Artificial-intelligence system surfs web to improve its performance

November 10 2016, by Larry Hardesty

Information extraction — or automatically classifying data items stored as plain text — is a major topic of artificial-intelligence research. Credit: MIT News

Of the vast wealth of information unlocked by the Internet, most is plain text. The data necessary to answer myriad questions—about, say, the

correlations between the industrial use of certain chemicals and incidents of disease, or between patterns of news coverage and voter-poll results—may all be online. But extracting it from plain text and organizing it for quantitative analysis may be prohibitively time consuming.

Information extraction—or automatically classifying data items stored as plain text—is thus a major topic of artificial-intelligence research. Last week, at the Association for Computational Linguistics' Conference on Empirical Methods on Natural Language Processing, researchers from MIT's Computer Science and Artificial Intelligence Laboratory won a best-paper award for a new approach to information extraction that turns conventional machine learning on its head.

Most machine-learning systems work by combing through training examples and looking for patterns that correspond to classifications provided by human annotators. For instance, humans might label parts of speech in a set of texts, and the machine-learning system will try to identify patterns that resolve ambiguities—for instance, when "her" is a direct object and when it's an adjective.

Typically, computer scientists will try to feed their machine-learning systems as much training data as possible. That generally increases the chances that a system will be able to handle difficult problems.

In their new paper, by contrast, the MIT researchers train their system on scanty data—because in the scenario they're investigating, that's usually all that's available. But then they find the limited information an easy problem to solve.

"In information extraction, traditionally, in natural-language processing, you are given an article and you need to do whatever it takes to extract correctly from this article," says Regina Barzilay, the Delta Electronics

Professor of Electrical Engineering and Computer Science and senior author on the new paper. "That's very different from what you or I would do. When you're reading an article that you can't understand, you're going to go on the web and find one that you can understand."

Confidence boost

Essentially, the researchers' new system does the same thing. A machine-learning system will generally assign each of its classifications a confidence score, which is a measure of the statistical likelihood that the classification is correct, given the patterns discerned in the training data. With the researchers' new system, if the confidence score is too low, the system automatically generates a web search query designed to pull up texts likely to contain the data it's trying to extract.

It then attempts to extract the relevant data from one of the new texts and reconciles the results with those of its initial extraction. If the confidence score remains too low, it moves on to the next text pulled up by the search string, and so on.

"The base extractor isn't changing," says Adam Yala, a graduate student in the MIT Department of Electrical Engineering and Computer Science (EECS) and one of the coauthors on the new paper. "You're going to find articles that are easier for that extractor to understand. So you have something that's a very weak extractor, and you just find data that fits it automatically from the web." Joining Yala and Barzilay on the paper is first author Karthik Narasimhan, also a graduate student in EECS.

Remarkably, every decision the system makes is the result of machine learning. The system learns how to generate search queries, gauge the likelihood that a new text is relevant to its extraction task, and determine the best strategy for fusing the results of multiple attempts at extraction.

Just the facts

In experiments, the researchers applied their system to two extraction tasks. One was the collection of data on mass shootings in the U.S., which is an essential resource for any epidemiological study of the effects of gun-control measures. The other was the collection of similar data on instances of food contamination. The system was trained separately for each task.

In the first case—the database of mass shootings—the system was asked to extract the name of the shooter, the location of the shooting, the number of people wounded, and the number of people killed. In the food-contamination case, it extracted food type, type of contaminant, and location. In each case, the system was trained on about 300 documents.

From those documents, it learned clusters of search terms that tended to be associated with the data items it was trying to extract. For instance, the names of mass shooters were correlated with terms like "police," "identified," "arrested," and "charged." During training, for each article the system was asked to analyze, it pulled up, on average, another nine or 10 news articles from the web.

The researchers compared their system's performance to that of several extractors trained using more conventional machine-learning techniques. For every data item extracted in both tasks, the new system outperformed its predecessors, usually by about 10 percent.

More information: Paper: "Improving information extraction by acquiring external evidence with reinforcement learning" arxiv.org/pdf/1603.07954v3.pdf

Provided by Massachusetts Institute of Technology