High-Performance Semi-Supervised Learning using Discriminatively Constrained Generative Models

Gregory Druck GDRUCK@CS.UMASS.EDU
Andrew McCallum MCCALLUM@CS.UMASS.EDU
University of Massachusetts Amherst, Amherst, MA 01003

Abstract

We develop a semi-supervised learning method that constrains the posterior distribution of latent variables under a generative model to satisfy a rich set of feature expectation constraints estimated with labeled data. This approach encourages the generative model to discover latent structure that is relevant to a prediction task. We estimate parameters with a coordinate ascent algorithm, one step of which involves training a discriminative log-linear model with an embedded generative model. This hybrid model can be used for test time prediction. Unlike other high-performance semi-supervised methods, the proposed algorithm converges to a stationary point of a single objective function, and affords additional flexibility, for example to use different latent and output spaces. We conduct experiments on three sequence labeling tasks, achieving the best reported results on two of them, and showing promising results on CoNLL03 NER.

1. Introduction

Semi-supervised learning aims to leverage unlabeled data to improve parameter estimation. Generative probabilistic models can easily incorporate unlabeled data into parameter estimation by maximizing the marginal log-likelihood (Nigam et al., 2006). However, if the generative model is misspecified, unlabeled data may degrade performance on the task of interest (Cozman & Cohen, 2006).

Discriminative methods can be advantageous because they do not expend effort modeling input variables and allow the inclusion of features that would violate generative model independence assumptions. However, unlabeled data has no effect on the conditional log-likelihood, so discriminative semi-supervised learning typically relies on additional regularization (Grandvalet & Bengio, 2006).

There has been interest in combining the advantages of generative and discriminative learning. Multi-conditional learning (MCL) (McCallum et al., 2006) and the approach of Lasserre at al. (2006) aim to interpolate between generative and discriminative parameter estimation. Additionally, several high-performance semi-supervised learning methods involve using generative models as features in a discriminative model (Koo et al., 2008), with the most successful methods additionally encouraging generative models to learn latent structure that is relevant for the task of interest (Suzuki & Isozaki, 2008).

Concurrently there has been interest in semi-supervised learning with side constraints (Mann & McCallum, 2008; Chang et al., 2007; Bellare et al., 2009; Liang et al., 2009; Ganchev et al., 2009).

In this paper we propose a new method for semi-supervised learning that combines the two approaches and can be written as a single objective function. The method uses a rich set of discriminative expectation constraints estimated with labeled data to guide generative model estimation. Parameters are estimated with a coordinate ascent algorithm in which one step consists of estimating the parameters of a discriminative log-linear model with a generative model embedded in its potential functions. This hybrid model can be used for test time prediction. Because the generative model is coupled with the task through expectation constraints, the method affords the flexibility to use a generative model whose latent variables are different than the task output variables. For example, using multiple latent states per label may allow the generative model to discover label sub-structure.

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

Unlike the methods of Ando and Zhang (2005) and Suzuki and Isozaki (2008), the approach can be written as a single objective function, and the optimization algorithm converges to a stationary point of a penalized marginal log-likelihood under a generative model.

We apply this method to sequence labeling with a hidden Markov model (HMM) as the generative model. We present experimental results on three information extraction tasks: *Cora* research paper reference information extraction, *Apartment* listing information extraction, and CoNLL 2003 *named entity recognition* (NER). We attain, to the best of our knowledge, the best reported results on *Cora* and *Apartments*. In initial experiments on CoNLL03, we improve by as much as 3.0% F1 over a fully supervised CRF (a 19.7% relative error reduction). Finally, we show that increasing the number of HMM latent variables encourages the discovery of semantically coherent label sub-structure.

2. Approach

In this section we describe our approach to semi-supervised learning in detail. We first review the *Alternating Projections* (AP) framework (Bellare et al., 2009) (or equivalently the *Posterior Regularization* (PR) framework (Ganchev et al., 2009)). We then describe a novel application of this framework to encourage generative models to discover relevant latent representations using a large and rich set of feature expectation constraints estimated with labeled data.

2.1. Constraining Generative Models

Suppose we have a generative model $p(\mathbf{x}, \mathbf{z}; \theta)$ of *input* variables $\mathbf{x}$ and *latent* variables $\mathbf{z}$ with parameters θ. Given an empirical distribution over input variables $\tilde{p}(\mathbf{x})$, we can estimate parameters θ to maximize the marginal log-likelihood of the data

$$\mathcal{L}_M(\theta) = \mathrm{E}_{\tilde{p}(\mathbf{x})}[\log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)] + \log p(\theta),$$

where $p(\theta)$ is a prior on parameters. Although $\mathcal{L}_M(\theta)$ is not convex, the EM algorithm (Dempster et al., 1977) optimizes a lower bound on $\mathcal{L}_M(\theta)$ and, with regularity assumptions, converges to a stationary point.

