Index and analogy: A footnote to the theory of signs

Derek Robinson

Analogical or similarity-based reasoning uses a database of stored examples to draw probabilistic inferences directly, without rule induction or (statistical or 'neural') learning. Novel instances are classified and outcomes are predicted on demand, based on prior experience of similar situations. In pattern recognition this approach is known as 'nearest neighbors'; despite its appealing simplicity, it is rarely employed due to the belief that finding the closest match entails the examination of all records in memory. This paper shows that, on the contrary, mere indexing (database inversion) can confer the desired 'best matching' capability, turning any collection of data into an efficient, high-capacity associative memory system.

1. Introduction

In a venerable, oft-repeated metaphor, nature is likened to a book. To this trope I would add another – of the mind as an index to the book that is the world. I mean this quite literally: the way that minds work, and the way the index works, are fundamentally the same.

The creature in the world must be able to discriminate what is going on around it and take steps appropriate to the situations so apprehended. The processes of recognition and action are in large part automatic and involuntary, requiring no deliberation, calculation or 'consciousness' (except in the sense of being awake and aware of the surroundings). The immediate circumstances 'bring to mind' or directly invoke any relevant knowledge associated with similar situations encountered in the past, in this way preparing the creature for what may happen next and allowing it to position itself at some strategic advantage. This is the classical doctrine of associationism, the theory of mind first essayed by Aristotle and still, with only small emendation, the cornerstone of scientific psychology. It is a theory of learning as 'animal habit' – how things come to be associated in memory more or less strongly in proportion to their proximity in space or time, the frequency of their co-occurrence, and the magnitude of their consequences for the creature.

Speculative thinkers have long sought a plausible physical basis for associative memory in the brain, though it is a matter of some indifference to the project of understanding intelligence (and of replicating its operations in a machine or a computer program) what the actual

Rivista di Linguistica, 7.2 (1995), pp. 249-272 *(ricevuto nell'ottobre 1994)*

biological mechanisms of learned association are. We can be fairly confident that any such neuroanatomical revelations as may be forthcoming in the next millennium will not contradict the classical account of learning. That Aristotle located the seat of intelligence "in the region of the heart" and viewed the brain as a heat-sink whose function was to "cool the blood" does not affect the perspicuity and enduring worth of his ideas about memory, mental association, language, reason, the formation of concepts and the foundations of prudent belief.

In the mathematical theory of probability, we have an abstract, quantitative model of empirical learning from experience – what Laplace called "common sense reduced to calculus" – that owes nothing to biological fact beyond the fact that living beings comprehended it. In the last fifty years, under the impetus of postwar advances in electronics and computing, there has emerged a new discipline of pattern recognition, conceived triunely as a branch of mathematical statistics, a branch of engineering, and a branch of theoretical biology, whose express aim is to reproduce in automata 'how we know universals' (Pitts & McCulloch 1947).

The field of artificial intelligence (AI) grew up alongside pattern recognition and information theory but early on repudiated 'bottom-up' empiricism for a rationalist, top-down, rule-based approach to cognition having a strong symbolic logic flavor. AI was much preoccupied with devising, by introspection from the stratagems or heuristics people use in solving puzzles and playing games, methods for carrying out automated deduction through the formal manipulation of symbol strings. The last decade has seen a dramatic shift in AI and the allied cognitive sciences of theoretical psychology and linguistics back to a more empiricist, statistical approach, following widespread disillusionment with the idea that formal rules and representations will ever be able to encompass 'what minds do' in dealing with the inherent uncertainty, ambiguity and variety of phenomena making up quotidian reality.

It goes without saying that the moment-by-moment operations of an automaton must be unambiguous and 'formal', otherwise it cannot run. We should like to know what is the minimum of a priori structure which would permit a synthetic 'intelligence' to learn things by itself from the raw data of experience? Classical associationism identified the prerequisites of intelligent, purposive action in a special kind of memory system, organized so that what is encountered in the world ineluctably conjures the appropriate 'consequential region' of memory, wherein is found a digest of what is most important to know about things 'like' the present case.

2. *Similarity*

To realize this kind of associative memory or analogical reasoning capability in software requires only an ability to measure the similarity of patterns, to find the 'nearest neighbors' of an input pattern and make these the basis for statistical generalization and prediction. The usual procedure has been to take the cosine of two n-dimensional vectors (each of n real-valued measurements) giving the classic Pearson's correlation coefficient as their angular 'distance'. Correlation was developed in the context of anthropometry (the comparison of physical quantities such as height, weight, arm-length and so forth across a population) and is ill-suited to discrete, categorical variables. For binary-valued attributes, marking the bare presence or absence of a feature (a discrete state or a quantized interval from the range of a continuous variable) simpler distance measures can be used, based on counting the number of attributes common to two feature-vectors and the number of places they differ. Here, similarity is taken to mean the same but in fewer than all respects; as William James wrote (1892), "Similarity, in compounds, is partial identity".

Statistical methods based on correlation (including linear regression and multivariate analysis) generally assume not just continuous quantities, but that these conform to a normal random distribution, occupying volumes of roughly the same size and shape over the sample space. These assumptions are seldom enough borne out by real-world data and represent a holdover from the days of hand computation, when the simplification given by assuming families of smooth, well-behaved, unimodal distributions having neat analytical formulae, which could be 'fitted' to the data by adjusting a couple of parameters (the mean and variance) was a practical necessity. So-called non-parametric estimation methods, apart from the histogram (1894) and moving-average methods for smoothing tabulated data, did not really appear until the 1950s, with the development of the closely related kernel and nearest neighbors techniques. (The nearest neighbors method takes a weighted average of a fixed number of data points closest to the 'probe'; the kernel method uses a fixed-width neighborhood, irrespective of how many data points fall within it (Silverman 1986)). With the computer, there has gradually come a less presumptive and constricting, more empirical and exploratory attitude toward statistical modeling, though ingrained statistical habits have been slow to change.

Nearest neighbor methods have the advantage over other (statistical, rule-based, or 'neural net') approaches to reasoning under uncertainty, of working directly from the data. There is no need for expensive off-line training, parameter-fitting or rule induction, heuristic search,

explicit 'knowledge representation' or subjectively assigned prior probabilities. Nearest neighbor approaches emerged concurrently in several different fields, reflecting no doubt the intuitive simplicity and naturalness of the idea once the means were at hand for storing and searching large databases of examples. In the AI subdiscipline of machine learning it turns up as memory-based reasoning, instance-based learning, reasoning-by-analogy, and case-based reasoning (Russell 1989). In remote sensing and spectroscopy, it is called 'library search'. And practically all the methods developed in neural networks and pattern recognition can be seen as attempts to achieve nearest neighbor look-up without having to compare an input pattern against all stored exemplars by exhaustive item-by-item search.

