Constrained Semi-supervised Learning in the Presence of Unanticipated Classes

Bhavana Bharat Dalvi
Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213
bbd@cs.cmu.edu

4th October 2015

Abstract

Traditional semi-supervised learning (SSL) techniques consider the missing labels of unlabeled datapoints as latent/unobserved variables, and model these variables, and the parameters of the model, using techniques like Expectation Maximization (EM). Such semi-supervised learning techniques are widely used for Automatic Knowledge Base Construction (AKBC) tasks.

We consider two extensions to traditional SSL methods which make it more suitable for a variety of AKBC tasks. First, we consider jointly assigning multiple labels to each instance, with a flexible scheme for encoding constraints between assigned labels: this makes it possible, for instance, to assign labels at multiple levels from a hierarchy. Second, we account for another type of latent variable, in the form of unobserved *classes*. In open-domain web-scale information extraction problems, it is an unrealistic assumption that the class ontology or topic hierarchy we are using is complete. Our proposed framework combines structural search for the best class hierarchy with SSL, reducing the semantic drift associated with erroneously grouping unanticipated classes with expected classes. Together, these extensions allow a single framework to handle a large number of knowledge extraction tasks, including macro-reading, noun-phrase classification, word sense disambiguation, alignment of KBs to wikipedia or on-line glossaries, and ontology extension.

To summarize, this thesis argues that many AKBC tasks which have previously been addressed separately can be viewed as instances of single abstract problem: multiview semi-supervised learning with an incomplete class hierarchy. In this thesis we present a generic EM framework for solving this abstract task.