Despite the simplicity of semi-supervised[1] learning with generative models, and some empirical success (Nigam et al., 2006), it is known that generative semi-supervised learning may degrade performance if the model is misspecified (Cozman & Cohen, 2006).

[1] In the semi-supervised case, some $\mathbf{z}$ are observed. We keep all $\mathbf{z}$ unobserved in our exposition because all supervision of p will come through the expectation constraints.

There has been recent interest in semi-supervised supervised learning with the aide of side constraints (Mann & McCallum, 2008; Chang et al., 2007; Liang et al., 2009; Ganchev et al., 2009). These methods can be used to address model misspecification by penalizing model parameter settings that do not respect the intended meaning of latent variables. That is, rather than maximizing $\mathcal{L}_M(\theta)$, we instead estimate generative model parameters θ by maximizing the following objective function

$$\mathcal{G}(\theta) = \mathcal{L}_M(\theta) - \mathrm{U}\Big(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{z}|\mathbf{x};\theta)}[\mathbf{f}(\mathbf{x}, \mathbf{z})]]\Big),$$

where U is a convex potential function that evaluates expectations of *constraint features* $\mathbf{f}$ under $p(\mathbf{z}|\mathbf{x}; \theta)$. This is a *generalized expectation* (GE) parameter estimation objective function (Mann & McCallum, 2008). Unfortunately, GE parameter estimation in structured output models typically requires computing marginal distributions over more variables than participate in model factors. Because our goal is large-scale semi-supervised learning, we seek a more efficient solution.

Following Bellare et al. (2009), we introduce an *auxiliary distribution* q and compute a variational approximation to $\mathcal{G}(\theta)$

$$\mathcal{O}(\theta, q) = \mathcal{L}_M(\theta) - \mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{D}(q(\mathbf{z}|\mathbf{x}) \;||\; p(\mathbf{z}|\mathbf{x}; \theta))] - \mathrm{U}(q),$$

where $\mathrm{U}(q) = \mathrm{U}\big(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[\mathbf{f}(\mathbf{x}, \mathbf{z})]]\big)$ and D is the KL-divergence. In this paper, we assume we have a vector of target expectations $\mathbf{b}$ for constraint features $\mathbf{f}$, and penalize the ℓ_2^2 distance between the model and target expectations, weighted by σ^2

$$\mathrm{U}(q) = \frac{\sigma^2}{2} ||\mathbf{b} - \mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[\mathbf{f}(\mathbf{x}, \mathbf{z})]]||_2^2.$$

We perform block coordinate ascent on $\mathcal{O}(\theta, q)$. The resulting algorithm can be equivalently viewed as a modified version of EM (Ganchev et al., 2009), or as alternating between two convex projections (Bellare et al., 2009). The steps are:

Information projection (modified E-step):

The likelihood term $\mathcal{L}_M(\theta)$ is constant with respect to q, so we need to solve

$$q^{t+1} = \operatorname*{argmin}_q \; \mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{D}(q(\mathbf{z}|\mathbf{x}) \;||\; p(\mathbf{z}|\mathbf{x}; \theta^t))] + \mathrm{U}(q)$$

This is a *generalized maximum entropy* problem (Dudik, 2007). The dual of this problem is

$$\lambda^{t+1} = \operatorname*{argmax}_\lambda \; \lambda \cdot \mathbf{b} - \mathrm{E}_{\tilde{p}(\mathbf{x})}[\log Z_{\lambda, \theta^t}(\mathbf{x})] - \frac{1}{2\sigma^2} ||\lambda||_2^2,$$

where $\log Z_{\lambda,\theta^t}(\mathbf{x})$ is the log-partition function

$$\log Z_{\lambda,\theta^t}(\mathbf{x}) = \sum_{\mathbf{z}} \exp(\lambda \cdot \mathbf{f}(\mathbf{x},\mathbf{z}) + \log p(\mathbf{z}|\mathbf{x};\theta^t)).$$

Consequently, q has an exponential family form

$$q^{t+1}(\mathbf{z}|\mathbf{x},\theta^t;\lambda) \propto \exp\left(\lambda \cdot \mathbf{f}(\mathbf{x},\mathbf{z}) + \log p(\mathbf{z}|\mathbf{x};\theta^t)\right).$$

When $\mathbf{b}$ comes from a labeled sample, the information projection is equivalent to maximizing the conditional log-likelihood under $q^{t+1}(\mathbf{z}|\mathbf{x},\theta^t;\lambda)$ with ℓ_2^2 regularization. Note that $\log Z_{\lambda,\theta^t}(\mathbf{x})$ can be computed efficiently as long as inference in the model implied by $\mathbf{f}$ is tractable and p and $\mathbf{f}$ factorize in the same way.

