaker communicates is among the words the hearer found, the game is a success, otherwise the game fails.
6. In case of a success, both the speaker and the hearer increase the counter for that word-node association by one.

7. In case of failure, counters are not updated, except when the word was not known beforehand. In that case, the word is added to the lexicon with an initial counter value of 1.

This algorithm results in the agents selecting the term to denote a given topic based on the maximum likelihood, $\max P(C|T)$, which is estimated as the number of successful uses of the term for that BMU, proportional to all of the successful uses of all the terms in that node. The likelihood is estimated for all the terms associated with the BMU and for those nodes adjacent to it, and the term with the highest likelihood is selected and uttered. If no term is found to be associated with the color or the neighborhood, a new term is invented. The hearer estimates the likelihood $P(C|T)$ in the same fashion for the BMU and the neighborhood in its own SOM finding the preferred terms (in order) for the given color.

We use the likelihood instead of a posterior probability since taking into account the a priori probability, the most frequent terms would always be preferred - regardless of the color they have been associated with before. In earlier work,[17,23] the associations between color terms and the map nodes were employed differently: Each color term – map node pair had an association weight which was changed according to the outcome of the language game: increased if the game was successful and decreased if the game was unsuccessful.

To evaluate the degree to which a language has emerged in the process, we define communication success ratio[23] (CS) as a measure which tells how often the communication is successful. It is given as the average number of successfully played games in 100 previous language games (or less, if 100 games have not been played yet).

4.1. *Experimental setup*

We implemented the simulation framework described earlier and conducted two experiments using different population sizes. In our experiments, each agent had a conceptual memory based on a self-organizing map. The size of the maps was 96 nodes. The neighborhood used was hexagonal and the maps were initialized randomly.

The maps were trained separately for each agent with different data sets, which were data taken from generated color pictures. The training data contained RGB values of eight different prototypical colors: black, blue, green, cyan, red, magenta, yellow and white. To make the distributions for each color less spiky, uniform noise was added independently to each of

the color channels. The noise level was set to 20%. The total length of the training data was 10,000 samples. The training data sets for each agent were slightly different, but generated in the same way. (See also Figure 3 where conceptual memories of two agents are shown.) A set of 400 color samples, created similarly to the training data, was used as the language game topics. These topics were not part of the training data. The emerging words were created in the simulation in the same way as in Ref. 17 and the set of words in the simulation can be considered open.

We ran three sets of experiments all of which consisted of 10,000 language games and 10 repetitions. In the first experiment, the population size was fixed to $N = 2$, and in the second experiment to $N = 4$. To experiment with a larger population size, we also did the similar simulation runs for agent population of $N = 10$. In these experiments, a game was considered a success if the word uttered by the speaker was also considered the best word by the hearer.

4.2. *Results*

Figure 2 shows the communication success for two, four and ten agents, each averaged over 10 simulation runs. In the two-agent case, the communication success rises rapidly to $CS = 0.8$ and then steadily up to $CS = 0.95$ during the $10,000$ simulated games. The communication success for four agents grows slower than in the previous experiment, but still increases up to $CS = 0.86$, where it seems to settle. The bigger population size, in the ten-agent case yields into considerably slower convergence, reaching approximately $CS = 0.8$ in 10,000 games.

All the language games are played pair-wise, i.e. only two agents of the whole population participate in each game, and other agents have access to the words only through subsequent language games with the same topic. This means that when the population size grows, the convergence to common vocabulary is considerably slower. More competing words for a given topic emerge, and it simply takes longer for each agent to see a representative subset of the topics.

Figure 3 shows the conceptual maps of the two agents in the first experiment. The colors denote the converged RGB values of the prototype vectors of the map. The map has organized well and transformations from one color to other are smooth. The eight prototypical colors used are more prominent, since they are represented more in the data than the intermediate colors that have resulted from added noise.

When comparing the figures, it is evident that for most prototypical

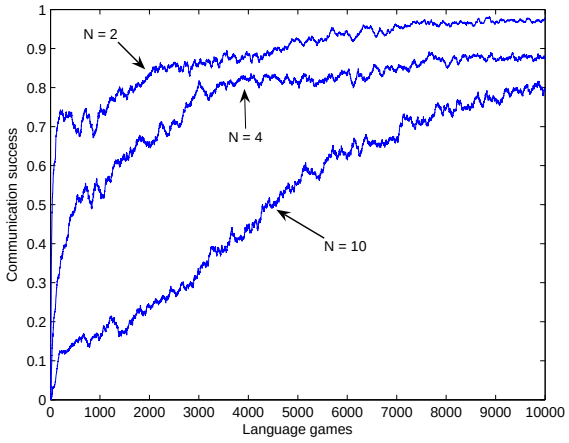

Fig. 2. Communication success for $N = 2$, $N = 4$ and $N = 10$ agents in the population.

colors there are one or two words that are preferred: *deci* for black or dark, *hihi* for blue, *fehe* for green, *hebe* for cyan, *defebe* and *gahefa* for red, *cede* for magenta, and *babi* and *dabide* for yellow. For white, the most common word used is *gedi*, but there are also competing labels for bluish white, pinkish white and so on because white covers a larger area in the space. The conceptual memories support the conclusion already visible in the communication success ratio – that a common vocabulary for the agents has emerged.

