Framework for Modeling Partial Conceptual Autonomy of Adaptive and Communicating Agents

Timo Honkela and Kevin Ilmari Hynna ({first}.{last}@hut.fi)
Laboratory of Computer and Information Science
Helsinki University of Technology, Espoo, Finland

Tarja Knuuttila (tarja.knuuttila@helsinki.fi)
Center for Activity Theory and Developmental Work Research
University of Helsinki, Helsinki, Finland

Appeared in the Proc. of CogSci2003, 25th Annual Meeting of Cognitive Science Society, Boston, MA, July 31-August 2, 2003.

Abstract

We develop a framework for discussing the degree of conceptual autonomy of natural and artificial agents. We claim that aspects related to learning and communication necessitate adaptive agents that are partially autonomous. We demonstrate how partial conceptual autonomy can be obtained through a self-organization process. The input for the agents consists of perceptions of the environment, expressions communicated by other agents as well as the recognized identities of other agents.

Agents and Communication

Agents communicate by sending and receiving messages. In the primitive case, all the agents share a common model of their environment (which implies a shared language), and messages between agents have fixed and common interpretations. At a more advanced level of multi-agent co-operation, each agent has its own model of the environment. Thus, each agent has a "subjective" interpretation for the relationship between the messages and the environment. These differences in the agents' models motivate the development of methods which provide the agents with the ability to learn, including learning to interpret messages from other agents. In this paper a general scheme for multi-agent communication is presented, along with a discussion of some applicational and methodological alternatives.

The basic elements of a generalized model of multi-agent communication are: the environment of the agents, the language used in communication, input-output functions between environment and expression, and the memory and processing mechanisms of the agents.

The agents can perceive their environment, they are part of it, and possibly they can change it. The environment may be a computerized representation, constructed, or natural. The borderlines of these domains may, of course, be vague. A natural environment, in particular, is ever-changing, and consists of various continuous phenomena.

In its simplest form, communication may be based on a fixed set of distinct signals. Here we consider the possibility of applying a natural, or near-natural language as the communication medium. The general properties of natural languages necessitate some capabilities that autonomous agents will need to have. These basic properties of natural languages and their interpretation include ambiguity, contextuality, open-endedness, vagueness, and subjectivity.

Brazdil and Muggleton (1991) show how to use symbolic inductive inference as a way of learning to relate terms in multiple agent communication. They have shown how to overcome language differences between agents automatically in a situation where the agents do not have the same predicate vocabulary. The system consists of a number of separate agents that communicate. Each agent has certain perceptive, communicative and reasoning capabilities, being able to (1) perceive a portion of the given, possibly simulated world, (2) accept facts and rules from another agent, (3) formulate queries and supply them other agents, (4) respond to queries formulated by other agents, (5) interpret answers provided by other agents, (6) induce rules on the basis of facts, and (7) integrate knowledge. Symbolic models of the environment are closely related to model-theoretic approaches used to define the semantics of formal languages. One obvious critique of such an approach lies in the observation that the meaning of an expression (query, response) in a natural domain is fuzzy and changing, biased by the particular context in which it occurs.

Ambiguity or vagueness, then, can be a "virtue" when the communication medium is used in an open and changing environment in which having a distinct and a priori determined symbol, or combination of symbols, would be difficult, or practically impossible. Finally, to ensure successful communication, both the sending and the receiving agent must share a similar enough framework of interpretation, and the message or the situation ("context") must contain enough information to activate a proper framework for the receiver.

Cognitive capabilities of agents

The basic cognitive functions of agents consist of their input (from the environment) and output (to expression) functions, as well as memory and processing mechanisms. The agents perceive their environment and have a model or representation of it. This representation may be

1. static or dynamical
2. given from outside or autonomously learned
3. common between agents or individual, the latter of which would make learning capabilities necessary
4. symbolic, non-symbolic or hybrid (Wermter and Sun 2000)

In this article, conceptual autonomy is based on the second point above. Accordingly, an agent is *conceptually autonomous if it learns its representation of the environment by itself*, where a concept is taken to be simply a means of specifying a relationship between language and world. By partial autonomy we mean a setting in which the learning process of an agent is influenced in some way by other agents. This influence can then serve as a basis for communication between agents. Thus, although each agent has an individual representation of the environment, the representations are related though the coordinating effect of communication between agents in situations where both agents have access to similar perceptions of the environment. A counter example to a conceptually autonomous agent would be a traditional expert system which has been preprogrammed and therefore is given a priori its representation of the environment.

