Associationism and Connectionism

John N. Williams

Research Centre for English and Applied Linguistics, University of Cambridge

In K.E. Brown (Ed.) Encyclopaedia of Language and Linguistics: Second Edition. Elsevier: Oxford. 2005.

Abstract

Connectionism is a powerful form of associationist learning theory that broadly reflects the computational style of the brain. Information is represented as distributed patterns of activity over densely interconnected networks of neuron-like processing units, and learning occurs through modifications of the strength of connections between units. Networks are trained on specific tasks and their performance is compared to that of humans. Models of conceptual representation, word reading, morphology, speech recognition, speech production, and syntax have displayed emergent rule-like behavior. The success of these models challenges the traditional separation of item memory and abstract, symbolic rule systems, and provides an empirically testable alternative to nativist approaches to learning.

Keywords

Parallel distributed processing, associative learning, distributed memory, aphasia, reading, morphology, speech recognition, speech production, syntactic processing,

constraint satisfaction, transition probability, statistical learning, nativism, empiricism, modularity, emergentism.

Article

Associationism maintains that all knowledge is represented in terms of associations between ideas, that complex ideas are built up from combinations of more primitive ideas, which, in accordance with empiricist philosophy, are ultimately derived from the senses. Connectionism is a more powerful associationist theory than its predecessors (Shanks, 1995), that seeks to model cognitive processes in a way that broadly reflects the computational style of the brain. The brain carries out millions of computations simultaneously in densely interconnected networks of neurons. Information is represented as distributed patterns of activity, and learning is achieved by modifying the strength of connections between neurons. Connectionist, or "parallel distributed processing" models attempt to capture these general properties in highly artificial neural networks. These are self-organising systems that discover for themselves the representations required to perform a specific task stipulated by the researcher. Their main interest lies in the fact that they display emergent rule-like behavior. The question is whether their learning trajectory, ultimate performance, and breakdown after damage mimic that of humans. Connectionist research provides an empirical means of evaluating whether associationist learning principles are sufficient to explain specific aspects of human behaviour.

The simplest kind of connectionist network consists of a single layer of neuron-like units in which each unit represents a pre-defined feature that can be activated by external input. Each unit is connected to every other unit by a connection whose strength, or 'weight', determines the amount of 'activation' that passes along it. During training, when two units are simultaneously active, the strength of the connection between them is increased. Through this associationist learning principle the network absorbs the correlational structure of the input, and comes to represent concepts as distributed patterns of activity over its processing units. It has been demonstrated that representations of multiple individual instances of experience can be encoded in a single network and retrieved from

partial cues by a process of pattern completion, while less specific cues prompt the retrieval of a prototype (McClelland & Rumelhart, 1985). Concepts can be thought of as "attractor basins" in a multi-dimensional space of possible network states, an idea that has proved useful in explaining semantic errors in dyslexia (Hinton & Shallice, 1991) and in modelling semantic priming (Masson, 1995).

More complex networks are required when the problem involves mapping between different kinds of representation; for example, converting from orthography to phonology in single word reading. Separate layers of units encode input and output representations and, typically, there is an intermediate layer of "hidden" units whose function is to represent the input in a form that can be mapped onto the output (it is these hidden units that increase the computational power of connectionist networks compared to earlier associationist theories). During training, for each input pattern, the degree of error on the output is used to modify connection weights within the network so as to reduce the error in future. Using this "back-propagation" learning algorithm the network's behaviour is gradually shaped until, ideally, it produces the correct output for the entire set of training examples. Of course, connectionist models would not be very interesting if all they did was store the input-output mappings they were trained on. What is critical is whether after training they exhibit human-like performance. Models of word reading have been particularly successful in this respect. They produce human-like responses to non-words, showing generalisation to items that they were not trained on, and they mimic subtle aspects of reading known words, such as the combined effects of regularity and frequency (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). This suggests that they extract knowledge of underlying regularities, rather than merely storing the specific items received during training. It is also claimed that when damaged (e.g. by removing some of the connections) they show similar patterns of behaviour to brain-damaged patients, and that different dyslexic syndromes can be simulated by damaging the network in different regions (Plaut et al., 1996).

Another area of intensive connectionist research has been in relation to the English past tense. In this case phonological representations of word stems are mapped onto

phonological representations of past tense forms. It has been claimed that these networks mimic the human developmental profiles for regular and irregular forms, and for sub-groups of irregulars (Plunkett & Marchman, 1993; Rumelhart & McClelland, 1986). They are also able to produce human-like past tenses for nonce forms (MacWhinney & Leinbach, 1991), despite containing no discernible rule-like representations. By damaging a network in different regions it is possible to simulate the kinds of dissociations between regular and irregular morphology found in brain-damaged patients (Joanisse & Seidenberg, 1999), paralleling the effects obtained in word reading.

