The Worrying Statistics of Connectionist Representation

Chris Thornton
Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QN
Email: Chris.Thornton@cogs.susx.ac.uk
Tel: (44)273 678856

December 14, 1994

CSRP 362

Abstract

The paper looks at how the hidden-vector cluster analyses associated with Elman and others seemed to provide a potentially important link between the symbolically-oriented level of analysis and the connectionist level of analysis — a link that might one day help to explain how higher mental processes are grounded in neural architectures. The paper goes on to reconsider the implications of these analyses in light of some recent work by Finch and Chater which shows that linguistically meaningful categories (of the type derived from hidden-vector analyses) are directly evidenced in the N-gram statistics of natural language. The implication of this work seems to be that hidden-vector analyses do not primarily address the link between the symbolic and connectionist levels of explanation but rather tell us something about the statistical properties of the training environments used. The consequences of this result for cognitive science are lightly sketched in.

1 Introduction

There has always been the hope that work in Artificial Intelligence (AI) would help to elucidate and extend the philosophical study of the mind. But, paradoxically, the interface between AI and philosophy appears to have become harder to negotiate as the years have gone by. In the early days links between AI and philosophical studies were readily apparent. AI researchers tended to construct programs that reflected their introspections about mental processes or, in some cases, verbal protocols given by human problem solvers [cf. 1]. This quite naturally produced systems populated with familiar landmarks.

But if the early research was relatively accessible to philosophical minds, developments in the field soon seemed to be carrying AI off into an 'outer space' remote from both introspective experience and philosophical conceptualisation. The effect was, perhaps, particularly noticeable in the area of vision research. Vision researchers of the 1960s, e.g., Roberts [2], were largely concerned with systems which sorted out neat, perspective drawings using rules whose good sense could easily be comprehended. cf. Waltz filtering [3]. But by the late 1970s researchers had begun to work with systems which dealt primarily in abstruse mathematical constructs having nothing obvious to do with the mind or the real world, cf. [4].

It was perhaps inevitable that there would be some kind of reaction to AI's gradual descent into high technology. And when it came, it came in a rush with the most obvious manifestation being John Searle's vigorous attack on 'strong AI'. Searle's aim was to demonstrate that AI-like systems could not possibly have intentional states (such as understanding) regardless of their level of performance.

Of course, the status of the argument was and is the subject of debate. But in attempting to demonstrate logical flaws in Searle's argument [cf. 5], AI researchers were, perhaps, missing the point. Whatever Searle demonstrated about the intentional properties of AI systems he did stimulate a much more thorough scrutiny of the functionalist assumptions underlying strong (and not so strong) AI. This scrutiny was, as Sloman pointed out at the time [6], urgently needed.

Functionalism is essentially the idea that intentional states have nothing to do with architectural substrates. Being a functionalist implies believing that all that is required for intentional states is the right *program*. The question of what machine (architectural substrate) the program is running on is assumed to be irrelevant. Functionalism legitimizes the sort of AI research which attempts to model mental processes on computers which do not even remotely resemble the brain. It suggests that the important thing is to capture the phenomenon at a computational level of description. Later on the algorithmic and implementation details can be worked out. Functionalism effectively urges a top-down program of research. The image is that of a 'triumphant cascade through Marr's three levels.' [7, p. 227]

But the validity of the functionalist stance is open to question. And, of course, it is not just a question of computer simulations of mental processes: the issue is much more general than that. It has to do with the validity of simulations in general. It has to do with abstractions and in particular, what happens (ie. how properties are affected) when one moves from a real phenomenon to an abstraction of it (i.e., a simulation, model or theory).

When we substitute some phenomenon X with a model of X, certain properties of X are carried over to the model and some are inevitably lost. Otherwise, what sort of 'abstraction' would it be? The model — if it is genuinely a model rather than a duplication — will abstract away certain characteristics of the original. If it is a *good* model it will abstract away the 'less significant' characteristics of X and leave behind the 'essential details'. But the point is, any properties associated with characteristics which *are* abstracted away will necessarily be absent in the model. The point is sometimes illustrated using the 'typhoon' example. When we substitute a typhoon with a simulation of a typhoon, the property of wetness is lost (unless it is a *very* realistic simulation). But the feedback property[1] which plays a central role in keeping the typhoon going is retained.

A more close-to-home example is the 'Sunday market'. If we construct a simulation of a Sunday market then we will have certain symbolic entities representing the people involved, the objects for sale and the sums of money that change hands. Let us say that for some reason the market is fairly unstable (prices rising and falling wildly). The market then has the property of 'instability.' But so does the simulation of the market. The physical 'weightiness' of the money that is exchanged, on the other hand, is lost as we move from the real market to the simulation. If we substitute the simulation with a high-level theory which merely abstracts out the basic equations of supply and demand then we will lose both 'instability' and 'weightiness'.