Moment projection (M-step):

The potential $\mathrm{U}(q)$ is constant with respect to θ, so we need to solve

$$\begin{aligned}\theta^{t+1} &= \operatorname*{argmax}_{\theta}\ \mathcal{L}_M(\theta) - \mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{D}(q^{t+1}(\mathbf{z}|\mathbf{x}) \mid\mid p(\mathbf{z}|\mathbf{x};\theta))] \\ &= \operatorname*{argmax}_{\theta}\ \mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{q^{t+1}(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x},\mathbf{z};\theta)]] + \log p(\theta)\end{aligned}$$

This step is equivalent to the M-Step in EM, but with the distribution over latent variables provided by q^{t+1}, rather than p^t. If p is a simple directed graphical model such as an HMM, this has a closed-form solution.

Note that $\mathcal{O}(\theta, q)$ is a lower bound on the marginal log-likelihood of the generative model $\mathcal{L}_M(\theta)$. Intuitively, $\mathcal{O}(\theta, q)$ can be viewed as a penalized marginal log-likelihood. Although $\mathcal{O}(\theta, q)$ is not convex, each projection step is convex, and each step increases $\mathcal{O}(\theta, q)$. With regularity assumptions, this algorithm converges to a stationary point of $\mathcal{O}(\theta, q)$ (Ganchev et al., 2009).

2.2. Constraints

Most previous applications of methods for semi-supervised learning with constraints (Mann & McCallum, 2008; Chang et al., 2007; Ganchev et al., 2009; Liang et al., 2009) used a small number of constraints. In contrast, we use a large and rich set of constraint feature functions $\mathbf{f}$. Specifically, we define $\mathbf{f}$ as the same set of features we would use when specifying the parameterization of a discriminative log-linear model or conditional random field (CRF) (Lafferty et al., 2001). Because $\mathbf{f}$ specifies the structure of the dual form of q, here q is a feature-rich CRF that is capable of making accurate predictions in the absence of p.

We next need to estimate target expectations $\mathbf{b}$ for the constraint features. In previous applications of methods for semi-supervised learning with constraints, target expectations were typically estimated using a combination of prior knowledge and simple heuristics.

In this paper, we use labeled data to estimate target expectations $\mathbf{b}$. Importantly, note that in order to address generative model misspecification we need to constrain the posterior distributions for both labeled and unlabeled examples, as otherwise $q(\mathbf{z}|\mathbf{x}) = p(\mathbf{z}|\mathbf{x};\theta)$ for unlabeled $\mathbf{x}$. Consequently, we need to estimate target expectations $\mathbf{b}$ that are appropriate for both labeled and unlabeled data. We propose two target expectation estimation methods.

labeled sample expectations: The simplest method for estimating $\mathbf{b}$ is to use the *labeled sample expectations* $\mathbf{b}_l = \mathrm{E}_{\tilde{p}_l(\mathbf{x},\mathbf{z})}[\mathbf{f}(\mathbf{x},\mathbf{z})]$, where $\tilde{p}_l(\mathbf{x},\mathbf{z})$ is the empirical distribution over labeled data.

supervised model expectations: An alternative estimation method is to use the dual parametric form of a supervised model $q(\mathbf{z}|\mathbf{x};\lambda_{sup})$ (estimated with only the *labeled sample expectations* and $\tilde{p}_l(\mathbf{x})$) to "fill-in" expectations for unlabeled examples:

$$\lambda_{sup} = \operatorname*{argmax}_{\lambda}\ \lambda \cdot \mathbf{b}_l - \mathrm{E}_{\tilde{p}_l(\mathbf{x})}[\log Z_\lambda(\mathbf{x})] - \frac{1}{2\sigma^2}||\lambda||_2^2$$

$$\mathbf{b}_{sup} = \frac{n_l}{n_l+n_u}\mathbf{b}_l + \frac{n_u}{n_l+n_u}\mathrm{E}_{\tilde{p}_u(\mathbf{x})}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x};\lambda_{sup})}[\mathbf{f}(\mathbf{x},\mathbf{z})]],$$

where n_l and n_u are the number of labeled and unlabeled examples, respectively. Note that this estimation is only performed once, and afterwards $\mathbf{b}_{sup}$ is fixed while optimizing $\mathcal{O}(\theta, q)$.

