Book Reviews

Semi-Supervised Learning and Domain Adaptation in Natural Language Processing

Anders Søgaard
University of Copenhagen

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 21), 2013, x+93 pp; paperbound, ISBN 978-1-60845-985-8, $40.00; e-book, ISBN 978-1-60845-986-5, $30.00 or by subscription

Reviewed by
George Foster
National Research Council Canada

Classical machine learning makes at least two assumptions that are at odds with its application to natural language. First, it implicitly assumes there are enough data. This is rarely the case in NLP, where sparse data is the norm, especially for intermediate tasks like parsing that require artificial labeling. Second, it assumes that all examples are drawn from the same distribution. Language is of course not like this: rather than being an ordered landscape, it is a wildly varying one, rich with strange growths and prone to sudden monstrous blooms like micro-blogging.

Both these traits can cause a classically trained NLP system to behave poorly. To cope with data sparsity, a common strategy is semi-supervised learning, in which a small labeled data set is augmented by a larger amount of (typically more abundant) unlabeled data. To cope with domain differences between training and test data, adaptation techniques can be used to mitigate training data bias by exploiting whatever is known of the test domain. The link between these two topics is that what is known of the test domain often comes in the form of an unlabeled sample, and hence semi-supervised techniques constitute an important class of adaptation strategies.

Søgaard's book has at its core the intersection of these two important topics, although it also covers semi-supervised techniques without considering data bias, and techniques for handling bias that are not semi-supervised.

Before proceeding, I should declare a bias of another sort, which is that of a machine translation (MT) researcher toward a book that cites only two MT papers related to semi-supervised learning or domain adaptation (Habash 2008; Daumé III and Jagarlamudi 2011), neither of which is highly representative of the fairly substantial MT work on these topics. I have done my best to apply my neural adaptation faculty to the material in this book; non-MT readers might find it helpful to do the same with this review.

The book begins with two chapters of introductory material on machine learning and NLP that occupy half of its 80 pages (exclusive of bibliography). The main topics of semi-supervised learning and adaptation are then presented in approximately equal-sized portions, with adaptation split into two chapters covering techniques for known and unknown test domains. A short final chapter deals with evaluation in the presence of domain shift.

doi:10.1162/COLI_r_00188

© 2014 Association for Computational Linguistics

The introductory chapter plunges somewhat abruptly into a sketch of NLP as a machine learning problem, then emerges to motivate the core material of the book. Faced with a domain mismatch between training and test data, we are given the option of performing semi-supervised learning with an unlabeled test-domain sample, using either standard algorithms or specialized ones that address domain shift; or, if nothing is known about the test domain, using a learning approach that encourages robustness. Given the focus of the book, this catalog is entirely appropriate, but the scenarios it contemplates are special cases of a more general one in which both labeled and unlabeled training data are available for a collection of domains that may or may not include the test domain. (Incidentally, the 20-Newsgroups data used in the book for text-classification examples falls more naturally into this heterogeneous case than into the binary train/test split to which it is cast.) This section would have benefitted from an attempt to situate the approaches considered here within the more general setting, which is barely acknowledged throughout the book. Pointers to representative relevant work, both in NLP (Daumé III 2007; Finkel and Manning 2009) and machine learning (Ben-David et al. 2010; Dredze, Kulesza, and Crammer 2010), would also have been an asset.

Chapter 2 is a lengthy chapter that lays out basic techniques in a bid to make the book self-contained. This is always a tricky proposition, as it must tread a fine line between boring the expert and baffling the beginner. The author strikes a good balance here by emphasizing practical advice over theoretical completeness, and providing experimental results to underscore various points. Although a few passages would probably cause beginners to stumble, for instance one that invokes the concept of a generative story when explaining hidden Markov models (HMMs), with only a reference to Brown et al. (1993) for explanation, most of the material is very accessible, and there are many links into the literature. The expert reader will also not be unduly bored, and will find this chapter a quick and probably productive read. Although it clearly does not present *all* basic techniques in NLP, the chapter manages to cover a lot of ground while remaining coherent and for the most part not superfluous with respect to the remainder of the book.

To summarize the contents of this chapter briefly, after a discussion of some relevant assumptions about data (smoothness, i.i.d., coherence), and related empirical tests, three basic classification techniques—nearest neighbor, naive Bayes, and perceptron—are described and compared experimentally on a 20-Newsgroups text-classification task. Next, weighted versions of these algorithms are given (omitting how the weights get set). Then a section on unsupervised learning presents hierarchical and k-means clustering, along with a somewhat elliptical account of generalized EM. Finally, structured learning is introduced through HMM-based POS tagging—with a nod to conditional random fields and structured perceptrons—as well as transition- and graph-based dependency parsing.

Chapter 3 describes a variety of semi-supervised learning algorithms. A section on wrapper methods traces two branches of work that refine self-training, in which a learner labels some of the unlabeled material and then trains on its own output. One branch is based on co-training, where two learners trained on complementary features label data for each other. This generalizes to three learners in tri-training, and ultimately to an arbitrary number in multi-view approaches (Ganchev et al. 2008), although these latter are not mentioned in the book. The other branch requires a learner that can assign probabilities to outcomes, and exploits this to iterate soft labeling and training on weighted examples in an expectation maximization (EM)-like (or just plain EM) procedure. (The section that describes this method includes "CO-EM" in the title, but the text never seems to make it to that tantalizing destination.) Another briefly mentioned

wrapper-type technique is to exploit a clustering of the unlabeled data in order to generate additional features for supervised learning; this is related to recent work that uses neural nets to learn embeddings from unlabeled data (Collobert et al. 2011).

