ARTIFICIAL INTELLIGENCE

DARPA sets out to automate research

Crash program aims to teach computers to read journals and hatch new ideas

By **Jia You**

The physics Nobel laureate Frank Wilczek has famously predicted that in 100 years, the best physicist will be a machine. Now the U.S. Defense Advanced Research Projects Agency (DARPA) is working toward that vision in a different arena: cancer research. Last summer, the agency launched a $45 million program called Big Mechanism, aimed at developing computer systems that will read research papers, integrate the information into a computer model of cancer mechanisms, and frame new hypotheses for flesh-and-blood scientists (or even other robots) to test—all by the end of 2017.

Last week, 12 teams of computer scientists and biologists met in Washington, D.C., to take stock of progress on the challenge. Although some outside researchers question Big Mechanism's methodology, others applaud it—including artificial intelligence researcher Oren Etzioni of the Allen Institute for Artificial Intelligence in Seattle, Washington, who calls it "an outstanding program."

The program's manager, artificial intelligence researcher Paul Cohen, says its goal is to help scientists cope with complexity at a time when most read more and more narrowly. "Just when we need to understand highly connected systems as systems, our research methods force us to focus on little parts," Cohen says.

Big Mechanism, if it succeeds, could aid researchers studying complicated systems from climate science to military operations and poverty. But for now it focuses on cancer driven by mutations in the Ras gene family, which underlie about a third of all human cancers. Cancer biologists have established a rough road map of Ras-driven cancer pathways: sequences of interactions among proteins affecting cell replication and death. But they amount to what Cohen calls a "hairball" of intertwining causal relations. "We all recognize the need for a better system of organizing this tremendous amount of information, visualizing it, and representing it in a way that's accessible," says Frank McCormick, who directs the Ras initiative at the U.S. National Institutes of Health.

The Big Mechanism program will tackle the problem in three stages. First, machines will read literature on the cancer pathways and convert useful information into formal representations that they can understand. Then, they will integrate the pieces of knowledge into computational models of the cancer pathways. Finally, the system will produce explanations and predictions that can be tested with experiments. The teams are developing four systems capable of all three tasks.

The evaluation meeting focused on the first step, machine reading. Pharmaceutical companies already text-mine papers to glean information on interactions between genes and proteins for drug development, but Big Mechanism seeks to develop machines that read a paper more as scientists do: judging how it contributes to existing knowledge.

The teams worked on different pieces of this "deep reading" challenge. One team, led by computer scientist Ed Hovy of Carnegie Mellon University in Pittsburgh, Pennsylvania, focused on extracting details on experimental procedures and assigning different certainty values to statements such as "we demonstrate" and "we suggest." Another, led by computational linguist James Allen of the Florida Institute for Human and Machine Cognition, built systems for mapping the meaning of sentences and their relationships to one another.

The evaluation started small: Participating teams were given a rudimentary model of Ras cancer pathways and six paragraph-long passages. Their systems had to extract information from the texts, determine how the passages related to the model, and suggest appropriate revisions based on their reading.

Two teams came close to fully automating the process. The best performing machine-reading system extracted 40% of all the relevant information from the passages and correctly determined how each passage related to the model—an excellent start, Cohen says. The systems will face a more comprehensive evaluation in July, he says.

Also coming this summer, Cohen says, is a hackathon in which programmers will build a single reference model of Ras-driven cancer pathways to replace the multiple models the teams are now using. A coherent model, including details about where and how proteins in the Ras pathways interact, is key to enabling the computer to generate hypotheses, Cohen says.

"Just when we need to understand highly connected systems as systems, our research methods force us to focus on little parts."

Paul Cohen, DARPA

Building a system that actually produces scientific insight will not be easy, says computational biologist Larry Hunter of Smart Information Flow Technologies in Minneapolis, Minnesota, a co–principal investigator of one of the teams. The artificial intelligence community doesn't have a strong track record at building systems that can develop useful causal hypotheses, he says. But molecular biology is a good place to try, he says, because it's an area in which common sense plays a minor role; most of the knowledge is technical and available in textbooks and papers.

Other researchers question whether Big Mechanism takes the right approach to studying complex systems. "The Big Mechanism program is trying to map microscopic mechanisms, but complex systems are characterized by collective behaviors," says complex system researcher Yaneer Bar-Yam of the New England Complex Systems Institute in Cambridge, Massachusetts, who was not involved with the project. "The expectation that the accumulation of details will tell us what we want to know is not well justified."

Cohen is confident that the program will pay off, one way or another. "DARPA seeks revolutionary technology," he says. "Sometimes those technologies are turned into practice. Sometimes, they show the world what is possible." ■

PHOTO: DARPA