5. Conclusions and discussion

In this article, we have contrasted the self-organizing and Bayesian approaches to concept formation. We analyzed the approaches taking into account what assumptions are made, what the representation of concepts is like, what kind of inference takes place, and whether symbol grounding is addressed or not. We also built a model of our own for concept formation in a naming game as a combination of the approaches. In our experiments, a common vocabulary emerges in a population of agents using the probabilistic concept naming model based on likelihood.

This work employs a different model for the concept naming than earlier,[17] and presents the preliminary experiments. Future work includes more detailed study of the process of language convergence and more rigorous study of the characteristics of the emerging vocabulary: how coherent is

Fig. 3. The conceptual memories of the agents in the two-agent simulation. Only the most probable label for each node is shown.

the language use and in which degree polysemous and synonymous terms exist. Also, employing the GTM approach as an alternative to the SOM will be considered.

References

1. L. Steels, Perceptually grounded meaning creation, in *ICMAS96*, ed. M. Tokoro (AAAI Press, 1996).
2. S. Laurence and E. Margolis, Concepts and cognitive science, in *Concepts: Core Readings*, eds. E. Margolis and S. Laurence (MIT Press, Cambridge, MA, 1999)
3. J. Locke, *An essay concerning human understanding* (Oxford University Press, 1690/1975).
4. E. Rosch, Principles of categorization, in *Cognition and Categorization*, (Lawrence Erlbaum Associates, 1978) pp. 27–48.
5. P. M. Churchland, *A neurocomputational perspective: the nature of mind and the structure of science* (MIT Press, Cambridge, MA, USA, 1989).
6. P. S. Churchland and T. J. Sejnowski, *The Computational Brain* (MIT Press, Cambridge, MA, USA, 1992).
7. P. Gärdenfors, *Conceptual spaces: The Geometry of Thought* (MIT Press, 2000).
8. J. Piaget, *Child's conception of the world* (Routledge and Kegan Paul, 1928).
9. H. Ritter and T. Kohonen, *Biological Cybernetics* **61**, 241 (1989).
10. T. Kohonen, *Biological Cybernetics* **43**, 59 (1982).
11. T. Kohonen, *Self-Organizing Maps*, Series in Information Sciences, Vol. 30,

3rd. edn. (Springer, 2001).
12. T. Kohonen and R. Hari, *Trends Neurosci.* **22**, 135 (1999).
13. C. M. Bishop and C. K. I. Williams, *Neural Computation* **10**, 215 (1998).
14. T. Honkela, V. Pulkki and T. Kohonen, Contextual relations of words in grimm tales analyzed by self-organizing map, in *Proceedings of International Conference on Artificial Neural Networks, ICANN-95*, (EC2 et Cie, 1995).
15. P. Schyns, *Cognitive Science* **15**, 461 (1991).
16. J. Raitio, R. Vigário, J. Särelä and T. Honkela, Assessing similarity of emergent representations based on unsupervised learning, in *Proceedings of IJCNN 2004)*, (Budapest, Hungary, 2004).
17. T. Lindh-Knuutila, T. Honkela and K. Lagus, Simulating meaning negotiation using observational language games, in *Symbol grounding and beyond*, , Lecture Notes in Computer Science Vol. 4211 (Springer Berlin/Heidelberg, 2006).
18. T. Honkela, K. Hynnä, K. Lagus and J. Särelä, *Adaptive and Statistical Approaches in Conceptual Modeling*, Publications in Computer and Information Science A75, Helsinki University of Technology (2005).
19. J. Tenenbaum, Bayesian modeling of human concept learning, in *Advances in Neural Information Processing systems 11*, (MIT Press, 1999).
20. J. Tenenbaum and T. Griffiths, *Behavioral and brain sciences* **24**, 629 (2001).
21. M. Dowman, *A Bayesian Approach to Color Term Semantics*, Technical Report 528, Basser Department of Computer Science, University of Sydney (2001).
22. M. Dowman, *Cognitive Science* **31**, 99 (2007).
23. L. Steels and P. Vogt, Grounding adaptive language games in robotic agents, in *Proceedings of the Fourth European conference on Artificial Life*, (MIT Press, Camridge, MA and London, 1997).
24. L. Wittgenstein, *Philosophical Investigations* (Macmillan, 1963).

Acknowledgments

This work was supported by the Academy of Finland through the Finnish Centre of Excellence programme and the Finnish Graduate School of Language Technology. We also thank the anonymous reviewers for their invaluable comments.