Note that $\mathbf{b}_l$ (and consequently $\mathbf{b}_{sup}$) may be noisy. Although $\mathbf{b}_l$ approaches the true expectation vector as the size of the labeled sample increases, when data is sparse (i.e. in NLP problems) a very large labeled sample may be required to obtain accurate expectation estimates. Note also that there may be no q that exactly satisfies the labeled sample expectation constraints, while there is a q that exactly satisfies the supervised model expectation constraints. In theory these are not problems, as U does not require the constraints to be matched exactly.

In practice, we find learning with estimated expectations to be challenging. To illustrate this, we compare simple ℓ_2^2-regularized maximum entropy estimation (without the generative model p) using $\mathbf{b}_l$ and $\mathbf{b}_{sup}$ with 350 labeled and 5,000 unlabeled examples from the first split of the *Cora* data set (see Section 5.1). Using $\mathbf{b}_l$ yields test F_1 of 61.0%, whereas purely-supervised estimation gives 90.8% F_1. Using $\mathbf{b}_{sup}$ performs much better, yielding 90.3% F_1, though this is still worse than supervised.

To alleviate issues with noisy target expectations, we down-weight unlabeled data by modifying the empiri-

cal distribution $\tilde{p}(\mathbf{x})$ with hyper-parameter γ.

$$\tilde{p}(\mathbf{x};\gamma) = \frac{n_l \tilde{p}_l(\mathbf{x}) + \gamma n_u \tilde{p}_u(\mathbf{x})}{n_l + \gamma n_u}.$$

Note that γ is also used in the estimation of $\mathbf{b}_{sup}$.

Using unlabeled data weighting, maximum entropy estimation with $\mathbf{b}_l$ matches but does not outperform supervised estimation, yielding F_1 of 90.8% with $\gamma \leq 0.0001$. In contrast, with unlabeled data weight $\gamma = 0.1$ maximum entropy estimation with $\mathbf{b}_{sup}$ improves over supervised estimation, giving 91.1% F_1. This trend holds across multiple data sets. We always obtain better performance with $\mathbf{b}_{sup}$ than with $\mathbf{b}_l$, whether the resulting constraints are used in maximum entropy estimation or in estimation with $\mathcal{O}(\theta, q)$.

We additionally experimented with several methods to improve either the estimation of $\mathbf{b}_l$ or learning with $\mathbf{b}_l$. In an earlier version of this work, we were able to obtain better results by first initializing parameters to λ_{sup} and then performing five steps of optimization. However, we prefer not to use this solution because it does not follow from the objective function. Simple changes to the estimation problem such as putting a hard constraint on the norm of the parameter vector and penalizing constraint violations with a different norm fail to improve the above results with $\mathbf{b}_l$.

Note that *supervised model expectation* estimation can be viewed as one step of self-training, which justifies why it works with appropriate γ. We conjecture that using *labeled sample expectations* performs worse because there is no guarantee that q labels the labeled data correctly as $\mathrm{U}(q)$ decreases, unless γ is very small.

We plan to explore improved methods for estimation of and with $\mathbf{b}$ in future work. Two promising directions are smoothing $\mathbf{b}_l$ and using different regularization weights for each constraint. For the remainder of this paper, however, we estimate $\mathbf{b}$ with the *supervised model expectations* method. We also use unlabeled data weighting as described above. The complete objective function is then $\mathcal{O}(\theta, q)$ with $\tilde{p}(\mathbf{x};\gamma)$ substituted for $\tilde{p}(\mathbf{x})$, and $\mathbf{b}_{sup}$ substituted for $\mathbf{b}$. In experiments we tune γ using cross-validation.

2.3. Latent Variables and Labels

Thus far we have described both the generative $p(\mathbf{x}, \mathbf{z}; \theta)$ and auxiliary $q(\mathbf{z}|\mathbf{x})$ models as modeling latent variables $\mathbf{z}$. We now additionally introduce *label* variables, denoted by $\mathbf{y}$, which are the variables we would like to predict. We make a distinction between the two because the latent and label spaces, denoted Z and Y, may be different. The correspondence between labels and latent variables is encoded using the constraint features $\mathbf{f}$ and the structure of the generative model. In this paper we define the correspondence between latent variables and labels using either *one-to-one* or *many-to-one* maps $\phi : Z \rightarrow Y$. For example, three latent variables may map to one particular label. This setup, described in detail in Section 4, allows the generative model to discover label sub-structure.

2.4. Test Time Inference

There are several ways to use the learned (θ, q) for inference at test time.

- *Generative:* Use the generative model posterior $p(\mathbf{y}|\mathbf{x};\theta)$ for test time inference.
- *Gen./Discr.:* Use the parametric form of the final auxiliary distribution $q(\mathbf{y}|\mathbf{x},\theta;\lambda)$ for test time inference.
- *Transductive Gen./Discr.:* Re-estimate q and $\mathbf{b}$ at test time using the learned $p(\mathbf{y}, \mathbf{x};\theta)$.
- *Generative as features:* Use $p(\mathbf{y}, \mathbf{x};\theta)$ to generate features for a downstream discriminative model.