3. The Search for the Best Match

In statistical work, the naive (direct search) nearest neighbor and kernel methods, although computer-intensive, can be usefully applied and their mathematical fine-points investigated at length, because the aim is primarily to analyze and summarize collections of data gathered some time in the past, and it doesn't matter very much how long the analysis takes. Areas such as pattern recognition, computer vision, robot motion-planning and speech understanding on the other hand demand real-time performance, precluding any inference regime involving linear search through many thousands of stored examples to find the 'best match' (the 'k-nearest neighbors'), no matter how attractive or desirable this might be in principle.

Because of the importance of flexible matching to so many areas of applied science which involve analyzing very large data sets (for example, satellite remote sensing, geophysics and astronomy databases, medical imaging systems, geographical and legal and chemical information systems, protein data banks, corpus-based linguistics) computer scientists have devoted a great deal of effort to devising faster search algorithms, a field that by the late 1970s had come to be known as computational geometry (Preparata & Shamos 1985). Its practical import has been a motley of (in general rather complicated) data structures and algorithms, including grid-files, multidimensional search trees (k-d trees, quad-trees) and various hybrids of table-based, tree-based and probabilistic computed look-up (hashing) functions.

To their credit, the computational geometers separated the abstract problem of locating the nearest neighbors of a data point in n-dimensions (and the related problems of computing convex hulls, minimum spanning trees, planar intersections, and the like) from particular applica-

tions, such as object recognition, hidden line and surface removal, geometrical curve and surface fitting, global optimization, approximate string matching and 'data mining'. The methods of computational geometry unfortunately tend to be practical only for small-dimensional problems; with increasing dimensionality the amount of housekeeping needed to negotiate complex multidimensional data structures can quickly get out of hand.

Other researchers sought a solution in massive parallelism, designing novel computer architectures of thousands of simple processors connected in dynamic marker-passing networks. Scott Fahlman's "NETL" (Fahlman 1979) was, if not the first such proposal, notable in the influence it exerted on subsequent developments. Fahlman argued compellingly that only by means of 'best matching set intersection' would AI ever be able to step out of its toy-sized 'Blocks World' to tackle the very large database problem: how to amass and access a vast store of commonsense knowledge about the ten million things a young AI needs to know in order to deal intelligently with the real world. Only then would AI become a viable technology. He also argued, and many were convinced, that the only way to realize a best matching machine capable of real-time operation was as a massively parallel multiprocessor network.

An immediate consequence of the NETL manifesto was the Connection Machine project at MIT, which by 1984 had succeeded in constructing just such an AI supercomputer (Hillis 1985). The basic idea was 'divide and conquer': to split the search task between thousands of processors, each with its own memory, so that they could all search their own small portions of the database concurrently and so achieve a speed-up proportional to the number of processors (not counting the time required to sort the results). This is the 'memory-based reasoning' approach promoted by Stanfill & Waltz (1986) of Thinking Machines Corporation, the manufacturer of the Connection Machine.

Connectionist neural networks offered another, finer-grained approach to massively parallel computation, which, like the Connection Machine, was inspired in part by Fahlman's vision of parallel distributed semantic networks. Connectionism sought to overturn the prevailing opinion in AI that artificial neural nets had been proved a dead end by Minsky & Papert (1969). The connectionists had developed, they claimed, new learning methods that overcame the limitations of the earlier single-layer (of adjustable weights) networks, and would therefore enable neurocomputers to accomplish all the wonderful things that biological brains can do. Connectionism swept like a brush-fire through the dry tinder of cognitive science, establishing itself as the major alternative to the symbolic, rule-based paradigm (Rumelhart et al.

1986). Now, after close to a decade of rapid growth, many within the neural networks field are looking outside it to areas like coding and information theory, Bayesian statistics, clustering, projection pursuit, spline fitting and non-parametric regression – while some from outside connectionism are wondering whether neural networks weren't perhaps just statistics all along (Ripley 1993).

The methods of statistical pattern recognition and connectionist neural networks are, like those of computational geometry, so many attempts to realize best matching look-up without paying the cost of having to actually examine every item in memory. All parties have implicitly accepted Minsky & Papert's verdict that, on serial computers, "for large data sets with long word lengths [i.e., high-dimensional feature vectors] there are no practical alternatives to large searches that inspect large parts of the memory". The remainder of this essay concerns a strangely overlooked yet extraordinarily simple solution to the best match problem, which conclusively refutes the "gloomy conjecture" that nearest neighbor methods cannot be realized efficiently on serial machines. The great irony is that everyone already knows the solution; it has in fact been part of the Western intellectual heritage for some 700 years.

4. Index and Inference

In the familiar back-of-the-book index, each keyword has its own inverted file identifying the places where the word is to be found. Such inverted indexing predated the appearance of mechanically printed books by two hundred years. In 1247 Hugo de St. Caro and 500 monks produced the first complete concordance of the Bible, whose undertaking reflected a sudden self-awareness within manuscript culture of the formal properties of texts and the interior technics of reading, memorization and visualization. Between 1220 and 1280 there appeared in the monasteries of Europe a great proliferation of coding and filing tools and schemes, including numerical pagination and the alphabetization of word-lists (Carruthers 1990). No matter how obvious these ideas may now seem, it must be appreciated that they were invented things, and as such, although the prerequisite elements were everywhere at hand, nothing necessitated their creative synthesis or the eventual widespread adoption of this 13th century information technology. The conventional 'A-B-C' order of the phonetic alphabet had, after all, been in place for some two thousand years without (so far as we know) it having occurred to anyone to organize lists of words alphabetically (Illich & Sanders 1988). It is not so surprising therefore that for forty years

computer science could have missed seeing what was right at its fingertips – that the index provides the associative, analogical matching capability researchers had long searched for and despaired of finding.

On October 14, 1953, there was registered with the British Patent Office in London "an apparatus to assist the logical faculties" in performing differential diagnosis, called by its inventor, Dr. F.A. Nash, the 'Grouped Symbol Associator' or 'Logoscope'. A brief article by Dr. Nash describing his invention appeared the following April in *The Lancet* (Nash 1954). The Nash Logoscope was a sort of semiotic slide rule, comprising a set of removable vertical strips of card, each dedicated to a particular symptom and marked by a spectrum of transverse lines whose positions were in correspondence with a separate reference card listing nearly 400 diseases. Each vertical strip was a place-coded inverted file giving for each symptom a list of all diseases in which that diagnostic sign is known to occur. Given the set of symptoms presented by a patient, the clinician would align the appropriate strips with the register-card of disease categories so that immediately the best match (the disease having the greatest number of signs in common with those present) would be revealed as an unbroken or nearly unbroken line running horizontally across the vertical columns.

