ADAPTIVE AND STATISTICAL APPROACHES IN CONCEPTUAL MODELING

Timo Honkela, Kevin I. Hynnä,

Krista Lagus and Jaakko Särelä, editors

Helsinki University of Technology
Laboratory of Computer and Information Science
Neural Networks Research Centre
P.O. Box 5400
FIN-02015 HUT, FINLAND

January 2005

Helsinki University of Technology
Department of Computer Science and Engineering
Laboratory of Computer and Information Science
Report A75

Contents

Preface **1**

1 Adaptive and Statistical Approaches to Conceptual Modeling **3**

Timo Honkela, Krista Lagus and Jaakko Särelä . . . .

2 Emergence of Structure and Relations in Conceptual Representations **7**

Krista Lagus, Aarno Lehtola, Mikko Määttä and Sakari Virkki . . . .

2.1 Introduction 7

2.1.1 What is structure? 7

2.1.2 Definitions 8

2.1.3 Structure of the remaining chapter 10

2.2 Conceptual Structure in Psychology, Philosophy and Linguistics 11

2.2.1 Concepts, Structure and Relations 11

2.2.2 Theories of the Internal Structure of Concepts (Ri) 13

2.2.3 Formation of Conceptual Structure 15

2.2.4 Relations in the Conceptual Spaces Model 16

2.2.5 Deep Cases in Case Grammars 17

2.2.6 Summary of Relations in the Discussed Theories 19

2.3 The Emergence of Structure in Computational Methods 19

2.3.1 Suggested subtasks regarding the emergence of structure 20

2.3.2 The Self-Organizing Map 20

2.3.3 On the emergence of perceptual features using SOM and ICA 22

2.3.4 IVGA as a model of the differentiation of Gärdenfors' quality domains 23

2.3.5 A probabilistic model of embodied lexical development 23

2.4 Discovering Conceptual Relations from Texts 24

2.4.1 Required Qualities 26

2.4.2 Approaches 26

2.4.3 Summary of the Reviewed Relation Discovery Approaches 30

2.5 Conclusions 30

3 Assessing Similarity of Emergent Representations **34**

Juha Raitio, Ricardo Vigário, Jaakko Särelä and Timo Honkela . . . .

3.1 Connectionist networks and representation of content 34

3.2 Measuring the similarity of state space representations 35

3.3 Emergent, unsupervised representations and association 36

3.4 Methodology 36

3.4.1 Self-Organizing Maps 36

3.4.2 Associative Mappings with the SOM 37

3.5 Experiments 38

3.5.1 Data 38

3.5.2 Testing procedure 38

3.6 Results 39

3.6.1 Similarity of the emergent representations 39

3.6.2 Effects of scaling of the map capacity 41

3.7 Discussion 42

3.8 Conclusions 43

4 Modeling Multimodal Concepts **45**

Ville Tuulos, Jukka Perkiö and Timo Honkela . . . .

4.1 Introduction 45

4.2 Burden of Complexity 46

4.2.1 Computational models of mind 46

4.2.2 Representations and LOT 48

4.3 Perceptual Multimodality 49

4.4 Experiments 50

4.4.1 Measuring image similarities 50

4.4.2 ICA 50

4.4.3 ICA for image data 51

4.4.4 Creating the filter set 51

4.4.5 Using the filter outputs 52

4.4.6 Assessing the method 54

5 On Representation of Action within Real-World Situations **56**

Kevin I. Hynnä, Mathias Creutz, Tarja Knuuttila, and Timo Honkela . . .

5.1 Introduction 56

5.2 Ontological and epistemological assumptions 57
5.3 Embodied cognition approach 59
5.3.1 Sense-think-act cycle 59
5.3.2 Frame-of-reference problem 59
5.3.3 The action selection problem 60
5.3.4 Loosely coupled, parallel processes 61
5.3.5 Examples of emergent complex behavior 61
5.3.6 Scalability and self-awareness 62
5.3.7 Embodied meaning in Neural Theory of Language 63
5.4 Dynamical Systems Theory Approach 64
5.4.1 Dynamical Systems 65
5.4.2 Agents and Environments as Dynamical Systems 66
5.4.3 Conceptual Space Theory as a Dynamical System 67

6 Concept Learning by Formation of Regions **71**

Jan-Hendrik Schleimer, Mikko Berg, Jaakko Särelä, and Timo Honkela . . .

6.1 Introduction 71
6.1.1 Prototype theory 73
6.2 Possible implementations in conceptual spaces 74
6.2.1 K-means clustering 75
6.2.2 Density estimation 75
6.2.3 Hierarchical clustering 76
6.2.4 Bayesian mixture model 76
6.3 Clustering of color spaces into concepts 77
6.4 Discussion 79
6.5 Acknowledgements 80

Preface

This book is based on the seminar "Adaptive and Statistical Approaches to Conceptual Modeling" that was organized by the Laboratory of Computer and Information Science at Helsinki University of Technology during autumn 2003. The participants of the seminar were senior and young researchers who all had previous experience on either the conceptual modeling aspect or the methodological aspect of the topic. One of the starting points for the seminar was the long tradition on statistical machine learning research in the laboratory. (Kohonen, Oja, etc.) Another important point of view was the special emphasis on Peter Gärdenfors' theory on conceptual spaces (Gärdenfors, 2000).

Figure 1: Participants of the second miniconference of the seminar on 16th of January 2004. Standing, from the left: Vibhor Kumar, Rong Yang Zhi, Mikko Määttä, Ville Tuulos, Jukka Perkiö, Sakari Virkki, Mathias Creutz, Juha Raitio, Jaakko Särelä, Kevin I. Hynnä, Ricardo Vigário, Jan-Hendrik Schleimer, Harri Sulkava and Mikko Berg. Sitting, from the left: Krista Lagus, Peter Gärdenfors and Timo Honkela. Persons missing from the photograph: Tarja Knuuttila and Aarno Lehtola. Photograph by Aarno Lehtola.

Chapter 1

Adaptive and Statistical Approaches to Conceptual Modeling

Timo Honkela, Krista Lagus and Jaakko Särelä

Conceptual modeling is a task which has traditionally been conducted manually. In artificial intelligence, knowledge engineers have written descriptions of various domains using formalisms based on predicate logic and other symbolic representations such as semantic networks and rule-based systems. The development of expert systems in 1980s was a notable example of such efforts. As modern, related attempts, the Semantic Web and knowledge representation formalisms like extendable markup language (XML) can be mentioned.

It seems that the complexity and changing nature of most of the domains makes such formalisms problematic in many real-world applications. Our basic aim is to provide the means for a more or less automatic process of concept formation. This will facilitate both cost-effective development of knowledge-intensive systems as well as serve as a good basis for systems that can update themselves taking into account changes in the domain of interest. We also attempt to tackle some traditional issues in philosophy of language and epistemology. One of our initial contentions is that semantics cannot be adequately handled within a "Language of Thought" framework by Fodor (1975) and others.

The traditional symbolic approach has concentrated on the linguistic domain. Therefore, the models often lack the connection to the perceptual domain. A certain derivative of Platonic idealism has been in use: it has been assumed that knowledge can be represented as propositional structures that are based on static shared concepts. It has been commonplace to assume that there is a one-to-one correspondence between words and concepts (early Wittgenstein). Moreover, it is assumed that a concept refers unambiguously to a number of distinct objects or events in the reality. The individual differences are assumed to be small and explained as errors. A similar notion in linguistics is the distinction into competence and performance (Chomsky 1965). Due to the lacking link to the perceptual domain it has been natural to use static models. Traditional model based on symbolic representations lack, among other things, symbol grounding (cf, e.g., Harnad 1989).

A problem often neglected in symbolic knowledge representation tradition is subjectivity.

For us, it seems more and more evident that major portions of individual conceptual systems are learned. Due to the individual and cultural differences, e.g., in phoneme categorization and color naming, it is hard to believe that concepts could be modeled with static structures without making use of adaptive processes.

The ease with which humans classify and describe patterns often leads to the incorrect assumption that this capability is straightforward. There has been a large body of research on pattern recognition that has highlighted the complexity of our perceptual processes. For instance, to interpret a natural scene, a human being utilizes the information provided by 6 million cones and 120 million rods in the retina (Kalat 1995). Similarly, a complex analysis problem arises from the situation in which a digital image of several million picture elements need to be analyzed. Interpreting any pattern of the basic perception as an object is made difficult, for instance, by different lighting conditions, partial occlusion, pattern distortions. In his seminal work, Marr (1982) developed a computational theory of visual processing including the representation of objects to facilitate recognition.

Statistical and adaptive approaches have been used successfully to develop artificial pattern recognition systems (see, e.g., Schalkoff 1992, Theodiridis and Koutrombas 1999). There are computational models of many human perceptual processes including vision, speech and touching. Industrial applications include speech recognition, recognition of handwritten characters, computer vision for quality analysis and fault detection, image recognition (e.g. faces), and robot grasping.

The traditional approaches in artificial intelligence and conceptual has been based on the idea that the world consists of discrete objects (Platonic idealism). In contrast, the statistical and adaptive approach follows the Aristotelian empiricist tradition (Dreyfus and Dreyfus 1990). Maturana and Varela (1980) and Von Foerster (1981) point out that cognitive, living agents construct their description of the world, and this description consists of constructed categories such as objects and events along with their associated subcategories. Each of those constructions is subjective but at the same time their formation is based on the interaction with other agents as well as artefacts that reflect the structural characteristics of the constructions of other agents.

The statistical approaches have concentrated on modeling the perceptual processes. However, evidently the cognition includes both linguistic and perceptual skills. There have been attempts to create hybrid models that apply statistical methods on the lower-level perceptual processing but also allow a symbolic interpretation (see e.g. Wermter and Sun 2000). An attempt to provide a concise descriptive framework for the integration of the neural, conceptual and symbolic levels of representation is presented by Gärdenfors (2000).

In summary, we consider the development and application of adaptive and statistical methods for conceptual modeling to be particularly important. Probability theory is an excellent model for dealing with noisy and ambiguous phenomena, such as language. Probabilistic models of linguistic structure exist at every level (phonology, morphology, the lexicon, syntax, discourse). Furthermore psycholinguistic research has shown that probabilities play an important role throughout language comprehension, production and learning.

Statistical and probabilistic approaches are nowadays rather widely used in natural language processing (Manning and Schütze 1999). Specific examples include methods such as Bayesian methods for spam filtering, Latent Semantic Analysis (LSA) for information retrieval applications, and Hidden Markov Models for speech recognition.

In this publication, adaptive and statistical methods are considered within the area of semantics, in particular. An important aspect is emergence: how representations emerge through a learning or analysis process. Specific topics include emergence of structure (Chapter 2), similarity of emergent representations (Chapter 3), modeling of concepts that are based on multimodal domains (Chapter 4), representation of action (Chapter 5), and category learning (Chapter 6).

References

Chomsky, N.: *Aspects of the Theory of Syntax.* MIT Press, Cambridge, 1965

Harnad, S.: Minds, Machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, 1, 1989, pp. 5-25.

Baker, G.P. and Hacker, P.M.S.: *Language, Sense & Nonsense - A Critical Investigation into Modern Theories of Language.* Basil Blackwell, Oxford, England, 1984.

Dreyfus, H. and Dreyfus, S.: Making a Mind versus Modeling the Brain: Artificial Intelligence back at a Branch-Point. M.A. Boden (ed.), *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990. Published also in *Artificial Intelligence*, 117:1, 1988.

Fodor, J.A.: *The Language of Thought.* MIT Press/Bradford, Cambridge, Massachusetts, 1975.

Gärdenfors, P.: *Conceptual Spaces.* MIT Press, 2000.

van Gelder, T.: Why Distributed Representation is Inherently Non-Symbolic. G. Dorffner (ed.), *Konnektionismus in Artificial Intelligence und Kognitionsforschung.* Berlin: Springer-Verlag, 1990, pp. 58-66.

Hallett, G.L.: *Language and Truth.* Yale University Press, New Haven, 1988.

Harley, T.A.: *The Psychology of Language.* Psychology Press, 1995.

Manning, C.D. and Schütze, H.: *Statistical Natural Language Processing.* MIT Press, Cambridge, Massachusetts, 1999.

Marr, D.: *A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H.Freeman, San Fransisco, 1982.

Maturana, H. R. and Varela, F. J. *Autopoiesis and cognition: The realization of the living.* Reidel, Dordrecht, 1980.

Ritter, H. and Kohonen, T. : Self-Organizing Semantic Maps. *Biological Cybernetics*, 61, 1989, pp. 241-254.

Schalkoff, R.: *Pattern Recognition. Statistical, Structural and Neural Approaches.* John Wiley & Sons, 1992.

Theodiridis, S. and Koutrombas, K.: *Pattern Recognition.* Academic Press, 1999

Von Foerster, H. (1981). Notes on an epistemology for living things. *Observing Systems.* Intersystems Publications, pp. 257-271. Originally published in 1972 as BCL Report No 9.3., Biological Computer Laboratory, University of Illinois, Urbana.

Wermter, S. and Sun, R. (eds.). *Hybrid Neural Systems.* Springer, 2000.

Wildgen, W. and Mottron, L.: *Dynamische Sprachtheorie: Sprachbeschreibung und Spracherklärung nach den Prinzipien der Selbstorganisation und der Morphogenese.* Brockmeyer, Bochum, 1987.

Chapter 2

Emergence of Structure and Relations in Conceptual Representations

Krista Lagus, Aarno Lehtola, Mikko Määttä and Sakari Virkki

2.1 Introduction

Concepts have been studied with different aims and emphasis in many fields of science, from philosphy, psychology and linguistics to the very technical and practical fields. In the former fields the emphasis is on the internal cognitive representations in human minds: what are the representations like and how do they form. In the technical fields concepts and structures related to them are needed in many practical applications where information must be represented and exchanged.

We will review literature on concepts and their structure from four seemingly distant perspectives. The aim is first to clarify the concept of structure in the context of conceptual representations, and to create a unifying approach for the remaining chapter. Second to view how structures are explained and presented in selected psychological theories. Third to view how structures might emerge from natural or textual data in selected statistical and neural learning methods. Fourth to view how structures are discovered from textual sources for the purposes of practical engineering or language applications.

By taking a wide perspective on the question of concepts and their structure we wish to aid the cross-fertilization of ideas in different disciplines.

2.1.1 What is structure?

One can easily find many intuitively appealing definitions of structure in various sources, including the internet. There seems to be no common agreement on the concept of structure, which can be readily noted from the following list of sample definitions:

- ”Structure is sort of permanence that makes changes difficult.”
- ”Structure is a complex of events interacting to form a stable pattern.”

- "Structure is dually composed of schemas and resources."
- "Structure is the path of least resistance."
- "Structure is an entity formed by the influence the parts have on each other and on the whole."
- "Structure is defined as a locally regenerative pattern integrity of Universe."
- "The concept of "structure" is a complex one and I don't intend, for the moment, to explore its many depths and facets in great detail. One way of thinking about the concept, however, is to see it as a framework of rules and relationships, in the sense that all relationships are governed in some way by rules of behaviour."

While all the definitions bear some relevance regarding the subject, most of them are not elaborable for further use. In order to tackle "emergence of structure" instead of merely "structure" we needed more precise and structured conceptual tools. Van Aken (1978) has presented a structured set of "system concepts" that we find very useful in defining and clarifying the objective of this chapter.

2.1.2 Definitions

- **Element** — "An element is the smallest entity considered in an argument." (Aken, 1978)

  This definition brings some insight in the controversy between concepts, properties and attributes. It is all about the chosen level of abstraction and levels of detail. From a more detailed level attributes can be considered as concepts and from a higher level concepts can be considered as attributes or properties.

- **Set** — "A set is a collection of elements" (Aken, 1978).

- **Relation** — "Relation is a link, connection, ratio, proportion, act, transition, transaction, etc. that can connect elements."

  This (our) definition of relation is a tentative one and lacks the compactness of definitions of van Aken, but is sufficient here. From the example in Fig. 2.1 one can easily note that most of the relations are dualistic in nature. For example one can state that concept shapes "includes" concept circle, but concept circle "is member of" concept shapes. It is arguable whether relations "includes" and "is member of" are different aspects of one relation or two directed, distict relations. We take a dualistic view and consider "includes" and "is member of" as two different aspects of one relation. This is motivated by the fact that "includes" always carries "is member of" as it counterpart. This also means that the interpretation of a relation depends on from which end of the relation one reads it.

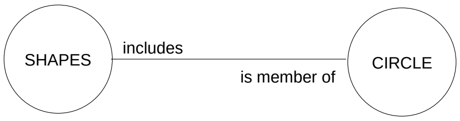

Figure 2.1: A sample relation.

There are also relations that are seemingly not affected by this dualism. For example relation "is equal to" is the same in both directions. Without going any further our main conclusion here is that from now on, relation is regarded as the constituent of structure.

- **System** — "A system S is a set E of elements with a set R of relations between the elements, R having the property that all elements of E are directly or indirectly related." (Aken, 1978)
- **Structure** — "The structure of a system S is the set R of relations of its elements with other elements. The internal structure Ri is the subset of R containing the relations between the elements of S. The external structure Re is the subset of R containing relations of S with elements outside S." (Aken, 1978)

  This definition of structure finally captures the objective of this chapter in an elaborate way. However, it is important to note that since the division of relations into internal and external relations depends on the object system, demand for discipline and preciseness in the definition of it (object system) is crucial. For example, if we select the *society* as an object system then relations between humans are considered internal, but if we take a *human* as an object system then relations between humans are considered external. Being precise with the object system clears a lot of confusion between internal and external relations. This definition of structure is used to evaluate selected psychological theories, selected technologies and text analysing methods in respect to structures in the Sections 2.2, 2.3, and 2.4.
- **Concept** — Traditionally concept is defined as follows: "Concept is constituted by two parts: its extension which consists of all objects belonging to the concept, and its intension which comprises all attributes shared by those objects."

  Surprisingly this definition totally ignores the context of the concept. We could not help thinking that also the context is a key constituent to interpret the meaning of a concept. Since concept is also a data structure we decided to apply the definitions above to define the concept of concept:

  "Concept is a data structure defined by its set E of elements, its internal relations Ri and its external relations Re that connect it to other concepts."

  As an example, one can think of elements E and relations Ri forming the properties and attributes of the concept, and other concepts and relations Re forming the context of the concept. But, as was noted in connection with the definition of Element, depending on the level of abstraction attributes can be considered as concepts and concepts can be considered as attributes or properties. This means that the internal relations Ri are also, in addition to Re, in fact relations between concepts. The relevance of this point will become apparent in Section 2.2.

  This is also a new way to define the concept of concept without the use of intension and extension. It also gives a role to internal relations Ri in the interpretation of concept: even if two entities (concepts) share the same attribute, the attribute can be shared in many different ways as described by the internal relations. Thus sharing the same attributes does not necessarily mean that the intensions are the same since internal relations can differ.

  This definition of concept is on the constructive-operational level. At this point and for the purpose of this chapter we try to avoid a more semantic and more teleological definition of concept of concept. This definition serves the purpose to understand the role of structures (ie. relations) within the concept of concept itself.

As an example of the application of the definitions above in Fig. 2.2 we present an EAR (entity-attribute-relationship) model of how the task of writing this chapter can be structured.

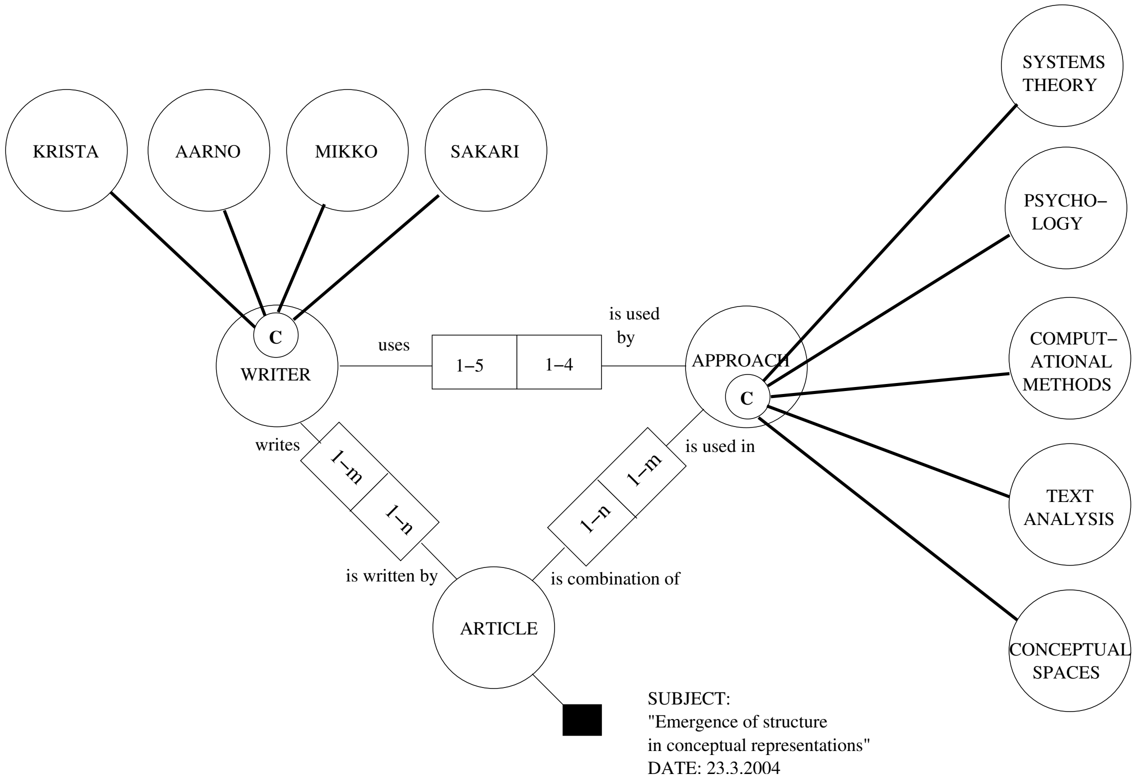

Figure 2.2: An example of conceptual structure depicted using the EAR formalism. A circle marks an entity (in practice often concepts), The letter **C** corresponds to classification, a thin line a lateral relation, a thick line abstraction/hierarchy relation (Re), the black box is attribute that describes the properties of the concept, and the boxes on the relations are attributes that describe the cardinality of the relation.

2.1.3 Structure of the remaining chapter

In Section 2.2, we review selected psychological and philosophical theories and look at how concepts and their relations are explained and presented in them. In particular, we wish to find out whether the theories consider the issue of emergence of structure, or whether the idea of emergence is at all compatible with the theory.

In Section 2.3, based on treatments in the previous sections, in particular Gärdenfors' model of conceptual spaces, we suggest a set of subtasks in the emergence of structure. We then review work on several (mainly unsupervised) learning methods, including SOM, ICA, IVGA, and probabilistic modeling, as potential models for performing various subtasks. We discuss how concepts, their internal structure and external relations might emerge autonomously from natural or textual data in these models.

In Section 2.4, we look at ontology discovery from text. The aim in this field is to derive what in the technical domain are called *ontologies*, that is, descriptions of concepts and their relations[1] Typically the concepts themselves are given, and the task is to discover relations between the concepts for the practical purposes of various language applications. Manual specification of concepts and their relationships is slow and expensive and there is a wide interest in developing machine learning tools for the task. Nevertheless currently it is common to utilize considerable amounts of heuristics to obtain the desired quality and coverage of the results.

Finally, we summarize our findings and contributions.

2.2 Conceptual Structure in Psychology, Philosophy and Linguistics

In this section we will discuss views of what concepts are. More particularly, in In Section 2.2.1 we look at views of what concepts are. Various theories of their internal structure are then briefly introduced in Section 2.2.2. In Section 2.2.3 we discuss how the theories treat the formation of the structure.

A more specific cognitive model, the Conceptual Spaces model of concepts and their structure is considered in Section 2.2.4. In Section 2.2.5 we will briefly discuss some linguistic theories of case frames that have a connection to external relations. Finally, we summarize our findings on relations in concept theories in Section 2.2.6.

