Dissertation Defense
Doctor of Philosophy in Information Science

“KEYPHRASIFICATION: SUMMARIZING TEXT INTO KEYPHRASES - USING NEURAL LANGUAGE GENERATION METHODS” by Rui Meng

Date: April 2, 2024
Time: 12:30 – 3:00 p.m.
Place: Room 502/Information Sciences Building, 135 North Bellefield Avenue, Pittsburgh PA 15260
Zoom Link: https://pitt.co1.qualtrics.com/jfe/form/SV_5mtRfPbyZ6dC4K2

Committee:
- Daqing He, Professor, School of Computing and Information, University of Pittsburgh
- Peter Brusilovsky, Professor, School of Computing and Information, University of Pittsburgh
- Paul Munro, Associate Professor, School of Computing and Information, University of Pittsburgh
- Cornelia Caragea, Professor, Department of Computer Science, University of Illinois at Chicago

Abstract:
Keyphrases encapsulate the core information of a text, acting as effective tools for organizing and retrieving extensive data. Their utility spans various applications, including information retrieval, document classification, and automatic summarization. Given the cost and limitations of manual keyphrase assignment, there has been a growing interest in automating this process. Traditional approaches to keyphrase assignment are categorized into extraction, which involves selecting phrases directly from the text, and tagging, where pre-defined tags are applied. Both methods often fail to address the complexity of natural language. For instance, a substantial fraction of keyphrases (8.9% to 21.5%) are absent from the source text and are missed by extraction methods. This observation highlights the need to reevaluate the paradigms within keyphrase studies and refine methodologies in automatic keyphrase prediction.

This dissertation introduces keyphrasification to formulate the task of keyphrase prediction. By developing a conceptual framework and defining essential properties, this work aims to deepen the understanding of keyphrase prediction and facilitate the development of more effective techniques. Furthermore, I propose a novel modeling approach, keyphrase generation, utilizing neural language generation to learn the mapping between texts and keyphrases directly from data to predict contextually relevant phrases in varied forms. The dissertation further presents various enhancements and mechanisms to refine this approach.

This work makes several pivotal contributions. It reformulates keyphrase prediction as a specialized form of summarization, thereby broadening the previous research scope. It innovates in automatic keyphrasification with a data-driven approach, employing neural networks to predict context-relevant phrases, overcoming the limitations of prior methodologies. Furthermore, the study explores a range of advanced language generation techniques, from basic to pre-trained
and large language models, making it a comprehensive investigation into the task of keyphrasification.