Discriminative Models for Semi-Supervised Natural Language Learning

Sajib Dasgupta and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{sajib,vince}@hlt.utdallas.edu

1 Discriminative vs. Generative Models

An interesting question surrounding semi-supervised learning for NLP is: should we use discriminative models or generative models? Despite the fact that generative models have been frequently employed in a semi-supervised setting since the early days of the statistical revolution in NLP, we advocate the use of discriminative models. The ability of discriminative models to handle complex, high-dimensional feature spaces and their strong theoretical guarantees have made them a very appealing alternative to their generative counterparts. Perhaps more importantly, discriminative models have been shown to offer competitive performance on a variety of sequential and structured learning tasks in NLP that are traditionally tackled via generative models , such as letter-to-phoneme conversion (Jiampojamarn et al., 2008), semantic role labeling (Toutanova et al., 2005), syntactic parsing (Taskar et al., 2004), language modeling (Roark et al., 2004), and machine translation (Liang et al., 2006). While generative models allow the seamless integration of prior knowledge, discriminative models seem to outperform generative models in a "no prior", agnostic learning setting. See Ng and Jordan (2002) and Toutanova (2006) for insightful comparisons of generative and discriminative models.

2 Discriminative EM?

A number of semi-supervised learning systems can bootstrap from small amounts of labeled data using discriminative learners, including self-training, co-training (Blum and Mitchell, 1998), and transductive SVM (Joachims, 1999). However, none of them seems to outperform the others across different domains, and each has its pros and cons. Self-training can be used in combination with any discriminative learning model, but it does not take into account the confidence associated with the label of each data point, for instance, by placing more weight on the (perfectly labeled) seeds than on the (presumably noisily labeled) bootstrapped data during the learning process. Co-training is a natural choice if the data possesses two independent, redundant feature splits. However, this conditional independence assumption is a fairly strict assumption and can rarely be satisfied in practice; worse still, it is typically not easy to determine the extent to which a dataset satisfies this assumption. Transductive SVM tends to learn better max-margin hyperplanes with the use of unlabeled data, but its optimization procedure is non-trivial and its performance tends to deteriorate if a sufficiently large amount of unlabeled data is used.

Recently, Brefeld and Scheffer (2004) have proposed a new semi-supervised learning technique, EM-SVM, which is interesting in that it incorporates a discriminative model in an EM setting. Unlike self-training, EM-SVM takes into account the confidence of the new labels, ensuring that the instances that are labeled with less confidence by the SVM have less impact on the training process than the confidently-labeled instances. So far, EM-SVM has been tested on text classification problems, outperforming transductive SVM. It would be interesting to see whether EM-SVM can beat existing semi-supervised learners for other NLP tasks.

Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, pages 84–85,
Boulder, Colorado, June 2009.

3 Effectiveness of Bootstrapping

How effective are the aforementioned semi-supervised learning systems in bootstrapping from small amounts of labeled data? While there are quite a few success stories reporting considerable performance gains over an inductive baseline (e.g., parsing (McClosky et al., 2008), coreference resolution (Ng and Cardie, 2003), and machine translation (Ueffing et al., 2007)), there are negative results too (see Pierce and Cardie (2001), He and Gildea (2006), Duh and Kirchhoff (2006)). Bootstrapping performance can be sensitive to the setting of the parameters of these semi-supervised learners (e.g., when to stop, how many instances to be added to the labeled data in each iteration). To date, however, researchers have relied on various heuristics for parameter selection, but what we need is a principled method for addressing this problem. Recently, McClosky et al. (2008) have characterized the conditions under which self-training would be effective for semi-supervised syntactic parsing. We believe that the NLP community needs to perform more research of this kind, which focuses on identifying the algorithm(s) that achieve good performance under a given setting (e.g., few initial seeds, large amounts of unlabeled data, complex feature space, skewed class distributions).

4 Domain Adaptation

Domain adaptation has recently become a popular research topic in the NLP community. Labeled data for one domain might be used to train a initial classifier for another (possibly related) domain, and then bootstrapping can be employed to learn new knowledge from the new domain (Blitzer et al., 2007). It would be interesting to see if we can come up with a similar semi-supervised learning model for projecting resources from a resource-rich language to a resource-scarce language.

References

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the ACL*.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT*.

Ulf Brefeld and Tobias Scheffer. 2004. Co-EM support vector learning. In *Proceedings of ICML*.

Kevin Duh and Katrin Kirchhoff. 2006. Lexicon acquisition for dialectal Arabic using transductive learning. In *Proceedings of EMNLP*.

Shan He and Daniel Gildea. 2006. Self-training and co-training for semantic role labeling.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2008. Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-08:HLT*.

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proceedings of ICML*.

Percy Liang, Alexandre Bouchard, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proceedings of the ACL*.

David McClosky, Eugene Charniak, and Mark Johnson. 2008. When is self-training effective for parsing? In *Proceedings of COLING*.

Vincent Ng and Claire Cardie. 2003. Weakly supervised natural language learning without redundant views. In *Proceedings of HLT-NAACL*.

Andrew Ng and Michael Jordan. 2002. On discriminative vs.generative classifiers: A comparison of logistic regression and Naive Bayes. In *Advances in NIPS*.

David Pierce and Claire Cardie. 2001. Limitations of co-training for natural language learning from large datasets. In *Proceedings of EMNLP*.

Brian Roark, Murat Saraclar, Michael Collins, and Mark Johnson. 2004. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proceedings of the ACL*.

Ben Taskar, Dan Klein, Michael Collins, Daphne Koller, and Christopher Manning. 2004. Max-margin parsing. In *Proceedings of EMNLP*.

Kristina Toutanova, Aria Haghighi, , and Christopher D. Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of the ACL*.

Kristina Toutanova. 2006. Competitive generative models with structure learning for NLP classification tasks. In *Proceedings of EMNLP*.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the ACL*.