Noname manuscript No.
(will be inserted by the editor)

Anders Søgaard: Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

Steven Abney

Received: date / Accepted: date

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing provides a survey of the state of the art, and specific practical information, for anyone confronting a mismatch between data available for training and data encountered at runtime. Three general approaches are discussed: *semi-supervised learning* methods that integrate labeled training data and unlabeled runtime data, *transfer learning* methods that attempt to correct sampling biases by comparing training data to runtime data, and *adversarial learning* and *meta-learning* methods that build in robustness to unknown bias when samples of runtime data are unavailable.

Applications that come under discussion include document classification, part of speech tagging, and dependency parsing. There is, unfortunately, no discussion of machine translation, but the techniques that Søgaard describes can certainly be applied to translation. Machine translation is no stranger to the divergence between training and runtime domains, or to the difficulty of obtaining labeled data and relative ease of obtaining unlabeled data. Domain adaptation in particular has received much attention in the recent machine translation literature.

The first half of the book consists of a brief introduction (Chapter 1) followed by a review of supervised and unsupervised learning and their applications in natural language processing (Chapter 2), focusing on classification. There are extended discussions of the nearest-neighbors classifier, naive Bayes, and the perceptron, including several variations on the perceptron, such as the averaged perceptron and passive-aggressive learning.

The treatment is not merely introductory, but focuses on quantification and evaluation of three learning assumptions that bear on semisupervised learning and domain adaptation. The first assumption is that test and training data are drawn from the same distribution; it is quantified by Kullback-Leibler

S. Abney
University of Michigan, Ann Arbor, Michigan, USA
E-mail: abney@umich.edu

divergence between training and test. The second assumption is that classes are compact, quantified as the mean squared distance of class members from the centroid. The third assumption is that classes are well separated, quantified as the mean squared distance between class centroids and the centroid for all the data. Classifier performance is evaluated in a way that specifically examines the sensitivity of different algorithms to violations of these assumptions. Naive Bayes appears particularly sensitive to divergence between training and test distributions, and to violation of class compactness, whereas the perceptron seems least sensitive. Later in the book, semi-supervised learning and domain adaptation algorithms are motivated in part as a means of addressing these learning assumption violations.

There is also a brief discussion of unsupervised learning, covering hierarchical clustering, k-means clustering, and the Expectation Maximization (EM) algorithm.

The second half of the book addresses the title topics, beginning in Chapter 3 with semi-supervised learning. The algorithms discussed include self-training, co-training, tri-training, the EM algorithm, which is introduced as a soft version of self-training, and three variations on the nearest-neighbors algorithm: label propagation, nearest-neighbor editing, and condensed nearest neighbors.

The final topic is domain adaptation, which is to say, learning under bias. When the bias is known (that is, when runtime data is available) transfer learning is applicable (Chapter 4). The basic idea is to use runtime data to detect and delete outliers in the training data, allowing one to select features, parameters, or instances based on their domain-independence. A soft generalization is to adopt a weighting function that makes the weighted training distribution more similar to the runtime distribution.

Chapter 5 presents adversarial learning and meta-learning as methods for increasing robustness to unknown biases, which arise when the learning algorithm is not provided with runtime data, but runtime data is expected to differ from training data. This is the typical case when developing any natural-language processing system: it is impossible to anticipate all domains in which users will employ the system. Because of target-domain differences, some features that are heavily weighted in training may be absent in the test data. To simulate the effect, one can delete random features at training time, forcing the learner to find redundant predictive features. This is the idea underlying adversarial learning. There is also a brief discussion of stacking and metalearning.

The book closes (Chapter 6) with a discussion of evaluation. Søgaard makes a case for using meta-analysis instead of simple performance scores, and presents an experiment showing that meta-analysis makes better predictions of performance on unseen data sets.

There have been two previous book-length treatments of semi-supervised learning [1,3], and a collection of papers on the topic [2]. Søgaard covers new material (domain adaptation in particular is not covered in the earlier books) and also provides a wealth of pointers into the recent literature in a fast-moving

field. But what really sets Søgaard's book apart is the amount of practical information it provides for conducting experiments. Python implementations are given, in the form of printed listings within the text, for almost all of the algorithms discussed, as well as recipes for using the implementations in the scikit-learn library. A number of experiments are presented, including performance evaluations for many of the learning methods, along with an unusual amount of detail for replicating the experiments, such as URLs of datasets and, in many cases, Python code.

This wealth of information is packed into a slender volume, comprising only 80 pages, excluding front material and bibliography. As an unavoidable result, the text is extremely terse. For example, KL divergence, Jensen-Shannon divergence, variance, and covariance matrices, are accorded one sentence each. For this reason, the book is not appropriate as an introduction for real beginners. But it has unique value as a condensation of key points for an advanced student who would like to get started on serious research in the area, or for an established researcher who would like to catch up with the current state of the art in semi-supervised learning and domain adaptation.

References

1. Steven Abney, *Semisupervised Learning for Computational Linguistics.* Chapman & Hall/CRC (2008)
2. Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (eds.), *Semi-Supervised Learning.* The MIT Press (2006)
3. Xiaojin Zhu and Andrew B. Goldberg, *Introduction to Semi-Supervised Learning.* Morgan & Claypool Publishers (2009)