2.2.1 Concepts, Structure and Relations

In modern cognitive psychology and mostly in philosophy, too, concepts are considered to be mental representations instead of, for example, abstract entities (Laurence & Margolis, 1999). Concepts can be thought of as mental representations that correspond to a class of objects in the world and store information about those objects (Murphy, 2002). This characterization focuses on the concepts of concrete objects and leaves out more abstract concepts whose referents can't be identified with objects. This is however the focus of most of modern psychological research of concepts (Murphy, 2002), so this chapter will also focus on these kinds of concepts.

The internal structure of concepts (Ri)

Most theories of concepts in psychology and philosophy assume that concepts have some kind of internal structure (Laurence & Margolis, 1999), so it is natural to ask what kind of structure concepts have. Laurence and Margolis (1999) differentiate two ways in which concepts can have structure. Although Laurence and Margolis talk explicitly about relations only in the second case, both of their models can be interpreted to claim that the internal structure of a concept is constituted by its (internal or Ri) relations to other concepts. These other concepts can also be called the attributes or properties of the concept in question, although attributes and properties are usually thought of as just concepts themselves.

[1] It should be noted that the use of the term "ontology" differs in the technical and philosophical domains.

The containment model. The first model Laurence and Margolis dub the Containment Model. According to this model, a concept is a structured complex of other concepts in the sense that it has some other concepts as its proper parts. For example, if the concept C has the concepts X, Y and Z as its proper parts, C can be said to be a structured concept. This kind of structure implies that whenever one entertains the concept C, one also has to entertain the concepts X, Y and Z. The relation between the concepts in this case, then, is the containment relation.

The inferential model. Laurence and Margolis call the second view the Inferential Model. In this case, a concept's structure is constituted by its inferential relations to other concepts. So, if one knows, for example, that concept C has a definition (C $=_{df}$ X, Y, Z), and one knows also that a certain object can be categorized as a C, then one can infer that the concepts X, Y and Z apply to the same object (this follows from the nature of definitions). Note that on this view, entertaining the complex concept doesn't entail entertaining the constituent concepts. One can perfectly well think about C:s without thinking about X:s, Y:s and Z:s.

Relations to other concepts (Re)

In addition to what were in Section 2.1.2 dubbed internal relations (Ri) that constitute a concept's structure, a concept can have relations to other concepts that are better described as external relations (Re). By external relations we mean the relations between concepts that are not the property or attribute relations "is-contained-in" or "can-be-inferred-from". In psychology, by far the most studied of the external relations are the taxonomic or hierarchical relations "is-a-superordinate-category-of" and "is-a-subordinate-category-of", so we will focus on them here.

The following description of the hierarchical structure of the human conceptual system is based on Murphy's (2002) exposition. The hierarchy consists of inclusion relations between categories or concepts. For example, the category of animal includes all mammals, which in turn includes all dogs and so forth. Psychologists usually consider one particular level of the conceptual hierarchy to be especially important. This level is called the basic level. The levels above the basic level are called superordinate levels and the levels below are called subordinate levels. Below are examples of concepts at the different levels:

The three levels of the conceptual hierarchy

1. Superordinate level (furniture)
2. basic level (chair)
3. subordinate level (an ergonomic working chair)

The basic level. The basic level is a privileged level in the sense that it is the level whose concepts people mostly use. Here are a few examples: People mostly talk about "chairs" and rarely about "furniture" or specific species of chairs when they are talking about a chair. Parents also use mostly basic level categories when talking to their children. Another example of the basic level effect is that when asked, people list many more common features for concepts at the basic level than for objects at the superordinate level, and only a few features more for concepts at the subordinate level.

Explanation of the basic level effects. Murphy (2002) explains the basic level effects in terms of the informativeness and distinctiveness of basic level concepts. He calls the explanation the differentiation explanation.

Informativeness means that the concept in question contains a lot of information about the category. Both basic level and subordinate level concepts usually contain a lot of information that applies to all (or most of the) category members, but the superordinate category doesn't. When one knows an object to be a piece of furniture, one doesn't know much about its features (chairs, beds and lamps are quite different).

Distinctiveness refers to the fact that basic level categories don't have many of the same features that other categories at the same level have. Chairs and lamps (both are furniture) don't share many features. In contrast the subordinate categories do have a lot of features in common, so they are harder to distinguish.

Put together, informativeness and distinctiveness explain why people usually prefer the basic level concepts.

2.2.2 Theories of the Internal Structure of Concepts (Ri)

Next we will present brief summaries of the major theories of the internal structure of concepts that philosophers and psychologists have considered.

The classical or definition theory

Until relatively recently, the idea that the mental representations of concepts are definitions was widely accepted in both philosophy and psychology. Murphy (2002) claims that this view was implicitly assumed in psychological concept learning studies from the 1920's to the 1970's. Laurence and Margolis (1999) summarize this view as the claim that concepts encode a set of necessary and sufficient conditions for the concepts' application.

Criticism of the definition theory. In the 1970's, this view became unpopular in both psychology and philosophy. Laurence and Margolis (1999) give six main reasons for the demise of the definitional or classical view of concepts. These can be divided into philosophical and empirical (psychological) reasons.

The philosophical objections are:

- Very few definitions have been found. It has been extremely difficult to find definitions for most of the concepts people use by analyzing the concepts.
- The analytic/synthetic distinction is not principled. Quine (1953) has argued that in scientific theories there is no principled way to make the distinction between analytically true (true in virtue of meaning alone) and synthetically true (true in virtue of something else than meaning, experience, for example) statements. Because definitions are considered to be analytically true statements, Quine's critique has been interpreted by many to imply that there is no principled way to form definitions, either, and hence concepts can't be definitions.
- Concept possession can't be a matter of knowing a description (or any description for that matter), because people often have erroneous beliefs about the entities that fall

under a concept, and sometimes they are completely ignorant about those entities' properties. In spite of this, in many such cases one would be inclined to say that these persons nevertheless possess the concepts in question.

The empirical objections are:

- Differences in the complexity of concepts don't show in psychological processes. For example, the processing of allegedly structurally complex concepts doesn't take longer than allegedly simple concepts, which has been demonstrated in reaction time studies.
- The boundaries of categories are not sharp. The definition theory predicts that the boundaries of concepts' extensions (the class of entities that fall under the concept) should be sharp: No fuzziness is allowed. This is, however, not the case. People are uncertain whether or not certain borderline cases belong to a category or not.
- Some members of a category are more typical than others. The definition theory also predicts that belonging to a category is an all-or-nothing matter. A large number of psychological studies show that people generally take some members of a category to be better examples of the category than others. These phenomena are called typicality effects.

The prototype theory

Mainly for the last two reasons the definition theory has been abandoned in experimental psychology of concepts. The dominant views in the last two or three decades have been the prototype theory and the exemplar theory. According to Murphy (2002), in the prototype theory a concept is a mental representation that is composed of a list of features. Thus, a prototype resembles a definition in many ways. The crucial difference is that the features that a prototype contains are not considered to be necessary for the concept. Each feature has a weight value that reflects its importance in the category. In fact, the weight reflects how common the feature is among the category members.

The exemplar theory

The exemplar view differs from the prototype view in that no summary representation is assumed. Instead, the category is represented by multiple exemplars of the category members. It is noteworthy that the exemplars themselves are not summary representations of encountered individuals. According to the exemplar theory, people store actual, separate encounters with a certain dog, for example. (Murphy, 2002)

The knowledge approach

A third major strand in recent psychological study of conceps is the knowledge approach, also known as the theory theory (Murphy, 2002). According to Murphy (2002) the knowledge approach is not a full-fledged theory of concepts. Rather, the researchers emphasize the importance of background knowledge to various conceptual phenomena.

For example, Wisniewski & Medin (1994) studied the effect of background knowledge on feature analysis of stimuli. They found that the subjects clearly analyzed the stimuli (pictures drawn by children) differently depending on the information they got prior to the presentation of the stimuli. One group of subjects were told that the pictures were made by either creative or noncreative children, whereas another group were told only that the pictures were made by two groups of children. The first subjects listed many more abstract features of the pictures than the second subjects.

This study highlights the difficulty of the experimental research of human conceptual representations. It also sheds doubt on some other experimental results in the field, such as those about different numbers of features shared on different category levels, discussed earlier in Section 2.2.1.

Conceptual atomism

The above theories all assume that the mental representations of categories are structured in some way or another. In the literature, there is one notable exception to this consensus. Conceptual atomism is the view that most lexical concepts don't have internal structure. Lexical concepts are concepts that usually correspond to lexicalized morphemes in natural languages (Laurence & Margolis, 1999). Conceptual atomism has been most prominently defended by Jerry Fodor (1998). Fodor's alternative to the dominant theories is that lexical concepts have no internal structure, hence they are atoms or primitives. Fodor arrives at this conclusion because he sees it as the only alternative that is left after the other theories have been shown to be incorrect. First of all, he agrees with much of the criticism mounted against the definition theory, and second of all, he claims that the concepts of prototype theory (Fodor seems to consider exemplar theory to be a variant of the prototype theory) aren't *compositional*, which means very informally that they can't be combined in the right way to form complex concepts. The main problem with prototypes is that the prototypes of complex concepts can't be derived from the prototypes of constituent concepts in any lawful way, which in turn is required to explain (among other things) the productivity of language and thought.

2.2.3 Formation of Conceptual Structure

The question of how conceptual structure emerges or is formed can be viewed from the perspective of internal or external relations. So two questions about the formation of structure can be asked. The first question concerns how, given a certain set of features or concepts, these features or concepts are combined (by forming relations between the concepts) to form new concepts, i.e. how the internal relations of a concept are created. This is the process normally referred to as concept learning. The second question concerns the formation of a concept's external relations.

Concept learning

The first question is the most studied one in the psychology of concepts. Laurence and Margolis (1999) describe the process of concept learning as the process of assembling or combining together the features of the concept. The learner notes which features go together in the world and builds the concept out of these correlations.

The concept learning experiment. The above view is compatible with the description that Murphy (2002) gives of the typical concept learning experiment. According to

Murphy, the concept learning experiment is a situation in which the subject of the experiment is presented with a series of stimuli, often simple geometric shapes that vary on a few dimension like shape and color. Usually the task of the subject is to categorize the stimuli into two categories. After each stimulus, the subject makes a categorization decision and gets feedback from the experimenter concerning the correctness of the decision. This pattern is repeated until the subject can categorize the stimuli correctly or a certain number of trials is reached, in which case the subject either has learned the concept or hasn't. Fodor (1981) emphasizes that the subject makes hypotheses about the features that are relevant for the concept, tests these hypotheses and adjusts them according to feedback.

Category construction. The noteworthy detail in the concept learning experiment is that the subject receives feedback on her categorization decisions. This is, however, usually not the case in real-life concept acquisition (Murphy, 2002). This unsupervised concept learning situation Murphy (2002) dubs category construction. In psychology, this is a much less studied strand of concept acquisition than supervised learning.

The formation of external relations

We now turn to the third question presented at the beginning of this section, the question of how the relations between concepts that are external to the structure of a concept are formed. In Section 2.2.1 it was noted that psychologists have mostly studied the external relations associated with conceptual hierarchies. Next we will give a suggestion of how these relations could be formed based on Murphy's (2002) description of the conceptual hierarchy.

Formation of hierarchical relations. Murphy claims that psychological studies favor the view of hierarchy representation according to which hierarchical relations are not explicitly represented. According to this view, concept a is the superordinate concept of concept b if the features of a are a subset of the features of b. Thus, a subordinate concept has all the features of its superordinate plus some others. So maybe super-/subordinate relations don't have to be learned explicitly. It could be enough to learn the concepts separately, and when the relations are needed they can be constructed from the structure of the concepts.

2.2.4 Relations in the Conceptual Spaces Model

Gärdenfors's model of conceptual spaces (Gärdenfors, 2000) cannot be reduced to any of the standard philosophical theories of concepts. However, as he points out, several aspects of the Prototype theory of concepts can be explained with his theory of properties. The theory proposes a conceptual level of representation between a neural level and a symbolic level, and suggests that the connection between the conceptual and the neural levels is mediated by prototypes.

Central terms regarding the Conceptual level

- **Quality dimensions**. Quality dimensions are geometric or ordered representations of qualities or possible values of attributes. As examples of quality dimensions closely connected to the sensory system Gärdenfors mentions temperature, weight, brightness, and pitch. An example of a more abstract dimension is physical or social force that may be exerted. By virtue of the geometrical representation, the dimensions provide a way for making judgements of similarity and difference by measuring distances. The similarity judgements in turn give rise to an ordering relation among stimuli. Gärdenfors emphasizes the phenomenal or psychological rather than scientific nature of these dimensions: "When the dimensions are seen as cognitive entities-that is, when the goal is to explain naturally occurring cognitive processes-their geometrical structure should not be derived from scientific theories that attempt to give a 'realistic' description of the world, but from psycho-physiological measurements that determine how our phenomenal spaces are structured."

- **Domain** is a set of integral quality dimensions that are separable from all other dimensions. For example, colour is a domain consisting of the integral dimensions hue, chromaticity, and brightness that are presumably separable from other quality dimensions. As another example, the shapes of physical objects form a domain.
- **Property** is a well-behaved region in a single domain. For example, the colour blue is a property. Dynamic properties are ones that describe actions, and in their representation the dimension of force (physical or social) might be instrumental. Similarity judgements can be made also in the space of dynamic properties, for example, walking is more similar to running than to throwing. Gärdenfors mentions that properties may, in principle, also be functional, e.g., relating to the uses or affordances of objects, although this possibility is not discussed in detail.
- **Concept** is a well-behaved region in several domains, as opposed to a property that is a region in a single domain. For example, apple is a concept. Not all domains are involved for all concepts. Moreover, each domain may have an associated salience weight for a particular concept.
- **A particular instance** of a stimulus corresponds to a point in a conceptual space.

On the origins of domains and quality dimensions

According to Gärdenfors, some domains and dimensions are needed to begin with. By learning, new dimensions can be added. Some quality dimensions are culturally determined. As an example, Gärdenfors mentions time which in some cultures is seen as circular, while in others, linear. Some dimensions are introduced by science, such as Newton's differentiation of weight from mass.

Gärdenfors considers examples of neural network methods that could be utilized to represent quality dimensions, among them the SOM and the MDS.

2.2.5 Deep Cases in Case Grammars

Since the previous theories have mainly concerned with the constituency of concepts (Ri) or their hierarchical relations, we wish to examine more closely what is meant by the external relation, that is the relation between separate concepts in (computational) linguistic theory.

Deep cases in case grammars address external/context-sensitive semantic relations of entities. Deep cases are assigned to linguistic entities extracted from sentence analysis to mark-up their semantic roles with respect to each other or the sentence wholeness. If the analysis is based on dependency grammar the semantic relations are between word centered entities. Bruce (1975) reviews several case systems and their deep cases. One of the most famous is Charles Fillmore's case system, which is suitable for describing events and which consists of the following deep cases:

- Agent — the instigator of event
- Counter-Agent — the force or resistance against which the action is carried out
- Object — the entity that moves or changes or whose position or existence is in consideration
- Result — the entity that comes into existence as a result of the action
- Instrument — the stimulus or immediate physical cause of an event
- Source — the place from which something moves
- Goal — the place to which something moves
- Experience — the entity which receives or accepts or experiences or undergoes the effect of an action

Fillmore's case system can explain sentences like "John opened the door with a chisel", in which John constitutes an agent, door serves as an object and chisel as an instrument. The main verb "opened" refers to the event with respect to which the relations are found. Joseph Grimes has developed an even more sophisticated case system with thirteen deep cases for discourse analysis.

Roger Schank's case system is a famous one, as well. His formulation implies a conceptual structure built out of actions and their role fillers. Such actions include primitive acts like moving of body parts (MOVE), building of thought (MBUILD), transfer a physical object (PTRANS), and transfer of mental information (MTRANS). Primitive acts together with the conceptual cases are regarded as the components of meaning representations. These representations are claimed to be unique in the sense that irrespective of the original language of surface sentences, if their semantics are equal, also the conceptual structures are equal. The same principle is in use in so called interlingual machine translation systems, that use an intermediate semantic language for translation. A successful example of the use of deep cases is in the Mu system (Hutchins 1995) that was accepted for operational use in 1986 to translate abstracts for Japanese Information Center for Science and Technology. It has an extensive case system of over 40 cases.

2.2.6 Summary of Relations in the Discussed Theories

Theory	Elements	Internal relations	External relations
Definition theory	Concepts[1]	Feature relations: Is-contained-in, Can-be-inferred-from	Not addressed
Prototype theory	Concepts[1]	Feature relations: Is-contained-in, Can-be-inferred-from	Hierarchical relations, e.g. Is-a-superordinate-category-of
Exemplar theory	Exemplars?	Is-an-exemplar-of-category-X?	This issue is not really addressed
Knowledge approach	Concepts?[1]	The distinction between internal and external relations is even more fuzzy.	All sorts of relations between concepts
Conceptual atomism	Concepts[1]	Non-existent. Relations to other concepts don't constitute a concept's structure.	This issue is not really addressed
Conceptual spaces	Quality dimensions, Domains, Properties, Concepts	Property is a region in one domain. Concept is a region in several domains.	This issue is not really addressed
Case grammars	Word-centered concepts	Not addressed	Deep cases, e.g. Agent, Counter-agent, Object, Result, Source, Goal, Experience

1) It seems that no distinction is made between concepts, features, attributes, properties etc. at least in definition and prototype theories and in conceptual atomism.

2.3 The Emergence of Structure in Computational Methods

There exist a large variety of computational methods that can be applied to learning models for a system based on the examination of data produced by the system (cf. e.g. Haykin 1999; Gelman & al, 1995). The human conceptual system can be treated as such a system. Perceptual data, such as natural images and speech, has been used mostly for modeling the emergence of low-level perceptual features. Reaching higher-level processes by learning from data while starting from unconstrained perceptual data sets and very general model families has nevertheless proven hard (for theoretical considerations for the reasons for this, see Tuulos et al, 2004). Therefore it is common to use e.g. artificial agent simulations, or lately, robots for examining models that operate also on a higher conceptual level. Another approach to modelling higher-level processes is to start with text data as input, suitably preprocessed; Section 2.4 examines this approach for the learning of (external) relations.

Table 2.1: Subtasks and potential methods in the emergence of structure.

Tasks	Potential Methods	Examples
Differentiation into domains	IVGA, MICA	-
Formation of features / quality dimensions	SOM, ICA	X X
Categorization or clustering	SOM, other clustering methods	X X
Which domains or features to connect a concept to (the Attribute relation)	Probabilistic evidence accounting	X
Hierarchical relations: super / subordinate category	SOM, hierarchical clustering	X
Other external relations	N/A	(See Sec. 2.4)

In this section we will consider the emergence of structure mainly in the context of Prototype theory of concepts and Gärdenfors' conceptual spaces, both discussed in Section 2.2. We will apply some of the terminology introduced in Section 2.1.

2.3.1 Suggested subtasks regarding the emergence of structure

The Table 2.3.1 outlines the subtasks in the formation of structured models as we currently perceive them. Moreover, connections are made between each task and methods that have been, or can in principle be, utilized for solving that task. Some of the methods will be discussed using examples from the literature. The list of subtasks is tentative, and the set of connections to methods is not complete, but rather a sample of connections examined in this chapter.

We will now look at some neural networks from the point of view of conceptual emergence, and in particular emergence of relations. We ask the question, whether a particular model type could, in principle, serve as an implementation where relations might emerge, given the right kind of input. By emergence it is meant here that the relations arise in a non-trivial manner from the properties of the data set by the utilization of a particular general-purpose statistical learning algorithm, such as a neural network.

2.3.2 The Self-Organizing Map

The Self-Organizing Map (Kohonen 1982; 2001) is a neural network method that utilizes unsupervised learning for obtaining an ordered representation of a large data set. A SOM consists of a set of prototypes in the input space and a (typically 2-dimensional) lattice of fixed connections that defines the neighbours for each prototype in the output space. During learning the prototypes move in the input space so that they sample the input signal space in an orderly fashion, roughly approximating the density of the samples in the input space. As a result, the prototypes reflect common patterns in the input data. Moreover, they form an ordered representation of the data set in the output space, the map lattice: any two neighbouring prototypes are generally very similar. While moving on the map lattice, the properties of the models (prototypes) and of the corresponding input data change gradually. The ordering of the map depends on the choice and weighting of the input features, and the statistical properties and dependencies in the data set given as input.

SOM as a model of the internal and external relations of a concept

In a seminal article on the use of SOM for semantic modelling, Ritter & Kohonen (1989) organized words with the SOM using as data three-word sentences generated from an artificial grammar. Later (Honkela et al, 1995) applied a similar method for the text of Grimm tales. We will highlight some of the results of the latter experiment. The information collected for each word consisted of its averaged context (+-1 words) in the whole data set. The word representations were then organized on a two-dimensional SOM, resulting in a map where an implicit ordering of syntactic and semantic word classes could be observed. For example, in the noun region there was a subregion for humans (Hans, woman, man, king, child, son, daughter, mother, father, wife), and next to it in the "other" category a region of pronouns referring to humans (she, he, they, we, I, you).

It thus appears that on a SOM organized using a suitable data set and a suitable feature selection, a concept can be viewed to correspond to a map region spanned by one or more model vectors. One can also observe that the hierarchical relationship, e.g. the "superordinate-category-of", is in some cases represented implicitly, as regions on the map.

It does not, however, seem reasonable to assume that a single SOM could be the representation of the totality of concepts. Already Ritter & Kohonen (1989) emphasize that their work is just a demonstration of the potential of SOMs, and that "Any realistic brain maps would need a much more complicated, probably hierarchical model." One reason for this becomes clearer when one takes notice that each prototype of a SOM has internal relations to an identical set of properties. In contrast, many concepts seem to require internal relations to altogether different domains (consider e.g. "book", "society", and "to escape"). This kind of per-concept feature selection is not implemented in the SOM, but must be implemented in some additional way.

Another reason why a single SOM is not sufficient for representing all concepts and their relations comes from the examination of external relations: It would appear that varying suitably the input features and their weighting, any binary relation could in principle emerge and be represented as a similarity relation on the map. However, there are two problems: (1) The relations are not named, i.e. "part-of" is not distinguishable from "subordinate-category-of" and (2) the number of relations that can be represented by neighbourhood connections is very limited, whereas the number of relations a concept can have is very large. There exist possibilities for solving the first problem (such as the use of additional relation maps) but we are not aware of works that have examined such solutions. The second problem can be alleviated with the use of several SOMs that have been ordered based on different sets of input features (or using different feature weightings).

As suggested by Gärdenfors (2000), it seems more appropriate to consider the output space of the SOM as a computational model of a single domain: it is an ordered representation that is made up by integral "dimensions". Kohonen points out in (Kohonen, 1990) that topological closeness and connectivity of representations alleviates the "property inheritance" problem found in semantics and artificial intelligence. He also gives several examples of sensory-level maps: acoustic (tonotopic) maps, phonemes of speech, colors (hue and saturation), all of which have been produced by analyzing particular natural signals. Such might serve as models for Gärdenfors' domains.

Some of the desired qualities with respect to Gärdenfors' domains are as follows: The SOM is able to represent efficiently a possibly sparse data set that contains statistical dependencies between its input features. The output space of the SOM is able to contract, stretch and bend in the input space, roughly following the distribution of the data. Moreover, other than two-dimensional lattices and different neighbourhood topologies may be applied for the SOM. Also differing metrics may be utilized. It is not clear, however, how a specific metric would be derived solely from observing a set of data.

The main challenge we see in the application of the SOMs is how to choose the input features for obtaining a particular, interesting quality domain. This question is outside the scope of the SOM algorithm, and must be determined prior to its application. Note that this is a question regarding the structure of the model that is not included in the concept of "relation", but exists on a lower level of the model.

The modeling of causal action schemas on a hierarchy of SOMs

In (Chaput 2003) a hierarchy of self-organizing maps is used to model causal action schemas, that is, to learn causal relations between actions performed by the agent and states of the world before and after the action. An example of such schema is “Door is open / I close door / Door is closed”.

There exists a predefined set of primitive actions that the agent can take. For each action, a specific action SOM is taught. The input to an action SOM consists of the context (the complete state of environment before the action) and the result (the state after the action). Any particular input is gated to the corresponding action SOM whose action was just performed. These action SOMs form the first layer of maps. On the second layer of SOMs, each SOM obtains as input the output “fingerprint” of activations of a whole SOM below it.