Put another way, a conceptually autonomous agent adapts its representations through unsupervised learning, whereas a partially conceptually autonomous agent adapts through supervised, self-supervised or reinforced learning. A strictly non-autonomous agent therefore does not learn based on the concept of adaptation. (Though one could perhaps consider as an exception a conceptually non-autonomous system which adapts through rote learning.) This idea of partial conceptual autonomy and its relation to agent communication is developed further in the section on communication and context sharing below.

We have emphasized the role of learning in determining the relative autonomy of an agent. But we take a broad view of learning, including any adaptation of internal representations (i.e., concepts) that may occur as a result of interactions with other agents or the environment. As Kirsh (2001) points out, in ecological systems each component of the system has a causal influence on the other. In the biological world organisms interact with their environment and with other organisms, who, of course, also tend to be part of each other's environment, the whole system of components being interdependent and interlocked. The result is a highly complex system displaying attractors, instabilities and cycles typical of dynamical systems. (Kirsh 2001) Finally, we can mention the criticisms of an autonomous approach found in Alterman (1997). He emphasizes aspects of collaboration and distributed cognition in multi-agent models, and even argues that the roles of the user, designer, and builder of the model preclude the possibility of complete autonomy for the models. Although we accept that the role of the modeller must be taken into account, the topic of meta-theoretical design principles is beyond the scope of the current paper.

Subjectivity and the vocabulary problem

In the field of information retrieval, Furnas et al. (1987) have found that in spontaneous word choices for objects in five domains, two people favored the same term with less than 20% probability. Bates (1986) has shown that different indexers, well trained in an indexing scheme, might assign index terms for a given document differently. It has also been observed that even the same indexer might use different terms for the same document at different times.

Moore and Carling (1988) state: "Languages are in some respect like maps. If each of us sees the world from our particular perspective, then an individual's language is, in a sense, like a map of their world. Trying to understand another person is like trying to read a map, their map, a map of the world from their perspective."

Thus, the language of an agent or individual can be idiosyncratic and based on the subjective experiences of the individual. As an example related to the vocabulary problem, two persons may have a different conceptual or terminological "density" of the topic under consideration. A layman, for instance, is likely to describe a phenomenon in general terms whereas an expert uses more specific terms. This aspect of partial conceptual autonomy will be discussed later in this paper in more detail.

Communication and Context Sharing

The general model of communicating agents consists of three input modalities:

- expressions used in communication (shortened as "E")
- contexts ("C")
- agents' identifications ("A")

An agent can process any channel of input alone if the other two are missing, or it can associate two channels (e.g., the "A" channel is only present when the "E" is in use), or even all three of the modality domains.

Alternatives of abstract situations

Next we'll consider the most relevant modality combinations one by one. We will make a distinction between input combinations that occur during a learning phase and input-output combinations that occur during a communication phase. Such a separation is not strictly necessary but it simplifies the description of the model. In the learning phase the main input combinations are:

- C as input: formation of a representation of the context domain
- C+E as input: association of context patterns with expressions
- C+E+A as input: associating context-symbol associations with an agent identification, or more specifically, associating all three input "channels".

Secondary input combinations may also be listed for completeness:

- E as input,
- A as input, and
- E+A as input.

In the communication phase the following input-output combinations are the most relevant.

