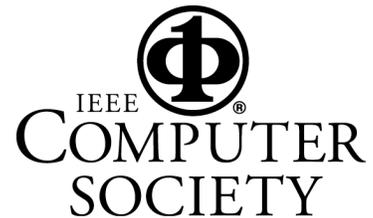


PRINCETON ACM / IEEE-CS CHAPTERS  
MARCH 2012 JOINT MEETING



## Washing Machines, Heavy Water, and the Web: Representation and Processing of Multi-Word Expressions

Multi-Word Expressions (MWEs) are a very difficult problem for natural language processing systems. There are many categories of MWEs, including idioms, phrasal verbs, compound nominals, and named entities. This talk will describe work on developing algorithms to recognize and classify such expressions.

Some MWEs are idiosyncratic interpretations of a group of words: “cat’s meow,” “pain in the neck,” “mixed bag.” Other MWEs can be organized more systematically. The idea is to have categories (semantic types) that are associated with different expressions. This categorization allows us to describe how an “operating system” is like a “washing machine,” and how “heavy water” differs from “hard water.” It is important for many software applications (such as web search engines) to recognize and understand MWEs. Moreover, expressions need to be treated in different ways depending on the application and the semantic type the phrase belongs to.

The talk will discuss an approach to ranking MWEs based on different types of linguistic support. This approach has been combined with a better statistical model of the linguistic phenomena. We will discuss how this compares with other approaches such as mutual information.

One special category of MWEs is named entities. This talk will look at a comparison of the performance of different classifier programs. We found that the agreement rate between two state-of-the-art classifiers ranged from 34% to 58% on the major classes of named entities (Person, Organization, and Location). This is despite reports in the literature of accuracy rates above 90%. We will conclude with suggestions for improving performance and a unit test for assessing accuracy.

**Robert Krovetz** is President of Lexical Research, a company doing research and development in natural language processing. Dr. Krovetz has lead a successful technology transfer project for the NEC Research Institute. He was a Senior Research Scientist at Ask Jeeves, and Principal Natural Language Engineer at CodeRyte. He is currently consulting for the Educational Testing Service. Dr. Krovetz received his PhD from the University of Massachusetts at Amherst.

Date:	Thursday, March 15, 2012, 8:00 pm. (Refreshments and networking at 7:30 pm.)
Place:	Small Auditorium, Room CS 105 Computer Science Building, Princeton University
Information:	Dennis Mancl (908) 582-7086, Jan Buzydlowski (610) 902-8343
On-line info:	<a href="http://PrincetonACM.acm.org">http://PrincetonACM.acm.org</a>

All Princeton ACM / IEEE-CS meetings are open to the public. Students and their parents are welcome. There is no admission charge, and refreshments are served.

A pre-meeting dinner is held at 6:00 p.m. at Ruby Tuesday’s Restaurant on Route 1. Please send email to [princetonacm@acm.org](mailto:princetonacm@acm.org) in advance if you plan to attend the dinner.

### Princeton ACM calendar

Mar. 15

Washing Machines, Heavy Water, and the Web: Representation and Processing of Multi-Word Expressions,” Robert Krovetz, Lexical Research

Apr. 19

P2P Security and Malware, Roberto Rojas-Cessa, NJIT.

May 17

“Strange Sounds in a Familiar World: An Electronic Music Performance,” Joo Won Park, Community College of Philadelphia. (Dinner meeting – at a restaurant to be announced)

All meetings are held at the Computer Science Building at Princeton University at 8:00 pm unless otherwise noted.