Computer and Information Sciences
Spring 2013 Colloquium

Analyzing the World's Languages

Ryan McDonald

Google

Wednesday, May 1, 11am, Wachman 447

Abstract:
Building computer programs that can analyze and distill human language embodies the field of natural language processing (NLP). In the past decade, a number of significant achievements have been made due to the availability of large training resources coupled with scalable machine learning algorithms. This can be seen in the improved quality of search engines as well as successes such as Google Translate, IBM's Watson and Apple's Siri. Despite this success, NLP technologies are still limited in a number of respects. One particularly glaring shortcoming is that most NLP technologies are robust only for English. This is largely due to limited training resources for the long tail of the world's languages. As English comprises only a fraction of the world's speakers, this is problematic. In this talk, I will describe recent advances in building multilingual language technologies. In particular, I will focus on algorithms that learn using weak constraints that can be derived from multilingual knowledge sources that are already in existence today. These constraints can be used as partial supervision in structured learning to construct analyzers for a diverse set of languages. The resulting system significantly pushes the state-of-the-art in multilingual syntactic and semantic analysis. I will conclude with thoughts on the future of such technologies, particularly with respect to their application to downstream technologies like automatic translation and knowledge curation.

Bio:
Dr. Ryan McDonald is a Research Scientist at Google. He received a Ph.D. from the University of Pennsylvania and a Hon.B.Sc. from the University of Toronto. Ryan’s thesis focused on the problem of syntactic dependency parsing. His work allowed complex linguistic constructions to be modeled in a direct and tractable way, which enabled parsers that are both efficient and accurate. In 2008 he wrote a book on the subject entitled 'Dependency Parsing'. Since joining Google, Ryan has continued to work on syntactic analysis, in particular, extending statistical models learned on resource rich languages, like English, to resource poor languages. Ryan’s research also addresses how these systems can be used to improve the quality of a number of important user-facing technologies, such as search, machine translation, and sentiment analysis.