In this paper, we use the *Generative* and *Gen./Discr.* methods. *Transductive Gen./Discr.* may provide better performance, but is less desirable because it requires re-training at test time. In future work, we plan to explore *Generative as features*, as this would allow the discriminative model to control the contribution of the generative model through parameter estimation.

2.5. Discussion

Intuitively, this method works by combining the strengths of generative and discriminative learning. The generative model $p(\mathbf{x}, \mathbf{z};\theta)$ benefits from q because we expect q's posterior distributions to be more accurate than $p(\mathbf{z}|\mathbf{x};\theta)$. Consequently, we expect this method to outperform maximum likelihood estimation of $p(\mathbf{x}, \mathbf{z};\theta)$ with EM.

Next, note that q only has parameters for features observed in the labeled data, whereas the generative model also includes parameters for features that only occur in unlabeled data. Therefore, the generative model posteriors $p(\mathbf{z}|\mathbf{x};\theta)$ help q to address sparsity.

Finally, we note that the way in which the generative model probabilities are incorporated into q, essentially as features without parameters, results in an interesting parameter estimation problem. Although we do not expect p to match the accuracy of q, we expect p to be accurate for particularly "easy" examples or

some subset of the latent variables Z. In these cases, q does not need to expend much modeling effort, as the constraints may be close to being satisfied with the generative posteriors alone. In contrast, the generative model will also make many mistakes, and q will need to compensate for these mistakes in order to satisfy the constraints. Consequently, q will spend most of its modeling effort on the more difficult examples or latent variables. The fact that the discriminative model does not waste effort modeling the "easy" labels may be beneficial in the same way that not wasting effort modeling input variables is considered beneficial.

3. Related Work

The proposed method is an application of the *Alternating Projections* (AP) (Bellare et al., 2009) or *Posterior Regularization* (PR) (Ganchev et al., 2009) framework for learning with expectation constraints. Ganchev et al. (2009) show that PR can be viewed as an approximation to the *measurements* framework (Liang et al., 2009) and is more efficient than *Generalized Expectation* estimation (Mann & McCallum, 2008). *Constraint-driven Learning* (Chang et al., 2007) can be viewed as a "hard" approximation to PR. Unlike prior work with these methods, in this paper we use a large and rich set of feature expectation constraints that are estimated using available labeled data.

The AP method is closely related to the work of Suzuki and Isozaki (2008), but has some important advantages. The *JESS-CM* method augments CRF potential functions with features of the log probability of output variables under a generative model, $f(\mathbf{x}, y_i, y_{i+1}) = \log p(y_i, y_{i+1}|\mathbf{x}; \theta)$. The iterative learning algorithm first re-estimates CRF parameters by maximizing the log-likelihood of the labeled data. Next, the generative models are re-estimated using unlabeled data with *Maximum Discriminant Functions sum* (MDF) estimation. MDF is used in place of marginal likelihood to ensure generative model predictions are relevant to the task. MDF estimates θ by maximizing $\mathrm{E}_{\tilde{p}(\mathbf{x})}[\sum_{\mathbf{y}} \exp(\lambda \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}, \theta))]$ with an EM-like algorithm. This method achieves impressive empirical results on natural language processing tasks. Recall that our approach can be written as a single objective function, and our optimization converges to a stationary point of the penalized marginal log-likelihood. In contrast, the JESS-CM method consists of two separate objective functions (maximum likelihood and MDF), and there is no guarantee that alternating between the two objective functions will lead to convergence. Additionally, the JESS-CM method assumes that the output spaces of the generative and discriminative models are the same. We provide additional comparison with this method in Section 5.3.

Ando and Zhang (2005) propose a semi-supervised learning method that also uses unlabeled data and discriminative information to generate new features for supervised learning. The first step of the algorithm is Alternating Structure Optimization (ASO). ASO uses the parameter vectors of *auxiliary problem* models to find a "predictive" low-dimensional feature representation. This representation is then used to generate additional features for a discriminative model. As with JESS-CM, the complete approach cannot be written as a single objective function, and several extensions are required to attain impressive empirical results.

The multi-view learning framework of Ganchev et al. (2009) uses expectation constraints to encourage models of multiple views of the data to agree. In this paper, we make no multi-view assumption, as there is overlap between the generative and discriminative model feature sets. Ignoring this issue, our method is also distinct from an application of the multi-view framework with generative and discriminative views. With our method, the generative model posterior probabilities appear directly in the potential functions of the discriminative auxiliary model during parameter estimation, in contrast to assisting in labeling unlabeled data for discriminative model parameter estimation. We believe that forcing the discriminative model to compensate for the mistakes of the generative model is beneficial, as discussed in Section 2.5.