The point, then, is that there is always a wasting process when we move from real phenomena to models, and on to higher-level abstractions and theories. Properties of the original phenomenon are lost or filtered out by the abstraction process. Exactly which properties are filtered out by a particular abstraction is a contingent issue which cannot be decided in general. And, of course, this applies directly

[1] Warm air being sucked in, releasing latent heat as it rises, causing more lift and more warm air to be sucked in.

to the case where we attempt to build computer simulations of mental processes. In simulating mental processes we are simply trying to construct abstractions of the original phenomenon.[2] No matter how accurate our abstraction, some properties of the original phenomenon will necessarily be lost. This is, after all, the essence of abstraction.

Since a computer simulation is just another form of abstraction, and since abstraction necessarily wastes properties, computer simulations of mental processes potentially lose some of the properties of real mental processes. The implication is that Searle was essentially correct: architecture may *not* be irrelevant. The properties we are actually interested in (understanding, belief etc.) may be to do with characteristics of the substrate.

2 Connectionism and the need for good grounding

Searle's attack on strong AI seems to have reflected the onset of a general worry about AI's tendency to ignore the need for 'good grounding'. Certainly, in the years following the attack new approaches began to emerge which were centrally concerned with issues of architecture and environment. Examples include the develop of geneticism (genetic algorithms, classifier systems etc.), the development of the reactive systems movement stemming from Brooks work on robot creatures [8; 9] and finally, of course, connectionism, a paradigm whose stated aim was to take the low-level architecture of the brain seriously [10].

The emergence of these new approaches seems, in retrospect, to have been something of a mixed blessing for those on the philosophical side of the fence who, like Searle, believed AI to have over-extended the functionalist position. Though the new approaches tended to have a better grounding in architecture or environment, they typically had no better grounding in philosophical approaches than did the abstruse engineering-oriented AI of the classical period.

Thus, though the emergence of connectionism etc. provided a response to the grounding problem, it did very little to help the interface between AI and philosophy. The natural kinds of the new approaches were just as remote from philosophical inquiry as were the natural kinds of Vision systems from the 1970s. A philosopher of the mind might reasonably be expected to be interested in the question, say, of when a particular being can be said to have a belief. But the connectionist researcher would be likely to be far more interested in the shape of an error surface. On a pessimistic reading, then, the only value of the new approaches — from the philosophical point of view — was that Dennett's triumphant cascade had been turned into a two-way street. Models could now be expected to bubble up from below (bottom-up from connectionism) or trickle-down from above (top-down from classical AI). Unfortunately, 'bubble-up' looked like it was going to be just as elusive as 'trickle-down'.

3 Enter the Hidden-Vector Analyses

Given this background, and the relatively poor prospects for a completion of the triumphant cascade (in either direction), the emergence in the mid 1980s of techniques for anylysing the representational properties of connectionist networks was a welcome innovation. The first widely-published usage of these techniques was in Sejnowski and Rosenberg's [11] work on the NETtalk system. These two researchers showed how a cluster analysis of the hidden-vectors of a backpropagation network (trained to convert text to speech) showed up linguistically meaningful groupings. In particular, the analysis

[2] Their computational nature makes *no* difference whatsoever — except to computerphobes.

showed how the network had constructed an internal hierarchy which flagged linguistically important distinctions (e.g., consonant versus vowel.) The newsworthiness of this work was founded on the fact that these distinctions were not given to the network a priori. Rather they were learned directly from the data.

Sejnowski and Rosenberg's hidden-vector analysis method soon became part of the standard toolkit of the connectionist researcher. Recently, it has been used to particularly good effect by Elman who showed how a copy-back network trained to do word-prediction (given only a diet of raw English sentences), formed an internal hierarchy that captured lexical and semantic categories [12].

For anyone dreaming wistfully of a bottom-up, 'reverse-cascade', this new work by Elman and others looked very promising. The notion that the behaviours of connectionist systems embodied tacit rules was fairly well accepted especially in light of Rumelhart and McClelland's work on the learning of past tenses [13]. But with the new hidden-vector analyses one could now say much more precisely what form the terms of these rules might take. In effect, the hidden-vector analyses provided an initial step-up on the reverse cascade. It built a small bridgehead that connected the mushy and remote world of low-level connectionism (a world of 'weights', 'activation values', 'links', 'units', 'energy levels' etc.) with the rather more tractable world of symbols and class definitions.

Andy Clark was quick to see the potential of this new method. In discussing the implications of hidden-vector analyses, he suggested that 'a fully interpreted cluster-analysis . . . constitutes the nearest connectionist analogue to a classical competence theory.' [14] Of course, by this date, cluster analysis had only managed to 'reify' fairly primitive types of class definition (eg. lexical categories) but it was easy to imagine how it might be possible to build one cluster-analysis on top of another and perhaps produce in the end a chain of connections linking the classical world with the connectionist. This would constitute a reverse-cascade likely to satisfy all customers. It would certainly satisfy those emphasising the need for firm grounding. It would also satisfy those wanting to know how higher-level mental objects such as rules and concepts correspond to lower-level neural processes.

