Jason Baldridge

Cutting Corpus Costs: Machine Learning and Annotation

Creating linguistically informed annotations on top of primary speech or textual data is a core task in language documentation. However, it is usually prohibitively expensive to obtain them for large quantities of material. In recent years, the field of computational linguistics has utilized many robust and highly accurate machine learning paradigms for producing systems which can automatically analyze linguistic input for a variety of tasks. There is great promise in the possibility of using such systems to speed up the annotation process.

The most straightforward scenario is semi-automation: a system is used to automatically annotate a corpus and then its output is manually corrected by a human expert. A flip-side to this scenario is active learning. With active learning, a machine learner is used to identify some of the hardest decisions and only these are given to a human expert to label. The choices made by the human can then be given back to the machine learner so that it improves its own accuracy, and as a result, its utility for semi-automated annotation. Active learning thus seeks to have the human consider only the cases that will be maximally informative for the machine learner, rather than requesting them to also review all the simple and redundant ones.

In this talk, I will give an overview of different annotation scenarios and highlight the promise and limitations of active learning and semi-automated annotation for reducing the cost of creating annotated corpora.