Machine Learning for Natural Language Processing, Information Extraction and Text Mining

Dan Roth
Computer Science Department
The University of Illinois at Urbana Champaign

The study of the computational processes underlying comprehension and generation of natural language is an important scientific and engineering task. Instilling machines with abilities that allow them interact to intelligently with humans depends on increasing levels of natural language comprehension (and generation), although shallow levels of understanding can already successfully support challenging applications such as intelligent information extraction, automatic translation, summarization and others.

It is generally accepted today that a statistical machine learning component must have a central role in supporting natural language related tasks; learning processes have multiple roles, from knowledge acquisition to supporting non-brittleness when facing previously unseen situations. A significant amount of work has been devoted in the last few years to developing statistics based learning methods for these tasks, with considerable success.

This short course will introduce some of the central learning frameworks and techniques that have emerged in this field and found applications in several areas in text processing. We will present the main theoretical paradigms used in natural language processing – learning theoretic, probabilistic, and information theoretic – the relations between them, and the main algorithmic techniques developed within these paradigms. Building on a brief theoretical introduction we will introduce key algorithmic techniques for classification and structured prediction in the context of NLP and information extraction tasks. We will also discuss issues such as feature extraction and training paradigms, and address some of the issues involved in using these techniques in real world NLP applications.