Language processing also has an obvious sequential component. To deal with this, "simple recurrent networks" are taught to predict each successive item in a sequence using a record of the hidden unit activations to previous events. Such networks do not only learn to reproduce sequences, but they also extract the transition probabilities between items across the entire training set. There are simple recurrent network models of single word production that make human-like speech errors (Dell, Juliano, & Govindjee, 1993), and models of speech recognition that use phoneme transition probabilities to help locate word boundaries in continuous speech (Christiansen, Allen, & Seidenberg, 1998; Elman, 1990). Models of syntactic processing use transition probabilities between words to form internal representations that reflect word classes (Elman, 1990). Such networks can also make structure-dependent predictions over long distances, and show sensitivity to embedding and recursion, suggesting that they could form the basis of a connectionist approach to syntactic processing (Elman, 1993).

Connectionism challenges a number of assumptions of traditional, "classical", approaches to studying language and the mind. The classical approach sees rule knowledge as being represented in symbolic form, whilst exceptions to rules are stored in memory (Pinker, 1994; Ullman, 2001). Connectionist models produce rule-like behaviour, and yet they contain no symbolic representations. Rule-like behaviour emerges as a consequence of the way in which individual instances of experience are stored in a single memory system. The connectionist challenge has provoked sustained debate (beginning with Fodor & Pylyshyn, 1988), focussing recently on the necessity for

symbolic, or algebraic, rules (cf. Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Seidenberg & Elman, 1999), and the separation between rule knowledge and item memory, particularly in relation to the English past tense (Pinker & Ullman, 2002 and commentary). The reason for this heated debate is that connectionism challenges the classical assumptions about modularity, domain-specificity, and innateness (as expressed in Pinker, 1994). The same neural architectures and learning algorithms are applied to linguistic and non-linguistic domains alike, and no information is encoded in connectionist networks that is not ultimately derived from the input. Issues such as the poverty of the stimulus, universals, critical periods and robustness have to be approached from novel perspectives (Elman et al., 1996).

Connectionism has inspired new ways of thinking about language processing and learning. It emphasises the statistical properties of the input, and how complex behaviours are the outcome of multiple, competing constraints. This has motivated empirical work looking at human abilities to learn by tracking statistics (Saffran, 2001), and has provided the theoretical foundation for interactionist and "constraint satisfaction" approaches to sentence processing and learning (MacDonald, Pearlmutter, & Seidenberg, 1994; MacWhinney, 1987). Whether or nor connectionism in its purest form will be able to explain all aspects of language processing and learning is an empirical question. But connectionism has stimulated debate and focussed empirical research on fundamental issues in the language sciences.

For a general introduction to connectionism, including reprints of seminal articles, see Ellis & Humphreys (1999). For reviews of language-related work see the special issue of *Cognitive Science* (1999, Vol. 23. No. 4) and Plaut (2003).

Bibliography

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes, 13*(2-3), 221-268.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149-195.

Ellis, R., & Humphreys, G. (1999). *Connectionist Psychology: A text with readings*. Hove: Psychology Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Elman, J. L., Bates, E. A., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development*. Cambridge, MA.: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98*, 74-95.

Joanisse, M., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Science, 96*, 7592-7592.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676-703.

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 249-308). Hillsdale, N.J.: Lawrence Erlbaum Associates.

MacWhinney, B., & Leinbach, J. (1991). Implementations Are Not Conceptualizations - Revising the Verb Learning-Model. *Cognition, 40*(1-2), 121-157.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning in 7-month-old infants. *Science, 283*, 77-80.

Masson, M. E. J. (1995). A distributed model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 3-23.

McClelland, J., & Rumelhart, D. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159-188.

Pinker, S. (1994). *The Language Instinct*. Harmondsworth: Allen Lane.

Plaut, D. C. (2003). Connectionist modeling of language: Examples and implications. In M. T. Banich & M. Mack (Eds.), *Mind, Brain, and Language* (pp. 143-167). Mahwah, N.J.: Lawrence Erlbaum Associates.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56-115.

Plunkett, K., & Marchman, V. A. (1993). From rote learning to system building: the acquisition of morphology in children and connectionist nets. *Cognition, 48*, 21-69.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2). Cambridge, Massachusetts: MIT Press.

Saffran, J. R. (2001). Words in a sea of sounds: the output of infant statistical learning. *Cognition, 81*, 149-169.

Seidenberg, M. S., & Elman, J. L. (1999). Do infants learn grammar with algebra or statistics? *Science, 284*, 435-436.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523-569.

Shanks, D. R. (1995). *The Psychology of Associative Learning*. Cambridge: Cambridge University Press.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews: Neuroscience, 2*, 717-726.