Computer Science Colloquium

Learning With Changing Language Data

Mark Dredze
Johns Hopkins University and Human Language Technology Center of Excellence

Abstract: The information revolution has produced huge quantities of knowledge in the form of natural human language. This explosion of data has pushed natural language processing (NLP) research towards empirical data driven methods, which rely on statistical machine learning. This effort has produced numerous high quality tools for processing language, including knowledge extraction, information organization and automated translation of numerous languages. With more data and better statistical methods, the state of the art advances. Behind the success of this statistical movement is a reliance on statistical methods that are susceptible to changes in data, a particular problem for language data which naturally transitions between topical domains, genres, formats, dialects and languages. High performing systems fail with even subtle changes to language input, like a change in the topic domain.

This talk will survey recent NLP successes in tackling complex natural language problems as well as challenges posed by changes in language data. I will present several approaches to solving domain change challenges that adapt a learned statistical model between one source domain and a new different target domain. I describe Confidence Weighted Learning, a streaming machine learning algorithm designed for the types of data distributions common in language tasks. I show how Confidence Weighted Learning both improves learning in NLP tasks and can be applied to confront the challenges associated with data shifts.

BIO: Mark Dredze is as an Assistant Research Professor in the department of Computer Science and a Senior Research Scientist at the Human Language Technology Center of Excellence at The Johns Hopkins University. His research interests include machine learning, natural language processing and intelligent user interfaces. His focus is on novel applications of machine learning to solve language processing challenges as well as applications of machine learning and natural language processing to support intelligent user interfaces for information management. He earned his PhD from the University of Pennsylvania and has worked at Google, IBM and Microsoft.

Friday, December 4, 2009, 2:00 pm
Mathematics and Science Center: W301

Mathematics and Computer Science
Emory University