Quadrianto et al. (2009) propose a transductive learning algorithm that encourages the model to have similar training and testing data feature expectations. In future work, we could use a similar formulation to constrain the posterior distributions for unlabeled data, rather than estimating target expectations directly.

This method also has advantages over previously proposed generative / discriminative hybrid methods. Unlike multi-conditional learning (McCallum et al., 2006), we may choose different generative and discriminative models. Unlike the method of Lasserre et al. (2006), our constraints make statements about expectations, rather than statements about parameters, which can be difficult to interpret.

Although we may avoid misspecification by designing a better model, finding the correct generative model is infeasible for most problems. An especially difficult part of generative model design is modeling correlations between input features. With the proposed method, we may avoid this using a rich set of expectation constraints on overlapping features.

4. Application to Sequence Labeling

In this paper we apply our approach to sequence labeling. The generative model $p(\mathbf{x}, \mathbf{z}; \theta)$ is a standard first-order HMM. The constraint features $\mathbf{f}$ specify a first-order CRF, i.e. each f considers two labels and the input $f(\mathbf{x}, y_i, y_{i+1})$. The potential function U is

$$\mathrm{U}(q) = \left|\left|\mathbf{b} - \mathrm{E}_{\tilde{p}(\mathbf{x};\gamma)}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[\mathbf{f}(\mathbf{x}, \phi(\mathbf{z}))]]\right|\right|_2^2,$$

where ϕ is either a *one-to-one* or *many-to-one* map between latent variables and labels. We leave learning the appropriate number of latent states to future work and simply provide each label with a constant number of associated latent states. Note that increasing the number of latent states by a factor of n increases the time for inference by a factor of n^2. However, the inference steps are straightforwardly parallelizable.

5. Experiments

We conduct experiments on *Cora* research paper information extraction, *Apartment* listing information extraction, and *Named-Entity Recognition* (NER).

HMM baseline setup: We use m latent variables per label, as described in Section 4. We denote an HMM with m latent variables per label HMMm. We estimate parameters with the EM algorithm, using available labeled data. In EM experiments with HMMm such that $m > 1$, we incorporate labeled instances by running a constrained forward-backward algorithm that requires the latent variable z at each position to be one that maps to the label $z : \phi(z) = y_i$. HMM states emit lowercase words for NER, and mixed case words for *Cora* and *Apartments*. We use a symmetric Dirichlet prior to smooth HMM emission and transition multinomials. Finally, we weight the contribution of unlabeled data in the M-step with γ. The Dirichlet prior parameter α and unlabeled data weight γ are selected using 5-fold cross-validation for *Cora* and *Apartments* experiments, and using a development set for *NER*.

CRF baseline setup: We estimate CRF parameters with supervised maximum likelihood (ML) and semi-supervised self-training (ST) methods. We use a Gaussian prior on parameters with variance σ^2 and, for ST, weight the unlabeled data with a parameter γ. We select γ and σ^2 using five-fold cross-validation.

AP setup: We setup HMMs and tune α, σ^2, and the unlabeled data weight γ as described above.

5.1. Cora Experiments

The *Cora* data set consists of research paper references, and the task is perform segmentation using 14 BibTeX-like labels such as author and booktitle. We use up to 350 references as labeled data, 150 references for the test set, and 5,000 unlabeled references. The results are averaged over 5 random 80:20 splits.

Table 1 displays the results, evaluated using F_1. Note that HMM1-CRF AP and HMM2-CRF AP always outperform CRF ML and CRF ST, and HMM1 AP and HMM2 AP always outperform HMMs estimated using EM. In summary, the AP method provides both more accurate generative and discriminative models.

Note that in several cases HMM2 AP outperforms HMM1 AP while CRF-HMM1 AP outperforms CRF-HMM2 AP. We performed detailed error analysis to understand this phenomenon. We find that in general, the posterior distributions provided by HMM2 are more peaked than HMM1, meaning that the log probabilities that appear in q's potential functions have more influence. Because the HMMs provide lower accuracy than discriminative methods, the overall effect is a decrease in accuracy. A method to address this might involve adding a temperature on the HMM probabilities in the objective function.

In all experiments HMM2 outperforms HMM1. We have found that HMMs with more latent variables are often able to discover coherent label sub-structure. To best illustrate this point, we conduct an additional experiment using HMM4 AP. Figure 1 depicts discovered label sub-structure for the author and title labels. The figure shows the 10 most probable words emitted from each latent variable state, as well as high probability transitions among each label's latent variable states.

