UNIVERSITY of GUELPH

COLLEGE of ENGINEERING AND PHYSICAL SCIENCES

SCHOOL OF COMPUTER SCIENCE

MSc Seminar

Thursday September 13, 2018 at 10:00AM in Reynolds, Room 1101

A Comparative Study of Keyword and Keyphrases Extraction from Scientific Articles for Foodborne Diseases

Jingjing Wang

Advisor: Dr. Fei Song

Advisory Committee: Dr. Rozita Dara

ABSTRACT:

Taxonomy is a general principle of scientific classification by sorting thing into categories based on similarities and differences. It is a semantic hierarchy containing the concepts and hierarchical parent-child relations. Taxonomies have many benefits. They describe vital objects in its domain in which objects can be classified and the relationships between them can be easily interpreted. For instance, Natural Language Processing (NLP) applies taxonomic similarity to resolve syntactic and semantic ambiguity. In addition, developers use taxonomies to discover the optimized solution for model transformation of Model-Driven Software Development. Furthermore, taxonomy can be used to semantically integrate and share data among various disjointed systems (i.e. semantic interoperability).

Many taxonomies have been created for biological concepts or disease. However, research on generating a taxonomy in domain of foodborne disease outbreak were poorly discovered. In the USA, there are approximately 250 - 350 million people suffering from acute gastroenteritis annually in which 1/4 to 1/3 of them were caused by foodborne diseases. Health Canada demonstrated that raw or undercooked ingredient result in certain bacteria and parasites are responsible for such foodborne outbreaks, such as E.coli, Salmonella etc. As a matter of fact, the direct cause of the outbreaks can be multifarious. Taxonomy, in this case, can help the public and researchers better understand the primary culprits behind the outbreaks. They can also be used to integrate concepts and data in disjointed databases and systems.

Taxonomies can be built manually by the subject matter experts. Another approach that has gained popularity in the past decade is the use of machine learning and statistical methods. The objective of this study is focused on keywords and keyphrase extraction, which is a foundational process in creating a taxonomy.

This seminar will explore the motivations and obstacles in terms of creating a taxonomy and extracting keywords from the research articles on foodborne disease outbreak. We utilized different supervised methods such as Term Frequency*Inverse Document Frequency (TF*IDF), Latent Dirichlet Allocation (LDA) and unsupervised approaches such as Keyphrase Extraction Algorithm KEA, Convolutional Neural Networks (CNN) to extract keywords and compare the results.