The output of the first map is processed to identify and collect a subset of the SOM units that have at least one data point, and transform them to schemas. In transforming a unit to a schema, value tresholding is applied to the context and result features to collect the ones that are sufficiently strong and to ignore others. The schemas learned are considered as synthetic states of the world (representing an action schema that was just performed). At this point the original action maps can be discarded and their resources reused. Next, a new set of action maps is created, with input that consists of both the initial inputs (primitives describing world state) as well as the synthetic input obtained from the previous maps (history information). This allows the learning of complex actions and their effects. The authors report exellent results of a simulation of Drescher’s “Microworld”: in addition to replicating the earlier results of earlier experiments that did not utilize neural representations, it is found that there emerge features that represent the beginnings of persistent-object -concept.

In summary, the paper shows how by the applicatin of a hierarchy of SOMs in a particular way leads to the non-trivial emergence of more powerful, higher-level abstractions. By looking at actions, their pre-and post-contexts, emerge not only more complex aggregate actions but also the beginnings of the concept of object persistence. The structural assumptions made regarding connectivity of the maps seem domain-general, but the emergence of such a structure is not discussed.

2.3.3 On the emergence of perceptual features using SOM and ICA

A large body of work has been carried out on modelling the emergence of cortical features on various areas of the cortex. We will only mention two approaches. Miikkulainen et al. have studied extensively the modelling of visual cortex using layers of connected sheets of neurons (a set of hierarchically organized self-organizing maps that have been enhanced with properties such as lateral connections, receptive fields, on-off channels, spiking neurons, and delay adaptation) that receive as input natural image data (e.g., Choe 1998; Bednar 2003). Also here the utilization of particularly connected hierarchy of maps is essential.

Hyvärinen et al. have applied variants of Independent Component Analysis (ICA) for modelling various functions of the visual cortex, including the emergence of features carrying out contour coding (see e.g. Hoyer & Hyvärinen, 2002). ICA is an unsupervised statistical method that models its input data in terms of linear combination of some hidden (latent) variables, while maximizing mutual independence of the hidden variables. In general, such latent variables can be considered to correspond to primitive features or quality dimensions that are rather closely connected with the sensory modalities.

2.3.4 IVGA as a model of the differentiation of Gärdenfors' quality domains

Gärdenfors' model left open the question regarding how quality domains might emerge, and we address the issue here.

Independent Variable Group Analysis (IVGA) (Lagus et al, 2001) is an unsupervised data analysis method that groups input features into subsets by minimizing the statistical dependencies between the subsets of features. To achieve this the data is in fact modeled using many different feature groupings, and an efficient search algorithm is applied to find a good feature grouping. The algorithm was evaluated on a set of natural images encoded using low-level visual features such as gray level and fourier features.

A particular feature group and its respective model can be considered to correspond to Gärdenfors' quality domain in the sense that the dimensions (features) in a group are statistically dependent (integral), but as independent as possible (separable) from other feature groups. Moreover, each feature group may be modeled using a completely different method. The only requirement is that a cost function must be definable that measures both model accuracy and model complexity.

IVGA provides an unsupervised, computationally feasible principle for explaining how the separation of the quality domains could in principle emerge during a combination of evolution (selection of a particular feature grouping) and individual learning (modeling of the data using a particular feature grouping). However, further experiments are needed to examine the plausibility of this hypothesis. An argument against the sufficiency of this approach would be that both data-directed (without feedback, unsupervised) and goal-directed (feedback-driven) effects are needed to produce the quality domains humans have. However, the same argument can be applied in general against the idea of unsupervised learning of concepts, their attributes or relations.

2.3.5 A probabilistic model of embodied lexical development

An implemented computational model of embodied lexical development for learning action verbs is presented in (Bailey, 1997). The authors stress the importance of embodiment in language acquisition, and address the question of how does a child learn to label his/her own actions, a task which the authors consider to be central in grounding language. The model consists of an embodied system, namely a robot, capable of performing various actions using its hands, like pushing or yanking.

A concept is represented by a Feature structure and a link to Execution schema. The Execution schema, implemented by a Petri Net, consists of a network of consecutive actions that control hand movement. Prior to execution, the feature values from the feature structure are passed to the execution schema affecting e.g. hand position.

The internal discrete features found in the feature structure are fixed by the designer, including motor parameter features such as elbow joint (fixed;extended), posture (values: grasp;palm;index), acceleration (low;medium;high) and direction (up;down;left;right), as well as world state features such as object shape (cube;button).

A feature structure for an actual hand motion (instance) consists of a set of features each with probability 1 (certainty). The feature structure for a particular concept, then, consists of a subset of the features present in the instance, as well as a probability distribution for each feature. Concepts are learned by merging instances of a word used in its motor context. At the same time some features may be dropped from a concept's feature structure if they are not informative enough; that is by observing the peakedness of the probability distribution of the feature for that particular concept.

The authors review results of a small experiment where 50 random executions of the execution schema SLIDE were generated and labelled by an informant as push, pull, and slide (for sideways motion). They report that the merging algorithm collapsed the 12 instances of “push” into a single sense, likewise for the 9 instances of “pull”. For the 4 instances of “slide”, the algorithm differentiated the senses of leftward and rightward slide. Some of the features that were present for an instance were abstracted away in the concepts, e.g. the feature acceleration for ”push”.

In summary, the system learns how many senses (concepts) per verb there are, which features to use for a verb sense, and what are their probability distributions. The internal features (the subclass of concepts that Gärdenfors calls properties) are not learned, it is assumed that they were formed prior to language acquisition, e.g. due to being embodied in the world. By fixing the features a stable basis is obtained for parameter passing between motor actions and concepts. While the learning in this example is very limited, we consider it relevant for completeness of the treatment of the emergence of structured models, in part because it exemplifies learning in an embodied system.

2.4 Discovering Conceptual Relations from Texts

In this review we concentrate on the discovery of conceptual relations when there is already knowledge about the concepts (i.e. what concepts there are and what are their properties). Whereas the approaches in Section 2.3, were attempts at neural or cognitive modelling of concepts, the examples in this section are intended as technical solutions to the problem of how to discover and represent relational information. Learning is utilized to ensure completeness of the result, and to avoid laborious manual coding.

Gruber gives a compact definition that ”An ontology is a formal, explicit specification of a shared conceptualisation” (Gruber 1993). The word formal has very important connotation for us, as we necessitate that the models are machine readable and suitable for automatic processing. Ontologies have been widely recognised as important for sharing conceptualisations in electronic commerce and business intelligence applications (Fensel

2001). However, manual specification of ontologies is slow and expensive and there is a wide interest in developing human assisted machine learning tools for the tasks. Starting from the 1990's automatic discovery of ontologies has been researched in several institutes. A popular approach has been to use natural language texts as a starting point. A typical approach may include following phases:

1. Tokenising including morphological analysis and associating lexical knowledge
2. Syntactic analysis (e.g. shallow parsing, dependency parsing)
3. Recognising concepts
4. Establishing concept taxonomy
5. Discovering non-taxonomic relationships
6. Manual editing and pruning of the outcomes (this may be interlaced with the earlier steps)

The phases 1 and 2 have been widely researched during the last two decades and there are several solutions available. The step 3 has also been elaborated considerably. In this review we concentrate on steps 4 and 5, that is, the discovery of concept taxonomies and non-taxonomic relationships from texts when there is already knowledge about the concepts.

There are several types of ontologies, such as *domain ontologies* (e.g. electronic, medical, clothing etc. domain), *metadata ontologies* (e.g. Dublin Core initiative for describing content of on-line info sources, www.dublincore.org), *generic* or *common sense ontologies* (CYC, meteorology, colours etc.), *representational ontologies* (e.g. frame ontology), *method and task ontologies* (e.g. workflow management definitions), etc (Fensel 2002).

The reviewed approaches focus on domain ontologies of rather narrow topic, such as models for corporate information in business news (Byrd & Ravin 1999), tourist services and telecommunication services (Maedche & Staab 2000a & 2000b), extracting models from corpora of cooking recipes and travelling (Faure & al 1998, Faure & Nédellec 1999), and biological relationships discovery from texts (Palakal 2002).

The survey revealed that the notion of a relation might vary much alongside with the application requirements. Are there discovered new relationships to add into an ontology model or are there discovered instances of relations that are pre-specified in an ontology? For instance, if we are doing the first activity, the sentence "the president of Finland, Mrs. Halonen, met with ..." would introduce to an ontological world model a new relation type: a country has-a-president who is a human. However, if we are only interested in recognising instances of relations, our world model should contain a definition of a president and a method how to recognise that relationship from texts. After having those, we could recognise the relation instance *president(Finland,Halonen)*.

There is also variance in the types of relations that are discovered. The IBM approach mostly concentrates on recognising instances of binary relationships between concepts (Byrd & Ravin 1999). The Univ. of Karlsruhe solution concentrates on discovering non-taxonomic binary relations to a domain ontology (Maedche & Staab 2000a & 2000b). The Asium system of Université Paris Sud finds syntactico-semantic n-ary relations of verbs to concepts/headwords, called case frames, and builds a concept hierarchy (Faure & al 1998, Faure & Nédellec 1999). The multi-level text mining method of Purdue University extracts hierarchical relations and non-taxonomic directional relations in the domain of biology (Palakal 2002). All the approaches regard concepts as points in the domain space. There is no consideration to their internal structure like attributes.

2.4.1 Required Qualities

We consider the following properties important for evaluating the approaches.

Precision and recall. It is better to discover some erroneous relations than to miss very central ones. Later checking and pruning of the results of automatic discovery is in any case necessary, before taking automatically discovered models into production use. Thus the goal usually is to achieve high recall even while sacrificing some precision.

Linguistic generality. The input language should not be restricted in its syntax or vocabulary. The domain specificity for the solution comes from the used lexico-semantic information and from the initial concept model, which both may be restricted by their domain.

Language independence. The overall approach should be as independent as possible of the language in the texts. This applies particularly to the later semantics centred phases of the discovery. Although there are language specific grammars involved, the formalisms and algorithms should be language independent.

2.4.2 Approaches

Company information modelling at IBM

IBM has developed the text mining system *Textract* that recognises instances of binary relations between concepts and entities in texts (Byrd & Ravin 1999). Predefined patterns are used to indentify *named relations* and statistical analysis of co-occurrences of concepts to postulate so far *unnamed relations*. The analysis starts with the identification of domain terms, proper names as well as their type (place, person, organisation etc.), abbreviations, and other special single words. After this tokenising the system analyses input with specially-built finite state automata specified by the patterns. The system does not require full syntactic analysis, it is enough that piecewise matches are found. There are three classes of patterns:

1. Patterns anchoring the positions of the both concepts and specify the discovery of the relation name. As an example, the pattern *"PERSON, ... of ORGANISATION"* recognises the new CEO relation $< Louis\ V.\ Gerstner : CEO : International\ Business\ Machines >$ from the text excerpt "Today, Gerstner, the CEO of IBM, announced that the company ...".

2. Patterns anchoring the positions of both concepts and assigning a predefined relation name.

3. Patterns anchoring the position of one of the concepts and fixing the relation name. When the pattern *"ORG MAKE-VP ... "* is applied to "... IBM, which manufactures computing equipment ...",it yields $< \mathit{International\ Business\ Machines} : \mathit{make} : \mathit{computing\ equipment} >$.

The process is not precise enough and there is need to prune the intermediate results. To do this, there are filters for rejecting bad candidates:

1. frequency filters, to reject, for example, a relation name that occurs only once and may therefore be just an accidental string of words;
2. lexical and morphological filters, to require, for example, that the lexical head of the verb phrase be a verb of manufacturing or selling;
3. selectional restrictions, to require a place name, for example, as the first concept in a location relation;
4. other filters like co-ordination censors, and length filters.

The approach involves additional heuristics, for instance, to categorise unknown entities based on the patterns where they have been recognised. For instance, an entity may be upgraded to a PERSON, if it is in a CEO-of relationship with an organisation. Co-occurrence with unnamed relations adds thrust to a named relation candidate.

Non-taxonomic binary relation discovery at Univ. of Karlsruhe

University of Karlsruhe has developed a system called Ontology Learning Environment for discovering ontologies from texts. It includes a parsing mechanism for acquiring concept taxonomy from a domain-specific dictionary and an algorithm for discovering non-taxonomic relations from texts (Maedche & Staab 2000a & 2000b). Human co-operation is involved in the pruning phase of the discovery process. The system has been tested in the restricted domains of accommodation services, tourist sights and telecommunication services. Next we outline the process of non-taxonomic relations discovery from texts.

The linguistic text analysis is based on the SMES (Saarbrücken Message Extraction System) software, which includes a tokenizer, a lexical analysis component, and a chunk parser, that makes first phrasal and then sentence level dependence analysis. Syntactic dependency relations coincide rather closely with semantic relations holding between the same entities. The dependence parser still returns many constituent trees that are not related within or across sentence boundaries. For instance, the parser does not attach prepositional phrases in any way and it does not handle anaphora. Three heuristic correlations have been added to ensure recall: (1) NP-PP-heuristics attaches all prepositional phrases to adjacent noun phrases, (2) Sentence-heuristics relates all concepts contained in one sentence if other criteria fail, and (3) title-heuristics links in titles with all concepts in the overall document. The last heuristics has been found very effective in handling hotel and tourist sight descriptions.

The previous processing steps have produced a set of candidate concept pairs. The discovery of generic relations uses a shopping basket mining algorithm (Srikant & Agrawal 1997). The found relation hypotheses are treated as shopping basket associations. Generalisations are done according to concept hierarchies and/or lexico-semantic categories. The algorithm expects a set of transactions T where each transaction t_i consists of a set of items $t_i = \{a_{i,j} | j = 1 \ldots m_i, a_{i,j} \in C\}$ which all belong to a set of concepts C. The algorithm computes association rules $X_k \Rightarrow Y_k$, in which $X_k, Y_k \subset C$ and $X_k \cap Y_k = \{\}$, and for which measures for *support* and c*onfidence* exceed user-defined thresholds (see Formulas 1 and 2). The associations are determined at the right level of a taxonomy, which is defined by the relation $H \subset C \times C$, and which may have been derived from a domain-specific dictionary or defined manually. First, each transaction is extended to include each ancestor of a particular item. Thus a transaction will contain items $t'_i = t_i \bigcup \{a_{i,l} | (a_{i,j}, a_{i,l}) \in H\}$. Secondly, there is calculated support for all association rules $X_k \Rightarrow Y_k$, where $|X_k| = |Y_k| = 1$. Thirdly, there is calculated confidence for all rules exceeding the user-defined support threshold in second step. Finally, those rules, which exceed the user-defined confidence threshold, are applied the ancestral pruning rule. This means pruning of those association rules that have ancestral rules with higher or equal confidence and support.

$$support(X_k \Rightarrow Y_k) = \frac{|\{t_i | X_k \cup Y_k \subseteq t_i\}|}{n} \tag{2.1}$$

$$confidence(X_k \Rightarrow Y_k) = \frac{|\{t_i | X_k \cup Y_k \subseteq t_i\}|}{|\{t_i | X_k \subseteq t_i\}|} \tag{2.2}$$

From the sentences "*Mecklenburg*'s most beautiful *hotel* is located in Rostock. A *hairdresser* in our *hotel* is a special service for our guests. The hotel Mercure offers *balconies* with direct *access* to the beach. All *rooms* have *TV*, telephone, modem and minibar", Words with concept references in italics. Four concept pairs, among many others, can be derived with knowledge from the domain lexicon: (area, hotel), (hairdresser, hotel), (balcony, access), (room, television). Similar analysis is done for a larger part of a corpus and the shopping basket algorithm is executed. The final result has two of the relations replaced by their ancestral relations found using the domain-specific taxonomy. The resulting non-taxonomic relations are (area, accommodation) with confidence 0.38 and support 0.04, (room, furnishing) with confidence 0.39 and support 0.034, (accommodation, address) with confidence 0.34 and support 0.05, and (restaurant, accommodation) with confidence 0.33 and support 0.02.

Asium system of Univ. Paris Sud and discovery of n-ary case frames and concept hierarchies

Université Paris Sud has researched discovery of syntactico-semantic relations of verbs and concept hierarchies from technical documents. Its Asium system is a co-operative machine learning system for learning case frames and for clustering concepts and learning their hierarchies from technical documents. The source documents are characterised by specific domain, limited vocabulary, restricted polysemy and verbs being mostly concrete and action verbs (Faure & al 1998, Faure & Nédellec 1999). The case frames are verb centric templates that denote which subordinates a verb must or may have. For a verb a template may reflect either its syntactic relationships (e.g. the syntactic functions subject and object) or semantic relationships (like the role of vehicle for the verb travel). The discovered knowledge is meant for a writer-assistant program that would help the technical editors at Dassault Aviation Company to produce technical documentation in a controlled way.

The format of frames is simply: <*verb*> (<*role*><*concept*><*optionality*>)$*$ In the learning process the roles can be taken from the syntactic dependence structure of the input and headwords of subordinates can be considered as concepts. When there are recognised more syntactically similar instances of the verb, new more abstract concepts are formed to replace the original ones.

When there are learned new generalised frames, the system is simultaneously building a hierarchical ontology of new concepts, which it uses while doing its generalisations. As an example, if there are recognised frames <*to travel*> <*subject*> <*Bart*> <*by*> <*boat*> and <*to travel*> <*subject*> <*David*> <*by*> <*train*>, there can be generalised the frame <*to travel*> <*subject*> <*Human*> <*by*> <*Vehicle*>.

Considering conceptual clustering Faure & al criticise the applicability of vector based learning methods for their task, such as (Cheeseman & Stutz 1996), as the size of the vectors of headword-frequencies would grow very high due to the large vocabulary and the vectors would be very sparse. As well, they doubt the applicability of FOL based learning methods like (Bisson 1992), because for them the semantic classes may have more than one super-class, to express different viewpoints on the same objects.

Faure & al present their own learning algorithm for finding concepts and their hierarchical relations. The inputs are the initial clusters, which are sets of words associated with the frequency of the corresponding syntactic verb frame in the corpus. The simplest version is called Asium-Best, which computes the distances between all pairs of clusters and compares them to a threshold. Clusters are merged if there are too close to form a new one. The process stops when all computed distances exceed the threshold. The learned clusters are displayed to the user for validation and labelling. The algorithm Asium-Pyramid applies a cleverer strategy for choosing the pair of clusters to process at a time and strives for forming a pyramid hierarchy between the concepts (i.e., a DAG that is possible to represent as a tree with no lines crossing each other). In both versions of the algorithm the distance between clusters is the proportion of common headwords between the two clusters, balanced by the relative frequency of the instantiated verb frames in the cluster. Formula 2.3 defines this *distance function between head word clusters C1 and C2.* $freq(C, w)$ is the frequency of head word w in cluster C and $nh(C)$ is the number of head words in cluster C. The distance varies between 0 for equal clusters and 1 for disjoint clusters.

$$d(C1, C2) = 1 - \frac{(\sum_{w_i \in C1 \cap C2} freq(C1, w_i)) \frac{nh(C1 \cap C2)}{nh(C1)} + (\sum_{w_i \in C1 \cap C2} freq(C2, w_i)) \frac{nh(C1 \cap C2)}{nh(C2)}}{\sum_{w_i \in C1} freq(C1, w_i) + \sum_{w_i \in C2} freq(C2, w_i)} \tag{2.3}$$

Biological relationships mining at Purdue Univ.

At Purdue University there has been developed a multi-level text mining method for extracting biological relationships from texts documents (Palakal & al 2002). The approach involves object identification, reference resolution, synonym discovery, extracting object-object relationships. Technical solution is based on Hidden Markow Models, domain dictionaries, and N-Gram models. Experiments with a corpus of around thousand Medline abstracts found 53 relations from which 43 were correct. The method includes the following phases: (1) extracting biological concept names, (2) grouping concept synonyms, and (3) extracting concept relations.

The goal of extracting biological concept names is to recognise concepts that denote genes, proteins, cell types, organisms, RBNA, chemicals, diseases, drugs etc. Concept name detection is based on using domain dictionaries for identifying known concepts, N-Gram models to resolve concept name ambiguity, and Hidden Markov models (HMM) to identify unknown concepts based on term suffices. A domain dictionary may state, e.g., that class protein consists of protein, kinase, enolase, antigen, cytokeratin, amelogenin and vimentin.

N-gram model is a simple Markov model where the probability of a word w_1 being in the position n is $\prod_{k=1}^{n} \frac{count(w_{k-N+1}^{k-1} w_k)}{count(w_{k-N+1}^{k-1})}$. Word probability is assumed to depend on the previous N words. $count(w_{k-N+1}^{k-1} w_k)$ is the number of times the previous N words are followed by w_k and $count(w_{k-N+1}^{k-1})$ is the total number of times the previous N words occur. For the disambiguation is taken the phrase data that was obtained using the N-Gram training process. Probabilities are analysed for each class given a phrase. An HMM is used to classify words that are abbreviations composed of less than six characters. Separate set of abbreviation dictionaries is used in training.

Extracting concept relations considers two types of relations. The first is non-taxonomic directional relations, like in "protein A inhibits protein B". The second type is hierarchical relations, like in "brain is part of the nervous system". Directional relations are mined using HMMs and hierarchical ones using a data mining algorithm developed for genetic analysis. The statistical nature of the extraction method allows for finding new relations that would not be found in a rule-based system. Combining hierarchical relations with directional relations creates more complex relations than just binary ones. For example, a directional binary relation (e.g. between proteins) could be associated with another concept (e.g. disease) using a common hierarchical relation that the concepts in the directional relation share.

2.4.3 Summary of the Reviewed Relation Discovery Approaches

A summary of the findings is presented in Table 2.2.

2.5 Conclusions

It seems that many aspects of the concept learning process can be modelled by the existing learning methods such as neural networks. The SOM demonstrates how unsupervised category construction could take place by noticing correlations in the input. Also, the hierarchical relations can be modeled implicitly as regions of a SOM map. Latent features can be derived also using other methods, such as the ICA.

The necessity of structured, possibly hierarchical representations was examined and argued for. Some aspects of the emergence of such a structured model were addressed using examples from literature, including the emergence of conceptual dimensions or features, the differentiation into domains, categorization of sample data, and feature/domain selection in concept formation. Also hierarchical representations and some of their benefits were observed. However, the discussed examples utilized different learning approaches and solve only partial problems. It remains to be seen how a complete conceptual system might be assembled from such parts, and to which degree the assemblage could emerge from properties of natural signals, communication, resource limitations of the brain etc.

Table 2.2: Types of relations found in the discussed discovery approaches.

Approaches	**Entities**	**Relations (E = external, I = internal)**	**Methodologies**
Company information modelling at IBM	Concepts (postulated along domain terms & names)	Instances of binary relations (E); relations (E, named and unnamed)	Pattern matching, statistical analysis, heuristic pruning rules
Non-taxonomic binary relation discovery at Univ. of Karlsruhe	Concepts (domain-specific dictionaries as source)	Generalised non-taxonomic binary relations (E)	Linguistic dependency relations + heuristics ⇒ initial hypothesis; Shopping basket DM algorithm
Asium system of Univ. Paris Sud and discovery of n-ary case frames and concept hierarchies	Concepts	Generalised case frames (E); Concept hierarchy (I/E); KD for HLP	Linguistic dependency relations of verbs yield initial frames; Generalisations along concept hierarchies; Concept clustering algorithm
Biological relationships mining at Purdue Univ.	Concepts	Non-taxonomic directional relations (E); Hierarchical relations (I/E)	HMM; DM algorithm for generic analysis; Deducing of more complex relations from simple ones

The methods discussed in Section 2.3 were general learning mechanisms designed for learning models for continuous-valued data, such as is found in the natural world (although also discrete inputs may be successfully analysed). In contrast, the methods in Section 2.4 are more oriented towards models where both inputs and outputs are discrete, which may indeed be appropriate for the treatment of external relations. Moreover, heuristics and prior information are readily utilized e.g. in the preprocessing of the language data, the design of patterns to be matched, etc.