Finally, we note that Peng and McCallum (2004) attained 91.5% F1 with a supervised CRF. CRF-HMM1 AP gives 92.6% F1, which to the best of our knowledge is the best reported result on this task.

5.2. Apartments Experiments

The *Apartments* task is to segment Craigslist apartment listings with 11 labels such as rent, features, and neighborhood. This data has been used by several researchers (Chang et al., 2007; Mann & McCallum, 2008; Liang et al., 2009) to evaluate methods for learning with constraints from prior knowledge. We applied AP to this data with HMM1, 100 labeled listings, and 1000 unlabeled listings, and obtained 83.0% token accuracy, as shown in Table 2. To the best of our knowledge, this is the best reported result for this task. All previous methods listed in Table 2 use additional resources, namely prior knowledge constraints and extra distributional similarity features, in addition to unlabeled data. In contrast, we use only unlabeled data.

Table 1. Cora results. Bold values indicate the best performing method, while underlined values indicate the best performing generative model. Using AP provides the best performing generative and discriminative models in all experiments. 92.6% F1 is, to the best of our knowledge, the best reported result on this task.

	HMM1	HMM2	CRF		AP with HMM1		AP with HMM2	
labeled	EM	EM	ML	ST	HMM1	HMM1-CRF	HMM2	HMM2-CRF
50	42.1	45.9	78.2	78.5	47.9	81.0	<u>64.4</u>	**81.2**
100	47.3	55.3	84.7	85.0	50.7	**86.7**	<u>67.1</u>	86.2
200	51.0	65.2	88.5	88.7	52.1	**90.7**	<u>71.9</u>	90.5
350	52.3	68.1	91.4	91.4	53.5	**92.6**	<u>73.9</u>	91.5

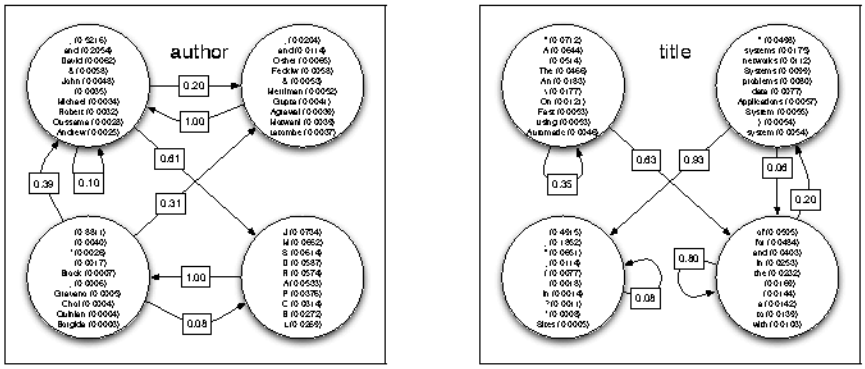

Figure 1. Label sub-structure discovered by HMM4 AP on *Cora* data for author and title labels. For the author the states roughly correspond to "and/first name", "last name", "punctuation", and "initial". For title, the states roughly correspond to "start of title", "terminology", "punctuation", and "function words".

Table 2. Apartments results with 100 labeled instances and 1000 unlabeled instances. We attain the best reported results on this task without additional resources.

method	resources	accuracy
(CCR07)	constraints	81.7
(MM08)	constraints, clusters	80.5
(LJK09)	constraints, clusters	82.5
HMM1 EM	unlabeled data	77.4
CRF ML	none	76.6
CRF ST	unlabeled data	76.9
HMM1 AP	unlabeled data	78.4
HMM1-CRF AP	unlabeled data	**83.0**

5.3. CoNLL03 NER Experiments

Finally, we present early results for CoNLL 2003 NER, in which the task is to recognize person, location, organization, and miscellaneous entities. We use a first-order CRF with the same feature templates as Suzuki and Isozaki (2008), and HMM1 and HMM2. We use the provided 17M words of unlabeled data.

Table 3 provides the results. AP improves F_1 by 2.4-3% over CRF ML. CRF-HMM2 gives slightly higher F_1 than CRF-HMM1. HMM2 is also more accurate than HMM1, improving from 64.6% to 67.8% F_1. HMM2 often discovers interpretable label sub-structure, for example "first name" (*john, david, de, bill, michael*) and "last name" (*clinton, yeltsin, dole, netanyahu, arafat*) sub-labels for the *person* entity, and, for the *miscellaneous* entity, "nationality" (*european, french, german, iraqi, british*) and "event" (*u.k., war, u.s., cup, index*). Note that the model discovers that *miscellaneous* is composed of several distinct entity categories.

