

T4: Adaptive Learning: From Supervised to Active Learning of Statistical Models for Natural Language and Speech Processing

Giuseppe RICCARDI

AT&T Labs-Research
180 Park Ave
Florham Park
NJ 07932-0971
USA
dsp3@research.att.com

Dilek HAKKANI-TUR

AT&T Labs-Research
180 Park Ave
Florham Park
NJ 07932-0971
USA
dtur@research.att.com

Gokhan TUR

AT&T Labs-Research
180 Park Ave
Florham Park
NJ 07932-0971
USA
gtur@research.att.com

Abstract

In the nineties there has been a large body of research work on data-driven algorithms that have been successfully applied to Natural Language and Speech Processing (NLSP). The fundamental assumption of most of these algorithms is that statistical models are trained and tested by drawing random samples from corpora, assuming that samples are i.i.d (independent and identically distributed). This leads to models that are by design suited for stationary channels. The stationarity assumption is not acceptable in the case of adaptive systems or when training from massive and heterogeneous amount of data.

The basic algorithms for supervised training of statistical models (e.g. stochastic parsers or stochastic language models) have been proposed and refined over the last decade. These modeling techniques belong to the class of passive supervised learning algorithms. Passive supervised approaches are very effective, while they require a large amount of data. However, passive learning lacks the ability to track non-stationarities in speech and language. Moreover, when a large amount of data is available it is not efficient to “blindly” sample and annotate data to train model statistics and parameters. Adaptive systems require predictions (e.g. word prediction) to be made on-line and “active” selection of the data to train statistical models.

In order to address the on-line learning nature of adaptive systems “active” learning and unsupervised” algorithms are borrowed from the machine learning literature. Active learning specifically makes efficient use of data by doing “selective sampling”. Unsupervised learning allows for exploiting massive amounts of data that has not been annotated.

In this tutorial we will review most approaches to training stationary and adaptive machines. We will cover the main statistical algorithms to train data driven model for NLSP. We will provide the theory and motivation for the different learning approaches. Finally we will report on the application of these techniques in the context of NLSP (e.g. parsing and stochastic language modeling).