What seems to be lacking from both the psychological and philosophical approaches and the modelling approaches is the inclusion of holistic sensory information in the conceptual representation. By this we mean that it is often not enough to know the feature decomposition of the shape of a concept, for example. What is needed is a stored memory of the holistic form of the object, or perhaps a visual exemplar of the object's shape. Similarly for actions. It is unclear how this could be represented in, for example, Gärdenfors' conceptual spaces. A possible solution might be the storage of a small number of remembered exemplars that are not decomposed into features, along with the corresponding prototypes that connect the neural and conceptual level.

References

Bailey, D., Feldman, J., Narayanan, S., Lakoff, G. (1997). Modeling Embodied Lexical Development. In Proceedings of the 19th Cognitive Science Society Conference, pp 19-24.

Bisson, G. (1992). Learning in FOL with a Similarity Measure. In: Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, California, pp 12-16.

Bruce, B. (1975): Case Systems for Natural Language. Artificial Intelligence, vol 6, pp 327-360.

Byrd, R, Ravin, Y (1999). Identifying and extracting relations from text, In: NLDB'99 — 4th International Conference on Applications of Natural Language to Information Systems.

Chaput, H.H., Kuipers, B., Miikkulainen, R. (2003). Constructive learning: A Neural Implementation of the Schema Mechanism. In Proceedings of WSOM'03 — Workshop on Self-Organizing Maps. Kitakyushu, Japan.

Cheeseman, P., Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results, In: Advances in Knowledge Discovery and Data Mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, & Ramasamy Uthurusamy (Editors), AAAI Press/MIT Press.

Faure D., Nédellec C., Rouveirol C. (1998). Acquisition of Semantic Knowledge using Machine learning methods: The System ASIUM, Technical report number ICS-TR-88-16.

Faure D., Nédellec C. (1999). Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the system ASIUM. In: Dieter Fensel, Rudi Studer (editors), 11th European Workshop EKAW'99, Springer-Verlag, pp. 329-334.

Fensel, D. (2001). Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce. Springer Verlag, Berlin Heidelberg, 138 p.

Fodor, J.A. (1981). The present status of the innateness controversy. In Fodor, J.A, Representations, 1981. Cambridge, Massachusetts: MIT Press.

Fodor, J.A. (1998). Concepts: Where cognitive science went wrong. Oxford: Oxford University Press.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995). Bayesian Data Analysis. Chapman & Hall/CRC.

Gruber, T.R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, vol 5, no 2, pp. 199-220.

Gärdenfors, P. (2000). Conceptual spaces: The geometry of thought. The MIT Press.

Hanson, R., Stutz, J., Cheeseman, P. (1991): Bayesian Classification Theory, Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch, May 1991.

Haykin, S. (1999). Neural networks — A Comprehensive Foundation. Prentice Hall.

Honkela, T., Pulkki, V., Kohonen, T. (1995). Contextual Relations of Words in Grimm Tales Analyzed by Self-Organizing Map. Proceedings of ICANN-95, International Conference on Artificial Neural Networks, pp. 3-7, Paris, EC2 et Cie.

Hoyer, P., Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. Vision Research, vol. 42, no. 12, pp. 1593-1605.

Hutchins, W. J. (1995): Machine Translation: A Brief History. In: E.F.K.Koerner and R.E.Asher (eds), Concise history of the language sciences: from the Sumerians to the cognitivists, Oxford, Pergamon Press, pp. 431-445.

Ikeda, N., Hagiwara, M. (1999). A Novel Knowledge Representation (Area Representation) and Its Implementation by Neural Network. Systems and Computers in Japan, vol. 30, no 13, pp. 34-42.

Kohonen, T. (1990). Internal representations and associative memory. Parallel Processing in Neural Systems and Computers, pp. 177-182.

Kohonen, T. (2001). Self-Organizing Maps. Springer. 3rd extendend edition.

Lagus, K., Alhoniemi, E., Valpola, H. (2001). Independent Variable Group Analysis. In Proceedings of ICANN'01-International Conference on Artificial Neural Networks, pp. 203-210.

Laurence, S., Margolis, E. (1999). Concepts and Cognitive Science. In Margolis, E., & Laurence, S. (Eds.), Concepts: Core Readings. Cambridge, Massachusetts: MIT Press.

Maedche, A., Staab, S (2000a). Discovering Conceptual Relations from Text. In: Proceedings of ECAI2000, pp 321-325.

Maedche, A., Staab, S. (2000b). Mining Ontologies from Text. In: R.Dieng & O Corby. EKAW-2000 — 12th International Conference on Knowledge Engineering and Knowledge Management. October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer.

Maynard, D., Sophia, A. (1999). Term Extraction using a Similarity-based Approach. In: Recent Advances in Computational Terminology, John Benjamins.

Murphy, G. (2002). The Big Book of Concepts. Cambridge, Massachustetts: MIT Press.

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R., Rhodes, S. (2002). A Multi-level Text Mining Method to Extract Biological Relationships. In: Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'02).

Quine, W. (1953). Two dogmas of empiricism. In From a logical point of view, 1980. Harvard University Press.

Ritter, H., Kohonen, T. (1989). Self-organizing semantic maps. Biological Cybernetics, vol. 61, no 4, pp. 241-254.

Romacker, M., Markert, M., Hahn, U. (1999). Lean semantic interpretation. In: Proceedings of IJCAI-99, pp. 868-875.

Srikant, R., Agrawall, R. (1997). Mining Generalized Association Rules. In: Future Generation Computer Systems, pp. 161-180.

Tuulos, V., Perkiö, J. & Honkela, T. (2004). On Modelling Multimodal Concepts. In: Honkela, T., Hynnä, K., Lagus, K. & Särelä, K. (Eds.), Adaptive and Statistical Approaches in Conceptual Modeling, pp. 45-55.

Van Aken, J.E. (1978). On the control of complex industrial organizations. Boston: Martinus Nijhoff Publishing.

Wermter, S., Sun, R., Eds. (2000). Hybrid Neural Systems. Lecture Notes in Artificial Intelligence, 1778, Springer, Berlin.

Chapter 3

Assessing Similarity of Emergent Representations

Juha Raitio, Ricardo Vigário, Jaakko Särelä and Timo Honkela

Introduction[1]

According to the connectionist view, mental states consist of activations of neural units in a connectionist network. We consider the similarity of representations that emerge in an unsupervised, self-organization process of neural lattices when exposed to color spectrum stimuli. Self-Organizing Maps (SOM) are trained with color spectrum input, using various vectorial encodings for representation of the input. Further, the SOM is used as a heteroassociative mapping to associate color spectrum with color names. Recall of association between the spectra and colors is assessed, and it is shown that the SOM learns representations for both stimuli and color names, and is able to associate them successfully. The resulting organization is compared through correlation of the activation patterns of the neural maps when responding to color spectrum stimuli. Experiments show that the emerged representations for stimuli are similar with respect to the partitioning-of-activation-space measure almost independently of the encoding used for input representation. This adds new evidence in favour of the usability of the state space semantics.

3.1 Connectionist networks and representation of content

The state of a connectionist network is the momentary activation levels of neurons in the network configuration [3]. A particular state may occur as a response to stimuli. Then the stimuli has a state space representation in the space spanned by the possible activations of the neurons in the network. Vice versa, any pattern of activations in the network may represent some, perhaps latent, information. According to the connectionist view, mental states are instantiated by these activations of neural units [3]. Therefore connectionists have been puzzled with a criterion for determining when activations in two connectionist

[1]Parts of this chapter have been published in the Proceedings of IJCNN'04, International Joint Conference on Neural Networks.

networks have similar content – or even, when they are representing exactly the same mental state.

Fodor and Lepore [5] argue that a connectionist theory of mind cannot give a satisfactory account of different individuals being in the same mental state, for identity of networks is a suffient condition for identity of content, but this condition will never be satisfied in practice. Laakso and Cottrell [11] note the same problem in their statement: *If connectionism is to be an adequate theory of mind, we must have a theory of representation for neural networks that allows for individual differencies in weighting and architecture while preserving sameness of content.*

In this article, we consider a method for comparing the similarity of representations in connectionist networks, and examine the possibilities for exploiting it for comparing emergent representations in unsupervised learning networks. We report the results of applying this method as a similarity measure for representations emerging in Self-Organizing Maps.

3.2 Measuring the similarity of state space representations

A straightforward way of measuring the similarity of the state space representations in a network, or between two networks having the same configuration, is to measure the distance between the activation levels of the neurons. In this *position-in-activation-space* view of similarity [3], the proximity of the state space representations are clearly dependent on the positions of activation. It is unclear however, how two networks with different number of neurons could be compared according to this view, for common distance measures are only defined for vectors of equal lengths.

Identifying content with characteristic groupings of activation patterns was proposed by Churchland [4]. He claims that people react to the world in a similar way, because their activation spaces are similarly partitioned. Laakso and Cottrell acknowlegde this as an evident solution, for it allows different individuals to represent the same latent information without having identical networks.

Adopting this *partitioning-of-activation-space* view to similarity of representations, Laakso and Cottrell [11] propose that content is associated with relative positions in the partitioning of the activation space. The momentary representations should then be compared by each representation's location relative to other possible activations in the same network.

Further, Laakso and Cottrell [11] develop a method for assessing the similarity of representations in two networks by comparing their partitionings through correlating the distances between all pairs of activation patterns in each network:

1. Collect the activation patterns evoked by inputs and compute all possible distances between these representations.
2. Compute the correlation between the distances between representations in one state space and distances between representations in the other state space.

The distances effectively capture the structure of representational space and eliminate the need to match the dimensions of the two spaces.

Laakso and Cottrell [11] test their measure and present two experiments that demonstrate MLP networks that learn to classify colors based on spectral stimuli. Reported results show that

1. MLP networks that were given differently encoded spectra as input, learn internal representations in the hidden layer that are quite similar by the measure, and that networks receiving identical stimuli learn nearly identical representations,
2. MLP networks having different number of neurons in the hidden layer, thus not sharing the same activation state space, were found to build nearly identical representations by the measure.

In computing the similarity of the distances between points in two representational spaces, Laakso and Cottrell [11] provide a *partitioning-of-activation-space* criterion for semantic similarity that answers the challenge Fodor and Lepore [5] place on state space semantics.

3.3 Emergent, unsupervised representations and association

The Laakso and Cottrell experiments are based on supervised learning to associate color names with color spectra. This, we think, is not a particularly plausible approach (though maybe intentionally simplified) in the context where they present it:

- encodings representing different sensory organs of animal species,
- different numbers of neurons in the hidden layer varying according to individual and cross-species differences in brain capacity and
- the color names being symbols to be identified.

We believe that, more realistically, color spectra stimuli as a physiological input to a connectionistic system emerge as unsupervised, latent representations, irrespectively of whether the colors have known symbols or not (answerme:could we refer here to some known brain maps of this nature?). Adopting an unsupervised learning approach, no color names or ready content are needed for the formation of meaningful representations of the stimuli.

If colors (names) are to have a representation in a network, it could have emerged independently of the physiological spectra stimuli. Also, we believe, an association between a color symbol (name) and spectrum stimuli could grow unsupervised by their simultaneous excitation in a learning connectionist network.

The *partitioning-of-activation-space* criterion of Laakso and Cottrell can be generally applied to measure the similarity between any two neural representations. In the following we introduce tools to study this criterion and to repeat their experiments in the unsupervised learning framework.

3.4 Methodology

3.4.1 Self-Organizing Maps

The set of input samples to a connectionist network is described by a real vector $\mathbf{x}_j \in R^n$ where j is the index of the sample. Each node in the Self-Organizing Map (SOM) contains a model vector $\mathbf{m}_i \in R^n$, which has the same number of elements as the input vectors.

The nodes of the map form an array with a definite topology each having a location vector $\mathbf{r_i}$. The array is often a two dimensional rectangular grid.

In the learning phase, the self-organizing algorithm creates an ordered mapping from the input space to the map array as a repetition of the following tasks [9] at each discrete-time step $t = 0, 1, 2, \dots$:

1. An input vector $\mathbf{x}(t)$ is compared with all the model vectors $\mathbf{m}_i(t)$. The best-matching unit (BMU) c on the map, i.e. the node, whose model vector $\mathbf{m}_c(t)$ is most similar to the input vector in some metric (e.g. Euclidean) is identified:
$$c = \underset{i}{\operatorname{argmin}}\{\|\mathbf{x}(t) - \mathbf{m}_i(t)\|\}. \tag{3.1}$$
This best matching unit is often call the winner.

2. The model vectors of the winner and a number of its neighbouring nodes in the array are changed towards the input vector according to
$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)\left[\mathbf{x}(t) - \mathbf{m}_i(t)\right], \tag{3.2}$$
where $h_{ci}(t)$ is the so-called neighborhood function that has higher values for nodes that are topographically close to the BMU in the map array, and smaller values for nodes that are distant. For convergence of learning it is necessary that $h_{ci}(t) \to 0$, as $t \to \infty$. One such function, written in terms of a Gaussian is
$$h_{ci}(t) = \alpha(t) \cdot \exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right), \quad \text{where} \tag{3.3}$$
$\alpha(t)$ and $\sigma(t)$ decrease monotonically in time.

The net outcome of the adaptation process is that ordered values for the $\mathbf{m}_i(t)$ emerge over the array. Initial values of the $\mathbf{m}_i(0)$ can be arbitrary. The basic properties of this ordering are that the distribution of the model vectors tends to approximate the density of the input vectors $\mathbf{x}(t)$, and that the organization of model vectors in the array is such that the mapping tends to preserve the topology of the input space.

The output or the activation of the SOM, as a response to stimuli $\mathbf{x}(t)$, is the excitation of the BMU and its neighbouring neurons $h_{ci}(t)$ (3.3) , where α and σ have some fixed values determined by the application. This is referred to as postsynaptic activation in [10].

A detailed description about the selection of the parameters, variants of the map, and many other aspects have been covered in [9]. Perhaps the most typical interpretation of the SOM is to consider it as an artificial neural network model of the brain [10], especially of the experimentally found ordered "maps" in the cortex. The Self-Organizing Map can also be viewed as a model of unsupervised statistical machine learning, as an adaptive knowledge representation scheme, as a statistical tool for multivariate analysis, or as a tool for data mining and visualization [7].

3.4.2 Associative Mappings with the SOM

Assume two input patterns $\mathbf{x}^{(A)} \in R^{n_1}$ and $\mathbf{x}^{(B)} \in R^{n_2}$ are concatenated to form a single input vector $\mathbf{x}^{(AB)} \in R^{n_1+n_2}$. $\mathbf{x}^{(A)}$ and $\mathbf{x}^{(B)}$ may encode some information A and

B presented simultaneously to the SOM. Now the model vectors $\mathbf{m}_i$ have components corresponding to A and B, respectively:

$$\mathbf{m}_i = \begin{bmatrix} \mathbf{m}_i^{(A)} \\ \mathbf{m}_i^{(B)} \end{bmatrix}. \tag{3.4}$$

During training the SOM builds an association between A and B. To evoke this association, the BMU c is defined on the basis of $\mathbf{m}_i^{(A)}$ and $\mathbf{x}^{(A)}$ only, then an estimate of $\mathbf{x}^{(B)}$, in the sense of the SOM mapping, is obtained as the vector $\mathbf{m}_c^{(B)}$. This recall of the $\mathbf{m}_c^{(B)}$ is referred to as *associative mapping* by Kohonen [9]. An example of utilising *associative mapping* are the early versions of the "Phonetic Typewriter".

If the magnitudes of the vector components of the pattern $\mathbf{x}^{(A)}$ are large compared to the magnitudes of $\mathbf{x}^{(B)}$, then component $\mathbf{x}^{(B)}$ of the input has in general little significance in choosing the BMU in (3.1). Consequently, the organization of the map is not affected by $\mathbf{x}^{(B)}$, but the SOM learns to approximate $\mathbf{x}^{(B)}$ in the SOM neighborhood of $\mathbf{x}^{(A)}$. The special case, where $\mathbf{x}^{(B)}$ is not used in finding the BMU at all is referred to as *heteroassociative mapping* [9].

3.5 Experiments

3.5.1 Data

In order to compare our results with those of Laakso and Cottrell presented for the MLP [11], we prepared the spectrophotometer measurements [2] of the *Munsell book of color: matte finish collection* [1] in the same manner. This resulted in 640 patterns of color spectrum $\mathbf{x}_j^{(S)}$, $(j = 0, 1, 2 \ldots, 640)$ consisting of colors red, yellow, green, blue and purple with hue values 2.5, 5, 7.5 and 10. The pattern is a 12-dimensional vector, where each component represents the reflectance intensity of a color chip measured at 25 nm intervals from 400 nm to 700 nm ranging from 0 to 4095. These spectrum patterns were further encoded as described in [11] into *binary*, *real*, *gaussian* and *sequential* representations $\mathbf{x}^{(S_b)}, \mathbf{x}^{(S_r)}, \mathbf{x}^{(S_g)}, \mathbf{x}^{(S_s)}$ having dimensions 96, 12, 60 and 3, respectively The symbol of the colors R, Y, G, B or P of the spectrum input was encoded into the binary vectors

$$\mathbf{x}_j^{(C)} = \begin{cases} [1\,0\,0\,0\,0]^{\mathrm{T}}, & \text{if the symbol is R} \\ [0\,1\,0\,0\,0]^{\mathrm{T}}, & \text{if the symbol is Y} \\ [0\,0\,1\,0\,0]^{\mathrm{T}}, & \text{if the symbol is G} \\ [0\,0\,0\,1\,0]^{\mathrm{T}}, & \text{if the symbol is B} \\ [0\,0\,0\,0\,1]^{\mathrm{T}}, & \text{if the symbol is P} \end{cases} \tag{3.5}$$

Every sixth of the patterns was taken in the holdout set and the rest were used as the training set.

(answerme:should we open the spectrum encodings more?)

3.5.2 Testing procedure

To study the effects of encoding, a sample of five SOMs, each with random initial values for model vectors, were trained for the four encodings of the color spectrum. Each SOM was configured to use 13x9 neurons in hexagonal lattice and the Gaussian neighborhood

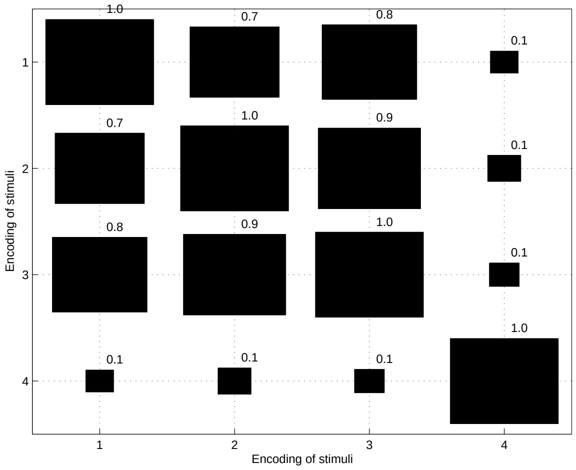

Figure 3.1: Representations for the stimuli are similar except for the sequential encoding. The Hinton diagram displays correlations between pairwise distances of the input patterns of different encodings. The area of a box is proportional to the correlation. Black boxes indicate significant correlation (p-value < 0.05). Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

function. To find the effect of the map size, maps of sizes 3x2, 5x3, 8x5, 10x8, 13x9, 15x11, 18x14, 20x15, 22x18 and 25x20 were trained additionally for the binary and sequential encodings.

Following the partitioning-of-activation-space criterion (Sec. 3.2), Euclidean distances of the activations on each map were computed for every pair of the test input patterns. Pearson correlations and their p-values were then computed for these distances. The activation of a neuron was computed with a Gaussian neighborhood function [9], where the radius was set to 1/10 of the smaller of the dimensions of the map lattice.

For the emergence of an association between spectrum input $\mathbf{x}^{(S)}$ and color symbol input $\mathbf{x}^{(C)}$, these were concatenated to form a single input vector during training (Sec. 3.4.2). The color symbol part of the input was not used in finding the BMU. After training, the map units were labelled with the color symbol, whose component had the highest value in the color part $\mathbf{m}_i^{(C)}$ of the model vectors. This *strongest association* for each test pattern was compared with the respective color symbol of the test pattern. The performance of recall of the color symbol was recorded.

SOM Toolbox [12] has been utilized throughout this study when working with the SOM.

3.6 Results

3.6.1 Similarity of the emergent representations

First we want to get an understanding on how similar the input representations resulting from the four encodings of the color spectra are. For this purpose the correlations between the distances between every input pattern pair is computed as described in Sec. 3.2. Correlations are strong except for the sequential encoding, where the pattern distances only correlate weakly with distances of the other encodings (Fig. 3.1). The strong correlations indicate that the respective encodings have preserved the relative distances between the

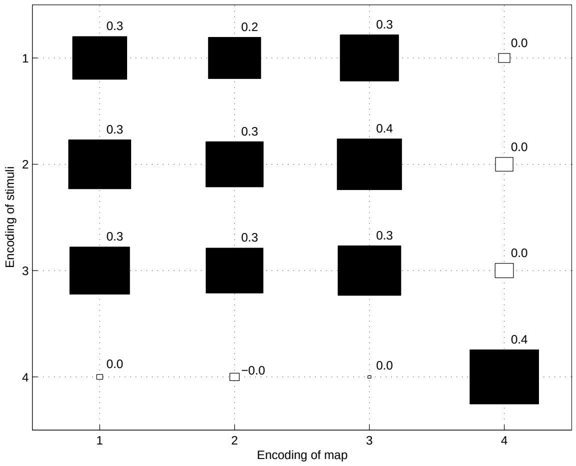

Figure 3.2: Emerged representations for the stimuli in the maps are similar to the representations of the stimuli to some degree except for the sequential encoding. The Hinton diagram displays the mean correlations between distances between input patterns for each encoding and distances between activations of five networks trained on each encoding. Black boxes indicate mean p-value less than 0.05. Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

patterns to a large degree. In these encodings like spectrum samples have close distances and different samples have longer distances. The reason for the weak correlations with the sequential encoded patterns is the peculiarity [11] of the sequential encoding itself that hardly reflects the distances between the original physical spectrum patterns.

Next we examine the similarity between the representation that has emerged in the map for the input stimuli and the representation of the stimuli itself, the encoding. For this purpose the distances between the activations of the map evoked by each input pair are computed. For the five samples of maps trained on each encoding, these are found to be similar to some degree to input representations across encodings (Fig. 3.2). Only the sequential encoded stimuli is not similar to any other representation. As the SOM forms a non-linear mapping that tends to preserve the topology of the input space, there is an expected similarity between the distances of the input pattern and the distances of their representations in the map — in proportion to the similarity of the stimuli encodings themselves (Fig. 3.1).

Finally we compare the emerged organization of representations in the maps with respect to the *partitioning-of-activation-space* view of similarity. The distances between the activations of the maps as response to the spectrum stimuli, do correlate irrespectively of the encoding that the map was trained with, except for the maps that were trained with the sequential encoded stimuli (Fig. 3.3). Those representations correlate only with other maps trained using the sequential encoding.

The representations are similar, where they correlate, to the extent that the same color spectra activate approximately equal positions in the maps relative to activations stimulated by other spectrum samples. If one compares individual responses between two maps they seem to have no relation at all. This is due to the degrees of freedom available to the organization during training. The SOM may take different directions in the organization process depending on the initial values or other randomness in the training phase. Still, the SOM partitions the activation space roughly the same way for the encodings that

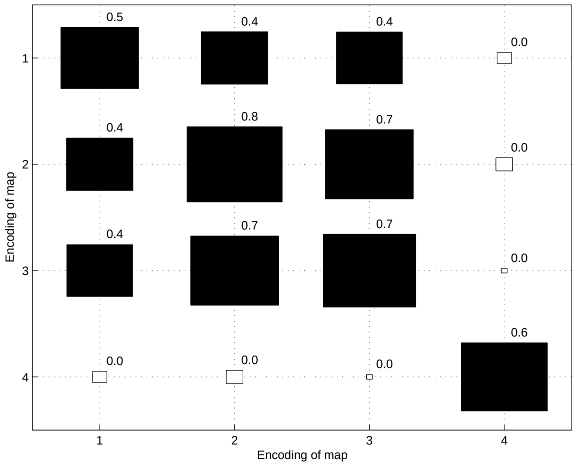

Figure 3.3: The emergent representations in the maps are similar irrespectively of the encoding the map was trained with, except for the maps that were trained with the sequential encoded stimuli. The Hinton diagram displays the mean correlation between activations of five networks trained on each encoding and five networks trained on each other encoding. The area of a box is proportional to the correlation. Black boxes indicate mean p-value less than 0.05. Numbering of the encodings: 1 for binary, 2 for real, 3 for gaussian and 4 for sequential.

