The Practice of Crowdsourcing: Things to Know about Using Humans and Machines for Labeling

Omar Alonso
Microsoft

Abstract

Many data science applications that use machine learning techniques depend on humans providing the initial data set so algorithms can process the rest or to evaluate the performance of such algorithms. Not only can labeled data for training and evaluation be collected faster, cheaper, and easier than ever before, but we now see the emergence of novel infrastructure that combines computations performed by humans and machines. Building these labeling pipelines remain difficult and these difficulties need to be addressed by practitioners and researchers to advance the state of the art. In this talk, I'll outline things that work in practice and describe a number of trade-offs when designing and implementing computation systems that use humans and machines.

Biography

Omar is a Principal Data Scientist Lead at Microsoft in Silicon Valley where he works on the intersection of social networks, temporal information, knowledge graphs, and human computation. He has shipped many features for Bing and other Microsoft properties. He is the co-chair for the new IR system-oriented conference, called DESIRES.