Unsupervised NLP aims to discover meaningful structure in unannotated text, such as parts-of-speech, morphological segmentation, or syntactic structure. Unsupervised systems improve when researchers incorporate knowledge to bias learning to better capture characteristics of the desired structure. Contrastive estimation (CE; Smith and Eisner, 2005) is a general approach to unsupervised learning with a particular way of incorporating knowledge. CE increases the likelihood of the observations at the expense of those in a particular neighborhood of each observation. The neighborhood typically contains corrupted versions of the observations.

In this talk, we generalize CE in two ways that allow us to add more knowledge to unsupervised learning. In particular, we augment CE with two types of cost functions, one on observations and one on output structures. The first allows the modeler to specify not only the set of corrupted inputs for each observation, but also how bad each one is. The second lets us specify preferences on desired output structures, regardless of the input sentence. We evaluate our approach, which we call cost-augmented contrastive estimation (CCE), on unsupervised part-of-speech tagging of five languages from the PASCAL 2012 shared task. We find that CCE improves over both standard CE and strong benchmarks from the shared task.

This is joint work with Mohit Bansal.