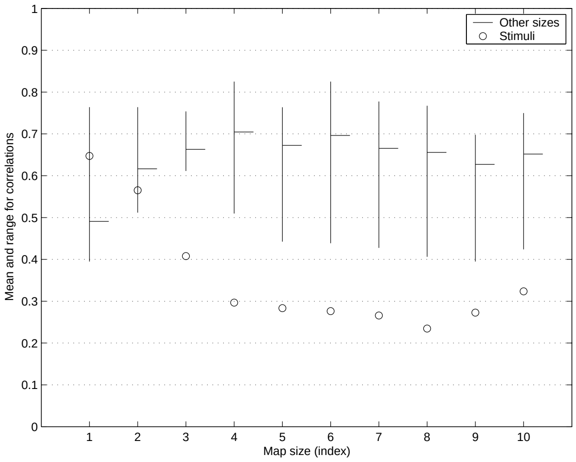

Figure 3.4: Correlations between the representations in the maps of different size trained on the real encoding, and their correlation to input patterns. Map sizes are 3x2, 5x3, 8x5, 10x8, 9x13, 15x13, 18x14, 22x16 and 25x20 (fixme:update sizes).

result in similar input representations. It is worth noting that the correlations seem to be stronger between the activations than between the activations and stimuli (Fig. 3.2).

3.6.2 Effects of scaling of the map capacity

Like in L&C, Figures 3.4 and 3.5 agree that there should be a much weaker relation between the stimulus and the coded representations than across representations. More interestingly, when the map size is small, i.e., there are not enough degrees of freedom to account for the complexity of the data to be coded, the best it can do is to get close to reproducing the input, hence the poor results found as well for different sizes of the maps

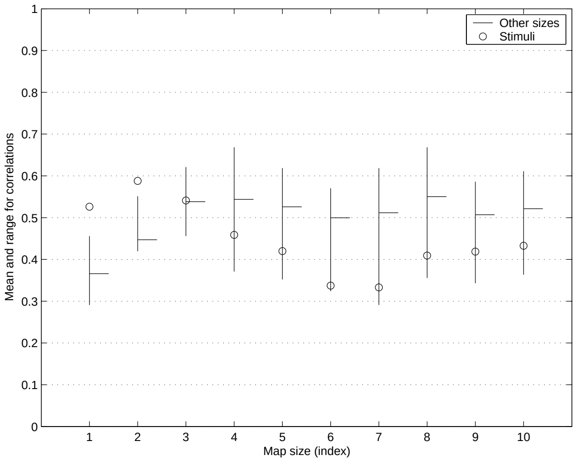

Figure 3.5: Correlations between the representations in the maps of different size trained on the sequential encoding, and their correlation to input patterns. Map sizes are 3x2, 5x3, 8x5, 10x8, 9x13, 15x13, 18x14, 22x16 and 25x20 (fixme:update sizes).

— metaphorically, it would correspond to being able to simply reproduce the inputs in a 'parrot-like' manner.

When the degrees of freedom increase, the map representation is able to reach a 'meaningful' coding of the inputs, in such a way that formation of the internal semantics occurs, hence getting more distant from the inputs, but better structured. Using a similar analogy as in the above, one could say that the map is already capable of understanding the meaning of what it is producing. After reaching a certain degree of complexity, any increase of map size can only help to refine the structuring.

Figure 3.5 shows that, if the input encoding is 'unnatural', it is quite expected that, without a clear external constraint on the representation, i.e., supervision, all maps can not reach the desired representation, hence staying at the level of simply reproducing as much as possible of the input pattern. We could say that these maps have not found any significant internal structure, or content, in the stimuli.

3.7 Discussion

We have studied the relationship between continuous (perceptual) domain and discrete (symbolic, linguistic) domain in supervised learning framework (see also [8]). In particular, we have considered how different encodings or representations of the input data influence concept formation process.

Figures 3.4 and 3.5 show that there is a much weaker relationship between the stimulus and the coded representations than across representations. This was also reported in [11]. More interestingly, when the map size is small, i.e., there are not enough degrees of freedom to account for the complexity of the data to be coded, the best it can do is to get close to reproducing the input. This is the reason for the poor results found for small sizes of the maps — metaphorically, it would correspond to being able to simply reproduce the inputs in a 'parrot-like' manner.

When the degrees of freedom increase, the map representation is able to reach 'meaningful' coding of the inputs, in such a way that formation of the internal semantics occurs, hence getting more distant from the inputs, but better structured. Using a similar analogy as in the above, one could say metaphorically that the map is capable of understanding the meaning of what it is producing. After reaching a certain degree of complexity, any increase of map size can only help refining the structuring.

Figure 3.5 shows that, if the input encoding is 'unnatural', it can be expected that without a clear external constrain to the representation, i.e., supervision, all maps can not reach the desired representation. The maps then stay in the level of simply reproducing as much as possible the input pattern. We could say that these maps have not found any significant internal structure, content, in the stimuli.

The measure of similarity presented in [11] is easily transposable to unsupervised mapping. We still find it to be a very useful one. Emergent representations follow a similar path as supervised codings, as different systems (e.g. varying sizes of maps) reach similar formation of the core content.

We have shown that supervision is not needed in order to gain meaningful representations regardless of the input encoding if the encoding can be considered 'natural'. Of course, raw input may not always be sufficient source for meaning conceptual organization but some external or secondary information is necessary. However, we claim that the statistical characteristics of the primary input data is a reasonable starting point for the formation of conceptual structures.

3.8 Conclusions

The motivation behind the present paper was to examine Laakso and Cottrell findings regarding measures of similarity between representations [11], in emergent, i.e. unsupervised environments. We observed the following:

1. the SOM learns representations both for stimuli and color symbols and is able to associate them successfully,

2. application of the partitioning-of-action-space criterion for measuring the similarity of the latent representations for the stimuli show that the representation are alike almost independently of the encoding used for input.

The discovered usability of this criterion for the emergent representations, adds new support in favour of the state space semantic view of mind, and gives a counter example against the challenges Fodor and Lepore [5] have placed on the connectionist theory.

Acknowledgement

The authors would like to thank the participants of the seminar on Statistical and adaptive approaches to conceptual modeling organized by the Laboratory of Computer and Information Science at Helsinki University of Technology.

Bibliography

[1] Anonymous. *Munsell book of color: matte finish collection*. Munsell Color, Baltimore, 1976.

[2] Anonymous. Joensuu spectra databases. `http://cs.joensuu.fi/spectral/databases/`, 2003.

[3] Paul M. Churchland. Some reductive strategies in cognitive neurobiology. In *A neurocomputational perspective: the nature of mind and the structure of science*, pages 279–309. MIT Press/Bradford Books, Cambridge, MA, 1986.

[4] Paul M. Churchland. Learning and conceptual change. In *A neurocomputational perspective: the nature of mind and the structure of science*, pages 231–253. MIT Press/Bradford Books, Cambridge, MA, 1989.

[5] J. A. Fodor and E. Lepore. Paul Churchland and state space semantics. In R. N. McCauley, editor, *The Churchlands and their critics*. Blackwell, 1996.

[6] C.L. Hardin. *Color for Philosophers*. Hackett Publishing Company, Indianapolis/Cambridge, extended edition, 1995.

[7] T. Honkela. Self-Organizing Maps in Natural Language Processing. Helsinki University of Technology, PhD thesis, 1997.

[8] T. Honkela. Self-organizing maps in symbol processing. In *Hybrid neural systems*, pages 348–362. Springer, New York, NY, USA, 2000.

[9] Teuvo Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences. Springer-Verlag, 3rd edition, 2001.

[10] Teuvo Kohonen and Riitta Hari. Where the abstract feature maps of the brain might come from. *Trends Neurosci.*, 22:135–139, 1999.

[11] Aarre Laakso and Garrison Cottrell. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76, 2000.

[12] Juha Vesanto, Johan Himberg, and Esa Alhoniemi. SOM Toolbox for Matlab 5. Publications in Computer and Information Science A57, Helsinki University of Technology, Espoo, Finland, 2000.

[13] S. Zeki and L. Marini. Three cortical stages of colour processing in the human brain. *Brain*, 121:1669–1685, 1998.

Chapter 4

Modeling Multimodal Concepts

Ville Tuulos, Jukka Perkiö and Timo Honkela

4.1 Introduction

There have been attempts to learn concepts based on text corpora: it is assumed that the statistical analysis of the co-occurrence relationships of the words in the texts would reveal some structures of the external world, at least partly following Language of Thought hypothesis. This appears to succeed to a certain, however limited extent. What appears to be even more important is the possibility to conduct crossmodal learning. In particular, for human beings the development of a large number of concepts is based on visual domain.

Use of multimodal concepts is motivated partly by computational complexity of natural phenomena which involve human cognition either directly or indirectly. Implications of resulting burden of complexity suggests taking a practical bottom-up approach.

We outline our experimental approach to model image similarity using independent component analysis and motivate its use to construct multimodal concepts.

By a widely accepted view the brain is a semantic engine. It captures invariances in perceptual data and associates them with some corresponding internal representations. Representations are flexible so they may accommodate to changing situation. With these representations the brain performs sophisticated inferences eventually leading to actions.

There is a two-fold motivation to study this process where the sub-personal representations seem to have an utmost importance. Firstly, we want to understand and describe the cognitive process within the relevant context. Secondly, we would like to mimic the process, typically with a computational method, and thus be able to handle automatically certain tasks where the brain seem to perform particularly well. However, there is no reason to believe *a priori* that these distinct problems could be solved with the same method simultaneously.

This paper focuses on the second problem of automating processes of conceptual learning based on visual input. The first part of the paper gives some philosophical background to our approach. We point out some reasons why we see that the two problems are not actually the same. We go through some central difficulties in the long tradition of modeling the mind computationally[1] which are mainly seen to be caused by a burden of complexity.

[1]For brevity terms "modeling" and "models" refer to *computational* modeling and models from now on.

The second part of this paper introduces our initial experiments with multimodal concept creation. We give some motivation for using varying modalities to model concepts instead of relying solely on natural language. The bottom-up approach we take is motivated partly by technical reasons of the first section, partly by ideas of perceptual symbol systems.

4.2 Burden of Complexity

For the point of view of modeling, the notion of "mind" can be considered misleading. Folk-psychological concepts like "beliefs", "intentions" and "desires" do not have clear correspondence to functions or states of the brain. Therefore, one can state that their importance for computational models is doubtful, or at least those concepts are only useful in characterizing some meta-level or emergent properties of the system. In an extreme eliminativistic stand even the concept of "concept" might be seen only as a convenient abstraction for humans lacking any relevance *per se* for the models. We may see products of folk psychology as special artifacts of one specific model, namely the human brain. Thus, by focusing on these properties we try to model the symptoms, not the cause.

This is just one instance of a constantly recurring theme in this paper: How to handle complexity. One common sense approach is to focus on facts which mostly constraint the problem at hand. Looking at the past focusing on the folk-psychological concepts seem to have made models more complex, not simpler, mostly due to their vagueness. The next section tries to show why these models may lack descriptive power as well.

4.2.1 Computational models of mind

Let us consider the following situation. You are given a problem $\mathcal{P}$ to solve. Here, $\mathcal{P}$ represents an arbitrary task where $\mathcal{P}$ involves interaction with concepts of the human world. Its level of abstraction may vary from modeling of "conceptual spaces" to recognizing potential aunts among other people in a forest. The idea is here that most of the "natural" problems, where human cognition is involved, employ the delicate machinery of human mind quite thoroughly. Thus, recognizing aunts can be considered to be effectively as difficult problem (to a model) as classifying documents by subject or telling jokes. The idea is analogous to complexity classes (*PSPACE,NPSPACE,PTIME* etc.) in theory of computation[2]. We believe that this is a result of tightly interwoven subsystems of the brain as well as their peculiar structure.

The idea is presented for this discussion to get the right mind-set about $\mathcal{P}$. It should not be read as a metaphysical claim about the inherent nature of the world.

Although we know that humans perform well on task $\mathcal{P}$, our primary task is not to understand why it is so. Let us suppose here that dumb mimicking or simulation is easier than gaining true understanding. If the task is extremely regular and its domain restricted, we may be able to construct a set of static rules to solve the problem, even without looking how the humans solve the same task. If the domain is not restricted enough, we probably face the frame problem as we cannot identify and explicate all the relevant facts with respect to $\mathcal{P}$. On the other hand, if the task is not regular enough but changing over time, a rule-based system lacks flexibility to accommodate to the changing situation. Variants of this approach are most widely used since practically all computer

[2] As with other hard problems, they may be approximated with varying success, but in essence they are as difficult.

programs are still written in a procedural or object-oriented manner which shares many properties with rule-based systems.

Connectionistic models are often usen when a task is too hard to model "by hand". They should solve the dilemma of complexity by changing their behavior based on invariances in the input data. Practically all connectionistic models are based on a non-structured homogeneous graph (self-organizing maps, multilayer perceptrons, etc.). Due to their graphical structure connectionistic models are extremely flexible. It is well known that multilayer perceptrons may approximate an arbitrary function. As their structure (model) is not strongly restricted to any particular kinds of invariances, parameter estimation process have to make some assumptions on the data[3] so that the problem would be computable. Thus, the burden of the complexity is effectively moved to the preprocessing of the data (feature selection) and selection of the model parameters.

It is not difficult to build an artificial neural network model to capture some invariances of the data set but guiding its behavior towards useful and non-arbitrary results is extremely non-trivial. This is a result of having too flexible a model family. Moreover, the effect of certain assumptions in the model and its parameter estimation process are hard to analyze.

The use of connectionistic methods is often argued to be data driven, saying that the input data is supposed to restrict the model to solve the particular problem at hand. This claim is analogous to an extreme *tabula rasa* viewpoint of human mind which hardly has any empirical evidence. Learning by induction is also effectively the same idea. All of these share the same difficulty that the data by itself contains typically arbitrary invariances so one can support almost any claim by looking at the data from the right direction. However, if one were to look for something, it would be possible to use the data to verify whether the search is on the right track. One can either explicitly express what she is looking for or one can trust on the model assumptions. The latter seldom produces wanted results, if the assumptions are not carefully and explicitly chosen.

One crucial feature of human mind is generativity – humans can easily produce infinite amounts of varying written symbols or utterances. Compared to the other models of mind, connectionistic models usually lack this feature. This makes especially evaluation more difficult as we will see later on.

It is no surprise that graphical models such as Bayesian networks have proven to be more useful in many practical applications than their connectionistic counterparts from a particular point of view. Namely, by letting the user specify the properties (assumptions) of the problem explicitly by choosing appropriate graph, the user may better embed her domain-specific knowledge to the system. In case that the user does not have any *a priori* knowledge, she may select some general (simplified) graph structure which is suitable for the problem at hand, like Naive-Bayes model for classification. The complexity has not been magically diminished but now the user has an opportunity to use her knowledge to restrict the problem. Again, the user could let the computer go through a part of exponential amount of possibilities, ranking the models according to some measure. The results would probably tell more about the part of the space and scoring methods used than about the actual $\mathcal{P}$. Therefore, it can also be stated that a connectionist model may be more faithful to a phenomenon under consideration while the user may also introduce erroneous or too limiting assumptions in the case of graphical models.

Metaphorically, the complexity of human mind propagates to the model through $\mathcal{P}$. No

[3] E.g. similarity measure

matter which model you use, you have to somehow handle the complexity or accept arbitrary results. The idea is that you cannot avoid this situation as long as your problem is in connection with the human cognition which inherently seems to "contaminate" the problem with complexity.

This leads to the dichotomy between descriptive and simulating models. Well-working computational models are bad descriptions of a phenomenon as they have captured the complexity of it and thus they cannot be described concisely. On the other hand good descriptions lead often to folk psychology – they are concise, but rough approximations and thus simulate badly the actual phenomenon. Descriptive models are inevitably reductionistic.

The third option would be a simple (deterministic) model producing complex behavior. This idea is actually not far-fetched as phenomena like that are found all over in the nature. However, these processes are almost always irreversible so looking at the data does not tell almost anything about the model, maybe except some sporadic invariances.

The dichotomy has also some implications to evaluation of the models. Simulating models should be easy to evaluate. If they can predict a phenomenon well *a priori* they have captured its essential features. Similarly, solving given $\mathcal{P}$ adequately should prove the model appropriate from a pragmatic point of view. Regrettably many $\mathcal{P}$ are so complex (e.g. natural language processing) that even formalization of an explicit prediction task may be too difficult.

Here it is relevant to ask how to evaluate a descriptive model. By definition, a descriptive model tries to show concisely and understandably essential features of $\mathcal{P}$. Yet, by definition $\mathcal{P}$ is complex, showing complex behavior, so it is practically impossible to just "see" whether the description is correct or not. Thus, the model is not falsifiable. If the model was generative, one could use it to generate some behavior (data) and then compare this generated behavior to the behavior produced by $\mathcal{P}$. This is typically easier than comparing the models directly. Still this approach needs a measure or another model to evaluate similarity between behaviors, introducing yet another problematic issue.

The account described above explains why descriptive connectionistic models are brittle – they can "explain" phenomena *post hoc* due to their flexibility but in a certain way they are too simplistic to be falsifiable. Seldom being generative, they make reliable evaluation almost impossible with respect to a complex problem.

This gives us also motivation for being pragmatic with modeling. Being descriptive inherently restricts complexity, as we must keep our description understandable. Moreover it is difficult to be generative and descriptive at the same time without losing much in the model likelihood. In the end, computer hardware is so different from the brain that following the same principles might lead to sub-optimal solutions. We believe that well working practical models solving actual problems may eventually increase our knowledge about the brain even more than unfalsifiable descriptive ones.

4.2.2 Representations and LOT

As described in the very first paragraphs of this paper, representations (schemata, concept spaces) have a central role in the modern (cognitive) psychology. Developments are reflected back and forth between philosophy, cognitive science and computational intelligence.

We define representations to be the part in the hypothetical cognitive processes where perceptions get transduced after some processing by perceptual systems. Atomic constituents of representational system are symbolic in the sense that their actual form is arbitrary, like words in the language. We do not claim that their implementation or form would be similar to the symbolic systems in the traditional sense. We do not either make any claims about structure or functioning of the system.

In the light of the previous section, constructing reliably a system like the above is not easy. In the descriptive point of view, hypotheses about representation systems do not give many restrictions on the system structure. One can ask whether it is possible to automatically "learn" a representation system. As the system is symbolic, it would be tempting to use some symbolic data in abundance, e.g. written language to model it. This would lead to a language of thought (LOT) hypothesis giving natural language a central or even a definite role in the human cognition.

The LOT hypothesis introduces several well-known problems:

1. Grounding: How arbitrary symbols lead to semantics? Perhaps the most famous exemplification of this problem is Searle's Chinese Room argument.
2. Transduction: How perceptions lead to symbols?
3. Bootstrapping: How to learn language without representations? How to learn representations without language?
4. Pragmatics: Does language have any meaning without use?
5. Evolution: If language is of central importance, how do other animals survive without it?

Following the Occam's Razor it's questionable whether this hypothesis could be correct. At least we may ask whether some alternative hypothesis leads to more effective models. Maybe the LOT hypothesis is a Gordian Knot.

4.3 Perceptual Multimodality

We have an intuition that behind the words there is something more abstract, let us call them concepts, which bring together different modalities (haptic, kinetic, visual etc). Yet, simple co-occurrence statistics of words alone do not seem to coincide well enough with our introspective ideas of what concepts are, partly due to technical reasons described in the first section, partly due to problems of the LOT[4]. Previous attempts to capture hidden variables behind text have produced varying results Instead of trying to build more elaborate models on text, we take a more bottom-up approach.

To be able to evaluate our model we take a practical problem. Given a set of different real-life data sets of different types (modalities), we try to model their co-occurrences so that one modality can be retrieved using another. In practice, we need a model of similarity for each modality and a model of similarity between the modalities.

Following our experimental setting, let us choose two modalities, i.e., text and images. Given an image of a tiger the model should return texts about tiger. Given an example

[4]If LOT hypothesis is not exactly correct, it is quite improbable that written language would reflect conceptual system with any satisfiable precision – refer also to Von Foerster.

document about tigers, the model should return images of tigers. Some implementations with these modalities (text and images) exist already using probabilistic approach. Our setting models image similarity with Independent Component Analysis (ICA) [2], which to our knowledge is not used before in this kind of an image retrieval task. Multinomial Principal Component Analysis (MPCA) is used for modeling language.

One can easily see that the experimental setting has no descriptive power with respect to the brain. However, we believe that systems combining different perceptual modalities may exhibit behavior which closely resembles results of human inference in certain tasks. Use of perceptual input (like images) directly in inference in spite of dubious transduction to some symbolic representation has some useful properties:

- Problems of grounding and transduction diminish. According to [1] concepts in perceptual systems may be seen as simulators which reproduce neural activations similar to those which typically co-occur with external phenomena behind the concept. Correspondingly, commonly co-occurring perceptions form a concept. Naturally the correspondence is not quite this simple in practice.
- Discrimination between syntax and semantics is not as crucial as it has often considered to be. In a LOT-based view syntax and semantics are orthogonal properties. If concepts are seen as multimodal co-occurrences of perceptions the distinction becomes more blurred. In cognitive linguistics, this conclusion has been made, e.g., by Langacker [7].
- Perceptual multimodality seems to suit well with the current understanding of evolution. Different modalities may have developed somehow separately at different pace. This stand agrees well with Dennett's Pandemonium-model for competing modules of mind.
- Neurophysiological evidence from fMRI imaging with certain tasks. See again[1] for details.

For more detailed discussion on benefits of perceptual symbol systems, see[1]. The above reasons are motivating also with respect to the first section: They introduce some structure to the model (separate models for different modalities with clear characteristics). It also provides a simpler explanation for the phenomenon of interest than a LOT-based approach. In the following section, we describe our setting to evaluate whether the model works also in practice.

4.4 Experiments

4.4.1 Measuring image similarities

Now we have the problem of measuring similarities between different images. That is a very widely researched area and there is a vast number of different statistical techniques that may be used. We use ICA as the framework for image processing that we need.

4.4.2 ICA

Independent component analysis (ICA) is a statistical technique to find hidden factors from data. The observed data is assumed to be a linear or non-linear mixture of some latent variables i.e. hidden factors. ICA is a generative model and it assumes latent variables to be non-Gaussian and mutually independent. The basic ICA model in its linear form is following

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_i a_i s_i, \tag{4.1}$$

where $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)^T$ is the observed data, $\mathbf{A}$ is the linear mixing matrix and $\mathbf{s} = (s_1, s_2, s_3, \ldots, s_n)^T$ is the latent data, which is also called the components.

Now the task is to find the mixing matrix $\mathbf{A}$ so that the components $\mathbf{s}$ can be estimated as

$$\mathbf{s} = \mathbf{A}^{-1}\mathbf{x}. \tag{4.2}$$

As it was mentioned independent components can be estimated also for non-linear mixtures but for our purposes we only need the linear version of ICA. In our context also the number of components and the number "observations" is always the same.

There are many different algorithms to estimate the independent components. One of the fastest and most widely used is the FastICA algorithm developed in [3].

4.4.3 ICA for image data

As it was mentioned ICA assumes some hidden factors in the data and that those factors are somehow more characteristic to the data than the observed data in itself. For image processing the feature extraction is a fundamental problem and it seems very intuitive to assume that in an image data there might be also some hidden factors that could be used as features to discriminate between images. It has been shown that the receptive fields of V1 simple cells resemble the independent components estimated from natural images [5]. That makes it even more attractive to use ICA for feature extraction. ICA has been used for feature extraction [4] but normally it has been done to many different images collectively and not to single images.

The application of ICA to image data is normally done by sampling small windows from the image and then the independent components are estimated from these samples. There are questions related to the nature of the sampling window e.g. its size, form and its functional form. Common choice but probably not the best one is to use a square sampling window of size 12^2 to 16^2 pixels. That approach have some drawbacks and a better choice is to use a round smoothly decaying window [6].

