NJIT | INSTITUTE FOR DATA SCIENCE

Department of Data Science

Data Science Seminar Series

Towards More Intelligent Extraction of Information from Documents

Xinya Du, Ph.D.

Postdoctoral Researcher
University of Illinois Urbana-Champaign

Date:	Monday, March 7th, 2022
Time:	11:00 AM – 12:00 PM EST
Location:	Zoom Virtual Room
Web Link:	Zoom Meeting Room Link

Large amounts of text are written and published daily. As a result, applications such as reading through the documents to automatically extract useful and structured information from the text have become increasingly needed for people's efficient absorption of information. They are essential for applications such as answering user questions, information retrieval, and knowledge base population.

In this talk, I will focus on the challenges of finding and organizing information about events and introduce my research on leveraging knowledge and reasoning for document-level information extraction. In the first part, I'll introduce methods for better modeling the knowledge from context: (1) generative learning of output structures that better model the dependency between extracted events to enable more coherent extraction of information (i.e., event A happening in the earlier part of the document is usually correlated with event B in the later part). (2) How to utilize information retrieval to enable memory-based learning with even longer context.

In the second part, to better access relevant external knowledge encoded in large models for reducing the cost of human annotations, we propose a new question-answering formulation for the extraction problem. I will conclude by outlining a research agenda for building the next generation of efficient and intelligent machine reading systems with close to human-level reasoning capabilities.

Xinya Du is a Postdoctoral Research Associate at the University of Illinois at Urbana-Champaign working with Prof. Heng Ji. He earned a Ph.D. degree in Computer Science from Cornell University, advised by Prof. Claire Cardie. Before Cornell, he received a bachelor's degree in Computer Science from Shanghai Jiao Tong University. His research is on natural language processing, especially methods that leverage knowledge & reasoning skills for document-level information extraction. His work has been published in leading NLP conferences such as ACL, EMNLP, NAACL and has been covered by major media like New Scientist. He has received awards including the CDAC Spotlight Rising Star award and SJTU National Scholarship.