- C as input, E as output: the agent names the "object" it has been "shown".
- E as input, best-matching instance of a list with C as output: the agent points out from a list an "object" that best matches the expression.
- C+A as input: E as output: the agent names the "object" taking into account the receiving agent of the message.
- E+A as input, best-matching instance of a list with C as output: the agent points out from a list an "object" that best matches the expression taking into account the agent that expressed itself.
- E+C as input, E as output: the agent evaluates whether it would use the same expression as the description of the "object".
- E+C as input, A as output: the agent specifies which agent is the likely utterer of the expression in the particular case.

Example: color naming

A practical example can be given related to the domain of colors. Naming colors in particular has certain invariant features as well as a potentially large number of borderline cases in which two subjects often name the same color differently. The alternatives considered above give rise to the following possible situation when two subjects are comparing their expressions on some color that both can perceive. Strictly speaking, the different visual points of view should be taken into account; one essential factor in the Talking Heads project (Steels and Kaplan 2002). If two subjects state the same expression then the situation is unproblematic. If they use different expressions the following options are possible: one can agree that the expression used by the other subject is a viable alternative, e.g., a piece of furniture is considered sky blue by one and plain blue by the other, but they agree that both expressions can be used, or they can simply disagree on the applicability of each other's expressions. An additional level of increased realism and complexity to the model can be obtained by considering the fuzziness of the color naming process. The borderlines of three-dimensional domains of color features for each color name are not crisp. Similarly, the degree of agreement or disagreement on the use of a color term can be considered from within the framework of fuzzy set theory (Zadeh 1965). Another point of view can be obtained by taking into account the distinction between active and passive vocabulary. One agent may name one color with a certain term but is ready to accept alternative expressions denoting the same color (e.g. 'dark salmon' versus 'rosy brown').

The previous discussion handled a situation in which both subjects experienced the same (or approximately the same) color perception and were comparing their associated color terms. However, if the perceptual input is missing for one or both of the agents, there is no direct source of evidence to check the agreement on naming. Thus, if one agent expresses a color name the other agent interprets this name within its own scheme. However, there may be other, indirect evidence on which to base (dis)agreement. For example, the color names can be expressed in the context of a symbol-level context, e.g., with associated nouns. The two subjects can check whether they share the conception of the color name 'sienna' by comparing whether they agree upon the nouns that this adjective can readily qualify. Similarly, the agents may compare the conceptually neighboring color terms within some, implicit similarity metrics scheme. For instance, one agent might state that the color 'light coral' is between the colors 'salmon' and 'rosy brown'. MacWhinney (1989) mentions that the prototype theory fails to place sufficient emphasis on these kinds of relations between concepts. MacWhinney also points out that prototype theory has not cov-

ered the issue of how concepts develop over time in language acquisition and language change, and, moreover, it does not provide a theory of representation. MacWhinney has presented a model of emergence in language based on the SOM (MacWhinney 1997). Gärdenfors (2000) has presented a detailed account on the motivation for the use of the SOM in modeling conceptual spaces. This topic will be discussed in more detail in the following section. Hardin (1993) presents a detailed account on philosophical aspects of color based on the evidence gained in the natural sciences.

From color naming to societies of agents

An additional aspect of modeling is obtained if the agent takes into account in its internal interpretation the utterer of the color expression. For example, if one agent uses a particular color term in an unusual manner, the other agent can learn this specific relation and use this naming convention while communicating with that particular agent. In general, this phenomenon is called meaning negotiation. The phenomenon can be considered as something that occurs between two communicating individual agents. In addition, this consideration is important while comparing the naming conventions between two more or less isolated agent communities.

Agent Learning and Self-Organizing Maps

One novel approach to modelling the learning, interpretation and use of natural language has been to develop systems that simulate the learning process. In the area of machine learning, for example, methods of generalization, i.e. inductive inference, have been implemented. However, many of these models are based on symbolic representations of rules which runs into the difficulty of grounding the symbols somehow into the continuous and changing perceptual domains. The self-organizing map (SOM) (Kohonen 1982, 2001) is an information processing model which is often considered as an artificial neural network model. There exists quite a lot of neurophysiological evidence to support the idea that the self-organizing map captures some of the fundamental processing principles of the brain, especially of the experimentally found topologically ordered "maps" in the cortex (Kohonen 1993, Kohonen and Hari 1999). The evidence for localization of linguistic functions in the brain is discussed, e.g., in (Caramazza et al. 1994). They cite studies in which category-specific deficits of selective impairment including proper names and geographical names have been reported (McKenna and Warrington 1978). These findings and the maps presented, e.g., in (Honkela et al. 1995, Kaski et al. 1996). exhibit a striking resemblance.