4 Finch, Chater and the statistics of English

Unfortunately, some recent work by Finch and Chater [15] seems to suggest that such a cascade of connections — were it ever to be derived — might not tell us any more than we could have found out using GOFSA — good, old-fashioned statistical analysis. Finch and Chater's work uses cluster analysis like the work done by Sejnowski, Rosenberg, and Elman. But instead of using it to analyse the representational properties of hidden-vector spaces they used it to analyse the statistical properties of ordinary English text. The main gist of their results is that the 'linguistically meaningful' hierarchical structures which can be obtained by clustering the hidden-vectors of, say, an Elman-style, copy-back network trained to do word-prediction [12], can also be obtained by a fairly straightforward statistical anylysis of a large corpus of English sentences.

Finch and Chater have carried out a whole range of experiments using a large corpus of text derived from electronic 'news' discussion groups. They tried various approaches and the details are described in their various papers [15, 16; 17]. The method they used involved sampling N-gram statistics and then using cluster analysis to discover groupings of words with similar probability distributions. The analysis produced groupings and structures which have a very close correspondence to known syntactic and semantic categories. In other words they were able to obtain cluster analyses that closely resembled and in some cases improved upon the analyses obtained by Elman.

4.1 A type-1 theory for copy-back networks?

What should we make of this work? Finch and Chater's own view is that statistical analysis provides us with a better understanding of the performance and behaviour of certain sorts of networks (e.g., Elman, copy-back networks). They conclude that their statistical work shows that the 'copy-back scheme is sampling these [N-gram] statistics successfully.' The go on to say that 'these results suggest that the hidden unit patterns that recurrent neural networks develop can be viewed as reflecting quite directly the statistical structure of the sequences learnt.' [17]

By showing that the internal structures formed by copy-back word-prediction networks closely resemble the structures derived from a particular statistical analysis, they have effectively shown that the networks are sampling the relevant statistic. In a sense, they have provided a type-1 theory [18] for the behaviour of these networks. The theory says that the network is performing a particular computation and it characterizes this computation without making any reference to implementation issues.

For those who want to believe that architecture and grounding are important, this is clearly a worrying demonstration since it seems to eliminate the 'ground' altogether. Surely, if all an Elman network is doing is sampling a certain statistic then its 'networkness' cannot be the origin of significant properties. A functionalist stance towards such networks, then, would seem to be perfectly appropriate. On the other hand it might be argued that any retreat into functionalism *must* be premature. The statistical work in question has only looked at one particular domain (natural language) and has produced results which seem to bear directly on only one type of network (the Elman copy-back net). Our assumptions about the importance of grounding and our hopes for the reverse cascade may then turn out — when other systems are analysed more carefully — to be be fully justified.

5 Is it statistics all the way up?

However things go for the 'grounding' issue, one thing is clear: Finch and Chater's work suggests that we should review our attitude to the value of statistical analysis. Classical AI made practically no use whatsoever of it. New approaches such as reactivism and alife-ism have also tended to largely ignore its potential. Connectionism has used it to a certain degree but typically only for the purposes of analysing the behaviour of models. Finch and Chater's work suggests that it can play a much more direct role in our attempt to understand the nature of concepts and classes. Of course, all that has been shown to date is that certain linguistic classes show up directly in the N-gram statistics of ordinary text. But the implication is that we may be able to find statistical justifications (explanations) for classes in all sorts of domains by producing statistical analyses of the relevant data.

Naturally, the big question is, how *many* domains? I.e., how general is the method of statistical class-recovery likely to be? We cannot hope to answer this without more empirical work but some progress can be made purely by rational argument. Let us assume, for now, (1) that for any concept there is an associated class and (2) that the set of all classes divides up into natural classes — whose boundaries are evidenced in the world — and artificial classes — whose boundaries are arbitrary.[3] One way for a class to be evidenced in the world is via statistical regularities of the type discovered by Finch and Chater's method. But how general is this? Are all classes evidenced this way? Or just some subset? If so, which subset?

Finch and Chater directed their efforts towards a relatively limited statistical analysis. In particular

[3] An example of an arbitrary class would be something like 'the class of all blue things that have been within six feet of someone singing Amazing Grace.'

they concentrated on analysing 5-gram statistics of text. As they note, 'an N-gram is an ordered sequence of N symbols. The frequencies of occurrence of each N-gram in a continuous stream of data constitutes the N-gram statistics of the data set.' [15]. Their aim was to look at the number of times that particular words were observed to appear as the last-but-one, last-but-two, next-but-one and next-but-two neighbours of every other, commonly occurring word. As we have seen, this approach was able to show up important class divisions. However, it is far from being the only way of applying statistical analysis to text.