The logoscope was essentially a graphical or panoramic index which, by the simple expedient of breaking apart a conventional book index into its constituent inverted files and representing the information in them as binary vectors, enabled parallel search with any chosen subset of keys, to identify at a glance the item or items having most terms in common with the query. It was a parallel 'best match machine', a cardboard AI (though the field of artificial intelligence did not yet exist) and in fact the first medical 'expert system'. As with the origin of indexing in the 13th century, nothing compelled the invention of logoscopy in the 20th. In principle, it could have been developed at any time since antiquity; but for all the notice it attracted, it might as well not have been invented at all. Perhaps because of the transparency of its mechanism, because it was neither obviously statistical nor an electronic 'black box', or (most likely) because it first appeared in the pages of a medical journal, it was ignored.

Dr. Nash again tried to bring his Logoscope to the attention of his scientific peers at the landmark 'Mechanisation of Thought Processes' conference held in England at the Teddington National Physical Laboratory in 1958. This was the second meeting to be convened on the subject of artificial intelligence and many of the key figures in cybernetics, information theory, and the nascent fields of neural networks and symbolic AI were present (Teddington 1959). No one, apparently, noticed that the logoscope was in fact an instructable associative memory

of the type that researchers like Donald MacKay, A.M. Uttley, W.K. Taylor, Frank Rosenblatt, W. Grey Walter and Ross Ashby were then attempting to reproduce in analogue electronic 'nerve nets'.

In logoscopy, the keywords are symptoms and the page numbers are diseases, but the logic is much more general. The keys could as easily be fragments of graphical shapes – letterforms, chemical structure diagrams, machine parts – and the resulting index would be a pattern recognition system which, presented with an unknown figure, would break it apart into its constituent features, find their corresponding inverted files and identify from its database the most similar letters or molecules or objects known to it, ranked by their proximity to the input pattern. The keys could be physical characters or measurements observed in different species of plant or animal or microbe, so that an unknown specimen immediately invokes, from a taxonomic index so configured, the known species it most nearly resembles. Or they could be anthropometric variables, as with Bertillon's system of criminal identification, in use from the 1880s to the First World War – but how was the investigator to use bertillonage, or for that matter the rival British system of 'dactyloscopy' (fingerprinting) to locate the records closest to a set of measurements? By means of a card index, accessed by numerical codes cunningly derived from different categories and grades of the measured attributes (Thorwald 1965).

5. Semiosis

The logoscope makes explicit the connection between the bibliographic index and diagnostic reasoning. It furnishes the middle term, as it were, of a syllogism – once one comprehends the mechanism of the logoscope, immediately any intellectual operation which can be regarded as diagnostic can be seen as (and implemented as) an index. Perception, for example, is a kind of diagnostic inference – how we get from the 'proximal threads' (the combined activities of nerve cells in the sensory array) to the 'distal knots' (the objects, solid and persistent, existing 'out there' in the world) – and so too by extension is the entire project of scientific understanding: to infer or reconstruct why the world is as it is and not otherwise (Campbell 1966).

In many situations, medical diagnosis and legal reasoning among them, attempts to consciously dictate general rules are foredoomed because they disregard the complex congeries of particulars and contextual factors whose appraisal constitutes precisely what we mean by 'judgement'. Rather, medical and ethical reasoning are based on the resemblance of present cases to known precedents (Jonsen & Toulmin

1988). Such 'case based reasoning' is termed casuistry, which in the 17th century (like the related terms 'rhetoric' and 'sophistry') acquired the stigma it still carries. One can discern an historical parallel in the odium with which statistical or empirical approaches to language are held by generative grammarians espousing a 'Cartesian linguistics' of pristine, geometrical rationality, which views all the messiness of real speech as 'merely performative' and, as such, not of interest. It is the position of this paper that rules are not needed at all if there exists a store of specific examples from which inference may be drawn 'abductively' on a case-by-case basis; and further, that this is how we do in fact reason, notwithstanding our willingness, when pressed, to make up 'rules' – that is to say, rationalizations – on the spot from examples and counter-examples, to justify our actions and choices after the fact.

Medical diagnosis is paradigmatic of the unconscious processes of recognition, recollection and interpretation underlying conscious experience, which Polanyi (1966) termed the 'tacit dimension' of knowledge: how is it "that we can know more than we can tell"? How can we tell a computer 'what things are' if we cannot 'tell', in words, how we know what they are? We recognize a face or a flower in the same way a diagnostician recognizes a fibroma or fibrillation, by means of their signs – but without, in general, being able to tell precisely (so that a computer could carry out the procedure) what the signs are. The signs are subordinate to the things: we perceive objects and relations and events, we do not perceive 'sense data'. In learning a new skill, the novice self-consciously attends to the individual constituents of the act, and the performance is to that degree awkward and inexpert. Only when the details have been fused and disappear into the meaning or purpose of the actions does the skill become artless and fluent, 'second nature'. Only a second-language learner or a linguist need learn the rules of grammar – rules are like training wheels, which may help one get started but then must be abandoned and the reliance on them unlearned if one is to progress. It is the same with artificial intelligence (Dreyfus & Dreyfus 1986). Rules were to provide a shortcut, a way to get AI started, but they ended up creating a cripple incapable of crossing a room unaided, distinguishing a smile from a frown, or understanding the words in this sentence.

In its older meaning, semiotics was defined as "that branch of pathology concerned with symptoms". We arrive at the interpretation of words spoken or written in essentially the same way that a clinician arrives at a diagnosis and decides on a course of treatment (Blois 1984). Signs are inherently ambiguous, underdetermined, polysemous. Any particular sign could be a sign of many different things. They are 'clues' (Ginzburg 1983) and in negotiating and comprehending the world we act as detectives. Perception and understanding are forensic sciences: from

numerous scattered, individually weak and undecidable pieces of evidence, we must reconstruct their most probable causes. In mathematical physics this would be termed an 'inverse problem' – ill-posed, admitting of a vast number of possible solutions which cannot be narrowed to a single best hypothesis without importing additional presumptions not strictly warranted by the facts. Overwhelmingly, the presumption we impose on the world, which enables it to 'make sense', is our prior experience of it and the animal faith that things will continue to be much the same as we have known them to be in the past. By the likeness of things present to things past, we infer the likelihood of their possible consequences in the future – in David Hume's words (1748): "From causes which appear similar, we expect similar effects".