Knowledge representation aspects

The self-organizing map can be viewed as a model of unsupervised machine learning and also, importantly in this context, as an adaptive knowledge representation scheme. The traditional knowledge representation formalisms (semantic networks, frame systems, predicate logic, to provide some examples) are static and the reference relations of their elements are determined by a human. Those formalisms are based on the tacit assumption that the relationship between natural language and world is one-to-one: the world consists of objects and the relationships between the objects, and these objects and relationships have straightforward correspondences to the elements of language. Moreover the representations are "programmed", not learned through experience.

The self-organizing map is described in the following using agent terminology slightly adapted from (Lagus et al. 1996). Consider a collection or system of agents which must learn to carry out very different tasks. Let us assume that the system may assign different tasks to different agents of the collection that the agents are able to learn from what they do. Each new task is given to the agent that can best complete the task. Since the agents learn, and since they receive tasks that they can do well, they become even more competent in those tasks. This is a model of specialization by competitive learning. If the agents are interconnected in such a way that also the neighbors of the agent carrying out a task are allowed to learn some of the task, the system slowly becomes ordered so that agents near each other have similar abilities, and the abilities change slowly and smoothly over the whole system. This is the general principle of the self-organizing map.

Self-organizing map algorithm

In the following, the self-organizing map algorithm is described in some more detail based on (Kohonen 2001). Assume that some sample data sets have to be mapped onto a two-dimensional array, a sample set is described by a real vector $x(t) \in R^n$ where t is the index of the sample, or the discrete-time coordinate. In setting up the Self-Organizing Map, we first assign to each unit in the array a parameter vector $m_i(t) \in R^n$ called the codebook vector, which has the same number of elements as the input vector $x(t)$. The initial values of the parameters (components of $m_i(t)$) can be selected at random. The process described below changes these parameters.

The "image" of an input item on the map is defined to be the unit whose $m_i(t)$ best matches with $x(t)$ according to some metric. The self-organizing algorithm that creates the ordered mapping can be described as a repetition of the following basic tasks:

1. An input vector $x(t)$ is compared with all the

codebook vectors $m_i(t)$.

2. The best-matching unit, BMU, on the map, i.e., the unit where the parameter vector is most similar to the input vector in some metric is identified.

3. The codebook vectors of the BMU and a number of its neighboring units in the array are changed incrementally according to the learning principle specified below. (Kohonen 2001)

The basic idea in the self-organizing map is that for each input sample vector, $x(t)$, the parameters of the BMU and units in its neighborhood are changed closer to $x(t)$. For different $x(t)$ these changes may be contradictory, but the net outcome of the process is that ordered values for the $m_i(t)$ are finally obtained over the array which capture the distribution of the $x(t)$. If the number of input vectors is not large compared with the number of codebook vectors (map units), the set of input vectors must be presented many times reiteratively. As mentioned above, the codebook vectors may initially have random values, but they can also be selected in an ordered way. Adaptation of the codebook vectors in the learning process takes place according to the following equation:

$$m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)]$$

for each $i \in N_c(t)$, where t is the discrete-time index of the variables, the factor $\alpha(t) \in [0, 1]$ is a scalar that defines the relative size of the learning step, and $N_c(t)$ specifies the neighborhood around the BMU in the map array. At the beginning of the learning process the radius of the neighborhood is fairly large, but it shrinks during learning. This enforces a more global ordering of the map at the beginning of training, whereas towards the end, as the radius gets smaller, the codebook vectors of the units in the map become more fine-tuned. The factor $\alpha(t)$ decreases during learning.

