The Hearts of Symbols: Why Symbol Grounding is Irrelevant

David S. Touretzky
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Upon closer examination, and depending on who you read, "symbol grounding" turns out to be either the induction of trivial sensory predicates or the relabeling of a large portion of intelligent behavior as "transduction." Neither activity shows much promise for advancing our understanding of intelligence, although symbol grounding does have some utility in philosophical debates. The proper concern for symbol processing researchers, both connectionist and classical, is to construct and manipulate symbols, not to ground them.

Introduction

A variety of activities are taking place today under the banner of "symbol grounding." They include philosophical arguments about how meaning enters a symbol system, and computer models that perform pattern association or induction of sensory predicates.

Harnad (1990) introduced symbol grounding as a solution to Searle's Chinese room problem (Searle, 1980). Briefly, Harnad argued that at least some of an agent's symbols must be grounded in sensory predicates in order that its mental structures acquire intrinsic meaning, by being causally connected to things in the world. When Searle, sitting in the Chinese room, manipulates formal symbols whose meanings are unknown to him, he is a pure symbol processor. Since the symbols are ungrounded, the Searle/room system as a whole makes no contact with the world. It does not *refer* to anything. Searle-as-symbol-processor can therefore be said to not understand Chinese, despite the fact that his reponses might seem perfectly sensible to a human reader of Chinese.

If Searle also had transducers available to him that transformed sensory events into symbols in the "right" way, then the Searle/room system *would* understand Chinese, even though Searle personally did not. But as Harnad (1992) points out, the transduction problem is much harder than it first appears.

What we have so far is a materialist solution to a philosophical problem: explaining how an agent's symbols acquire meaning. They acquire it by being in a causal relationship to aspects of the physical world. So, for example, the symbol `red` can *mean* "red" to an agent because certain frequencies of light cause a systematic, recognizable response in the agent's visual transducers, which in turn affects the agent's behavior. Objects lacking a capacity for systematic, causal interactions with the world, such as rocks, or computer programs running without sensors or effectors, do not have meaningful symbols. But external observers are always free to pick out components of these objects and *assign* meanings to them. This is what we do when we use a computer.

Cognitive scientists interested in symbol manipulation have not been troubled by the meaninglessness of their computer models, as this is a purely philosophical point. Nonetheless, recent references to the symbol grounding problem have implied that if symbols were grounded, something good would come of it. Harnad (1992) speculates that "the failure of AI" might be due to the ungroundedness of its symbols. It seems to me more likely that the limitations of the classical AI approach are due to its using the wrong kinds of symbols.

How could groundedness be the source of power for a representational system? This suggestion is perhaps a result of the tendency on the part of some authors to conflate symbol grounding with other inductive phenomena, such as category formation. Perceptual predicates such as "red" or "striped" can be directly grounded in sensory processes, but conceptual categories such as "horse" cannot.

If we take concepts to be complex symbol structures, then we must ask what level of meaning isolated symbols encode. An intelligent system certainly has to be able to construct categories representing the natural kinds in its environment, but the relationship of categories such as "horse" to sensory phenomena is so abstract that the groundedness of the base level primitives seems irrelevant.

Why worry about grounding of symbols when we don't know yet what symbols are? The symbols in our heads could perhaps be arbitrary syntactic tokens like Lisp atoms or the marks on a Turing machine's tape, but they are not necessarily so. Recent work in connectionist symbol processing suggests other alternatives. For example, Elman (1989) showed that symbols in a recurrent network are represented differently depending on the context in which they appear. Elman's symbols differ from Turing machine symbols, which have a fixed form independent of context. Another interesting property of symbols in backprop-trained connectionist networks is that they are emergent, manufactured objects. A "symbol" is a cluster of points in a continuous, evolving state space.

When arguing about the groundedness of symbols, it's important to remind ourselves that the symbols in our heads remain unknown entities. The space of symbolic representations is very rich, and we have only managed to explore a small portion of it. The Church-Turing hypothesis tells us that all points in this space are equivalent, but this oft-repeated observation is irrelevant to cognitive science. It ignores such essential constraints as neural implementability, cognitive resource limits, and real-time performance demands. So it is a mistake to think that any notion of symbol will do, and that effective symbol manipulation is an already solved problem.

At a higher level of structure, consider possible representations for "What if the dog had wanted to chase a red ball?" There are today a variety of competing linguistic theories suggesting representations for this sentence. Besides the obvious syntactic issues, the sentence also requires decisions to be made about modeling of mental states of other agents, and about reasoning with conditional or counterfactual statements. Research has proceeded on these topics for decades, and progress on computer models has not been noticeably impeded by failure to ground the concepts "dog," "red," and "ball" in sensory predicates.

