



COMPUTATIONAL

LINGUISTICS +

NLP

**lunch
seminar
series**

GHC 6501 • 12:00_{PM}

November 3, 2014

LUNCH WILL BE PROVIDED

Wei Xu

University of Pennsylvania

Modeling Lexically Divergent Paraphrases in Twitter (and Shakespeare!)

Paraphrases are alternative linguistic expressions of the same meaning. Identifying paraphrases is fundamental to many natural language processing tasks and has been extensively studied for the standard contemporary English. In this talk I will present MULTIP (Multi-instance Learning Paraphrase Model), a joint word-sentence alignment model suited to identify paraphrases within the noisy user-generated texts on Twitter. By reducing the dependence on word overlap as evidence of paraphrase, our approach identifies more lexically divergent expressions with equivalent meaning. For experiments, we constructed a Twitter Paraphrase Corpus of about 19,000 sentences using a novel and efficient crowdsourcing methodology. Our new approach improves the state-of-the-art performance of a method that combines a latent space model with a feature-based supervised classifier. I will also present findings on paraphrasing between standard English and Shakespearean styles.

This is joint work with Chris Callison-Burch (UPenn), Bill Dolan (MSR), Alan Ritter (OSU), Yangfeng Ji (GaTech), Colin Cherry (NRC) and Ralph Grishman (NYU).

Wei Xu is a postdoc in Computer and Information Science Department at University of Pennsylvania, working with Chris Callison-Burch. Her research focuses on paraphrases, social media and information extraction. She received her PhD in Computer Science from New York University. During her PhD, she visited University of Washington for two years and interned at Microsoft Research, ETS and Amazon.

