Rich Prior Knowledge in Learning for Reducing Annotation Cost

21 May 2012

Presenters:

João Graça, Gregory Druck, Kuzman Ganchev

Tutorial Programme/Overview

The cost of creating language resources can be reduced by correcting the errors of or directly using an automatically labeled corpus, rather than annotating the corpus from scratch. These automatic labels could come from either a rule-based or trained system. This tutorial describes how to quickly train effective systems for this purpose by leveraging prior knowledge and unannotated data. For example, we could use our linguistic knowledge of likely syntactic dependencies, unannotated sentences, and the described methods to quickly train an initial parser without needing to provide any complete parses.

Specifically, we survey four general frameworks for expressing and learning with prior knowledge about output variables. These frameworks are Constraint-Driven Learning (UIUC), Posterior Regularization (UPenn), Generalized Expectation Criteria (UMass Amherst), and Learning from Measurements (UC Berkley). We explain how the frameworks are connected and discuss the trade-offs between them. We also survey applications that have been explored in the literature, including applications to grammar and part-of-speech induction, word alignment, information extraction, text classification, and multi-view learning. Prior knowledge used in these applications ranges from structural information that cannot be efficiently encoded in the model, to labeled features, to knowledge of some incomplete and noisy labellings. These applications also address several different problem settings, including unsupervised, lightly supervised, and semi-supervised, and utilize both generative and discriminative models. The diversity of tasks, types of prior knowledge, and problem settings explored demonstrate the generality of these approaches, and suggest that they will become an important tool for researchers in natural language processing. Additionally, we review work in active solicitation of prior knowledge, for example active learning by labeling features instead of complete examples.

The tutorial will provide the audience with the theoretical background to understand why these methods have been so effective, as well as practical guidance on how to apply them. Specifically, we discuss issues that come up in implementation, and describe a toolkit that provides "out-of-the-box" support for the applications described in the tutorial, and is extensible to other applications and new types of prior knowledge.

Tutorial Description/Outline/Contents

We possess a wealth of prior knowledge about most prediction problems, and particularly so for many of the fundamental tasks in natural language processing. Clearly we should be able to use such information to reduce annotation cost. Though we could use prior knowledge to build rule-based systems, machine learning allows generalization. Unfortunately, it is often difficult to make use of prior knowledge during learning, as it typically does not come in the form of labeled examples, may be difficult to encode as a prior on parameters in a Bayesian setting, and may be impossible to incorporate into a tractable model. Instead, we usually have prior knowledge about the values of output variables. For example, linguistic knowledge or an out-of-domain parser may provide the locations of likely syntactic dependencies for grammar induction. Motivated by the prospect of being able to naturally leverage such knowledge, four different groups have recently developed similar, general frameworks for expressing and learning with side information about output variables. The tutorial surveys this work and gives the attendee some practical tips on how to use it.

Introduction:
- Introduction to different types of prior knowledge about NLP problems
- Limitations of previous methods for incorporating prior knowledge, including Bayesian and heuristic approaches
- Motivation for constraining the output variables directly

Examples and Demos

Recent Frameworks for Learning with Prior Knowledge:
- Brief theoretical overview of and discussion of connections between:
  - Learning from Measurements (University of California, Berkeley)
  - Generalized Expectation (University of Massachusetts, Amherst)
  - Posterior Regularization (University of Pennsylvania)
  - Constraint Driven Learning (University of Illinois, Urbana-Champaign)

Coffee Break

Applications:
- Unstructured problems:
  - Document Classification: labeled features, multi-view learning
- Sequence problems:
  - Information Extraction: labeled features, multi-view learning, long-range dependencies
  - Word Alignment: bijectivity, symmetry
  - POS Tagging: posterior sparsity
- Tree problems:
  - Dependency Parsing: linguistic knowledge, noisy labels, posterior sparsity
- Active / interactive training: active feature labeling for sequence problems, interactive training

Coffee Break

Implementation:
- Guidance on implementation
- Description and walk-through of existing software packages

Closing Remarks and Discussion