Partial Conceptual Autonomy through Self-Organization

Consider that an agent is to denote an interval of a single continuous input parameter using a limited number of symbols, and that these symbols are then used in communication between agents. In a trivial case, two agents would have the same denotations for the symbols, i.e. the limits of the intervals corresponding to each symbol would be identical. If the "experience" of the agents is acquired from differing sources, then their conceptualizations may very well differ.

One may then ask how to deal with this kind of discrepancy. The following section describes self-organizing maps and a model of their use in this task. The key idea is to provide the means for each agent to associate continuous-valued parameter spaces to sets of symbols, and furthermore, to "be aware" of the differences in this association and to learn those differences explicitly. These kinds of abilities are especially required by highly autonomous agents that need to communicate using an open set of symbols or constructs in the agent language.

The self-organizing map is especially suitable for the central processing element of autonomous agents because of the following reasons:

- The self-organizing map algorithm modifies its internal presentation, i.e., the codebook vectors, according to the external input which enables the adaptation of the agents.

- The self-organizing map is able to process natural language input to form "semantic maps" (Ritter and Kohonen 1989) Natural language interpretation using self-organizing maps has further been examined, e.g., by Miikkulainen (1993), Scholtes (1993), Honkela and Vepsäl'äinen (1991) and Honkela et al. (1995). In (Honkela et al. 1995), the input data consisted of a set of English translation of Grimm fairy tales. Word trigrams were used as input vectors taking the encoded representations of three subsequent words from the preprocessed text. Summarizing the results of the statistical analysis conducted by the map algorithm, all verbs were located in the top section of the map whereas the nouns occupied the opposite side of the map. Among the nouns, inanimate and animate nouns formed areas of their own. Similar results can be obtained using other clustering algorithms.

- Symbols and continuous variables may be combined in the input, and are associated by the self-organizing map (Honkela 1991). Continuous variables may be quantized, and a symbolic interpretation can be given for each section in the possibly very high-dimensional space of perceptual variables (Honkela 2000).

- Because the self-organizing map is based on unsupervised learning, processing external input without any prior classifications is possible (Kohonen 2001). The autonomous agent may form an individual model of the environment and of the relation between the expressions of the language and the environment.

- If the interpretation of the messages needs to be identical among the agents, the self-organizing map enables creating a model of the relation between the environment and the expressions of the language used by the other agents. In addition, generalizations of these relations can be formed (Honkela 1993).

Discussion

We have presented a framework for considering the level of conceptual autonomy of communicating agents. Based on the framework various practical applications can be built. The use of the self-organizing map algorithm was considered as one option for modeling the internal conceptual and adaptive processes that were presented.

The conceptual spaces of partially autonomous agents were earlier discussed in the context of color names. Different situations related to "subjective" variance in naming the colors were studied. The issue of the cultural and societal levels of conceptual spaces of agents is of great importance when practical political and societal phenomena are considered. Instead of the names of colors, the differences in the conceptual spaces can be related to expressions such as 'democracy', 'freedom', 'equality', and 'terror', etc. Whether the scientific community dealing with the study of societies of agents and conceptual spaces will ever be able to contribute to solving such problems in the future remains to be seen.

References

Alterman, R. (1997). *Rethinking Autonomy.* Technical Report CS-97-195, Computer Science Department, Brandeis University. (Appeared also in Minds and Machines, 10:1 15-30, 2000)

Bates, M. J. (1986). Subject access in online catalog: a design model. *Journal of the American Society of Information Science*, 37(6):357-376.

Brazdil, P. & Muggleton, S. (1991). Learning to relate terms in a multiple agent environment. *Proc. of Machine Learning - EWSL-91*, pp. 424-439, Berlin, Springer-Verlag.

Brooks, R. (1991). Intelligence without reason. *Proc. of IJCAI-91, Intern. Joint Conference on Artificial Intelligence*, pp. 569–595.

Caramazza, A., Hillis, A., Leek, E.C., and Miozzo, M. (1994). The organization of lexical knowledge in the brain: Evidence from category- and modality-specific deficits. In *Mapping the mind: Domain specificity in cognition and culture*, Cambridge University Press, Cambridge, pp. 68-84.

