Computer equal to or better than humans at indexing science

December 1 2014, by David Tenenbaum

Credit: George Hodan/Public Domain

In 1997, IBM's Deep Blue computer beat chess wizard Gary Kasparov. This year, a computer system developed at the University of Wisconsin-Madison achieved something far more complex. It equaled or bested scientists at the complex task of extracting data from scientific publications and placing it in a database that catalogs the results of tens of thousands of individual studies.

"We demonstrated that the system was no worse than people on all the things we measured, and it was better in some categories," says Christopher Ré, who guided the software development for a project while a UW professor of computer science. "That's extremely exciting!"

The development, described in the current issue of *PLoS ONE*, marks a milestone in the quest to rapidly and precisely summarize, collate and index the vast output of scientists around the globe, says first author Shanan Peters, a professor of geoscience at UW-Madison.

Chess, however complex, is built on rigid rules; in any given situation, only certain moves are legal. The rules for scientific publication are less exact, and so extracting structured information from publications is a challenge for both humans and machines.

Peters and colleagues set up the face-off between PaleoDeepDive, their new machine reading system, and data that scientists had manually entered into the Paleobiology Database. This repository, compiled by hundreds of researchers, is the destination for data from paleontology studies funded by the National Science Foundation and other agencies internationally.

The knowledge produced by paleontologists is fragmented into hundreds of thousands of publications. Yet many research questions require what Peters calls a "synthetic approach: for example, how many species were on the planet at any given time?"

Despite 16 years of effort, the Paleobiology Database remains incomplete and a large amount of hard-earned field data remains locked in publications. Was it possible to automate and accelerate the process?

Teaming up with Ré, who is now at Stanford University, and UW-Madison computer science professor Miron Livny, the group built on the

DeepDive machine reading system and the HTCondor distributed job management system to create PaleoDeepDive. "We were lucky that Miron Livny brought the high throughput computing capabilities of the UW-Madison campus to bear," says Peters. "Getting started required a million hours of computer time."

Much like the people who assembled the Paleobiology Database, PaleoDeepDive inhales documents and extracts structured data, such as species names, time periods, and geographic locations. "We extracted the same data from the same documents and put it into the exact same structure as the human researchers, allowing us to rigorously evaluate the quality of our system, and the humans," Peters says.

Many organizations, including IBM and Google, are trying to extract meaning from natural language, but Ré says, "The thing that is different here is that we decided to pivot and look at the scientific literature, where the language is cleaner."

Instead of trying to divine the single correct meaning from any body of copy, the tactic was to "to look at the entire problem of extraction as a probabilistic problem," says Ré, who credits much of the heavy lifting to UW-Madison Ph.D. candidate Ce Zhang. "People had done pieces of that, but not the entire problem, end to end. This was the DeepDive advance."

Ré imagines a study containing the terms "Tyrannosaurus rex" and "Alberta, Canada." Is Alberta where the fossil was found, or where it is stored? Did the finder work there? Did the study actually focus on a fossil related to T.rex? Computers often have trouble deciphering even simple-sounding statements, Ré says. "We take a more relaxed approach: There is some chance that these two are related in this manner, and some chance they are related in that manner."

In these large-data tasks, PaleoDeepDive has a major advantage, Peters says. "Information that was manually entered into the Paleobiology Database by humans cannot be assessed or enhanced without going back to the library and re-examining original documents. Our machine system, on the other hand, can extend and improve results essentially on the fly as new information is added. It can also extract related information that may not have been in the original database, but that is critical to tackling new science questions, and do so on a huge scale."

Further advantages can result from improvements in the computer tools. "As we get more feedback and data, it will do a better job across the board," Peters says. "There are, potentially, systematic and wholesale improvements to the quality of all of the data."

Jacquelyn Crinion, assistant director of licensing and acquisitions services at the UW-Madison General Library System, says the volume of downloads of scientific papers from publishers threatened logjams in document delivery. "Publishers are not going to complain about usage, but about how hard their system is getting hit." Eventually, Elsevier gave the UW-Madison team broad access to 10,000 downloads per week.

As text- and data-mining takes off, Crinion says the library system and publishers will adapt. "Elsevier is very interested in this project; they see it as the future, and it might allow them to develop new products and ways to deliver service. The challenge for all of us is to provide specialized services for researchers while continuing to meet the core needs of the vast majority of our customers."

The Paleobiology Database has already generated hundreds of studies about the history of life, Peters says. "It's a very good example of the added scientific value provided by synthetic databases, where the whole truly is greater than the sum of its individual data parts."

Peters notes that many fields are being challenged to optimize usage of old findings and make streams of new data readily accessible.

Paleontology and geology are inseparably linked through the role that fossils have played in characterizing geologic sequences, Peters notes. "Ultimately, we hope to have the ability to create a computer system that can do almost immediately what many geologists and paleontologists try to do on a smaller scale over a lifetime: read a bunch of papers, arrange a bunch of facts, and relate them to one another in order to address big questions."

Provided by University of Wisconsin-Madison