BY KATHARINE MILLER

BRINGING LIGHT TO DARK DATA:

Using SNORKEL to Label Training Data

Unstructured data—sometimes called dark data—abounds in many domains, including biomedicine. It includes text, such as published scientific literature or physician notes, as well as tables, figures and images. Using computers and advanced machine-learning approaches, researchers are becoming adept at extracting valuable knowledge from dark data. But machine-learning algorithms often require large sets of labeled training data, which in many areas requires the efforts of domain experts. This is a serious bottleneck: "Making training data is so expensive that there are a lot of domains that could never afford to do it," says **Jason Fries, PhD**, a postdoctoral research fellow who is part of the Mobilize Center at Stanford University. Moreover, researchers who want to use machine learning find that creating labeled training data takes up a significant portion of their time.

To address that problem, Fries, **Alex Ratner**, **Steven Bach, PhD**, and others in **Chris Re**'s lab at Stanford are developing an application framework called Snorkel that can automatically generate labeled training data using sources of "weak supervision"—i.e., sources of rules that were not directly intended for the labeling purpose and for which there's no expectation that the labels will be perfect. For many tasks, Snorkel's results are surprisingly good. In recent work extracting mentions of diseases and chemical names from PubMed abstracts, for example, "Snorkel can train a model that performs as well as one trained on human-labeled data," Fries says.

Snorkel starts with a bunch of noisy rules—heuristics—for finding mentions of some concept, such as a disease. In biomedicine, these are often derived from ontologies, but they can come from other sources as well. Snorkel then automatically learns the accuracy of these heuristics as they generate labeled training data, and then uses that accuracy information to de-noise those labels. Under the hood, Snorkel is training what's called a generative model, and can be intuitively understood as having parallels to crowdsourcing algorithms, where the goal is to figure out which people do a better job than others at a particular task, and to take that accuracy into account in de-noising the data. Similarly, in Snorkel, if rules tend to agree with each other and cover a lot of data, Snorkel will trust them more than it will contrarian rules. But Snorkel has a significant advantage over crowdsourcing: "Instead of one person or a few people labeling a small subset, you have labeling functions that can scale to millions of samples," Fries says.

In the labeling of diseases, Snorkel actually has another advantage: It captures some of the inherent disagreement that exists around disease labels. In small human-labeled datasets, gold standard disease definitions are often imperfectly negotiated by a small group of people, Fries says. Snorkel provides a natural mechanism for learning in the presence of disagreement without resorting to manual adjudication.

In addition to extracting mentions of diseases and chemicals from PubMed abstracts, Snorkel can successfully extract relationships from the scientific literature, such as causal relations between genetic mutations and phenotypes.

For biomedicine, where ontologies are prevalent, Snorkel should prove particularly valuable, Fries notes. In their initial experiments, there is only a small gap between the quality of labels generated using Snorkel with weak supervision by ontologies and the quality of ordinary labeled data—and in some areas there is no gap. "That's a nice finding of the work," Fries says. ☐

DETAILS

For more information about Snorkel or to download this open-source application framework, visit http://hazyresearch.github.io/snorkel/

The Mobilize Center for Mobility Data Integration and Insight is an NIH Big Data to Knowledge (BD2K) Center of Excellence at Stanford University.