Nancy Ide, New York

Data and Annotation: The Impact of Big Data on Discourse Annotation (Abstract)

When large amounts of language data first became available thirty years ago, annotation of that data for linguistic phenomena of all kinds became a major activity in the field of Natural Language Processing (NLP) in order to enable development of statistical models.
Later, the rise of machine learning, coupled with the availability of language data of orders of magnitude greater size, further fueled the need to annotate linguistic phenomena in order to train robust language models. The impacts on the annotation of discourse-related phenomena over this period fall along two primary dimensions: the development of automatic means to annotate discourse structure, and the attempt to define appropriate and usable schemes for annotating discourse phenomena of all kinds, both manually and automatically.

This presentation will survey strategies for discourse annotation over the past thirty years and provide an assessment of the considerations for the use and further development of discourse annotation schemes and strategies in the light of historical factors.