SCHOOL OF COMPUTER SCIENCE
Special Lecture

Alan Ritter

Ohio State University

Weakly Supervised Information Extraction for the Social Web

Social media has recently challenged traditional news as the best source of information on current events. This user-generated content is highly disorganized; in order to address the problem of information overload, there is a pressing need to develop systems capable of extracting knowledge from massive user-generated text streams. However, the diversity of topics and nonstandard language variations present serious challenges for traditional information extraction technology.

To address these challenges, I will present a series of weakly-supervised approaches to information extraction. Weakly supervised learning has the potential to scale up to heterogeneous corpora, such as Twitter and the Web, by learning from large volumes of unlabeled text. These methods leverage large knowledge bases (KBs) as distant supervision rather than relying on small manually annotated datasets. Most previous work, however, has relied on the closed world assumption: all propositions missing from the KB are considered false. When information is missing, this leads to errors during learning. To relax the closed-world assumption, I will present a latent variable model that jointly reasons about information extraction and missing information. This approach provides a natural way to incorporate side-information from a missing data model, resulting in further performance improvements.

In addition, I will provide several examples of how these methods can be applied to organize the flood of information on social media and enable new applications. These include a system that continuously extracts a calendar of popular events in the near future (http://statuscalendar.com), an approach to extracting structured user profiles from social media content, and a weakly supervised method for detecting cybersecurity events (including denial-of-service attacks, data breaches and account hijacking).

Bio:

Alan Ritter is an assistant professor in Computer Science at Ohio State University. His main research areas are natural language processing, social media and machine learning. He completed his PhD in computer science at the University of Washington in 2013 and was a postdoctoral fellow in the Machine Learning Department at Carnegie Mellon University. He received an NSF CRII Award in 2015.

Monday, April 20
GHC 6115 10 AM

Host: Eduard Hovy

For Appointments: Mary Jo Bensasi