Capturing Global Coherence for Cross-source Information Extraction

Heng Ji
University of New York

Abstract: Information Extraction (IE) is a task of identifying facts, such as the attack/arrest events, people’s jobs, people’s whereabouts, merger and acquisition activity from news, patient diagnosis history from discharge summaries and experiment chains from scientific papers. Traditional IE techniques assess the ability to extract information from individual documents in isolation. However, users need to gather information which may be scattered among a variety of sources. These facts may be redundant, complementary, incorrect or ambiguously worded. Furthermore, the extracted information from a document may need to augment an existing Knowledge Base (KB). This requires the ability to link events, entities and associated relations in a document to KB entries and thus present many unique challenges. In this talk, I define several new extensions to state-of-the-art IE and systematically present the foundation, methodologies, algorithms, and implementations needed for more accurate, coherent, complete, concise, and most importantly, dynamic and resilient extraction capabilities.

More specifically, my talk aims to answer the following questions: Cross-document IE: how to ensure global coherence and commonality in topically-related documents to reduce uncertainty? I will present a case study of applying these new inference frameworks to resolve morphed and implicit information in data under active censorship. Cross-lingual IE: how to translate the extracted facts into another language accurately? I will present an information-aware machine translation (MT) approach. Cross-genre IE: how to adapt the methods from one formal genre to the other informal genre (e.g. tweets and discussion forums)? I will present a novel cross-genre propagation framework that combines Natural Language Processing and social cognitive theories. Cross-media IE: How to discover and fuse information from noisy data in multiple data modalities (e.g. text, speech, image and video)? I will present a new structured representation called Multimedia Information Networks.

Thursday, February 21, 2013, 4:00 pm
Mathematics and Science Center: W301