Independent components estimated from image data share some very nice properties. They are not only local in spatial domain but also they are local in frequency domain and orientation. In that they resemble Gabor filters that are widely used in image processing.

We use ICA to create a filter set for an image and then we apply this filter set on other images to estimate how similar those images are to the image from which the filter set was estimated.

4.4.4 Creating the filter set

We produce the filter set in the following manner:

1. We sample the image using a rectangular sample window to produce 256-dimensional vectors as we are dealing with gray-scale images.

2. Local mean is subtracted from the vectors.
3. The dimensionality of these vectors is reduced to n using PCA.
4. Independent components are estimated from this n-dimensional data.
5. The estimated components are projected back to the 256-dimensional space to produce the filter set.

Now we have a single image $\mathfrak{I}_k$ and a set of images

$$\mathfrak{I} = \bigcup_k \mathfrak{I}_k \tag{4.3}$$

and a filter set

$$F^k = \bigcup_i F_i^k, \tag{4.4}$$

where i denotes a single filter. We calculate for each image its specific filter set and we get a set of filter sets

$$\mathfrak{F} = \bigcup_k F^k, \tag{4.5}$$

which contains one filter set containing n filters for each image $\mathfrak{I}_k$.

As it happens, different images produce different filter sets and these filter sets can be used for discriminating between images. We could try comparing those filter sets directly. However a better solution is to compare the outputs of those filters applied to an image. Now we denote the output of filter set F^k on image $\mathfrak{I}_j$ with O_k

$$O_k = F^k(\mathfrak{I}_j). \tag{4.6}$$

Output of a filter set on an image is calculated simply by sliding the filter over the image and calculating a dot product between the filter and the image window below the filter. That produces a vector, which size depends on the way we do the sliding. In our case we slide the filter horizontally from left to right and from top to bottom using different step sizes. It seems that the step size is not very critical as long as it is small enough to capture local changes in the image e.g. from 1 to 8 pixels seem to be all right. Of course that depends on the image resolution.

4.4.5 Using the filter outputs

Now that we have the filter responses from a set of images including the image from which the filter set is calculated. We could compare these responses and try to estimate the similarities based on that. However if we calculate the probability distributions of these responses we can use different measures e.g. Kullback-Leibler divergence or Jensen-Shannon divergence for the comparison. We calculate for each filter set output O_k its empirical probability distributions that we denote with D_k. Now

$$D_k = \bigcup_i p_i^k, \tag{4.7}$$

where p_i^k is the distribution of a single filter F_i^k. In order to do that there are some technical questions like discretization etc. but they are rather trivial.

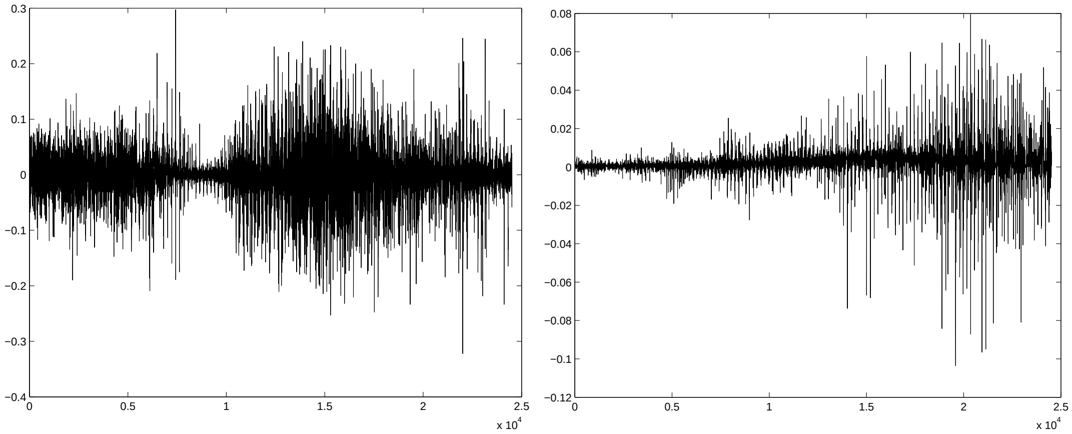

Figure 4.1: The output of same filter from two different images. The upper plot is from the same image that the filter was calculated from.

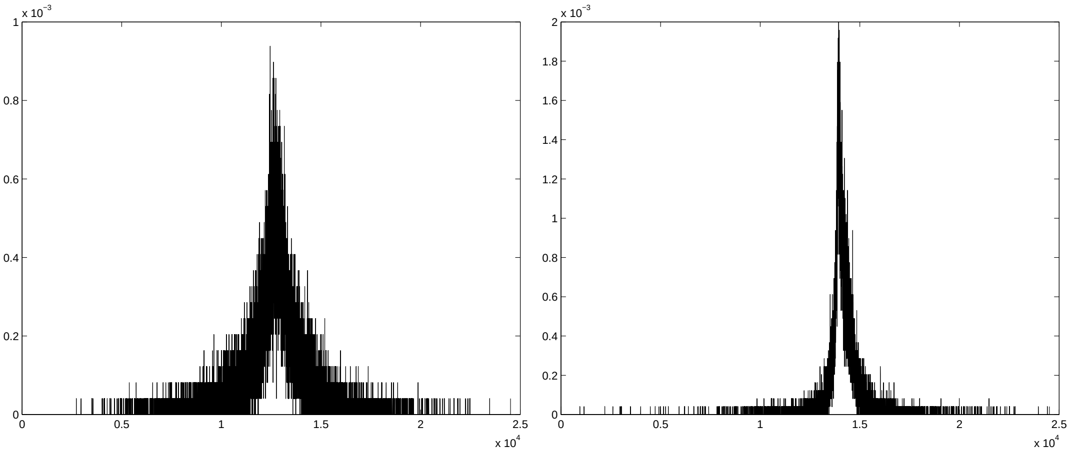

Figure 4.2: The probability distributions of the filter outputs in figure 4.1. The plots are in the same order as in figure 4.1.

Now that we have the distributions of the filter outputs we can compare those distributions directly. The Kullback-Leibler divergence of two discrete distributions is defined as

$$KL(p_1||p_2) = \sum_i p_1(x_i) \log \frac{p_1(x_i)}{p_2(x_i)} \tag{4.8}$$

as we can see it is not a symmetric operation but it is always ≥ 0 and it is 0 only if $p = q$. Furthermore the Kullback-Leibler -divergence can easily be done symmetric if we define

$$KL_{symm}(p_1||p_2) = KL(p_1||p_2) + KL(p_2||p_1). \tag{4.9}$$

That is still not a true metrics as it does not satisfy the triangle inequality. However that is not a problem in our case.

Another option would be to use Jensen-Shannon divergence, which is a true metrics if we take its square root. Jensen-Shannon divergence is defined as follows. Let us define the average of two distributions p_1 and p_2

$$\bar{p}_{12} = \frac{p_1(x) + p_2(x)}{2}, \forall x \in X. \tag{4.10}$$

Now Jensen-Shannon divergence of distributions p_1 and p_2 is defined as

$$JS(p_1||p_2) = \frac{KL(p_1||\bar{p}_{12}) + KL(p_2||\bar{p}_{12})}{2}. \tag{4.11}$$

As it was mentioned if we take the square root of Jensen-Shannon divergence it then is a true metrics in a sense that it satisfies non-negativity, reflexivity, symmetry and the triangle inequality. However in this context we prefer symmetric Kullback-Leibler divergence for its simplicity.

Now that we have calculated the Kullback-Leibler divergences between given distributions we can deduce the similarity between one image compared to others. In other words we produce a ranking telling which image is closest to a certain image. Our empirical tests are not yet finished but at the moment our approach looks quite promising.

4.4.6 Assessing the method

The method explained in this chapter can be seen as a query model to retrieve images. That is we want to retrieve images that are similar to one particular image. For our purposes that is sufficient as there are always more than one image involved in the process and we are mostly interested in seeing whether other images are similar to one particular image and which one is the most similar. The method explained does not produce any "real" distance information as all we can say is that one image is closer to another than some other image.

As it was mentioned we only produce a ranking according to similarity to one image. Intuitively it feels very hard to give any absolute metrics for distances between images the way human observers do on average. The empirical tests have not yet been performed completely but the preliminary results are encouraging.

Bibliography

[1] Barsalou, L.W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.

[2] Hyvärinen, A. Karhunen, J. and Oja, E.: *Independent component analysis*. John Wiley & sons, 2001

[3] Hyvärinen, A.: Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks*, 10(3):626-634, 1999.

[4] Hurri, J., Hyvärinen, A., Karhunen, J. and Oja, E.: *Image Feature Extraction Using Independent Component Analysis*. Helsinki University of Technology, 1996.

[5] Hyvärinen, A. and Hoyer, P.O.: Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705-1720, 2000.

[6] Hurri, J.: *Independent component analysis of image data*. Helsinki University of Technology, 1997.

[7] Langacker, R.W. *Foundations of Cognitive Grammar, Vol. I: Theoretical Prerequisites*. Stanford University Press, 1987.

Chapter 5

On Representation of Action within Real-World Situations

Kevin I. Hynnä, Mathias Creutz, Tarja Knuuttila, and Timo Honkela

5.1 Introduction

In traditional artificial intelligence, models of action have been considered in the context of planning and decision making (see, e.g., Levison 1996, Tiebaux et al. 1994). It has been commonplace to represent the action directives as rules that determine the actions of an agent. In real-world contexts in which the agent receives perceptual input and may have to act in a natural environment, the input data and the low-level action directives are not straightforwardly in symbolic form. For instance, if the plan for a robot in some hypothetical action description language would be `(GET SALT (FROM-PLACE TABLE))` the concepts[1] `SALT` and `TABLE` refer among other things to complex patterns in the visual scene. Moreover, the agent would need to have information of the position about its effectors and their relation to the physical item in some spatial coordinate systems in order to accomplish the actual grasping. This grounding problem has widely been acknowledged as a central problem for symbolic representations. Therefore, a link between the symbolic level and the perceptual or pattern level is needed.[2]

Wermter and Sun (2000) define hybrid neural systems as computational systems which are mainly based on artificial neural networks (often referred to as connectionist systems) but which also allow a symbolic interpretation or interaction with symbolic components. The artificial neural networks deal naturally with perceptual level. The recognition of the limitations of symbolic and connectionist approaches as such has lead into active research that combines or integrates these approaches within hybrid systems (consider, e.g., Dorffner (1997), Hilario (1995), Barnden and Holyoak (1994), Miikkulainen (1993), Reilly and Sharkey, Ritter and Kohonen (1989), and Wermter (1995).

Gärdenfors (2000) has developed a conceptual spaces theory (CST) that he provides as an

[1] Here we use the word 'concept' even though we consider concepts to be complex patterns in continuous multidimensional spaces and which cannot usually be represented symbolically.

[2] Even if we assume that we could build a successful cognitive agent model without the use of explicit symbolic descriptions, symbolic representations are presumably needed in communication with human beings. In principle, artificial agents could have a language, or rather a form of communication which is based on continuous representations.

alternative for plain symbolic representations of knowledge and as a higher level of representation of information processing than in connectionist systems. Conceptual spaces are spanned by a number of quality dimensions which can be used to describe, e.g., concept formation. Gärdenfors divides in his theory the levels of representation into symbolic, conceptual and subconceptual. As opposed to theories based solely on symbolic representations, he considers how to take into account symbol grounding and perceptual input in the concept formation process.

Gärdenfors (2000) does not, however, explain in detail where the quality dimensions are derived from and how, e.g., they are to be related to perception or behavior. Moreover, Gärdenfors' theory, in its current form, is focused on static objects and perceptual qualities. He characterizes one consequence of this restriction as follows: "another large class of properties are the functional properties that are often used for characterizing artifacts. For example, Vaina (1983) notes that when deciding whether an object is a "chair", the perceptual dimensions of the object, like those of shape, color, and weight, are largely irrelevant or at least extremely variable. Since I have focused on such variables in my description of conceptual spaces, the analysis of functional properties is an enigma for my theory." Gärdenfors also provides similar criticism on the ability of his theory to deal with action concepts.

Action is used to refer to:

1) Action words or concepts

2) Actions as behavior of a model, i.e., ability to make decisions or perform actions.

Investigation of (1) is ultimately useful only to the extent that it aids in our explanations of behavior, of both real and simulated agents. Nevertheless, there is an intimate, interrelated relationship between the ability to identify and classify other people's actions and the classification and performance of one's own actions, as evidenced by recent work concerning so-called 'mirror neurons'.

In this chapter, we bring together a number of recent trends from cognitive modelling that have arisen from concerns over the ability of a purely symbolic approach to ground symbols in any realistic, biologically plausible way. our main goal is to present a unified framework which is sensitive to the criticisms laid at symbolic approaches,, and which can be used to guide investigations involving robot or agent-based simulations.

After outlining and motivating the basic tenets of this framework we consider whether Gärdenfors' approach to conceptual modelling is in accordance with this framework. We conclude that it is possible to align Gärdenfors' approach with the framework outlined here, but in order to do so it is useful to consider Gärdenfors' approach as falling within the more general framework of Dynamical Systems Theory.

5.2 Ontological and epistemological assumptions

When we discuss the representation of action relevant to agents in real-world contexts, it is useful to characterize what we mean by the real world, i.e., what are our ontological assumptions. In addition to that, in the following we also outline the basic epistemological points of view. First of all, the world is a dynamic, continuous process. The continuity is a relevant point of view, for instance, when the symbolic or linguistic level is considered, in which matters are discretized. Cognitive agents perceive and conceptualize the world to consist of objects (persons, trains, houses, flowers, countries, etc.) and events (leaves falling, people running, earthquakes, television reporting an earthquake, etc.) and they use words and phrases, e.g., to communicate them in propositional form. Concepts reside in the minds (and even brains) but they are formed in a cultural and historical process which involves perception, communication and collaboration, activity, etc.

Maturana and Varela's theory of autopoietic systems provides a useful account on considering the nature of the agents that are responsible for conducting actions. Maturana (1978) describes autopoietic systems in the following manner: "There is a class of dynamic systems that are realized, as unities, as networks of productions (and disintegrations) of components that: (a) recursively participate through their interactions in the realization of the network of productions (and disintegrations) of components that produce them; and (b) by realizing its boundaries, constitute this network of productions (and disintegrations) of components as a unity in the space they specify and in which they exist. Francisco Varela and I called such systems autopoietic systems, and autopoietic organization their organization [...]. An autopoietic system that exists in physical space is a living system (or, more correctly, the physical space is the space that the components of living systems specify and in which they exist) [...]."

Maturana (1978) continues: "Autopoietic closure is the condition for autonomy in autopoietic systems in general. In living systems in particular, autopoietic closure is realized through a continuous structural change under conditions of continuous material interchange with the medium. Accordingly, since thermodynamics describes the constraints that the entities that specify the physical space impose on any system they may compose, autopoietic closure in living systems does not imply the violation of these constraints, but constitutes a particular mode of realization of autopoiesis in a space in which thermodynamic constraints are valid. As a result, a structurally plastic living system either operates as a structurally determined homeostatic system that maintains invariant its organization under conditions of continuous structural change, or it disintegrates."

In summary, autopoietic (living) systems are autonomous and they aim at keeping themselves alive but in their autonomy they are also constantly "in the world".

If we follow the theory of autopoietic systems the ontological status of objects and events as objective, distinct and independent of any observer can be questioned as Maturana (1988) does: "Since everything said is said by an observer to another observer, and since objects (entities, things) arise in language, we cannot operate with objects (entities or things) as if they existed outside the distinctions of distinctions that constitute them. Furthermore, as entities in language, objects are brought forth as explanatory elements in the explanation of the operational coherences of the happening of living in which languaging takes place. Without observers nothing exists, and with observers everything that exists exists in explanations."

Von Foerster (1972) has presented a similar account: "Objects and events are not primitive experiences. Objects and events are representations of relations. Since 'objects' and 'events' are not primary experiences and thus cannot claim to have absolute (objective) status, their interrelations, the 'environment' is a purely personal affair, whose constraints are anatomical or cultural factors. Moreover, the postulate of an 'external (objective) reality' disappears to give way to reality that is determined by modes of internal computations."

It is important to note that Maturana, Varela and Von Foerster do not wish, with their point of view, to deny the existence of the external world. They merely point out that cognitive, living agents construct their description of the world, and this description consists of constructed categories such as objects and events along with their associated subcategories. Each of those constructions is subjective but at the same time their formation is based on the interaction with other agents as well as artefacts that reflect the structural characteristics of the constructions of other agents.

5.3 Embodied cognition approach

In this chapter we study the embodiment of action in robotics, in particular. We refer extensively to Pfeifer and Scheir's work. Similar argumentations and results can be found, for instance, in Brooks (1991) and Steels (1995).

5.3.1 Sense-think-act cycle

Planning is the core of traditional artificial intelligence (AI). Based on a comparison of a representation of a goal state to the current state, an agent constructs a plan for moving from its current state to the goal state. A popular strategy has been means-end analysis (Newell and Simon, 1972). Means-end analysis requires a measure of distance between the current state and goal state. Operators [actions] are then chosen on the basis of an evaluation of how much their application will reduce the distance to the goal state. Operators are associated with certain preconditions, which have to be met in order for an operator to be chosen. Preconditions can be achieved by applying other operators [actions]. This produces a chain of subgoals that leads to the final goal state. The final goal state is thus attained by applying subgoal-directed operators in a sequence.

This kind of planning method seems intuitively plausible, but it is subject to a combinatorial explosion. For example, if there are 10 branching points in a plan, and at each branching point there are two possibilities, there will be 2^{10} or roughly 1000 different plans. Such planning systems have therefore not been very successful on real robots. (Pfeifer and Scheier, 1999)

Means-end analysis goes nicely with the classical idea that intelligent behavior is based on a sense-think-act cycle. The agent first perceives something (sense), it then processes what it has perceived (think) and finally executes an action (act). In terms of means-end analysis, the agent has to be able to *sense* its environment in order to determine its current state. The *think*ing corresponds to setting up a plan including subgoals; and finally, *act*ing consists in choosing appropriate operators.

5.3.2 Frame-of-reference problem

In their book "Understanding intelligence" Pfeifer and Scheier (1999) challenge the idea that intelligent behavior is based on a sense-think-act cycle. They set up a framework for embodied cognitive science with the goal of designing so-called complete agents. A complete agent is autonomous, i.e., it is independent of external control to a certain degree. Furthermore, it is self-sufficient, which means that it is capable of sustaining itself for extended periods of time without any human intermediary. It is embodied, i.e., realized as a physical robot or as a simulated agent, and it is situated, thus acquiring information about its environment solely through its own sensors in interaction with the environment. Complete artificial agents are inspired by natural agents, animals and humans, which are capable of surviving in the real world. Being "complete" means incorporating everything that is required to perform actual behavior. (By contrast, standard computer programs are not complete because they cannot behave in the real world.)

Pfeifer and Scheier argue that as far as the behavior, or the actions, of an agent is concerned, we have to be aware of the frame-of-reference problem: "We have to distinguish between the perspective of an observer looking at and agent and the perspective of the agent itself. In particular, descriptions of behavior from an observer's perspective must not be taken as the internal mechanisms underlying the behavior." Or put differently: "We must make a distinction between rational thought, which concerns the mechanism within the agent, and rational behavior, which pertains to the agent's interaction with the environment. Rational behavior is, of course, behavior and can thus be perceived by an observer. It is not necessary, in order for rational behavior to take place, to postulate goals and knowledge as being explicitly represented within the agent. In other words, rational thought is not a prerequisite for rational behavior."

Pfeifer and Scheier stress that the behavior of an agent is always the result of system-environment interaction. It cannot be explained on the basis of internal mechanisms only. They illustrate that the complexity that we as observers attribute to a particular behavior does not always indicate accurately the complexity of the underlying mechanisms. Experiments with very simple robots that merely react to stimuli in their environment have shown that rather complex behavior can *emerge*. In the study of robot communities, where robots can interact with other robots in their environment, nature-like behavior has been observed, such as bird-like flocking (REF) and ant-like heap building (REF).

5.3.3 The action selection problem

Pfeifer and Scheier further address the action selection problem: "Complete systems always have several behaviors in which they must engage. Some of the behaviors are compatible, others mutually exclusive. Because not all behaviors are compatible, a decision must be made as to which behaviors to engage in at each point in time...

The most straightforward solution to this problem is to assume that there is an internal module or representation for each observed behavior category. For example, if we observe that a rat (or a robot) is following a wall, we might postulate that it has an internal module or a representation for wall following. Such a representation is often called an action... [T]o control behavior under this assumption, you need a mechanism for which action to choose for execution at any given point in time...

The problem with this approach to behavior control is that the assumption of a straightforward, one-to-one mapping from a specific behavior to a specific internal action does not reflect what actually occurs in natural systems... There are two issues of which to be aware: First, the segmentation of an agent's behavior is observer-based and largely arbitrary... Second, it is not appropriate to conclude that for each of these behavioral segments there is an internal module."

As an example of different possible segmentations of a particular event Pfeifer and Scheier contrast the action "going to class" with the action sequence "getting up from chair", "moving left leg forward", "moving right leg forward", and so forth. These are observer-based categories chosen according to some criteria.

5.3.4 Loosely coupled, parallel processes

In contrast to earlier AI and cognitivistic approaches, Pfeifer and Scheier demonstrate mechanisms for behavior control that *do not require the existence of internal actions*. These mechanisms are based on the so-called subsumption architecture, which was first introduced by Brooks (1986). Subsumption is a method of decomposing a robot's control architecture into a set of task-achieving behaviors or competences. In contrast to the traditional approach, information from different sensory systems (e.g., vision, auditory) is *not* integrated to a central representation. The subsumption architecture consists in the incremental adding of task-achieving behaviors on top of each other. Implementations of such behaviors are called layers. Instead of having a single sequence-of-information flow – from perception to world modeling and planning to action (sense-think-act) – there are multiple paths, the layers, that are active in parallel. Each of these paths is concerned only with a small subtask of the robot's overall task, such as avoiding walls, circling around targets, or moving to a charging station. Each layer can function relatively independently without having to await instructions or results produced by other layers. (pp. 201–202)

The use of layers naturally leads to extendable designs in which new competences can simply be added to the already existing and functioning control system. At each level, there are sensory inputs and motor outputs. Higher levels, just like lower ones, can directly interact with the environment, without the need to go through lower levels. Subsumption combines robot design with evolutionary principles. Once a particular competence is in place and is working (such as moving around and avoiding obstacles) it should no longer be changed when new competences are added (such as approaching a light source).

However, a certain amount of interaction between layers is necessary, even though this interaction should be minimized in order to facilitate the design process and to achieve maximum incrementality and emergence of optimal behavior. Subsumption means that higher layers can subsume lower ones, i.e., suppress their input or inhibit their output. For instance, a chair-pushing robot has to stop avoiding obstacles when it has to push a chair.

5.3.5 Examples of emergent complex behavior

Wall-following robot

Pfeifer and Scheier describe several experiments where complex behavior emerges as a result of robot-environment interaction. In the following example, a neural network is used for controlling the robot. The robot lives in an "ecological niche", a closed environment with obstacles and light sources. The light sources are placed along the walls and the obstacles are distributed randomly over the open space. The robot has three kinds of sensors: collision, proximity and light. The sensors are distributed along the front half of the robot, and two are in the back. There are two wheels, each with one motor. The robot has a number of built-in reflexes. If it hits an obstacle, activating a collision sensor on its either side, it will back up a little and turn to the other side. Whenever it is sensing light on one side, it turns towards that side. If it senses no obstacles and no lights, it simply moves forward.

In the beginning, the robot will crash into obstacles often. However, over time it will "learn" the correlation between the activation of its collision sensors and its proximity sensors. This leads to avoidance behavior before the actual collision takes place. The robot is thus "anticipating" a collision and taking corrective actions in order to continue its forward moving behavior.

As a consequence of its built-in reflexes, the robot will approach light sources along the walls. However, as it approaches the wall, it turns away because it has learned to avoid obstacles. Now the turn-toward-target is activated again, and the robot wiggles its way along the wall. Over time wall-following behavior emerges. The robot will further associate light with lateral stimulation in its proximity sensors, and will continue to follow the wall, even though the light is switched off. If we assume that the light represents food, we could say that the robot has learned that food is normally located along walls. Because it has this knowledge, it follows the walls hoping to find food, even if it currently doesn't sense any. This is our interpretation as observers. All that happened within the robot was a change of weights in the neural network.