Table 3 also provides state-of-the-art semi-supervised results. Although our results are lower than Suzuki and Isozaki (2008) (SI08) and Ando and Zhang (2005) (AZ05), we obtain them with a simpler setup. SI08 use a second-order CRF with unsupported features, 79 HMMs that emit different features, 27M and 1G words of unlabeled data, and perform maximum F_1 estimation. AZ05 use a second-order model, 27M words of unlabeled data, require the specification of auxiliary problems, and use a modified ASO algorithm. See Section 3 for additional discussion.

6. Conclusion & Future Work

We presented a new method for semi-supervised learning that encourages generative models to discover relevant latent representations using expectation con-

Table 3. Early *CoNLL03 NER* F_1 results. We achieve large improvements over a supervised CRF with a simpler experimental setup than state-of-the-art methods.

method	**dev**	**test**
CRF ML	89.7	84.8
HMM1-CRF AP (17M)	92.4 (+2.7)	87.7 (+2.9)
HMM2-CRF AP (17M)	92.3 (+2.6)	87.8 (+3.0)
(AZ05) ASO (27M)	93.2	89.3
(SI08) CRF Max F_1	91.7	86.4
(SI08) JESS-CM (27M)	93.7 (+2.0)	89.4 (+3.0)
(SI08) JESS-CM (1G)	94.5 (+2.8)	89.9 (+3.5)

straints estimated with labeled data. We attained state-of-the-art results on two information extraction tasks, and promising results on NER. Future directions include pursuing state-of-the-art results on NER and other tasks, and further exploring methods for estimating, re-estimating, and learning with constraints.

Acknowledgments

We thank Kedar Bellare, Sameer Singh, Kuzman Ganchev, Sebastian Riedel, and the reviewers for helpful comments. This work was supported by the Center for Intelligent Information Retrieval and by SRI International subcontract #27-001338 and ARFL prime contract #FA8750-09-C-0181. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect those of the sponsor.

References

Ando, R. K. and Zhang, T. A high-performance semi-supervised learning method for text chunking. In *Proc. of Meeting of Assoc. for Computational Linguistics*, pp. 1–9, 2005.

Bellare, K., Druck, G., and McCallum, A. Alternating projections for learning with expectation constraints. In *Proc. of Conf. on Uncertainty in Artificial Intelligence*, 2009.

Chang, M., Ratinov, L., and Roth, D. Guiding semi-supervision with constraint-driven learning. In *Proc. of Meeting of Assoc. for Computational Linguistics*, pp. 280–287, 2007.

Cozman, F. and Cohen, I. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In *Semi-Supervised Learning*. MIT Press, 2006.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Ser. B*, 39(1):1–38, 1977.

Dudik, Miroslav. *Maximum entropy density estimation and modeling geographic distributions of species*. PhD thesis, Department of Computer Science, Princeton University, 2007.

Ganchev, K., Graça, J., Gillenwater, J., and Taskar, B. Posterior regularization for structured latent variable models. Technical Report MS-CIS-09-16, University of Pennsylvania Department of Computer and Information Science, 2009.

Grandvalet, Y. and Bengio, Y. Entropy regularization. In *Semi-Supervised Learning*. MIT Press, 2006.

Koo, T., Carreras, X., and Collins, M. Simple semi-supervised dependency parsing. In *Proc. of Meeting of Assoc. for Computational Linguistics*, pp. 595–603, 2008.

Lafferty, J., McCallum, A., and Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. Int'l. Conf. on Machine Learning*, pp. 282–289, 2001.

Lasserre, J. A., Bishop, C. M., and Minka, T. P. Principled hybrids of generative and discriminative models. In *Conf. on Computer Vision and Pattern Recognition*, pp. 87–94, 2006.

Liang, P., Jordan, M. I., and Klein, D. Learning from measurements in exponential families. In *Proc. Int'l. Conf. on Machine Learning*, pp. 641–648, 2009.

Mann, G. and McCallum, A. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proc. of Meeting of Assoc. for Computational Linguistics*, pp. 870–878, 2008.

McCallum, A., Pal, C., Druck, G., and Wang, X. Multi-conditional learning: Generative / discriminative training for clustering and classification. In *Proc. Conf. on A.I.*, pp. 433–439, 2006.

Nigam, K., McCallum, A., and Mitchell, T. Semi-supervised text classification using em. In *Semi-Supervised Learning*. MIT Press, 2006.

Peng, F. and Mccallum, A. Accurate information extraction from research papers using conditional random fields. In *Proc. of Human Language Technology Conf.*, pp. 329–336, 2004.

Quadrianto, N., Petterson, J., and Smola, A. Distribution matching for transduction. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Neural Information Processing Systems*, pp. 1500–1508. MIT Press, 2009.

Suzuki, J. and Isozaki, H. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proc. of Meeting of Assoc. for Computational Linguistics*, pp. 665–673, 2008.