Collective intelligence: it's all in the numbers

Karen Spärck Jones
Computer Laboratory, University of Cambridge
ksj@cam.ac.uk

AI has been an exporter of ideas to computing in general (neural networks, agents, though robotics is more complex). But AI is now embracing ideas from elsewhere that were initially scorned because they were thought to have nothing to do with modelling intelligence and, especially, human intelligence. These are the statistical and probabilistic approaches to information capture and use that have become particularly prominent in machine learning but have spread all over AI in the last two decades. Pattern recognition was accepted in particular areas, like machine vision, as a kind of technological fix. But statistical and probabilistic approaches are now mainstream.

Thus these approaches are proving remarkably successful as a means of transforming data, certainly into information and, perhaps, knowledge in many complex domains where intelligent action is sought, and thus of supporting robust system behaviour in these domains. In many applications, statistics and probability are used to to transform heterogeneous, multidimensional inputs to serve some dominant, unidimensional system purpose, for example to determine the next move for a mobile robot. However the rapid expansion of the information environment in the electronic world is presenting new challenges for these statistical and probabilistic techniques, even though they are ones preeminently suited to digesting bulk input. Specifically, the challenge is not just of scaling up to absorb more inputs bearing on some particular task, but that of making systems more powerful because they can serve, and choose dynamically between, multiple functions according to local context without higher-level guidance.

For example, the mechanisms underpinning current probabilistic information retrieval systems are designed to deliver texts that the user hopefully will like. They are not currently able, working from the same sources, and deploying the same generic probabilistic capabilities, to choose, autonomously, from a range of tasks. These retrieval systems cannot produce a range of distinct types of output, for instance texts, or extracted data nuggets, or summaries, or even translations, as seems to the system the most appropriate response to the user's current information state.

Information retrieval systems are not currently seen as AI systems. But whatever they cannot do now, we would like them to be intelligent information systems. What we want is a far larger capacity to move between task and task responsively, without depending on the top-level goal definition that is currently required to motivate a robot to choose subtasks like touching, or looking, to get from A to B.

However if specific task purposes are distinct, why should the processes serving them be similar or of the same general sort: what has playing the xylophone, for example, got to do

with reading a political biography (as alternative leisure acitivities), such that statistically-based probabilistic processing can drive both? Current AI systems already serve multiple purposes, but through distinct processes, and why should we think we can do better by homogenising the basic mechanisms they use?

The reasons for thinking this come from developments in natural language processing. Document retrieval was not seen formerly even as natural language processsing, let alone real AI. But the different language information processing tasks - text retrieval, information extraction, summarising, etc have turned out to share a significant need for statistical and probabilistic processing. These tasks are elements in a seamless web where operations worthy of the label 'intelligent' are often needed, and where symbolic processing may be needed, but where, even here, statistics and probability play a vital part. Choosing one sense of a word rather than another because the first is a lot more frequent than the second may be simplistic, but it works. As retrieval, along with speech research, showed, relative frequency, in occurrence or co-occurrence, is a lever that moves many language objects. Thus statistical data and probabilistic computation are now delivering goods for language processing, whether alone or in conjunction with symbolic procedures, in challenging tasks like question-asnwering and summarising.

There is no reason to suppose that other areas for perception and action are fundamentally different; and indeed statistics and probability are on the march everywhere, with the same sort of challenges for the future in developing these numerical technologies to support properly intergrated and therefore properly flexible systems. But the natural language case draws attention to a more interesting challenge for AI in the future.

This is by courtesy of the Web. Much of the stimulus to language processing research, with retrieval in the driving seat, has come from the Web and all the electronic stuff that the Internet and the Web and the ambient electronic world in general proffer. Since natural language figures largely in this stuff, language processing matters in itself. This is the prima facie reason for wanting more powerful, i.e. intelligent, language-based information systems. But the Web suggests that this may not be, indeed cannot be, an intelligent system of (one of) the kind(s) to which we have long aspired. This is because the Web is not a single language world, i.e. world of single language use even for any one language. It has to look like one world at the point where any particular user interacts with it. But that does not mean that the linguistically embodied information (or knowledge) that it contains, and the processes that internally operate on this, are in any sense unified, or coherent, or 'singular' in the sense in which an individual human being is ultimately singular.

That is to say, is there any reason to suppose that the information operations done within such gigantic systems, when they are intelligent (and the argument so far implies that to be effective they have to be intelligent in some way), will be intelligent in the ways that human processes are intelligent? This is not to replay the old arguments about what constitutes intelligence, or whether machines can ever be intelligent, or to talk about enhanced humans with chips in their arms drip feeding Web stuff into them. The issue is the quite different one considered, in particular forms, by anyone concerned with human population behaviour, like economists. We sometime use "social intelligence" to refer to individuals' responses to others, especially multiple others, like themselves. But there is quite another interpretation that applies to humans and, also, to the Web. This is social intelligence as the collective intelligent behaviour of multiple individuals.

Just as history, as a collection of events and states that happen, is one manifestation of social intelligence, so the Web, and future systems, will be manifestations of social intelligence.

Social intelligences are related to individual intelligences, but they are not the same, so there is no reason to suppose there mechanisms (below the global level labelled 'evolutionary') are the same, and that cognitive models for individual human intelligence carry over to the Web and its successors.

This is where numbers come in. Because the Web and its successors will be big they will have to be interpreted and processed using numbers: statistics and probability are the right way to approach them. Humans use numbers (frequencies) which explains why statistically-based language processing delivers them goods they find useful. But scaling up to vast and heterogenous future Webs is going to change the quantitative game so much the qualitative notion of intelligence will have to change. Even though humans cope with astonishing perceptual data volumes, the mechanisms that do this are designed to force the data through a single person filter. The Web is the emergent noisy roar of multiple filters that cannot, for example, be tidied away into an ontology. The challenge for AI for the future will be to figure out how to get 'mega-intelligence', how to do intelligent information processing without being bound by an individual's capabilities. This will not be needed for all applications: but the more stuff there is out there, of one sort or another, and not just language-based stuff, the more getting the benefit of mega-stuff will be an AI system requirement, along with the increasingly taxing requirement to make the result comprehensible to the individual human user.

When I began research on automatic classification a year after the Dartmouth Conference, this was not AI. There was no way I would dare to call it learning, though in learning's current form that is exactly what it was; and I would never refer to the word-usage data to which it was applied as memory, though that (very rudimentarily) was what it was. The forefather of AI, as then conceived, was Leibniz, with his symbolic T-box and A-box. Now there's another figure on the forefathers pedestal, alongside Leibniz, pointing forward: Thomas Bayes.