Symbol Grounding as Pattern Association

To date there have been no satisfactory models of symbol grounding. What we have been offered instead turns out, upon closer examination, to be mechanisms either for arbitrary pattern association, or for induction of crude sensory predicates.

The pattern association models are basically trying to show that intrinsically meaningless symbols can acquire sensory correlates, and that cross-modal correlations can be established. For example, Dorffner (1992) describes a system that learns to associate "visual" inputs (actually just binary vectors) with "acoustic" inputs (also binary vectors) serving as category labels. This model is concerned with the induction of conceptual structure from labeled examples. There is no attempt to discover nontrivial, perceptually invariant properties of the input, which is the role Harnad assigns to transducers.

Nenov (1991) utilizes more elaborate encodings for both visual and auditory input, but again the model is essentially a pattern associator.

Induction of Sensory Predicates

Another candidate for symbol grounding models are systems that induce sensory predicates from examples. The most recent and interesting is Regier's (1992) dissertation on learning spatial prepositions such as *into*. The input is a cartoon image, or sequence of images, of a landmark and a trajector. The model has pre-wired hardware for detecting properties such as center of mass and major axis orientation, and relationships such as spatial overlap, or the angle of the line joining the centers of mass of the landmark and trajector. The model also has units with tunable parameters that can be adjusted by backpropagation. So, for example, the predicate *above* can be learned as a set of soft constraints, one of which is that the line between the landmark and the trajector centers of mass should have an essentially vertical orientation.

"Into" and "above" are rich concepts with many metaphorical extensions, though they are grounded in spatial relationships. But Regier's model is not an example of symbol grounding because there are no symbols in it. Symbols are the things whose composition produces systematic, combinatorial representations. Regier has nothing to say about this; he merely shows us how to train "into" and "above" detectors: sensory predicates which output a 1 when the desired relationship exists in the input image. What Regier has built is a tunable transducer, of the sort Harnad requires to ground symbols in the physical world, but as a transducer it's terribly impoverished. Computer vision systems already perform much more complex interpretations of visual scenes, using real images rather than cartoons.

The real contribution of Regier's model, in my view, is that it is a working demonstration of the ideas of linguists who have been analyzing spatial prepositions for some years now. Specifically, it demonstrates the idea that a broad concept such as "above" can be captured by a set of soft constraints on multiple spatial properties, only some of which need be satisfied in any specific instance. The model also shows that these soft constraints can be learned by exposure to positive and negative examples. For more on the relationship of concepts to soft constraints see the discussion of radial categories in (Lakoff, 1987).

Transduction Means Cognition

"Transduction" is a somewhat misleading term, in that it implies a simple bottom-up mapping of sensations to symbols. Processing in the Dorffner and Regier models proceeds this way. The story, then, is that we have a layer of transducers at the bottom to associate physical phenomena with primitive concepts, and from there we proceed upward via symbolic composition to increasingly abstract concepts. Harnad gives the example of *zebra* defined as *horse* plus *striped.*

But transduction cannot be a purely bottom up process. The interpretation of visual scenes must be guided by knowledge about objects and expectations about the world, and is to some extent a problem-solving activity. It makes no sense to suppose that predicates like "horse" are constructed by observing countless images of horses and constructing *de novo* a complex model of how these objects give rise to the variety of images we perceive as horses. Far more reasonable to assume an innate theory of 3D space and 2D projections, with innate representations for surfaces, textures, objects and prominent protrusions (body parts), and so forth. The task for the transducer may still be to classify inputs as instances of *horse* or *dog* or *field of flowers*, but it requires a massive representational foundation.

The interesting question here is not how transducers get connected up to symbols, it's how a capable transducer could be constructed in the first place. This boils down to asking how to build a vision system. So in pursuing the symbol grounding issue, we have shifted our focus from disembodied cognition of the sort required to pass the Turing test to cognition as intelligent perception, the latter being the means for grounding the symbols used in the former. But have we learned anything useful about symbols?

The Hearts of Symbols

Humans use grounded symbols in a variety of ways not directly related to their origin as sensory predicates. These include abstract modes of reference and very complex spatial metaphors. Explaining the sensory roots of symbols will not open up some magic path to intelligence; it ignores the way symbols are really used. The crucial question about symbols is: What are the mechanisms that enable these abstract uses?

An example should make clear the irrelevance of grounding to human-like symbol manipulation. Consider the concept *heart*. Although one could construct an elaborate perceptual schema for hearts, with visual templates for arteries, veins, atria, and ventricles, the question cognitive scientists should be asking is how `heart` has also come to refer in certain abstract ways ("artichoke hearts," "the heart of the argument"), and how it functions simultaneously as a metaphor — at least in Western cultures — for certain aspects of mind, as in "In your heart, you know I'm right."