The index reverses or inverts the usual direction of logical implication which goes from cause to effect, whole to part, object to attribute, class to individual. What makes the inverse problem so hard is its indeterminacy: the path from cause to effect is one-way, but to get from an effect or sign to its possible causes is a one-to-many mapping and is therefore ill-defined. However, by considering many such one-to-many mappings at once and seeing where they converge or overlap, the 'circles of confusion' are drawn to a tight focus and the indeterminate can be readily determined. Best matching is a screening procedure which eliminates at outset the far greater number of irrelevant candidates sharing no properties or signs with the present case. To this is owed its great efficiency: no matter how many records there are in the database, only those having features in common with the query need be considered at all.

6. *Some Applications*

Best matching search using inverted files has been independently invented several times by researchers in different disciplines. Inverted indexing has been the backbone of information retrieval (IR) since the field's inception (around 1950) but it was wedded to a Boolean model of query formation based on the logical operators AND (set intersection), OR (set union) and NOT. Unfortunately, people are not very good at constructing or deciphering complexly nested logical expressions; moreover, Boolean search demands an intimate prior knowledge of the contents of the database if it is not to produce either a superfluity of spurious matches, or none at all. The first appearance of the best matching rule – which uses, instead of set intersection and union, set superposition to rank the items whose keywords coincide in the greatest number of places with the terms of the query – seems to have been in an

IR system from 1967. Peter Willett reinvented the algorithm in 1981 and has subsequently, in many publications, endeavored to bring it to the attention of information scientists (Willett 1988). Much of Willett's work has been in chemical information systems, which occupy a position midway between information retrieval and pattern recognition. Systems designed by Willett and his students enable chemists to sketch a partial chemical structure diagram at a computer terminal and immediately retrieve, from a large on-line database, the set of known molecules which are closest in structure to the unknown compound, together with all pertinent facts and citations concerning them (Willett 1987).

In 1978 Teuvo Kohonen (best known for his work on self-organizing neural networks) developed a technique called 'redundant hashed addressing' (RHA) for realizing content-addressable associative memory in software on serial computers (Kohonen 1980). A decade later, when neural nets were once again the fashion, he pointed out that RHA was a general and efficient way to realize any neural network design. Kohonen applied RHA to spelling correction by breaking up the words in a dictionary into redundant overlapping trigrams (e.g., _TR, TRI, RIG, IGR, GRA, RAM, AM_) and assigning to each trigram key an inverted file of the words in which it occurs. Independently, Peter Willett and his colleagues developed virtually the same trigram-based spelling correction scheme, which performs as well or better than systems constructed at much greater effort and expense as large, complex sets of hand-crafted stemming and lemmatization rules and long lists of exceptions (with which the English language is especially blessed).

N-gram partial matching reappears in high-speed search software for massive protein databases (Lipman & Pearson 1985). Sequences are broken up into overlapping segments (or 'w-mers') of from 4 to 16 amino acid symbols in length, which are used as keys into an inverted index of known protein molecules. Work has also been done, using a weighted nearest neighbors rule, on the problems of predicting the two-dimensional folding of protein molecules from one-dimensional data and identifying DNA promoter sequences (Cost & Salzberg 1993).

In machine vision research, best matching is being used (under the alias 'geometric hashing') to perform rapid identification of two-dimensional and three-dimensional objects (Grimson 1990). Earlier approaches to object recognition used sequential search to measure an unknown object against a set of stored 'models', one model at a time. Since a robot designed to operate freely in natural or domestic environments will need to be able to recognize and understand the functions of some ten thousand or more everyday objects, time-consuming linear search is clearly out of the question.

In the above applications, the developers and users of best matching software seem unaware that the same fundamental idea has been used (or could be of use) in many other disciplines. Despite its repeated reinvention, computer science as a whole still believes that efficient serial best matching is either not possible at all, or that it remains an open problem. Thus it is that a recently published collection of technical papers and tutorials on nearest neighbor methods in pattern recognition (Dasarathy 1991) could overlook indexing completely, while dealing at length with complicated and costly optimization procedures like 'branch and bound' search.

7. Applications to Linguistics

The signs of language, like the signs of things in the world, are in varying degree ambiguous. Each sign may partially satisfy many possible meanings or interpretations, and each state of affairs indicated through language could be expressed by many different combinations of signs. And while any individual sign (a word, a letter, a sound) may be by itself equivocal, overall the sense of an ensemble of signs (a text, a connected utterance) is, to the members of a linguistic community, decidable or decodable without conscious effort due to the many-ways overdetermination afforded by numerous redundant, overlapping clues.

If we may take orthography as typical of the kinds of coding and decoding accomplished in language, the success of nearest neighbor spelling correction without any explicit knowledge of word-roots or morphological rules suggests that people may be doing something rather similar in mastering language, and that the problem of language understanding by computer might be profitably tackled in much the same way. It could be objected that fixing spelling or typographic errors is a trivial problem; after all, automatic spell-checkers are now a standard feature of word processing software. But if we can associate words with fragments of words, we can also associate sounds and can readily create an index capable of 'learning' to convert text into speech. This would be considered by most a non-trivial task, and many within linguistics and psychology were first alerted to the promise of connectionist modeling by the ability of the NETtalk neural network (Sejnowski & Rosenberg 1987) to accomplish it. It is less widely known that NETtalk's success has been equaled or bettered by several other researchers (Stanfill & Waltz 1986, Wolpert 1990, Cost & Salzberg 1993, Lowe 1993) using nearest neighbor methods (none used best match indexing, which would of course produce equivalent results but more efficiently).

Less trivial still is the problem of semantic disambiguation: deter-

mining, of the several possible meanings of a word, which one was intended. In his recent book *The Science of Words*, George Miller (1991) accounts polysemy the hardest problem facing computational linguistics. This is due to the fundamental 'chicken-and-egg' circularity of meaning: the sense of a sentence can come only from the meanings of the words composing it, but the senses of these words are determined by the meaning of the sentence in which they appear. The problem was substantially solved by Karen Sparck Jones in her 1964 thesis *Synonymy and Semantic Classification* (Sparck Jones 1986) using inverted files of 'topics', such that determining the correct sense of content-words in a sentence required only superimposing their respective lists of topics to identify the topic common to all or most of them. The topics were ready-made, being the thematic terms associated with the words in the alphabetic index to Roget's *Thesaurus*. Thus, for example, while 'play' and 'game' in the sentence "Let's play a game" have each up to ten different senses (as given by their index entries in the thesaurus) only one sense, AMUSEMENT, is common to both, preventing such possible misreadings as "Let's <PERFORMANCE> a <RESOLUTE>". (Or, conceivably, "Let's <DRESS-UP> a <FOWL>".) This cannot by any means be considered a complete solution to the problems of understanding language, but it is a significant and too little appreciated step toward it.

