Abstract: Neural networks are becoming increasingly ubiquitous in the field of Natural Language Processing (NLP). For example, neural encoder-decoder models have achieved notable empirical success in machine translation. However, Neural networks still lack interpretability compared to other models such as SVM. Consequently, this poses a problem for developing new neural models for applications. For example, researchers can benefit from the understanding of the architecture and the underlying computational process to extend and improve these models. Moreover, even non-expert users often require justification for the model's prediction in many application scenarios, for example, movie recommendation. In this survey, the objective is to review the main techniques in interpretable NLP literature. One primary research direction is understanding what linguistic information neural models capture. Additionally, this survey attempts to explore several directions for future research.

Committee:

- Professor Jia Xu, mentor, Hunter College
- Professor Saad Mneimneh, Hunter College
- Professor Zhigang Zhu, The City College of New York