Colloquium

Towards Large Scale Open Domain Natural Language Processing

Gourab Kundu
University of Illinois

Abstract: Machine Learning and Inference methods are becoming ubiquitous a broad range of scientific advances and technologies rely on machine learning techniques. In particular, the big data revolution heavily depends on our ability to use statistical machine learning methods to make sense of the large amounts of data we have. Research in Natural Language Processing has both benefited and contributed to the advancement of machine learning and inference methods. However multiple problems still hinder the broad application of some of these methods. Performance Degradation of machine learning based systems in domains other than the training domain is one of the key problems hindering widespread deployment of these systems.

In this talk, I will present techniques for domain adaptation "on the fly", that allows adaptation to test domains using the same model from training domain. This is accomplished by transforming text from the test domain to look more like the training domain and running the same model from the training domain. This process of text adaptation treats the model as black box, thus makes the adaptation of complex pipelines of models easy and flexible. The next key challenge for machine learning is the processing of vast amounts of data in an efficient manner. Prediction problems for tools are often complicated, for natural language processing and other disciplines, making application of these tools to big data infeasible. The later part of the talk will focus on improving the scalability of machine learning tools with complex prediction component to meet the challenges of big data. I will show how it is possible to amortize the cost of prediction over the lifetime of any machine learning tool. Particularly, I will focus on amortizing integer linear programs which can represent a wide variety of prediction problems. I will present exact and approximate theorems for speeding up the solution time of new integer programs by reusing solutions of previously solved integer programs.

Gourab Kundu is a doctoral candidate in Computer Science Department of University of Illinois at Urbana-Champaign, supervised by Prof. Dan Roth. He has also worked in IBM research and Google for summer internships. He is broadly interested in all aspects of machine learning and natural language processing. He has publications in top tier natural language processing conferences along with a best student paper in CoNLL 2011.

Wednesday, March 5, 2014, 3:00 pm
Mathematics and Science Center: W201

Mathematics and Computer Science
Emory University