The heart of an argument is the portion which is central or vital. "Central" and "vital" are themselves metaphors for "essential to proper function," and hearts are both central in our body and vital to our functioning. So by viewing an argument as an object with spatially distributed components and an intrinsic function, we establish a metaphorical framework in which it makes sense to refer to the "heart" of the argument. The crucial role of spatial and force metaphors in human conceptual structure has been worked out in considerable detail by linguists such as Talmy, Lakoff, Langacker, and others. Let me just underscore the point that in daily conversation, "heart" is used almost exclusively metaphorically. The fact that hearts are also physical objects which can be recognized by a clever visual transducer is, for practical purposes, irrelevant.

How are people able to metaphorically extend physical concepts such as *heart* or *above*? Until we can answer this question, we will not have a satisfactory theory of symbols, grounded or otherwise. It is a common mistake to think of the symbols in our heads as equivalent to Lisp atoms or marks on a Turing machine's tape, where the basic operations are copying and composition. This is too low level a view. Composition is important, but the crucial operations we perform on symbols in our heads have to do with analogy, association, and reference.

An interesting property of metaphorical extensions is that they are not always mutually compatible. This can give rise to metaphorical puns, such as the phrase "broken-hearted artichoke," which might refer either to a real artichoke with a broken center, or an imaginary, animate artichoke[1] suffering an emotional trauma. It is difficult to imagine a scenario in which both interpretations could apply simultaneously, however. For further discussion of interactions among multiple metaphorical meanings, see Norvig (1988) or Lakoff and Turner (1989).

The Chinese Pen-Pal

As Harnad observes, Searle toiling in his Chinese room could in principle pass the Turing test in Chinese, meaning that symbolic interactions with the Searle/room system might be "indisinguishable from a lifelong correspondence with a real Chinese pen-pal." Searle's argument is that the Chinese room is a fraud, in that there is no "understanding" going on and no "meaning" to be found in the system. Harnad attributes this to the ungrounded quality of the symbols Searle is working with, but I would like to suggest another reason why the Chinese room without transducers would be troubling as a cognitive agent, even if it produced convincing correspondence.

First, consider what happens when we ask the pen-pal to "go to the window and tell me what you see." There can be no question that whatever answer Searle produces would be fraudulent. A robot can perhaps "see" if we relax the constraint that it be done with a retina and cortex, but since Searle's room is without transducers of any sort, there is no seeing going on there.

The Searle/room system's "seeing" may be fraudulent, but is its "thinking" fraudulent as well? Even if we decide that grounding is irrelevant, there are reasons to doubt whether this system can think if by "think" we mean "think in a roughly human-like way."[2] The Chinese pen-pal may be an unsatisfactory thinker because perception is such a large part of intelligence. If the pen-pal is without transducers, we are naturally led to doubt whether it has the mental faculties that follow from transduction, such as mental imagery, and the ability to construct and comprehend spatial metaphors.

Any agent that lacked these faculties could not think in a human-like way. An agent that possessed them,

[1] Perhaps cousin to the famous singing California raisins.

[2] This does not exclude thinking in an ape-like or dolphin-like way, which may be nearly as sophisticated as human thought. But it does exclude crude symbol manipulators such as pocket calculators, and crude sensor and effector equipped robotic devices, such as washing machines.

even if it lacked transducers, would be a much more satisfactory symbol manipulator than anything cognitive science has been able to construct to date. If external observers assign meanings to the the agent's symbols they will find that its computations produce meaningful symbolic results. For cognitive science, that's all that matters. Grounding should be left to the philosophers.

Acknowledgements

I thank Stevan Harnad for insightful discussions, and Mike Gasser and Georg Dorffner for catalyzing some pre-symposium interaction among the participants which contributed to the writing of this paper. This work was supported in part by a grant from the Fujitsu Corporation.

References

[1] Dorffner, G. (1992) Taxonomies and part-whole hierarchies in the acquisition of word meaning – a connectionist model. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 803-808. Hillsdale, NJ: Erlbaum.

[2] Elman, J. L. (1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7(2,3):195-225.

[3] Harnad, S. (1990) The symbol grounding problem. *Physica D* 42:335-346.

[4] Harnad, S. (1992) Connecting object to symbol in modeling cognition. In A. Clarke and R. Lutz (eds.), *Connectionism in Context*. Springer Verlag.

[5] Lakoff, G. (1987) *Women, Fire, and Dangerous Things*. University of Chicago Press.

[6] Lakoff, G., and Turner, M. (1989) *More than Cool Reason*. University of Chicago Press.

[7] Norvig, P. (1988) Multiple simultaneous interpretations of ambiguous sentences. *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, 291-297. Hillsdale, NJ: Erlbaum.

[8] Regier, T. (1992) *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. Doctoral dissertation, University of California, Berkeley.

[9] Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 417-424.