A final group of semi-supervised algorithms is specific to nearest-neighbor classification. Label propagation (Zhu and Ghahramani 2002) is a well-known iterative graph-based algorithm where neighbors vote on each node's label, with votes weighted by distance. A similar voting takes place in editing and condensation, which are methods for identifying a subset of prototypical data points (similar to support vectors in support vector machines) in order to speed up nearest-neighbor search. Unlabeled data can be used to improve this process by essentially providing greater resolution.

Chapter 4 is the first to grapple explicitly with domain mismatch, and begins by making a standard distinction between conditional and marginal distributions for inputs and outputs (Jiang and Zhai 2007). Here we are clearly limited to considering biased inputs, because there is only an unlabeled sample from the test domain. One strategy for exploiting this is to apply the semi-supervised methods from the previous chapter, which will work as-is if the mismatch between training and test domains is not too great. Otherwise, we can downweight instances in the labeled training set whose inputs are not close to those in the test-domain data; techniques for measuring distance include LMs and KL divergence. A similar approach can be applied to features, by comparing the training-data values of a feature across all examples (ignoring labels) to its test-domain values. A more sophisticated extension of this idea is structured correspondence learning (Blitzer, McDonald, and Pereira 2006), which automatically places test-domain features in correspondence with training-domain features.

Chapter 5 tackles the ambitious goal of learning "defensively" so as to be robust to domain shifts in the absence of any prior information about the test domain. A common effect is the out-of-vocabulary (OOV) problem, in which features do not appear in a new test domain. If these have high weights, test-domain performance can degrade badly, to the point where it would have been better to have left them out of training in the first place, allowing other features to take up the slack. (Note that for some versions of the OOV problem—such as in MT—things aren't this easy.) An interesting technique for countering feature OOVs, recently introduced to NLP by the author (Søgaard and Johannsen 2012), is adversarial learning, in which random subsets of features are removed during training. The chapter ends with a discussion of more traditional ensemble-based methods (voting, product of experts, stacking, meta-learning) that combine predictions from a set of diverse base classifiers in order to decrease variance.

The final chapter gives a short treatment of the problem of how to predict the performance of systems on new domains. The main, and most interesting, suggestion is meta-analysis, a technique widely used in fields such as medicine and psychology. The idea is to extrapolate from many experimental trials (typically in previously published work) that use the same method but different data, while calibrating for differences among the trials and making use of their internal estimates of variance. I am not sure this is a perfect fit with NLP for many reasons, such as that many papers do not include estimates of variance, and the tendency of NLP systems to be sensitive to undocumented "minor" configuration changes; but it is definitely a direction that bears further investigation.

To summarize, this is a monograph that focuses on the problem of optimizing an NLP system trained on a single domain for a test domain that is either unknown or represented by an unlabeled sample. When the unlabeled sample comes from the training domain, the task is plain semi-supervised learning, a case that is dealt with extensively. The book is written in an informal style, and sticks to basic machine learning concepts, which are illustrated with many examples involving text classification,

POS tagging, and dependency parsing. In my opinion, the most interesting chapters are the shorter ones toward the end. I recommend these to any NLP researcher interested in the perennial problem of making do with not enough of the right kind of data.

The printed version of the book has some imperfections that are mostly due to production quality. Many equations and graphs are fuzzy, and some graphs (e.g., Fig 3.3) are completely indecipherable due to their having been intended for color printing. An index would have been a boon, though its absence is partly compensated for by an extensive back-referenced bibliography. Finally, the text contains many large blocks of Python code, which are sometimes used in lieu of pseudo-code for describing algorithms. Although having real code is a bonus, it is harder to read than pseudo-code, and would have been much more useful to readers had it been moved out of the book and made available separately on-line.

References

Ben-David, Shai, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79:151–175.

Blitzer, John, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney.

Brown, Peter F., Stephen A. Della Pietra, Vincent Della J. Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312.

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Daumé III, Hal. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague.

Daumé III, Hal and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 407–412, Portland, OR.

Dredze, Mark, Alex Kulesza, and Koby Crammer. 2010. Multi-domain learning by confidence-weighted parameter combination. *Machine Learning*, 79:123–149.

Finkel, Jenny Rose and Christopher D. Manning. 2009. Hierarchical Bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, Boulder, CO.

Ganchev, K., J. Graça, J. Blitzer, and B. Taskar. 2008. Multi-view learning over structured and non-identical outputs. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 204–211, Helsinki.

Habash, Nizar. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 57–60, Columbus, OH.

Jiang, Jing and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 264–271, Prague.

Søgaard, Anders and Anders Johannsen. 2012. Robust learning in random subspaces: Equipping NLP for OOV effects. In *Proceedings of COLING 2012: Posters*, pages 1,171–1,180, Mumbai.

Zhu, Xiaojin and Zoubin Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, Pittsburgh, PA.

This book review was edited by Pierre Isabelle.

George Foster is a Senior Researcher in the Multilingual Text Processing Group at the National Research Council of Canada. His recent research has focused on adaptation and parameter estimation for machine translation. Foster's e-mail address is `george.foster@nrc-cnrc.gc.ca`.