SCHOOL OF COMPUTER SCIENCE
Faculty Candidate

Wei Xu
University of Pennsylvania

Learning and Generating Paraphrases from Twitter and Beyond

Why is it so difficult for computers to understand and generate natural language? The main challenge arises from the fact that human language is both rich and ambiguous. One way to address this challenge is to learn large-scale paraphrases -- language expressions that are worded differently but have (nearly) equivalent meanings -- from massive amounts of naturally occurring data.

First, I will highlight the utility of paraphrases to adapt statistical machine translation techniques for text-to-text generation tasks like text simplification or stylistic rewriting. Second, I will present a novel multi-instance learning model that can capture a broad range of paraphrases from Twitter's data stream, including synonyms, acronyms, misspellings, slang and colloquialisms, such as [has been sacked by | gets the boot from] and [oscar nom'd doc | Oscar-nominated documentary]. By jointly reasoning about relations between words and sentences, the model is able to learn more lexically diverse paraphrases than existing methods. I will hint at how similar models can be used for information extraction to leverage large knowledge bases as training source instead of human labeled data.

Bio:
Wei Xu is a post-doctoral researcher in the Computer and Information Science Department at the University of Pennsylvania. Her research interests are in paraphrases, social media, information extraction and, especially, the intersection of these areas. She has received her PhD in 2014 from New York University, and is a recipient of the MacCracken Fellowship. During her PhD, she visited University of Washington for two years. She is organizing the ACL Workshop on Noisy User-generated Text.

Friday, April 10
GHC 6115 10 AM

Host: Alan Black

For Appointments: Mary Jo Bensasi