Artificial intelligence

A better way to search through scientific papers

Get a neural network to do it for you

Science and technology Oct 19th 2017

ARTIFICIAL intelligence (AI) is not just for playing games. It also has important practical uses. One such is in Semantic Scholar, a system developed by researchers at the Allen Institute for Artificial Intelligence, in Seattle, for the purpose of ferreting out the scientific papers most relevant to a particular problem. This week Marie Hagman, the project's leader, and her colleagues have launched an updated version of the system. They have added 26m biomedical-research papers to the 12m previously contained in its database, and upgraded the way that the database's contents can be searched and correlated. Instead of relying on citations in other papers, or the frequency of recurring phrases to rank the relevance of papers, as it once did and rivals such as Google Scholar still do, the new version of Semantic

Scholar applies AI to try to understand the context of those phrases, and thus achieve better results.

Like most AI systems, the new Semantic Scholar relies on a neural network—a computer architecture inspired by the way real neurons connect to each other. Neural networks are able to learn tasks by trial-and-error. Ms Hagman's team wished to bend their network to the task of recognising scientific phrases and their contexts.

Unlike the latest Go-playing program, which works things out from first principles, Semantic Scholar still had to be trained how to perform its task. To do this Ms Hagman asked four medical researchers to annotate ten entire research papers and 67 isolated abstracts, which were to serve as fodder for the training process. The annotators read the papers and abstracts, and highlighted within them a total of about 7,000 medical "topics" (particular diseases, particular genes, particular proteins and so on). Between these topics they identified some 2,000 pairwise relationships, such as a particular gene encoding a particular protein, or being associated with a particular disease.

That done, they fed the results into the neural network, which, based on the context of a topic (ie, the words surrounding it in the various places it appears) and the pairwise relationships identified by the researchers, was able to find new topics and relationships to add to the hoard. The team then improved the network's performance by presenting it with previously unseen papers to annotate, and correcting its suggestions until it was able, without help, to annotate such papers correctly. It can now identify 368,071 topics (mentioned a total of 236,979,862 times) and 6,756,863 relationships in the 38m papers available to it.

The upshot is that both scholars and laymen can pull out clutches of papers on particular topics from the database, with a reasonable presumption that those papers are the ones most pertinent to their needs. Ms Hagman's personal interest in making this happen is that, 15 years ago, she did such a search herself, using the far-more-primitive technology then available. She had stomach ulcers, and was able to bring to her doctors' notice research showing (which is now common knowledge) that such ulcers are caused by infection with a particular bacterium, and are thus treatable using antibiotics. The hope is that an approach which employs AI will not merely help patients and their doctors by flagging up treatments in this way, but also assist in the development of new ones.