Seeing how computers 'think' helps humans stump machines and reveals AI weaknesses

August 6 2019

Credit: University of Maryland

The holy grail of artificial intelligence is a machine that truly

understands human language and interprets meaning from complex, nuanced passages. When IBM's Watson computer beat famed "Jeopardy!" champion Ken Jennings in 2011, it seemed as if that milestone had been met. However, anyone who has tried to have a conversation with virtual assistant Siri knows that computers have a long way to go to truly understand human language. To get better at understanding language, computer systems must train using questions that challenge them and reflect the full complexity of human language.

Researchers from the University of Maryland have figured out how to reliably create such questions through a human-computer collaboration, developing a dataset of more than 1,200 questions that, while easy for people to answer, stump the best computer answering systems today. The system that learns to master these questions will have a better understanding of language than any system currently in existence. The work is described in an article published in the 2019 issue of the journal *Transactions of the Association for Computational Linguistics*.

"Most question-answering computer systems don't explain why they answer the way they do, but our work helps us see what computers actually understand," said Jordan Boyd-Graber, associate professor of computer science at UMD and senior author of the paper. "In addition, we have produced a dataset to test on computers that will reveal if a computer language system is actually reading and doing the same sorts of processing that humans are able to do."

Most current work to improve question-answering programs uses either human authors or computers to generate questions. The inherent challenge in these approaches is that when humans write questions, they don't know what specific elements of their question are confusing to the computer. When computers write the questions, they either write formulaic, fill-in-the blank questions or make mistakes, sometimes generating nonsense.

To develop their novel approach of humans and computers working together to generate questions, Boyd-Graber and his team created a computer interface that reveals what a computer is "thinking" as a human writer types a question. The writer can then edit his or her question to exploit the computer's weaknesses.

In the new interface, a human author types a question while the computer's guesses appear in ranked order on the screen, and the words that led the computer to make its guesses are highlighted.

For example, if the author writes "What composer's Variations on a Theme by Haydn was inspired by Karl Ferdinand Pohl?" and the system correctly answers "Johannes Brahms," the interface highlights the words "Ferdinand Pohl" to show that this phrase led it to the answer. Using that information, the author can edit the question to make it more difficult for the computer without altering the question's meaning. In this example, the author replaced the name of the man who inspired Brahms, "Karl Ferdinand Pohl," with a description of his job, "the archivist of the Vienna Musikverein," and the computer was unable to answer correctly. However, expert human quiz game players could still easily answer the edited question correctly.

By working together, humans and computers reliably developed 1,213 computer-stumping questions that the researchers tested during a competition pitting experienced human players—from junior varsity high school trivia teams to "Jeopardy!" champions—against computers. Even the weakest human team defeated the strongest computer system.

"For three or four years, people have been aware that computer question-answering systems are very brittle and can be fooled very easily," said Shi Feng, a UMD computer science graduate student and a co-author of the paper. "But this is the first paper we are aware of that actually uses a machine to help humans break the model itself."

The researchers say these questions will serve not only as a new dataset for computer scientists to better understand where natural language processing fails, but also as a training dataset for developing improved machine learning algorithms. The questions revealed six different language phenomena that consistently stump computers.

These six phenomena fall into two categories. In the first category are linguistic phenomena: paraphrasing (such as saying "leap from a precipice" instead of "jump from a cliff"), distracting language or unexpected contexts (such as a reference to a political figure appearing in a clue about something unrelated to politics). The second category includes reasoning skills: clues that require logic and calculation, mental triangulation of elements in a question, or putting together multiple steps to form a conclusion.

"Humans are able to generalize more and to see deeper connections," Boyd-Graber said. "They don't have the limitless memory of computers, but they still have an advantage in being able to see the forest for the trees. Cataloguing the problems computers have helps us understand the issues we need to address, so that we can actually get computers to begin to see the forest through the trees and answer questions in the way humans do."

There is a long way to go before that happens added Boyd-Graber, who also has co-appointments at the University of Maryland Institute for Advanced Computer Studies (UMIACS) as well as UMD's College of Information Studies and Language Science Center. But this work provides an exciting new tool to help computer scientists achieve that goal.

"This paper is laying out a research agenda for the next several years so that we can actually get computers to answer questions well," he said.

More information: Eric Wallace et al, Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering, *Transactions of the Association for Computational Linguistics* (2019). DOI: 10.1162/tacl_a_00279

Provided by University of Maryland