Categorizing robot

In another experiment, a similar robot has the task of distinguishing between small and large objects. It should learn to collect small object with its gripper and bring them to its nest. Large objects are too large to fit into the gripper and they should be avoided. The robot has only two types of sensors: infrared sensors (for proximity) and wheel encoders. Seven processes are working in parallel: move-forward, avoid-obstacle, turn-toward-object, grasp-object, turn-away, and bring-to-nest.

To distinguish between objects, the robot develops a strategy to explore the objects in such a way that the sensor readings are correlated in time and correlations exist between different sensors. Circling is such a behavior. However, there is no simple one-to-one mapping between behavior and the underlying mechanism: The robot normally moves forward and when it senses an obstacle it avoids it by turning away. At the same time, if it senses stimulation in one of its lateral sensors, it turns slightly toward the object. (Such a reflex ensures that the robot has a tendency to be near objects that are generally more interesting than open spaces.) The interaction of these three processes leads to a behavior that can be called move-along-object or circling. Circling implies that the motor speeds are different for the left and the right wheel. The larger the difference, the smaller the object being circled. The robot can thus learn to categorize objects in its environment by interacting with them.

In addition to emerging behavior, this example illustrates that there is no special categorization module. Most current models study categorization in isolation, but Pfeifer and Scheier claim that categorization makes sense only with respect to the complete agent and involves sensory-motor coordination. By interacting with an object to be categorized the agent overcomes the object constancy problem, i.e., the problem of determining which parts of the input belong to one and the same object. The problem is hard, because the same object can lead to very different input patterns depending on the viewing angle, light conditions, noise in the sensors, etc. The problem is much reduced, if rather than passive perception active interaction with the environment is allowed. In this way, the agent can produce sensory input that is considerably easier to process.

5.3.6 Scalability and self-awareness

Whether the embodied cognition approach scales to human levels of intelligence is unknown. Pfeifer and Scheier feel that they have not currently found any in-principle reasons why their principles should cease to work at some level of complexity. However, they suggest that a distinction be made. Are we trying to understand human intelligence or are we trying to build an enormously complex robot, comparable in complexity to a human? If the former is the goal, they say that they have shown that relatively simple agents can be employed to study issues in human intelligence (e.g., category learning). The point is more about *understanding* than scale, and about how to employ models in the scientific process. But if the latter is the goal, i.e., building highly complex robots, we indeed face a scaling issue.

Furthermore, Pfeifer and Scheier speculate whether an artificial agent could eventually develop a sense of "self". Patterns and regularities in sensory data and behavior, though emergent, are objectively measurable. These regularities can potentially be observed and used by the agent itself. This might form the basis on which the agent could develop a sense of "self" that would be grounded in physical interactions rather than an abstract entity living in the agent's head. It would enable the agent to develop knowledge about its own sensory-motor setup and its relations to the world. It is possible that such a need is only present in social environments, where the agent has to be able to communicate about his own needs or capabilities.

5.3.7 Embodied meaning in Neural Theory of Language

In the previous section, we demonstrated that seemingly intelligent behavior need not be accompanied by any explicit internal conceptual representation. In this section, we demonstrate how the embodied cognitive approach can incorporate the notion of concepts and language.

Feldman and Narayanan (2004) propose a neural theory of language (NTL) that is based on embodiment. They claim that one should not expect to find brain areas specialized only for language or to find language processing confined to only a few areas of the human brain. The NTL assumption is that people understand narratives by subconsciously imaging (or simulating) the situation being described, e.g.:

"When asked to grasp, we enact it. When hearing or reading about grasping, we simulate grasping, being grasped, or watching someone grasp... The action of grasping has both a motor component (what you do in grasping) and various perceptual components (what it looks like for someone to grasp and what a graspable object looks like). There are other modalities involved as well, such as the somato-sensory component (what it feels like to grasp something or to be grasped yourself)."

The NTL approach to language suggests that the complex synergy that supports grasping *is* the core semantics of the word. If we accept this complex of neural circuits and behaviors as the core meaning of grasping, it remains to show how a word like "grasp" gets associated with the embodied concept. Apparently, in the approach proposed by Pfeifer and Scheier this would theoretically not be a problem. If an agent were to associate a certain behavior with hearing or uttering a certain linguistic expression, hearing that expression could in the future trigger the behavior, and the behavior could in turn trigger the uttering of the expression.

However, Feldman and Narayanan identify a problem, the correlation problem. A child (or a robot) must learn what features of the situation and of its actions the parent (or tutor) is talking about. Furthermore, different languages differ widely in the way that they label actions. In English there are quite many verbs denoting actions involving hands (e.g., grasp, grip, drop, tug, lift, nudge) and other languages make distinctions that English does not. But according to Feldmann and Narayanan, if the meaning of an action word were supposed to be the activity of a vast distributed network of neurons, there seems to be a *complexity barrier*. The key to solve this problem is *parameterization*. All people share the same neural circuitry and thus the same semantic potential. "A motor action such as grasping involves many coordinated neural firings, muscle contractions, etc., but we have no awareness of these details. What we can be aware of (and talk about) are certain parameters of action – force, direction, effector, posture, repetition, etc. The crucial hypothesis is that languages only label the action properties of which we can be aware. That is, there is a fixed set of embodied features that determine the semantic space for any set of concepts, such as motor actions... [C]ontrol of action can also be parameterized and thus made available to language learning."

By claiming that parameterization, and thus discretization, of the semantic space is crucial, Feldman and Narayanan indirectly take a stance on the action selection problem. The segmentation of an agent's behavior is indeed largely arbitrary, and different languages can use different categorizations, but there is a minimal set of action properties that we can be aware of. This makes learning of concepts and language possible in the first place.

The authors do not restrict themselves to single words describing concrete actions. They extend their theory to abstract words and entire sentences. In the NTL, abstract and metaphorical words derive their meanings from concrete words. There are, e.g., several metaphorical uses of the word "grasp": "grasp an idea", "grasp an opportunity" etc. NTL suggests that all understanding involves simulating or enacting the appropriate embodied experience. "This ability to simulate or imagine situations is a core component of human intelligence and is central to our model of language." The understanding of whole sentences is based on constructional composition according to the principles of Construction Grammar (Bergen and Chang 2002).

So far, we have discussed how an agent can learn concepts and language relating to its own actions. But how can an agent identify actions performed by other agents? Several studies on so called "mirror neurons" that have demonstrated the existence of a mirror system in the brains of apes and humans, such that the same neurons are activated both when an action is *observed* and *performed* (cf., e,g, Meltzoff and Decety 2003). Thus, the same neural circuitry can be used for observing and performing action, and furthermore *imagining* action. It does not seem far-fetched that a particular activation pattern of neurons could trigger a linguistic expression that is related to a particular kind of situation, and vice versa.

5.4 Dynamical Systems Theory Approach

We began this paper with a problem, namely, how do we extend Gärdenfors' conceptual space theory (CST) to account for the actions of agents. In particular, we considered the problem of modelling autonomous agents which are capable of acting and interacting adaptively and flexibly in some environment. Following Pfeifer and Scheier (1999) and Brooks (1991), methodological considerations suggest that a key requirement of any such models is that the agents be situated and embodied, i.e., they deal with the here and now of the real world around them, and they experience the real world directly via their bodily manifestations. In this section, we will attempt to make explicit the connection between CST and the modelling of the action of autonomous agents, and we will do so by suggesting that Gärdenfors' model be recast in terms of dynamical systems theory

(DST). Or, in other words, we will argue that sufficient similarity exists so as to warrant characterizing the former as an instance of the latter, and doing so allows us to quite naturally extend CST to the domain of actions.

A word of caution is in order. As noted by van Gelder (1998), dynamical modelling taken by itself "does not somehow automatically constitute an account of cognition. It is a highly general framework which must be adapted, supplemented, fine-tuned, etc., to apply to any particular cognitive phenomenon". In other words, the DST approach is simply a framework, a set of tools, as it were. Establishing that CST can be viewed from a DST perspective does not in and of itself accomplish very much. Demonstrating that this is a *useful* exercise is the hard part. In this paper, we only suggest why this might be the case. Accordingly, our strategy will be as follows: First, we will briefly introduce the dynamical systems framework. Then, we will show that this framework is in principle capable of accomodating the observations about the design of autonomous agents which have arisen from the field of robotics and were discussed in the previous section. Finally, we will discuss why it seems plausible to consider CST as an instantiation of DST. Here we can only introduce some of the main concepts which characterize the DST approach. There are numerous introductions to the more formal aspects of DST available, with one common source being Abraham and Shaw (1982). Bechtel (1998), Beer (1995), Elman (1998), Smolensky (1986), van Gelder (1998), all introduce and discuss some more or less methodological aspect of the DST approach as it applies to the modelling of cognition. Kelso (1995), and Thelen and Smith (1994) develop the DST approach to cognitive modelling, and van Gelder and Port (1995) serves as a standard overview of the DST approach which includes numerous papers involving applications of the DST approach to cognitive modelling. Although what follows is largely based on the presentation found in Beer (1995), it is a standard treatment of the material.

5.4.1 Dynamical Systems

Simply put, a dynamical system is a system that changes or evolves over time. A dynamical system can be characterized by a set of state variables, x, and an evolution equation (or dynamical law), F, that governs how the values of the state variables change over time. The set of all possible values of the state variables constitutes the system's state space. In a continuous-time dynamical system, the evolution equation is given in terms of a set of differential equations.[3] Often, change in the system depends on factors outside the system itself, which are usually referred to as parameters. If the evolution equation depends only upon the values of the state variables and the values of some set of fixed parameters, u, then the system is said to be autonomous. A non-autonomous dynamical system, by contrast, is one in which one or more parameters are allowed to vary in time. Time-varying parameters can be thought of as inputs into the system.

Note that this formal, mathematical notion of autonomy of a dynamical system should not be confused with what we mean by an autonomous agent, which was informally introduced above as an agent that is independent of external control but which, following Beer (1995), can now be specified further as "any embodied system designed to satisfy internal or external goals by its own actions while in continuous long-term interaction with the environment in which it is situated". As we shall see below, when describing agents (or systems) within DST, they may turn out to be non-autonomous in the mathematical sense,

[3] If time is considered discretely then difference equations are used. Here, we shall assume that time is continuous.

but still, intuitively at least, be considered autonomous as agents, per se (see below).

$$dx_1/dt = F_1(x_1; u_1) \tag{5.1}$$

$$dx_2/dt = F_2(x_2; u_2) \tag{5.2}$$

Let equations (1) and (2) give the evolution equations for two systems, s_1 and s_2 respectively. Equation (1) simply says that x_1, the state variable of system s_1, changes over time as some function of the (current) state variable itself, x_1, and some set of parameters, u_1. Equation (2) says the same thing for system s_2. A very important phenomena in DST occurs if the state variable of one system appears as a parameter in the evolution equation of another system, and vice versa. In this case, the two systems are said to be coupled. So, if x_2 is a parameter in equation (1), and x_1 is a parameter in (2), then systems s_1 and s_2 are coupled.

Given some initial state x_0, the sequence of states generated by the application of the evolution equation is called a *trajectory* of the system. In describing a dynamical system, one is usually interested in the qualitative long-term behavior of the system. The state of some systems will simply diverge to infinity (which we shall not consider further here), while others will eventually converge over time to a limit set, which is simply a set of points that is invariant with respect to the evolution equation, meaning that if the system ever enters this region of its state space, it will stay there indefinitely. Of particular interest are the *stable* limit sets or attractors. All trajectories passing nearby to an attractor converge to that attractor and the set of initial state points for which this holds is called the basin of attraction for that attractor. Four classes of attractors are usually distinguished based on qualities of the trajectories associated with each: point attractors, periodic attractors (or limit cycles), quasi-periodic attractors and chaotic attractors.

5.4.2 Agents and Environments as Dynamical Systems

Let us model an agent and its environment as two dynamical systems, A and E, respectively.

$$dx_A/dt = A(x_A; u_A), \tag{5.3}$$

$$dx_E/dt = E(x_E; u_E). \tag{5.4}$$

It is obvious that an agent and its environment are in constant interaction. Let $S(x_E)$ represent the agent's (sensory) inputs from its environment and let $M(x_A)$ represent its (motor) outputs to the environment. (3) and (4) then become:

$$dx_A/dt = A(x_A; S(x_E), u'_A), \tag{5.5}$$

$$dx_E/dt = E(x_E; M(x_A), u'_E). \tag{5.6}$$

where u'_A captures all of the agent's parameters u_A that do not come from the environment and u'_E represents all of the parameters of the environment u_E that do not depend on the agent. Note that S and M are defined rather broadly. S, for example, is intended to capture *all* of the effects of the environment E on agent A, regardless of whether these effects occur as part of what is normally regarded as 'sensory input'. Likewise M is meant to capture *all* effects agent A has on the environment E, whether or not they occur as a result of what would normally be considered the 'motor output' of an effector. Note that what we have just outlined is simply a *coupling* between an agent and its environment. Any action by the agent affects its environment in some way through M, which in turn affects the agent itself through the feedback from its environment via S. Likewise, the environment's effect on the agent through S are fed back through M to affect the environment itself. In other words, what we informally described as a constant interaction between an agent and its environment is straightforwardly implemented formally using the notion of coupling.

Since equations (5) and (6) are defined with respect to continuous real time, we have implicitly satisfied the criterion of situatedness mentioned above. What then of embodiment? As noted by Beer (see also Chapter 2 and van Gelder(1998)), the division between an agent and its environment is always somewhat arbitrary (e.g., is an artificial limb or a tool part of the agent or part of the environment?) It should be clear however that a dynamical systems approach is in principle capable of capturing (given some appropriate set of agent variables) the notion of embodiment as outlined above. We can make one final point about using a DST approach to capture the relationship between an agent and its environment. Although two coupled systems, such as A and E as given in (5) in (6) above, are non-autonomous in the formal sense mentioned earlier, intuitively, at least, the agent 'acts' independently of any external control. Furthermore, as Beer notes, formulating the agent-environment relationship in this way captures nicely one of the central themes of recent autonomous agent research, namely, the idea that an agent's behavior arises not simply from within the agent itself, but rather through its interaction with its environment. Note finally, that two non-autonomous coupled systems such as A and E taken *together* can be viewed as a single autonomous dynamical system U (whose state variables are the union of the state variables of A and E and whose evolution equation is modified accordingly). The question therefore is simply one of scale.

5.4.3 Conceptual Space Theory as a Dynamical System

Points of Contact

In this section, we present prima facie support for the similarity of the CST and DST approaches.

1. Emphasis on geometry and topology.

   "Here, I advocate a third form of representing information that is based on *geometrical* structures rather than symbols or connections among neurons." (Gärdenfors 2000, p. 2)

   A property is "a *region* of a conceptual space S, where 'region' should be understood as a spatial notion determined by the topology or geometry of S." (Gärdenfors 2000, p. 67)

   "The main difference among these theories and the one presented here is that I put geater emphasis on the geometrical structure of the concept representations." (Gärdenfors 2000, p. 105)

"Dynamicists... conceptualize systems *geometrically*, i.e., in terms of positions, distances, regions, and paths in a space of possible states. DST aims to understand structural properties of the *flow*, i.e., the entire range of possible paths." (van Gelder, p. 11)

2. Importance of similarity relations.

   "One notion that is severely downplayed in symbolic representations is that of *similarity*. Judgments of similarity are central for a large number of cognitive processes... the similarity of two objects can be defined via the distance between their representing points in a conceptual space..." (Gärdenfors 2000, p. 44)

   "[T]he domains we consider have a *metric* so that we can also talk about *distances* between points in the space." (Gärdenfors 2000, p. 71)

   "[D]ynamical systems in cognitive science might be defined as quantitative systems. Roughly, a system is quantitative when there are *distances* in state or time, such that these distances matter to behavior." (van Gelder, p. 7)

   "A system that is quantitative... is one whose states form a *space*; states are *positions* in that space, and behavior are paths or trajectories. Thus quantitative systems support a geometric perspective on system behavior, one of the hallmarks of a dynamical orientation." (van Gelder, p. 8)

Points of Disagreement

Gärdenfors considers but rejects identifying CST directly with DST approaches, in a one-to-one manner:

"Conceptual spaces are *static* in the sense that they only describe the structure of the representations. A full model of cognitive mechanisms not only includes the representational form, but also a description of the processes operating on the representations. A particular conceptual space is, in general, compatible with several types of processes, and it must therefore be complemented with a description of the *dynamics* of the representations to generate testable predictions." (Gärdenfors 2000, p. 31)

Literature

Abraham, R.H. and Shaw, C.D. (1982). Dynamics, the Geometry of Behavior. Benjamin Cummings, Reading, Massachusetts.

Badler, N.I., Bindiganavale, R., Allbeck, J. and Schuler, W., Zhao, L., Lee, S.-J., Shin, H., and Palmer, M. (2000). Parameterized Action Representation and Natural Language Instructions for Dynamic Behavior Modification of Embodied Agents. AAAI Spring Symposium 2000.
http://hms.upenn.edu/software/par/

Barnden, J.A. and Holyoak, K.J. (eds.) (1994). *Advances in connectionist and neural computation theory, volume 3.* Ablex Publishing Corporation.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist challenge in cognitive science. *Cognitive Science*, 22, 295-318.

Beer, R.D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(12):173–215.

Bergen, B. and Chang, N. (2002). *Embodied Construction Grammar in Simulation-Based Language Understanding.* Technical Report TR-02-004, International Computer Science Institute.

Brooks, R. A. (1991). Intelligence Without Representation. *Artificial Intelligence Journal*, vol. 47, pp. 139-159.

Dorffner, G. (1997). *Neural Networks and a New AI.* London, UK: Chapman and Hall.

Elman, J.L. (1998). Connectionism, artificial life, and dynamical systems: New approaches to old questions. W. Bechtel and G. Graham (eds.), *A Companion to Cognitive Science*, Basil Blackwood, Oxford.

Feldman, J. and Narayanan, S. (2004). Embodied Meaning in a Neural Theory of Language, *Brain and Language*, 89, 385-392.

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

Hilario, M. (1995). An Overview of Strategies for Neurosymbolic Integration. *Proceedings of the Workshop on Connectionist-Symbolic Integration: From Unified to Hybrid Approaches,* Montreal, pp. 1-6.

Kelso, J.A.S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior.* MIT Press, Cambridge, MA.

Kopp. L. (2003). *Natural Vision for Artificial Systems, Active Vision and Thought.* PhD Thesis, Lund University Cognitive Science.

Levison, L. (1996). *Connecting planning and acting via objectspecific reasoning.* PhD thesis, University of Pennsylvania.

Maturana, H.R. (1978). Biology of Language: The Epistemology of Reality. Miller, George A., and Elizabeth Lenneberg (eds.), *Psychology and Biology of Language and Thought: Essays in Honor of Eric Lenneberg*, New York: Academic Press, pp. 27-63.

Maturana, H.R. (1988). Ontology of Observing - The Biological Foundations of Self Consciousness and the Physical Domain of Existence. *Conference Workbook: Texts in Cybernetics*, Felton, CA: American Society for Cybernetics.

Meltzoff, A.N. and Decety, J. (2003). What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society*, 358:491-500.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing.* MIT Press, Cambridge, MA.

Newell, A. and Simon, H.A. (1972). *Human Problem Solving.* Prentice-Hall, Inc., Englewood Cliffs NJ.

Pfeifer, R., and Scheier, C. (1999). *Understanding intelligence.* MIT Press, Cambridge, MA.

Reilly, R. G. and Sharkey, N. E. (1992). *Connectionist Approaches to Natural Language Processing.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, vol. 61, pp. 241-254.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D.E. Rumelhart, J.L. McClelland and the PDP Research Group, *Parallel*

Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations. MIT Press/Bradford Books, Cambridge, MA, pp. 194-281.

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

Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA.

Vaina (1983). From shapes and movements to objects and actions. *Synthese*, 54:3-36.

van Gelder, T.J. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 1-14.

van Gelder, T.J. and Port, R. (1995). It's About Time: An Overview of the Dynamical Approach to Cognition. R. Port & T. van Gelder (ed.), *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge MA, pp. 1-43.

Wermter, S. (1995). *Hybrid Connectionist Natural Language Processing*. London, UK: Chapman and Hall, Thomson International.

Wermter, S. and Sun, R. (eds.) (2000). *Hybrid Neural Systems*. Springer.

Von Foerster, H. (1981). Notes on an epistemology for living things. *Observing Systems*. Intersystems Publications, pp. 257-271. Originally published in 1972 as BCL Report No 9.3., Biological Computer Laboratory, University of Illinois, Urbana.

Chapter 6

Concept Learning by Formation of Regions

Jan-Hendrik Schleimer, Mikko Berg, Jaakko Särelä, and Timo Honkela

6.1 Introduction

In this paper, we discuss the issue of conceptualization. The traditional view is that concepts are essentially linguistic. Recently, Gärdenfors has proposed a contradicting view where the concepts get associated to language terms, but essentially belong into other domain called conceptual spaces defined by quality dimensions. These dimensions form meaningful representations of the concept domains in hand and they should be formable by mappings from the sensory input and possibly from other more basic quality dimensions as well.

In the space spanned by the quality dimensions, natural concepts form convex regions. The borders of these regions can be hard or soft and can vary according to the context. In the present work, we have decided to code the regions by prototypes, so that instances closest to a particular prototype in the conceptual space form a region. In other words, the regions are defined by the Voronoi tessallations of the prototypes, which then define hard bordered regions. In the case of soft borders, the prototypes can consist of probabilistic density functions defining graded membership function for each point in the conceptual space.

In this paper, we mainly discuss the formation of the regions of concepts via different kinds of clustering approaches. Some discussion of the connections to the lower, connectionist level and to the higher, symbolic level are discussed in brief.

Intelligent systems generalize and compress the complex input they get from their perceptual organs. This is extremely necessary to survive in a complex and hostile world. Human beings have an exceptional capacity to utilize this process. We often rise from the basic regularities of the world to secondary, sometimes very abstract, models. This makes it possible to exploit even very distant (in time or place) similarities to make effective predictions of the state of the world. Another trait of humans is the capability to complex communication. Probably for robustness reasons, this communication occurs mainly using discrete symbols, words.

Until now, the biggest issue in artificial intelligence (AI) is arguably the relation between these very central traits: effective modeling of the world (accessed by sensory organs) and effective communication of the relevant parts of these models (language).

The traditional view, formulated by Newell & Simon [17] is that we are physical symbol manipulating systems. This is to claim that the models we have of the world, are essentially linguistic. The modern view relies on dynamic systems theory [10]. It claims that symbols emerge from dynamic interaction processes.

In the early 90's, the connectionist paradigm for AI gained popularity, mainly through two books [14, 11]. There it was argued that human information processing is mainly continuous not discrete. Furthermore, the essential feature of human intelligence is learning, thus making it a dynamic process rather that a static one. One drawback of most of the connectionist algorithms are their distributed knowledge representation, which does not allow explicit interpretation of the inference process. Therefore these systems are sometimes referred to as "blackboxes". A famous example is the NetTalk system from Seinowski and Rosenberg, a multi-layerd perceptron capable of reading English texts. The system was trained in a supervised manner with text as input and corresponding phonemes as output. Although achieving an accuracy of 95% the neural network did not extract rules for the decision making, that could be interpreted by linguistic processing. This shows the gap between the connectionist models and symbol manipulation systems.

Connectionism can be interpreted as a special case of associationism using ANN (Artificial Neural Networks). Gärdenfors [7] presents a new level on top of that trying to reach the symbolic level proceses that humans are naturally capable of. *Functional model* Gärdenfors states that conceptual spaces can be seen as a set of attractor points of dynamic systems. Yet, his model with the three levels: connectionism as the lowest, conceptual spaces in between and classical symbol manipulation as the highest level, retains the possibilities of classical symbol manipulation approaches to AI. Also Kelso [10] agrees that attractors of dynamic systems could be interpreted as prototypes.