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964-971.

Gärdenfors, P. (2000). Conceptual Spaces. MIT Press.

Hardin, C.L. (1993). *Color for Philosophers.* Expanded edition. Hackett Publishing Company.

Honkela, T. (1991). Interpreting imprecise expressions: Experiments with Kohonen's self-organizing maps and associative memory. T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, (eds.), *Artificial Neural Networks*, North-Holland, pp. I–897–902.

Honkela, T. (1993). Neural nets that discuss: A general model of communication based on self-organizing maps. Stan Gielen and Bert Kappen, (eds.), *Proc. of Intern. Conference on Artificial Neural Networks (ICANN-93), Amsterdam*, pp. 408–411, London, Springer-Verlag.

Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual Relations of Words in Grimm Tales Analyzed by Self-Organizing Map. *Proc. of Intern. Conference on Artificial Neural Networks, ICANN-95*, F. Fogelman-Soulié and P. Gallinari (eds.), EC2 et Cie, Paris, pp. 3-7.

Honkela, T. (1997). Learning to Understand - General Aspects of Using Self-Organizing Maps in Natural Language Processing. *Computing Anticipatory Systems*, D. Dubois (ed.), American Institute of Physics, Woodbury, New York, pp. 563-576.

Honkela, T. (2000). Self-Organizing Maps in Symbol Processing. *Hybrid Neural Systems*, Stefan Wermter, Ron Sun (eds.), Springer, Heidelberg, 2000, pp. 348-362.

Kaski, S., Honkela, T., Lagus, K. and Kohonen, T. (1996). Creating an order in digital libraries with self-organizing maps. *Proc. of WCNN'96, World Congress on Neural Networks*, Lawrence Erlbaum and INNS Press, Mahwah, NJ, pp. 814-817.

Kirsh, D., The Context of Work. *Human Computer Interaction*, 2001.

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

Kohonen, T. (1993). Physiological interpretation of the selforganizing map algorithm. *Neural Networks*, 6(7):895-905.

Kohonen, T., and Hari, R. (1999). Where the Abstract Feature Maps of the Brain Might Come from. *Trends in Neurosciences*, 22(3):135-139.

Kohonen, T. (2001). *Self-Organizing Maps.* Extended Edition. Springer, Berlin, Heidelberg.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. Simoudis, E., Han, J., and Fayyad, U., (eds.), *Proc. of the 2nd Intern. Conference on Knowledge Discovery and Data Mining*, pp. 238-243. AAAI Press, Menlo Park, CA.

MacWhinney, (1989). chapter Competition and Lexical Categorization. *Linguistic categorization*, Benjamins, New York.

MacWhinney, (1997). chapter Lexical Connectionism. *Cognitive approaches to language learning*, MIT Press.

McKenna, P. and Warrington, E. K. (1978). Category-specific naming preservation: A single case study. *Journal of Neurology, NeuroSurgery and Psychiatry*, 41:571-574.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory.* MIT Press, Cambridge, MA.

Moore & Carling (1988). *The Limitations of Language.* Macmillan Press, Houndmills.

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

Scholtes, J. (1993). *Neural Networks in Natural Language Processing and Information Retrieval.* PhD thesis, Universiteit van Amsterdam, Amsterdam, Netherlands.

Steels, L. (1995). Intelligence - dynamics and representations. L. Steels, (ed.), *The Biology and Technology of Intelligent Autonomous Agents*, Berlin, Springer-Verlag.

Steels, L. and Kaplan, F. (2002). Bootstrapping grounded word semantics. Briscoe, T., (ed.), *Linguistic evolution through language acquisition: formal and computational models*, pp. 53-73, Cambridge University Press. Cambridge.

Wermter, S. & Sun, R. (eds.) (2000). *Hybrid Neural Systems*. Lecture Notes in Computer Science, Springer.

Zadeh, L. (1965). Fuzzy sets, *Inf. Control*, 8, pp. 338-353.