Text Summarization and Categorization for Scientific and Health-Related Data

Arman Cohan
Department of Computer Science
Georgetown University
Washington DC
armancohan@gmail.com

Abstract

The increasing amount of unstructured health-related data has created a need for intelligent processing, summarizing, and categorizing these data to extract knowledge from them. My research goal in this dissertation is to develop Natural Language Processing (NLP) and Information Retrieval (IR) methods for better processing and understanding health-related textual information to promote health care and wellbeing of individuals.

In this dissertation I first focus on scientific literature as an important source of knowledge distribution in health care. It has become a challenge for researchers to keep up with the increasing rate at which scientific findings are published. To address this problem, I propose summarization methods using citation texts and discourse structure of the papers to provide a concise representation of important contributions of the papers. I also investigate methods to address the problem of citation inaccuracy by linking the citations to their related parts in the target paper, capturing their relevant context. In addition, I raise the problem of the inadequacy of current evaluation metrics for scientific document summarization and present a superior method based on semantic relevance in evaluating the summaries.

In the second part, I focus on other significant sources of health-related information including clinical notes and social media. I investigate categorization methods to address the critical problem of medical errors which are among leading causes of death worldwide. I demonstrate how we can effectively identify significant errors and harmful cases through medical narratives that could help prevent similar future problems. Mental health is another significant dimension of health and wellbeing that is sometimes overlooked. Suicide, the most serious challenge in mental health, accounts for approximately 1.4% of all deaths and approximately one person dies by suicide every 40 seconds. I investigate social media as a platform through which mental problems such as depression and self-harm can be investigated. I present both feature-rich and neural network methods for assessing the risk of depression, self-harm, and suicide to the individuals based on their general language expressed in social media.

PhD Advisor: Nazli Goharian
Committee members: Nazli Goharian, Ophir Frieder, Jimmy Lin, Calvin Newport, Nathan Schneider, Elad Yom-Tov
The dissertation is available at: `https://bit.ly/2yOfuiM`