Dissertation Defense

Doctor of Philosophy in Computer Science

“Techniques to Enhance Abstractive Summarization Model Training for Low Resource Domains” by Ahmed Magooda

Date: March 4, 2022
Time: 2:00PM – 4:00PM
Place: https://pitt.co1.qualtrics.com/jfe/form/SV_8HOly4np5bL7kJk

Committee:

• Diane Litman, Professor, Department of Computer Science, School of Computing and Information
• Adriana Kovashka, Assistant Professor, Department of Computer Science, School of Computing and Information
• Milos Hauskrecht, Professor, Department of Computer Science, School of Computing and Information
• Daqing He, Professor, Department of Informatics and Networked Systems, School of Computing and Information

Abstract: Nowadays, the amount of information accessible is growing exponentially, and it is becoming challenging to digest even the information available for a particular topic. Summarization can reduce the information into a handful of paragraphs, helping human readers digest much information in a few lines. Automatic summarization spans different summarization techniques (Abstractive, extractive, phrase extractive, etc.) depending on the granularity of text used and how summarization is carried out. Abstractive summarization specially aims to mimic how humans summarize, as it aims to summarize a large amount of text into a readable, comprehensive summary. However, the majority of prior studies focus on data-rich domains. The availability of large datasets allows researchers to develop, and train complicated and robust models efficiently. On the other hand, few studies focus on developing/improving models for scarce data domains. A typical practical issue that is rendered in such domains is model overfitting. As a step towards remedying these shortcomings, this thesis aims to enhance abstractive summarization models in low-resource settings by tackling three challenges.

1- Can we adapt widely used data augmentation/synthesis techniques to abstractive summarization to remedy the scarceness issue?
2- How can we benefit from domain transfer or pretraining, and what can be a helpful strategy to do it more efficiently?
3- Can we extract additional information from the data and thus use it more effectively?

The thesis tackles these challenges in two directions: (1-Manipulating data usage, and 2-Introducing changes to the models themselves).

This thesis first proposes new paraphrasing and template-based data synthesis (augmentation) models, novel techniques to synthesize new data for model training. We then introduced a variant of a recent data augmentation technique to be used in generative tasks. Additionally, we explored the utility of using curriculum learning to both improve pretraining and fine tuning. Finally, to overcome the third challenge, we propose integrating the summarization model into a multitask learning setting and using the same data multiple times to train the different tasks. We finally combine multiple techniques (i.e., multitask learning and data augmentation) to observe if the combination would be more helpful than each approach in isolation.