Yale University
Computer Science Talk

Beyond Sentences and Paragraphs: Towards Document and Multi-document Understanding

Host: Dragomir Radev

Arman Cohan

Monday – February 28, 2022
4:00 p.m.

Zoom Presentation

Abstract: During the past few years, there has been significant progress in natural language understanding, primarily due to the advancements in transfer learning methods and the increasing scale of pre-trained language models. However, the majority of progress has been made on tasks concerning short texts with sentences or paragraphs as the basic unit of analysis. Yet, many real-world natural language tasks require understanding full documents which includes learning effective representation of documents, resolving longer range dependencies, structure, and argumentation. Further, certain tasks require incorporating additional context from multiple related documents (e.g., understanding a scientific paper) and aggregating information across multiple documents. In this talk, I will discuss some of our recent works on addressing these challenges. I will first discuss general methods for document representation learning that help to achieve strong downstream performance on a variety of document-level tasks. Then I will focus on how we can have a general pre-trained language model that can process long documents. Using this framework, I will discuss extensions to multi-document natural language understanding for a variety of classification, extraction, and summarization tasks. I will also briefly discuss a few of our newly developed benchmarks from challenging domains that enable us to better measure progress on document natural language understanding.

Bio: Arman Cohan is a Research Scientist at the Allen Institute for AI (AI2) and an Affiliate Assistant Professor at the University of Washington. His broad research interest is developing natural language processing (NLP) methods for addressing information overload. This includes models and benchmarks for document and multi-document understanding, natural language generation and summarization, as well as information discovery and filtering. He is additionally interested in real-world interdisciplinary applications of NLP in the science and health domains. His research has been recognized with multiple awards, including a best paper award at EMNLP 2017, an honorable mention at COLING 2018, and the 2019 Harold N. Glassman Distinguished Doctoral Dissertation award.