Sticking with the basic idea of looking at N-gram statistics, we can envisage many variants of Finch and Chater's method. One might look, for example, at N-gram statistics for bigger or smaller values of N. One might look at statistics relating to 'holey' N-grams; ie. non-continuous sub-sequences of the original data stream. One might look at N-grams which are derived according to an algorithm or some other dynamic, selection criterion. Having discovered class boundaries at some given level of analysis, one might then look at the range of possible N-gram statistics that can be derived from a reconstructed data stream in which class labels have been substituted for class members. And this process might be repeated recursively through many levels of analysis.

The space of recognized statistical regularities is, then, rather large. There thus seems to be no *a priori* reason for ruling out the possibility that statistical regularities of the type discovered by Finch and Chater's method underpin all concepts, classes and natural kinds. If this is so then any cascade connecting higher-level mental objects with lower-level architectural substrates must be mediated via a chain of essentially statistical relationships. A central task, then, is to determine what these relationships might be, by what processes they can be *derived* from the world and how they can then be *represented* by cognitive agents. Without good answers to these questions we cannot finally decide what the implications of the statistical language analyses truly are.

6 Concluding comments

The main aim of the paper has been to note the way in which Finch and Chater's work undermines several central assumptions widely espoused in the connectionist community. The principle 'victim' is the assumption that hidden-vector cluster analyses could constitute the beginnings of a reverse cascade (a 'bubble-up'). However, a rather dark shadow is also cast over approaches which stress the importance of architecture, grounding and environment. The essential point to draw out, though, is the fact that cluster analyses such as Elman's do not primarily tell us anything about the networks in question but rather something statistical about the environment in which the networks were trained. This conclusion seems a little worrying at first. But if we dust off our attitude to statistical analyses and accept the premise that natural kinds are almost certainly rooted in statistical regularities it begins to seem much more positive.

References

[1] Newell, A. and Simon, H. (1963). The logic theory machine. In .E. Feigenbaum, Feldman and J. (Eds.), *Computers and Thought.* New York: McGraw-Hill.

[2] Roberts, L. (1965). Machine perception of three-dimensional solids. In J. Tippett, D. Berkowitz, L. Clapp, C. Keoster and A. Vanderburgh (Eds.), *Optical and Electro-Optical Information Processing.* Cambridge, Mass.: MIT Press.

[3] Waltz, D. (1975). Understanding line drawings of scenes with shadows. In P. Winston (Ed.), *The Psychology of Computer Vision* (pp. 19-92). Mcgraw-Hill.

[4] Scott, G. (1988). *Local and Global Interpretation of Moving Images.* Research Notes In Artificial Intelligence, London: Pitman.

[5] Thornton, C. (1985). A response to searle's thesis. *AISB Quarterly*, No. 52 (pp. 32-33).

[6] Sloman, A. (1985). Strong strong and weak strong AI. *AISB Quarterly*, No. 52 (pp. 26-31).

[7] Dennett, D. (1987). *The Intentional Stance.* Cambridge, Mass.: MIT Press.

[8] Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence, 47* (pp. 139-159).

[9] Brooks, R. (1991). Intelligence without reason. *Proceedings of the Twelth International Joint Conference on Artificial Intelligence* (pp. 569-595). San Mateo, California: Morgan Kaufman.

[10] Hinton, G. and Anderson, J. (Eds.) (1981). *Parallel Models of Associative Memory.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

[11] Sejnowski, T. and Rosenberg, C. (1987). Parallel networks that learn to pronounce english text. *Complex Systems, 1* (pp. 145-68).

[12] Elman, J. (1989). Representation and structure in connectionist models. CRL Technical Report 8903, San Diego: Center for Research in Language (UCLA).

[13] Rumelhart, D. and McClelland, J. (1986). On learning the past tenses of english verbs. In D. Rumelhart, J. McClelland and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vols I and II.* Cambridge, Mass.: MIT Press.

[14] Clark, A. (1990). Connectionism, competence, and explanation. In M.A. Boden (Ed.), *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

[15] Finch, S. and Chater, N. (1991). A hybrid approach to the automatic learning of linguistic cateogories. In S. Torrance (Ed.), *AISB Quarterly*, No. 78 (pp. 16-24).

[16] Finch, S. and Chater, N. (forthcoming). *Bootstrapping Syntactic Categories Using Statistical Methods.*

[17] Chater, N. and Conkey, P. (forthcoming). *Finding Linguistic Structure with Recurrent Neural Networks.*

[18] Marr, D. (1977). Artificial intelligence: a personal view. *Artificial Intelligence, 9* (pp. 37-48).