Raymond Mooney

Maximizing the Utility of Small Training Sets in Machine Learning

Statistical natural language processing methods typically require large annotated corpora for training. Assembling such large annotated corpora for less-studied languages is very difficult. A variety of machine learning techniques have been developed for improving the accuracy of models learned from small training sets. We review four such general approaches: 1) Ensemble methods, which construct and combine multiple, diverse hypotheses; 2) Active learning methods which select the most informative training examples for annotation; 3) Transfer learning methods that exploit previously learned knowledge for related problems; and 4) Semi-supervised learning methods that use a combination of labeled and unlabeled data. We present experimental results on a variety of problems demonstrating the ability of these methods to improve predictive accuracy when training data is limited.