The generic word-senses or topics associated with individual words can be augmented with their respective parts of speech to accomplish considerable syntactic disambiguation by lexical means, without using formal grammars. Currently, one of the most active research areas in computational linguistics is the construction of machine readable dictionaries (MRDs), a byproduct of the widespread adoption of computer indexing by lexicographers and dictionary publishers (Butler 1992). MRDs represent a compromise between formal (syntactic, semantic, morphological) rules and purely extensional approaches based on counting word collocation frequencies in huge databases of tagged text – the outstanding example being the CLAWS probabilistic parser (Garside et al. 1987) which uses the Lancaster-Oslo-Bergen Corpus to reliably achieve 97 percent correct parsing of raw text input.

Words do not appear in isolation, but in combination. Commonly recurring patterns are deemed cliches or idioms, and often must be considered 'of a piece' in order that the intended meaning be taken. The expression 'of a piece' is a good example – its sense is to a degree figurative, insofar as a literal reading, 'belonging or pertaining to a part', does not correspond in any straightforward way to 'taken-all-together'. Far more of language than one might at first suppose is owed to cliche. If vocabulary is enlarged to include common idioms and turns of phrase, the number of 'words' in everyday use would grow to many hundreds of

thousands, and much of what is now considered syntax would disappear into the lexicon. At the same time, the frequencies of the lexical items would be 'flattened' to approach a uniform distribution, which has certain advantages for textual data compression and discovering statistical regularities in language (Lynch & Rawson 1976, Bell et al. 1990).

Neural associative memories and feed-forward networks do not easily scale up to the kind of vast encyclopedic memory that language understanding requires. To deal with the millions of things that must be known in order to understand what is talked about in discourse, neural nets which connect every input attribute to every possible output category (or in multilayer nets, to every hidden unit, each connected to every output unit) are, simply on combinatorial grounds, infeasible. Further, training the connection strengths gets progressively slower with increasing network size. By contrast, text retrieval systems (such as library catalogues) based on inverted indexing commonly handle collections numbering in the millions of items (Witten et al. 1994). Best matching becomes more, not less efficient with increasing numbers of specialized features. Adding new examples to a neural network entails training it all over again from scratch, but an index 'learns' new relations immediately as they are added to the database.

8. Neural and Statistical Modeling

The importance of a feature in discrimination is a function of its frequency distribution across a population of individuals or classes (where too common means poor discrimination, while too rare is uneconomical), and secondly, its interactions with all other features. It is the latter which makes feature discovery a notoriously difficult combinatorial search problem (Lefkovitch 1993). If one already possesses a good feature set, or a good 'model', then learning and discrimination are trivial – most of what is considered learning in neural networks is mere calibration or parameter-fitting, the kind of iterative relaxation or adjustment procedures in use since the 1930s for fitting statistical and econometric models by, in effect, solving or inverting large systems of simultaneous equations (Deming 1943).

Finding a 'good representation', an optimal feature set, is in general so computationally onerous that (a) it is in the first place usually left to human judgement; or (b) the features chosen will typically be just the original measurement variables or a reduced set of principal components (orthogonal linear combinations of the original variables which compress the greatest amount of variance into the fewest terms at the cost of making certain presumptions about the data which may, if

unwarranted, irremediably lose or confound valuable information); or (c) neural learning methods like backpropagation may be used to generate 'inner representations' loosely analogous to principal components or factors – a set of black-box latent variables that needn't resemble anything we would ordinarily consider 'features'. Even where a neural net is relied upon to perform feature selection automatically, the net's designer must preselect the number of layers, the number of units per layer, their connectivity, and various gain and decay terms for the learning regime, which amounts to specifying the free parameters of a 'model' – which is just what the net was supposed to be doing for us!

A good feature set (however it was produced) should be reasonably 'well-balanced': the features should be of roughly the same frequency and variance, and they should as far as possible be uncorrelated with one another. A too-common feature can be discounted or penalized by assigning it a numeric weight inversely proportional to its frequency. Supposing that the data records are grouped into different classes of object or outcome, then the number of records in a given class which show a certain feature is divided by the total number of instances of that feature over all classes, giving the conditional probability that an occurrence of the feature belongs to a member of that class. This weight can be combined with other 'weights of evidence' associated with the other features present and possibly multiplied by the prior probability of the class (the ratio of within-class objects to all objects in the universe of discourse) to yield, finally, the posterior probability that an unknown object is an instance of the class. Comparing the posterior distribution of votes for each possible class allows identifying the candidate hypothesis which maximizes the likelihood that the unknown object comes from this ('most probable') class. That, in a nutshell, is the standard approach to inverse probability (statisticians are divided on the issue of 'priors': Bayesians feel justified in assigning prior probabilities subjectively, but likelihoodists think this is in poor taste. For their part, the likelihood camp will choose, from a range of standard random distributions, one whose parameters are fitted using an iterative estimation procedure (Hastie & Tibshirani 1990) – though the initial choice of 'error model' is really no less subjective here than in the Bayesian approach).

Alternatively, the feature set can be decorrelated and 'load-balanced' by adding new, higher-order terms (conjunctions of the input variables) to the model. Such higher-order features represent more-specific, narrower 'contexts', hence their incorporation serves to 'sparsify' the model, increasing the total number of features while ensuring that none is found in more than a small portion of the database. Attributes which are found together improbably often (and are therefore highly mutually correlated) are 'suspicious coincidences'. By identifying these

clusters and making them their own, distinct, higher-order features, the feature set becomes a 'good code' – the input patterns or codewords become widely separate and easily distinguishable, while the features are made statistically independent.

Independence is undoubtedly the single most important concept underlying probability theory and statistics (or for that matter, the numerical solution of systems of linear constraints). Statistical independence means that the product of the ('marginal') probabilities of variables considered individually will approximate their joint probability, P(A) x P(B) x P(C) = P(ABC), which is a prerequisite for valid inference (that is, computing the correct posterior distribution). Independence allows the linear superposition of probabilities, which in turn means that the product space (the multidimensional feature-space where individual records appear as points at coordinates given by the features) can be compressed into a number of independent subspaces or 'projections' without losing critical information or unduly distorting the distance relations between data objects. Only additive storage (A' + B' + C') is needed to represent or reconstruct the original multiplicative (A' x B' x C') product space, where A', B' and C' denote the number of states (cardinality) of the feature 'dimensions' A, B, C.

