UW Center for
Pattern Analysis and Machine Intelligence

Graduate Seminar Series

Identification of Informativeness in Text using Natural Language Stylometry

Speaker: Rushdi Shams, Ph.D.

Researcher, Computational Linguistics Lab, Western University

Date: Wednesday, November 5, 2014

Time: 2:00pm – 3:00 pm

Place: E5-5106

Abstract :

Today, written text is growing rapidly and over-abundantly. This plethora of texts is now widely used to develop and optimize statistical natural language processing (NLP) systems. Surprisingly, the use of more fragments of text to train these statistical NLP systems may not necessarily lead to improved performance. We hypothesize that those fragments that help the most with training are those that contain the desired information. Therefore, determining informativeness in text has become a central issue in our view of NLP. We take a different approach to this problem by considering the underlying theme of a linguistic theory known as the *Code Quantity Principle* that suggests that humans codify information in text by changing their writing style so that readers can retrieve this information more efficiently. In another vein, Stylometry is a modern method to analyze literary style. With this as background, we model text using a set of stylometric attributes to characterize variations in writing style present in it. We explore their effectiveness to determine informativeness in texts of different genres, viz., scientific papers, technical reports, emails and newspaper articles, that are selected from assorted domains like agriculture, physics, and biomedical science. In addition, the potential of stylometric attributes is also explored in some NLP application areas---including biomedical relation mining, automatic keyphrase indexing, spam classification, and text summarization---where performance improvement is both important and challenging. The success of the attributes in all these areas further highlights their usefulness.

WATERLOO
ENGINEERING

CPAMI