

Natural Language Learning using Both Labeled and Unlabeled Data

Hang LI, *Microsoft Research Asia, PRC*, E-mail: hangli@microsoft.com

Abstract: Recently, a new trend has arisen in the field of Natural Language Processing (NLP): the development of machine learning technologies that use both labeled and unlabeled data for training. For many NLP tasks, existing data are by their nature unlabeled and manually labeling them is prohibitively expensive. Effective utilization of both unlabeled and labeled data in learning is a challenging but important issue. Methods that have been proposed under this paradigm include co-training, EM learning, and transductive learning. In this talk, I will first make a brief survey on co-training. After that, I will introduce our work on bootstrapping including bilingual bootstrapping and collaborative bootstrapping.