Inverted indexing can turn any collection of data records into an associative memory system, such that the attributes of an unknown entity can quickly retrieve the set of instances closest to it. Once we have the nearest neighbors, their properties (class labels and class probabilities, or the predicted values of response variables) can be compared or averaged to make probabilistic inferences about corresponding properties of the present case. The stored exemplars serve to mediate between the signs or facts observed in the case at hand, and other properties, themselves unobserved, which have been found in similar cases in the past (just this is what is meant by 'analogy'). A great economy of representation can be realized by discarding the actual data records (or 'training set') and retaining only a statistical summary of the frequencies of properties of interest (the dependent or response variables) vis a vis the input (sometimes confusingly termed 'independent') variables or features. Indeed, such concision is the rationale for having 'classes' at all, rather than a nominalistic universe of only singular individuals. A class mediates between individuals and individuals, so that what is true of one member of the class will also, with high probability, be true of the other members as well. A class label can be considered just another property or predicate associated with an individual – it affords an inferential shortcut by collapsing sets of similar individuals into a single 'prototype', a statistical average or conflation of the traits appearing in different members of the class. There need be no single attribute common to all

members, as Wittgenstein pointed out in his discussion of family resemblances, and as cognitive anthropologists have confirmed in numerous recent studies of human categorization (Lakoff 1985).

Where there does not already exist an a priori classification, 'unsupervised learning' or clustering can be employed to discover (or impose) a segregation of the data samples into distinct types or classes based on their similarities and differences, to achieve the data compression or economy of representation which is the aim of symbolization. Clusters of similar objects can be reduced to a single 'latent class' or prototype vector, just as clusters of similar variables or measurements can be collapsed into a 'latent variable' or eigenvector (the group mean of a set of correlated signs whose coincidence across a population of individuals allows regarding them as measuring essentially the same thing).

Thus, rather than maintaining inverted files of pointers to numerous individual data records, a best matching index can associate, with its keys, lists of the classes in which a feature is known to occur, with a frequency count (or class conditional probability) telling what proportion of individuals possessing a certain feature belong to each class. This could be termed a 'conditional probability computer' (Uttley 1956) or a 'Bayes net'. It resembles a standard, single-layer neural network, except its weights are explicit probabilities and are calculated in a single pass through the training set. In neural nets on the other hand, the weights are not probabilities but rather the coefficients of a system of linear equations or inequalities found by iterative error-minimization requiring multiple passes through the data. Hence, 'supervised learning' in neural nets is coercive – weights can be made to conform to a classification or mapping which needn't reflect the intrinsic similarities or distances in the data.

For predicting real-valued quantities a neural network (viewed as an adaptive filter or linear predictor) can be computationally very efficient since it has only to sum the weights associated with the input variables to produce its response. But for classification tasks, where the output will be the identity of an unknown pattern's most probable class, each class will require its own 'neuron' with its own set of weights, and the outputs (summed weights per class) are typically thresholded to give a '1-of-n' place-valued output code. This requirement makes training slow and difficult, since there must be found a set of weights which will, for all input patterns over the n regression equations, produce n - 1 sums below threshold, and just one which is above the threshold value.

By contrast, the conditional probability computer produces posterior probabilities for every class, and the class having the highest probability is deemed the best match. This 'winner-take-all' decision rule incurs an extra sorting step (not needed in threshold nets), but even

so it is plain that if fewer classes are selected by each input feature, then less work is required to sum their weights and sort the results. Again we see that sparse coding – where each feature is specific to a small subset of all cases – confers a greater efficiency and economy overall, even though 'paradoxically' the number of features (hence the dimensionality of the problem space) has been increased. If features are chosen to be of approximately equal (and equally low) frequency over the database of examples, the cost of computing the appropriate output response or deciding an input pattern's class can be made roughly constant-time, irrespective of the number of training instances and categories. Searching ten thousand or ten million cases can be accomplished in the same time it takes to search a few hundred. But using threshold logic to produce a '1-of-10,000,000' output code, which is how a neural net would naively go about it, is hardly practical!

Sparse coding also makes finding a feasible set of weights by iterative relaxation easy and rapid, requiring only a small number of training cycles to converge to a solution. (Both neural net learning and maximum likelihood parameter-fitting can be viewed as relaxation methods.) For decorrelated, statistically-independent features, the simplest iterative error-correction scheme is numerically stable, provably convergent and very efficient (Hackbusch 1994, Shewchuk 1994). The great challenge in statistical or neural learning is therefore discovering a 'good code' or 'good representation' in the first place.

9. Analogical Modeling and Higher-Order Feature Selection

Royal Skousen, in his book *Analogical Modeling of Language* (Skousen 1989) has introduced a novel, principled approach to inference based on 'natural statistics', which gives a formal procedure for discovering optimally discriminating feature sets. In the analogical modeling methodology, sets of features (or contexts) are higher-order combinations or cross-terms of the original variables, chosen to be 'homogeneous' with respect to an output categorization. The homogeneity constraint guarantees that any potentially different context will be distinguishable; as Skousen says, it "represents the strongest statistical test possible". Each homogeneous context will have been selected so that its instances conform uniformly in behavior.

Inference proceeds by first identifying the 'analogical set' of an unknown instance, being the set of its applicable contexts for which homogeneity obtains. Each context has associated with it a frequency count of its known cases and the category to which they belong. Per each category implicated in the analogical set, the frequencies associated

with the selected contexts are tallied (using a special information theoretic rule) to give the posterior distribution of possible outcomes for that instance (Skousen 1992).

The analogical modeling method is thus stricter in its inference rule (based on homogeneity) than the best matching approach outlined in this paper; both methods however implicitly preserve the true n-dimensional 'neighborhoods' of the data, by replicating the individual training instances to all of their contexts or features. This means that the information contributed by an individual exemplar will be 'counted', in the evaluation of a novel case, in proportion to the number of homogeneous contexts they have in common. Enforcing strict homogeneity is quite costly, as it entails extracting all homogeneous contexts from the 2^n possible combinations of n input variables. Note however that the decomposition into homogeneous contexts need only be performed once, in an off-line compiling stage, and the resulting set of contexts can be partially ordered (in a lexicographic tree or lattice structure) for rapid look-up at runtime.

Analogical modeling is unique in making explicit a strong and principled criterion for feature selection which ensures that no pertinent higher-order information will be overlooked. While statisticians would acknowledge that to simply assume independence without verifying it is improper, and that taking higher-order interactions into account can impart valuable additional information which is otherwise thrown away, they have been prevented from exploiting higher-order statistics by the daunting combinatorial explosion of product terms, effectively prohibiting any complete account of the interactions of more than a dozen or so variables. However, search costs can be greatly reduced by considering only those combinations of variables which actually appear in the training data (Steeg & Robinson 1993).