Domains in conceptual spaces are an attempt to give functional and contextual focus for otherwise ambiguous symbolic level. One concept can be evaluated in several domains using different salience weights, where as properties are domain specific. Scale of the particular dimension in a domain is obtained using contrast classes. In another words, the continuous mapping to the subspace is performed within the boundaries of contextual extreme values. For example, what is considered to be (phenomenological) hot for bathing water is merely warm for coffee. In general, different abstractions are created with the corresponding *quality dimensions* having specific metrics. But although the explanation of how these domains and quality dimensions come about is not satisfactory answered in [7], it will not be further discussed in this article. Here we are taken them as given, as some of them result from innate biological structures with evolutionary background. This is of course not true to all dimensions that are more abstract and which can be learned.

In conceptual spaces, (natural) concepts are defined as (convex) regions[1] The convex spaces are necessarily result of Voronoi tessellations in Euclidean metrics. Voronoi tessellation partitions given space based on prototypical attractors. Clustering methods tackle the reverse problem, by defining regions which detect the prototypes.

The nature of a concept in conceptual spaces is

1. prototypical, coding of the structure

[1]According to Gärdenfors, natural concepts are the only concepts that can participate in inductive reasoning.

2. regional, geometric area instead of points (objects are very narrow concepts, perhaps even points), this makes the concepts vague or fuzzy, which relates to frame theory

In a sense prototype and frame theory are combined.

It still seems that there are a lot more reason for vagueness in concept formation and even more in communication them. There are thought to be 3 different entities in interaction: (1) cognitive concepts (including laws of psychology), (2) language and social interaction, and (3) phenomenal common world (including laws of physics). Connection between any of them is considered to be a source of impreciseness or fuzziness.

Next two sections discuss these two essential properties, one at a time. After that, in Sec. 6.2 we review two clustering methods, as well as discuss the possibility to combine similar concepts into more general concepts corresponding to larger regions in the conceptual space. Finally, in Sec. 6.3 we apply these clustering methods to divide a space with colour quality dimensions into concepts according to two images differing in characteristics.

6.1.1 Prototype theory

Prototype theory was formulated by Rosch and got started from findings relating to typicality (not yet having prototypical structure) among the category members. Findings of Rosch and Mervis [20] emphasized typicality as opposed to all category members representing the category equally. Rosch [19] found that there are more typical members that are learned faster and serve as cognitive reference. The membership was considered to be graded and it was shown not to result from frequency or familiarity of the particular test items. The correlations with frequencies turned out to be useful in many cases, but not being definitive. As an exception, chicken is frequent, but not typical bird. The results of Rosch & al. [21] supported this finding, but only when structural relations between items were held constant.

After that, the characterising properties were the target of the research. First, Wittgenstein's family resemblance rate was found to describe categories better. There were no explicit definitions, but similarities between individual group members, that could be modeled with locally similar cells. Second, exclusiveness (not total) was also proposed as typicality measure. Then the typicality would not only relate to the features of particular group, but also to the shortage of important features from other groups (contrast category). This is the phenomenon that Gärdenfors' [7] quality dimensions are explained to obtain their scaling. Contrast categories are difficult to verify empirically, because it would involve all the (other) categories. Third, it was found that broader knowledge structures and top-down processing play their part in this as well. For example functionalities can be inherited to sub-categories [20]. Barsalou [2] repeated the related experiments later.

The actual prototype theory was based on one summary representation of all the members, not as commonly misunderstood on the best match. Based on psychological experiments, Strauss [23] proposed a method, in which features of the prototype should be averaged if their distribution is small and counted distinctively if it is sparse. The counting was explained by subject's interpretation as qualitative differences, not on one continuos axis. There is an analogy to how Gäärdenfors' dimensions evolve from integral, having correlation, to distinct separable dimensions, for example when child learns to separate shape from colour. Feature correlations are method for applying prototypes and correlations alone are not sufficient for categorization. In terms of conceptual spaces, after arbitrary mapping, any two points in space can be close to each other. It has been claimed that people use hierarchical clusters. The intermediate groupings effect the typicalities, for example the statement that robin is a typical bird may be overlooking the fact that it is small, chirping, worm- or seed-eating tree bird [13].

Rosch [19] describes the vertical dimension of the structure as taxonomy of category relations. There is inclusiveness of subordinate (lower-level) through basic level into superordinate (higher-level). The basic level categories is a topic with much empirical research. Read more from [19]. The horizontal dimension is segmented structure without clear-cut boundaries. There is only the judgment for clearness of the case, the prototypicality.

There has been the idea of using probabilities to increase the accuracy of categorization and for example Churchland [5] uses term warranty for uncertainty of chosen prototype. Experiments of Ross and Murphy [22] showed that this was not actually accounted and turned the focus on preciseness of categorization.

There should be discussion about what extent can human cognition be modeled with prototypes or with ANN (Artificial Neural Networks) algorithms such as SOM and this discussion should be guided by psychological research, not ideological speculations. Some of such attempts find the limitations in the past are context model (started by Medin & Schaffer [15]) exemplar effect and human memory, and different models about the use of background knowledge (e.g. [16], read more from "Theory-Theory" in [12])

SOM vector as Roschian prototype represent summary of all the members of the cell, not the best match. The prototype theory doesn't provide any model for the process, representation or learning. It only presents constraints and a possibility to deal with abstractions without any context. One of such constraints or descriptions is that there is correlation structure of the neighbors in nature of family resemblance [19]. This is the exact way in which input of SOM map is connected, because it gives emphasis on retaining the local level structure. There is no explicit way to define how SOM creates the model vectors, because the process is a result from heuristic principles. Neither is there evidence there should be such for prototype theory. For example independent cue model [15] is only one ineffective implementation.

6.2 Possible implementations in conceptual spaces

Identifying concepts with regions in the space already adds the element of vagueness to the concept representation, because it subsumes objects $\boldsymbol{x} \in \mathbb{R}^d$ with a variety of different attributes as one concept.

Assuming that there is a further vague element, namely that objects do not utterly belong to concepts or putting it in a probabilistic way, there are varying probabilities with which different objects are explained by a concept. Then the hard margins of the regions, representing concepts, in the plainly geometric approach make it difficult to incorporate this vagueness. A possible solution is to define a probability distribution in the conceptual space, that itself corresponds to a concept.

Finding the regions can be solved by clustering methods, but it is as well necessary to infer how many clusters are needed, and in an dynamical environment, the decision whether to split or combine regions, respectively concepts, arises. This question can be partly solved by hierarchical clustering methods or moving to Bayesian versions of clustering algorithms that give evidence on the model complexity, e.g. the number of concepts needed.

In the following sections we discuss three methods for finding these regions and defining a vague concept in them. It is assumed that the objects, perceived in nature or encountered in a more abstract way in our mind, are represented in as points in a conceptual space [7].

6.2.1 K-means clustering

The *k-means clustering algorithm* [3] moves a chosen number of k cluster centers, so that they cover the whole data and thereby partitioning it for $i \in [1, k]$ into subsets $\mathcal{S}_i$, defined by their center $\boldsymbol{\mu}_i$ and containing the N_i nearest data points. It does it via minimizing the sum-of-squares error function,

$$E = \sum_{j=1}^{k} \sum_{n \in \mathcal{S}_i} ||\boldsymbol{x}_n - \boldsymbol{\mu}_i||^2 \tag{6.1}$$

but other distance measures can be used as well [the batch version of the algorithm has an update rule $\Delta\boldsymbol{\mu}_i = \eta(\boldsymbol{x}_n - \boldsymbol{\mu}_i)$ quite similar to SOM's only lacing the neighborhood function]. With the help of the mean vectors a Voronoi tessellation can be found, as used by Gärdenfors for concept representation.

The defined regions are vague representations of concepts. But if the euclidean distance is used to identify the k nearest neighbors or even a tessellation, than there are hard margin between concepts, which does not seem to be a natural representation.

6.2.2 Density estimation

As shown in [3] the k-mean algorithm can be regarded as a limit of the EM optimization of a *Gaussian mixture model* (MOG) with a common variance, when $\sigma^2 \to 0$. In a Gaussian mixture model the probability density of the data $p(\boldsymbol{x}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n)$ is modeled as a weighted sum of Gaussians

$$p(\boldsymbol{x}_n) = \sum_{i=1}^{k} p(\boldsymbol{x}_n|i)p(i). \tag{6.2}$$

with a soft max prior $p(i) = \frac{\exp(\gamma_i)}{\sum_j \exp(\gamma_j)}$ and $p(\boldsymbol{x}_n|i) \sim \mathcal{N}\boldsymbol{\mu}_i\sigma_i$. The negative log-likelihood of the data

$$-\log(p(\boldsymbol{x})) = -\sum_{n=1}^{N} \log \left\{ \sum_{i=1}^{k} p(\boldsymbol{x}_n|i)p(i) \right\} \tag{6.3}$$

can be used as an error function. Finding the minimum by setting the derivatives for $\boldsymbol{\mu}_i$, σ_i^2 and γ_i to zero and using the the Bayes' theorem to get the corresponding posterior $p(i|\boldsymbol{x}_n) = \frac{p(\boldsymbol{x}_n|i)p(i)}{p(\boldsymbol{x}_n)}$, the following updating rules can be derived

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_n p(i|\boldsymbol{x}_n)\boldsymbol{x}_n}{\sum_n p(i|\boldsymbol{x}_n)} \tag{6.4}$$

$$\hat{\sigma}_i^2 = \frac{\sum_n p(i|\boldsymbol{x}_n)||\boldsymbol{x}_n - \hat{\boldsymbol{\mu}}_i||}{d\sum_n p(i|\boldsymbol{x}_n)} \tag{6.5}$$

$$\hat{p}(i) = \frac{1}{N}\sum_n p(i|\boldsymbol{x}_n) \tag{6.6}$$

Due to the nonlinear dependencies in the equation a iterative update scheme is used to solve the problem. Starting with random initial values for the parameters and then calculating the posterior and the new parameter values. It can be shown that repeating this process will converge to a maximum likelihood solution.

Applying this algorithm to points in a conceptual space results in a probability density function that covers the structure of the points arrangement in the space. This distribution can be identified with a certain concept, where the mean vectors of the Gaussian mixture components are prototype like examples of them. The individual Gaussians can represent more detailed sub-concepts. But still remains the question of how many centers shall be used.

Another unsolved problem is that, when operating the algorithm on every object of the conceptual space one large MOG distribution will result and therefore only one concept. So one has to use the clustering in a hierarchical way. For example first tessellate in a crude way to find different concepts using the k-means algorithm and than find the distributions in the cluster with the help of a Gaussian mixture model.

6.2.3 Hierarchical clustering

In stead of applying the above mentioned clustering methods repeatedly one can utilize a *hierarchical clustering* in the first place. A possible class of methods are called single linkage algorithms for a detailed description see [18]. These algorithm start by treating every data point as one cluster and than combine the "most similar" according to the used metric. This is done repeatedly using minimum, maximum, the average distance or the distance of the centers of gravity[2] for comparing clusters containing more than one data point, and thereby creating a hierarchical structure.

The lower branches in the hierarchy can be cut away, meeting the concerns of difference only to a certain level of detail. But how is it then that a concept generating process in an intelligent system could find a level that is meaningful to use? There are two answers at hand: (i) just use any detail level for a start, and then, by a process similar to natural selection in living creatures or maximizing the model evidence in AI, it will turn out to be more useful to go into a more detailed version of the concepts or to thin them out and therefore have broader concepts; (ii) in a Bayesian version of the clustering algorithms, in spirit closer to density estimation, it is possible to combine the data likelihood with a prior distribution, representing the anticipation for the number of concepts needed, which can itself result from previous knowledge and experience in the world, and hence get a posterior probability distribution over the needed number of concepts.

6.2.4 Bayesian mixture model

Deriving concepts from available facts, e.g. sensory data and existing knowledge of the world - in this case represented in conceptual spaces, is an inferential task with statistical properties, resulting from the irregularities in the frequency of the data and the incertitude of the already gained knowledge, respectively.

A mathematical framework for describing statistical inference problems is the Bayesian statistics, where a basic idea is to interpret the probability of an event as the *degree of*

[2] this relates to discussions in prototype theory about which set member should be used as the representative

belief on the occurrence of that event. Learning the attributes $\boldsymbol{\theta}$ of a model structure $\mathcal{H}$ e.g. the shape and location of the gaussians forming the distribution associated with a concept, is achieved by combining prior knowledge, described by a distribution indicating the believe in certain facts, with new information from data $\boldsymbol{x}$, described by a likelihood of the data given the learned quantity and the model structure. A possibility to calculate the posterior distribution of the attributes, which combines old an new knowledge is given by Bayes' theorem

$$p(\boldsymbol{\theta}|\boldsymbol{x},\mathcal{H}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta},\mathcal{H})p(\boldsymbol{\theta}|\mathcal{H})}{p(\boldsymbol{x}|\mathcal{H})}, \tag{6.7}$$

with $p(\boldsymbol{x}|\mathcal{H}) = \int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{H})d\boldsymbol{\theta}$ being known as the *model evidence.* This integral over all possible parameter values is, for difficult distributions not always solvable, but maximizing it with respect to $\mathcal{H}$ would lead to more optimal model structures.

This calculation of the posterior can be conducted each time new data is available and if the posterior distribution of the former inference step is used as the prior in the next execution of the Bayes' rule, it will lead to an adaptive learning mechanism. An intelligent system acting in a new environment and starting to conceptualize from scratch might in some circumstances not have prior knowledge for the shape of concepts, and therefor the categorization of the new and unknown. Still it is possible to define *non-informative priors*, that do not influence the finding of the posterior for the attributes, but "let the data speak for its self".

As mentioned earlier, one can express the density estimation problem in the bayesian framework (see [1] for a detailed derivation). One advantage is that this treatment allows searching for optimal model structure, e.g. the number of gaussians in the mixture model, whereas this is not feasible in the ML solution (paragraph 6.2.2) without empirical regularization terms. This is due to the fact that the ML solution from the EM algorithm prefers more complex model structures, that fit better to the data.

The approach in [1] is from the structure of the algorithm related to EM, but utilizes a helpful technique in bayesian inference called variational learning. There the posterior distribution of the parameter, that is often complicated to calculate, due to the difficult integral in (6.7), is approximated by a distribution with desired properties. In the case where the best model structure should be determined the requirement is that the approximate model evindence needed to optimize the number of gaussian components can be obtained in closed form.

It should be mentioned that there are many other model selection techniques like bootstrapping [6], cross-validation, Markov-Chain-Monte Carlo sampling and Bayesian Information Criterion (BIC), see[8], which all somehow work in practice, but most of them are theoretically only justified for infinite data sets, whereas concepts can certainly emerge from only a view examples.

6.3 Clustering of color spaces into concepts

As a simple example the conceptualization of colors in two pictures, originating form a landscape in summer and winter, was studied. Choosing these pictures it can be expected that the process of conceptualization in our model depends on the encountered examples, a peculiarity of concept forming, that can be observed in the real world, e.g. considering various ethnic groups, that divide the color spectrum in to differently detailed colors [4, 9].

The color code for the pixel elements of the pictures is the hue-saturation-value color map, which is a intuitive representation for humans. The colors are coded with three numbers, firstly the *hue*, ranging from 0 to 360 degree in a circular arrangement and indicating the color type according to its wavelength, secondly the *saturation* or intensity between 0-100%, telling how grayish the color is and finally the *value* in percentage, that tells the brightness or the spread of wavelength. The hsv color space is redundant because there exists white and black for every color. Therefore, a color spindle instead of the cylinder in HSV model has been suggested. It is achieved by reducing the range of the saturation linearly as the intensity approaches 0 or 100 %. This modified color code has been used in the experiments and the intervals were scaled to unit.

A representative set of the data points for the summer and winter pictures can be seen in Figs. 6.1 and 6.2 respectively. The prototypes for the MOG model, i.e., the means of the Gaussians have been marked there with x's as well. As expected, the MOG model has used more resources that is, more prototypes to account for areas having more data points. Observing that they cover the distribution of the color samples quite well, the corresponding colors can be expected to cover the coloring, present in the picture, appropriately. But the results depend completely on how many initial mixture components are chosen.

Thus the clusters given by the EM algorithm were further combined to bigger clusters by the hierarchical linkage algorithm. The resulting colors as well as the hierarchy can be seem in Figs.. 6.1 and 6.2 for the summer and winter pictures respectively. Now one can see the grouping of different shades of white an brown to a more general concept of the color.

Definite differences in the prototype colors can be seen. While the clusters formed from the summer picture have several shades of green and dark gray, the colors in the winter picture are concentrated in lighter shades of gray and white.

[More inferences of the results are made, when we have the results of the spindle model. Now there are, for example, very dark colors that do not appear to be close to each others. This is due to the significant difference in the hue.]

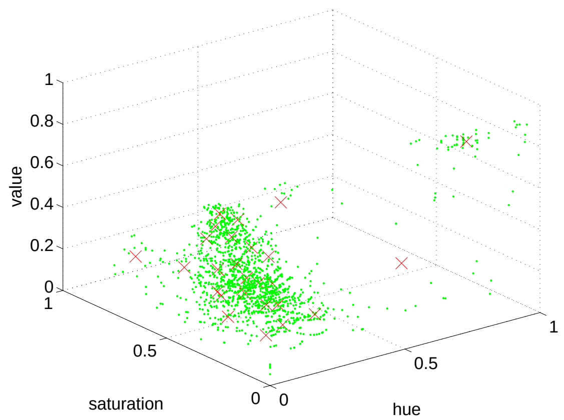

Figure 6.1: The color vectors of the summer picture with 27 centers for a mixture of Gaussian model after 30 iterations of training with EM algorithm

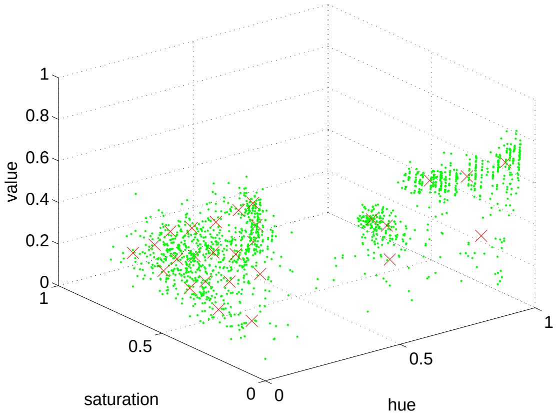

Figure 6.2: The color vectors of the winter picture with 27 centers for a mixture of Gaussian model after 30 iterations of training with EM algorithm

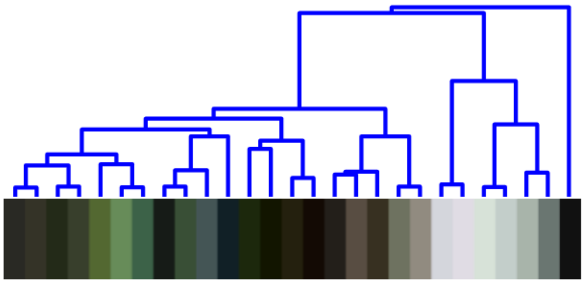

Figure 6.3: The 27 colors of the summer picture in a dendrogram

6.4 Discussion

In this paper, we issued some implementational aspects left open by Gärdenfors' Conceptual spaces [7]. We mainly discussed the formation of the concepts that is regions in a given conceptual space. The significance of these results to the understanding of actual implementation of human intelligence might be questionable or at least modest. However, the central contribution of this paper does not lie therein, but in simulation of the intelligence, that is AI project.

In this paper, we only payed attention to the categorisation in already acquired conceptual spaces. We now discuss in brief the connection of the conceptual level to the connectionist, namely the acquisition of the quality dimensions, and the symbolic levels, namely thought processes and language.

A natural way to connect the conceptual level to the basic sensory input level is provided by the connectionist approach. Then the quality dimensions are determined by the sensory input as well as possibly some other more basic quality dimensions using a flexible nonlinear mapping. However, Gärdenfors usually takes the quality dimensions as given, though clearly this cannot be true for all concepts. The principles guiding the learning are not easy to state as they should include at least, capacity constraints, generalization properties and finally, the relevance of different structures in the sensory data for the particular task the concepts are needed for.

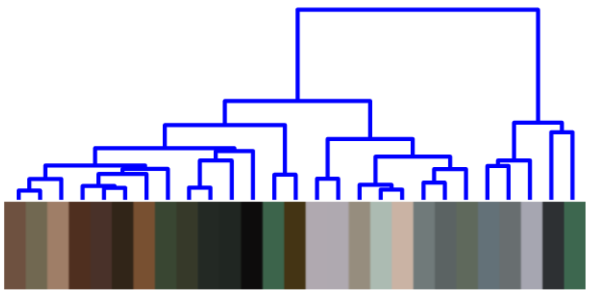

Figure 6.4: The 27 colors of the winter picture in a dendrogram

Furthermore, to really bridge the conceptual level to the symbolic level, one needs to explain the relation between the acquired concepts and language. We see it plausible to assume that language terms get associated to the regions in the conceptual spaces, that is concepts. Then concepts that get instantiated due to sensory input or voluntary thought processes may trigger the use of language, internally or in a speech act.

An other property of concepts, at least of those used by humans in their language, is their context sensitivity. Exemplary one could think of the different meaning of hot when going to sauna or having fever. Gärdenfors suggests that by a magnification or scaling of the quality dimensions (see the skin color example on page 119f of [7]) could amount to this property. In the bayesian framework context sensitivity can be achieved by the use of different priors, that modify the mean and variance of the gaussians to meant the contextual environment.

6.5 Acknowledgements

We would like to thank all the participants of the seminar of conceptual modeling at Helsinki University of Technology, Autumn 2003, for their valuable input, especially Mr. Janne Hukkinen and Mr. Karthikesh Raju. We would like to thank the Neural Networks Research Centre of Helsinki University of Technology for providing the infrastrucutre for working on this subject.

Bibliography

[1] H. Attias. A variational bayesian framework for graphical models. In T. et al Leen, editor, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, Cambridge, MA, 2000.

[2] L. Barsalou. Ideals, central tendency and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11:629–654, 1985.

[3] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[4] M.H. Bornstein. Color vision and color naming: A psychological hypothesis of cultural difference. *Psychological Bulletin*, 80:257–285, 1973.

[5] P.M. Churchland. *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA: MIT Press, 1989.

[6] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall, 1993.

[7] Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, 2000.

[8] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2 edition, 2003.

[9] C.L. Hardin. *Color for Philosophers*. Indianapolis/Cambridge: Hackett Publishing Company, expanded edition, 1993.

[10] J. A. S. Kelso. *Dynamic Patterns: The Self-Organization of Brain and Behavior (Complex Adaptive Systems)*. Cambridge: MIT press, 1995.

[11] T. Kohonen. *Self-Organization and Associative Memory*. Berlin, Heidelberg: Springer, 1984.

[12] S. Laurence and E. Margolis. *Concepts: Core Readings*, chapter 1, pages 3–81. Cambridge, MA: MIT Press, 1999.

[13] Barbara C. Malt and Edward E. Smith. Correlated properties in natural categories. *Journal of verbal learning and verbal behavior*, 23:250–269, 1984.

[14] J. L. McClelland, D. E. Rumelhart, et al., editors. *Parallel distributed processing: Volume1: Foundations*. Cambridge: MIT press, 1987.

[15] D. Medin and M. Schaffer. Context theory of classification learning. *Psychological Review*, 85:207–238, 1978.

[16] G. Murphy and D. Medin. The role of theories in concept coherence. *Psychological Review*, 92:289–316, 1985.

[17] A. Newell, H.A. Simon, and J.C. Shaw. Elements of a theory of human problem solving. *Psychological Review*, 65:151–166, 1958.

[18] F. James Rohlf. Single-link clustering algorithms. In P.R. Krishnaiah and L.N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 267–284. Elsevier Science Publishers, Amsterdam, The Netherlands, 1982.

[19] E. Rosch and B. Lloyd. *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum, 1978.

[20] E. Rosch and C. Mervis. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605, 1975.

[21] E. Rosch, C. Simpson, and R. Miller. Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2:491–502, 1976.

[22] B. Ross and G. Murphy. Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25:51–63, 1996.

[23] M. Strauss. Abstraction of prototypical information by adults and 10-month-old infants. *Journal of Experimental Psychology: Human Learning and Memory*, 5:618–632, 1979.