It is perhaps not necessary that the feature set be entirely homogeneous (although of course this is the ideal) in order to avail of information that naive linear models ignore. It is only necessary that the features be numerous enough and sparse enough to together permit unambiguous identification. The requisite sparsity can be achieved by forming higher-order cross-terms of the presenting variables, or by breaking up the ranges of continuous variables into discrete and possibly overlapping intervals or subranges. Variants of this strategy are currently receiving a lot of attention in signal processing (as sub-band and wavelet filtering), in approximation theory (localized radial basis functions) and in fuzzy logic and fuzzy control theory. The redundant n-grams of the Kohonen-Willett spelling correction scheme are examples of such higher order features – a given trigram ('ING' for example, or 'QUE') will have a joint probability significantly greater than predicted from the product of its

component marginal probabilities. The individual symbols are not independent; together they form a suspicious coincidence and provide discrimination over and above what they give alone. A good statistical model or a good code will take such interactions between variables into account; a poor model will not. A model which recognizes only first-order probabilities cannot help making wrong guesses about the likelihoods of their causes (unless of course the first-order terms are indeed the independent and identically distributed random variables presumed by standard mathematical statistics, upon which the validity of its inference procedures rests).

Sparse coding, which increases the number of variables, has been neglected for the apparently sensible reason that the goal of statistical modeling and data analysis is to reduce dimensionality, not expand it. Conventional dimensionality reduction methods, like principal components or factor analysis, will often produce unintelligible or uninterpretable latent variables (as do the 'inner representations' of multilayer neural nets): each synthesized component or factor is a weighted linear combination of all the input variables, whereas higher-order cross-term features represent specific, tightly focused facets of the training data. For large data sets, sparse coding realizes a significant combinatorial advantage over typically dense-coded linear models or (non-linear) neural networks. And it is interesting to note that the receptive fields of sensory neurons in vertebrate nervous systems realize just this kind of sparse, distributed 'population code' (Churchland & Sejnowski 1992).

The inferential shortcut and compression conferred by compiling a database of examples into a special-purpose predictor or classifier is bought at the cost of reduced flexibility and loss of information, compared with a pure, memory-intensive nearest neighbors or instance-based approach. For dedicated applications, statistical compilation can be advantageous. A fully indexed database of cases will of course allow generating such special-purpose views as needed, while still permitting using the same data for other purposes, or simply exploring the database interactively to discover unsuspected patterns or relationships and suggest new hypotheses (so-called data mining, which in the past was strictly beyond the statistical pale, since it was thought to compromise the analyst's objectivity – but is now indispensable if we are not to be overwhelmed by the growing 'data mountain').

10. Summary

Analogy allows "going beyond the information given" by implicating a set of known precedents related to the present case by the possession

of common features. The more things they have in common, the more similar they are and the more confident we can be that what is known to be true of this 'analogical set' is probably also true of the case at hand. In his *System of Logic*, John Stuart Mill (1843) put it this way: "If, after much observation of B, we find that it agrees with A in nine out of ten of its known properties, we may conclude with a probability of nine to one, that it will possess any derivative property of A as well."

Analogy offers a way around the vexing problem of 'small sample size', permitting inference in situations where there may exist only a single precedent or training instance. Thus it doesn't rely upon the frequentist assumption of a hypothetically infinite population of trials; neither does it demand, unreasonably, that we should always be able to furnish subjective estimates of prior probabilities. Analogy therefore provides a missing link in statistical inference: a plausible and objective justification for our readiness to give odds in situations which do not, on the face of it, seem subject to the law of large numbers. The existence of an analogical best matching capability, for example, puts paid to the Chomskian 'poverty of stimulus' argument for the necessity of a special innate language machine – the idea that a language learner (or an autonomous discovery procedure) is exposed to far too few examples of well-formed grammatical productions to allow learning the 'rules' of proper sentence construction solely from experience.

Implicitly, both conventional inverse probability methods and pattern recognizing neural networks rely on an indexical inversion which associates frequency information from a population of individual training cases with certain of their properties. This allows inference to proceed first abductively, going from the particulars of the present case to the analogical set of closest known exemplars, then back again, extrapolating from properties or consequences associated with those precedents to the present case. This twofold motion is masked however by the compilation involved in statistical learning procedures, where the sets-in-extension of instances sharing the same property are replaced by numeric weights or probabilities. Best matching, on the other hand, exposes the underlying logic, since it retains all the original data and aggregates them 'on the fly' only as needed to answer specific queries about any given derivative property (and not just those for which discriminant or regression functions have been previously synthesized).

The fundamental mechanism of analogical inference by object-attribute inversion is present already in the familiar bibliographic index, which as we have seen permits rapid multi-key search for the best matching exemplars and can therefore be regarded as a content-addressable associative memory. Thus we find there has long existed, latently, a formal and procedural theory of analogy, hidden in this

everyday artifact of literate culture. Moreover, this 'inverse theory' is very close to our intuitive, commonsense notions of similarity (things are similar in proportion to the number of things they have in common) and memory (we consider similar things to be in some sense close to one another in the mind).

Aristotle (called by Dante the "master of those who know") anticipated just this 'inverse' conception of formal analogy in the *Posterior Analytics*. The following passage from that book – which expresses in one sentence both the modus operandi of best matching set intersection and the still challenging problem of determining the optimal feature set or model – provides, I feel, a fitting coda to our investigations: "We must select attributes of this kind, up to the point where, although each of them has a wider extension than the subject, all together they have not; this will be the essence of the thing".

Address of the author

Derek Robinson: New Media Department, Ontario College of Art, 100 McCaul St., Toronto, Ontario M5T 1W1, Canada, e-mail: jutta@utirc.utoronto.ca

References

ARISTOTLE, *Posterior Analytics*.
BARNDORFF-NIELSEN O., JENSEN J. & KENDALL W., eds. (1993), *Networks and Chaos: Statistical and Probabilistic Aspects*, London, Chapman & Hall.
BLOIS M. (1984), *Information and Medicine*, Berkeley, University of California Press.
BELL T., CLEARY J.G. & WITTEN I. (1990), *Text Compression*, Englewood Cliffs, New Jersey, Prentice-Hall.
BUTLER C.S., ed. (1992), *Computers and Written Texts*, Oxford, Basil Blackwell.
CAMPBELL D.T. (1966), "Pattern matching as an essential in distal knowing", in HAMMOND 1966:81-106.
CARRUTHERS M. (1990), *The Book of Memory: A Study of Memory in Medieval Culture*, Cambridge, Cambridge University Press.
CHURCHLAND P. & SEJNOWSKI T.J. (1992), *The Computational Brain*, Cambridge, Massachusetts, MIT Press.
COST S. & SALZBERG S. (1993), "A weighted nearest neighbor algorithm for learning with symbolic features", *Machine Learning* 10:57-78.
DASARATHY B.V., ed. (1991), *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, Los Alamitos, IEEE Computer Society Press.
DEMING W.E. (1943), *Statistical Adjustment of Data*, New York, John Wiley & Sons.

DREYFUS H. & DREYFUS S. (1986), *Mind Over Machine*, New York, Free Press.
ECO U. & SEBEOK T., eds. (1983), *The Sign of Three*, Bloomington, Indiana, Indiana University Press.
ECO U. (1984), *Semiotics and the Philosophy of Language*, London, Macmillan.
FAHLMAN S.E. (1979), *NETL: A System for Representing and Using Real-World Knowledge*, Cambridge, Massachusetts, MIT Press.
GARSIDE R., LEECH G. & SAMPSON G., eds. (1987), *The Computational Analysis of English*, London, Longman.
GINZBURG C. (1983), "Clues: Morelli, Freud, and Sherlock Holmes", in Eco & SEBEOK 1983:81-118.
GRIMSON W.E.L. (1990), *Object Recognition by Computer: The Role of Geometric Constraints*, Cambridge, Massachusetts, MIT Press.
Hackbusch W. (1994), *Iterative Solution of Large Sparse Systems of Equations*, Berlin, Springer Verlag.
HAMMOND K.R., ed. (1966), *The Psychology of Egon Brunswik*, New York, Holt, Rhinehart & Winston.
HASTIE T. & TIBSHIRANI R. (1990), *Generalized Additive Models*, London, Chapman & Hall.
HILLIS D. (1985), *The Connection Machine*, Cambridge Massachusetts, MIT Press.
Hume D. (1748), *An Enquiry Concerning Human Understanding*.
ILLICH I. & SANDERS B. (1988), *ABC: The Alphabetization of the Popular Mind*, San Francisco, North Point Press.
JAMES W. (1892), *Psychology: Briefer Course*, New York, Henry Holt.
Jones A. & Churchhouse R.F., eds. (1976), *The Computer in Literary and Linguistic Studies*, Cardiff, University of Wales Press.
JONSEN A. & TOULMIN S. (1988), *The Abuse of Casuistry*, Berkeley, University of California Press.
KOHONEN T. (1980), *Content-Addressable Memories,* Berlin, Springer Verlag.
Lakoff G. (1985), *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*, Chicago, University of Chicago Press.
LEFKOVITCH L.P. (1993), *Optimal Set Covering for Biological Classification*, Ottawa, Agriculture Canada Research Program Services.
LIPMAN D.J. & PEARSON W.R. (1985), "Rapid and sensitive protein similarity searches", *Science* 227:1435-41.
LOWE D.G. (1993), "Similarity metric learning for a variable-kernel classifier", Vancouver, University of British Columbia Computer Science Department.
LYNCH M. & RAWSON S.D. (1976), "Equifrequent character strings: A novel text characterization method", in JONES & CHURCHHOUSE 1976:47-58.
MILL J.S. (1843), *A System of Logic*, London, Longman.
MILLER G.A. (1991), *The Science of Words*, New York, Scientific American Library.
MINSKY M. & PAPERT S. (1969), *Perceptrons: An Introduction to Computational Geometry*, Cambridge, Massachusetts, MIT Press.
NASH F.A. (1954), "Differential diagnosis: An apparatus to assist the logical faculties", *The Lancet*, April 24, 1954:874-875.
PITTS W. & MCCULLOCH W.S. (1947), "How we know universals: The perception of auditory and visual forms", *Mathematical Biophysics* 7:127-147.

POLANYI M. (1966), *The Tacit Dimension*, New York, Doubleday.
PREPARATA F. & SHAMOS M. (1985), *Computational Geometry*, Berlin, Springer Verlag.
RIPLEY B. (1993), "Statistical aspects of neural networks", in BARNDORFF-NIELSEN *et al.* 1993:40-123.
RUMELHART D. *et al.* (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Cambridge, Massachusetts, MIT Press.
RUSSELL S. (1989), *The Use of Knowledge in Analogy and Induction*, London, Pitman.
SEJNOWSKI T.J. & ROSENBERG C.R. (1987), "Parallel networks that learn to pronounce English text", *Complex Systems* 1:145-168.
SHANNON C.E. & MCCARTHY J., eds. (1956), *Automata Studies*, Princeton, Princeton University Press.
SHEWCHUK J.R. (1994), "An introduction to the conjugate gradient method without the agonizing pain", Technical Report CMU-CS-94-125, Pittsburgh, Carnegie-Mellon University Computer Science Department.
Silverman B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London, Chapman & Hall.
SKOUSEN R. (1989), *Analogical Modeling of Language*, Dordrecht, Kluwer.
SKOUSEN R. (1992), *Analogy and Structure*, Dordrecht, Kluwer.
SPARCK JONES K. (1986), *Synonymy and Semantic Classification*, Edinburgh, University of Edinburgh Press.
STANFILL C. & WALTZ D. (1986), "Toward memory-based reasoning", *Communications ACM* 29 (12):1213-28.
STEEG E.S. & ROBINSON D. (1993), "The efficient determination of higher-order features in protein sequence data (extended abstract)", *Proceedings, Workshop on AI and the Genome*, IJCAI-93, Chambery, France.
TEDDINGTON NATIONAL PHYSICAL LABORATORY (1959), *Mechanisation of Thought Processes*, London, H. M. Stationery Office.
THORWALD J. (1965), *The Century of the Detective*, New York, Harcourt, Brace & World.
UTTLEY A.M. (1956), "Conditional probability machines and conditioned reflexes", in SHANNON & MCCARTHY 1956:253-275.
WILLETT P. (1987), *Similarity and Clustering in Chemical Information Systems*, Letchworth, Research Studies Press.
WILLETT P. (1988), *Document Retrieval Systems*, London, Taylor Graham.
WITTEN I., MOFFAT A. & BELL T.C. (1994), *Managing Gigabytes: Compressing and Indexing Documents and Images*, New York, Van Nostrand Reinhold.
WOLPERT D. (1990), "Constructing a generalizer superior to NETtalk via a mathematical theory of generalization", *Neural Networks* 3:445-452.