Constraint-Driven Training of Complex Models Using MCMC

SAMEER SINGH
Computer Science, University of Massachusetts, Amherst MA 01003

GREGORY DRUCK
Yummly, Palo Alto CA 94301

ANDREW MCCALLUM
Computer Science, University of Massachusetts, Amherst MA 01003

Technical Report
Department of Computer Science
University of Massachusetts, Amherst

UMASS-CS-2012-032
October, 2012
Contact Email: sameer@cs.umass.edu

Abstract

Standard machine learning approaches require labeled data, and labeling data for each task, language, and domain of interest is not feasible. Consequently, there has been much interest in developing training algorithms that can leverage constraints from prior knowledge to augment or replace labeled data. Most previous work in this area assumes that there exist efficient inference algorithms for the model being trained. For many NLP tasks of interest, such as entity resolution, complex models that require approximate inference are advantageous. In this paper we study algorithms for training complex models using constraints from prior knowledge. We propose an MCMC-based approximation to Generalized Expectation (GE) training, and compare it to Constraint-Driven SampleRank (CDSR). Sequence labeling experiments demonstrate that MCMC GE closely approximates exact GE, and that GE can substantially outperform CDSR. We then apply these methods to train densely-connected citation resolution models. Both methods yield highly accurate models (up to 94% mean pairwise F_1) with only two simple constraints.

1 Introduction

Standard machine learning approaches require labeled data. However, labeling data for many tasks, domains, and languages of interest is prohibitively expensive and time-consuming, so there has been recent interest in developing training algorithms that can leverage alternate forms of supervision. Usually we have substantial prior knowledge about the task of interest, and this knowledge can often be naturally specified as constraints on feature expectations. For example, when the task is named entity recognition, we may know that roughly 80% of tokens that appear in a known list of last names should be labeled *person*. Methods that use such constraints for training have been successfully applied to classification and sequence labeling tasks (Mann and McCallum, 2007; Chang et al., 2007; Druck et al., 2009; Bellare et al., 2009; Ganchev et al., 2010), among others. Most previous work has used these methods in conjunction with exact inference. However, complex information extraction and natural language processing tasks, such as coreference resolution and relation extraction, can benefit from joint inference and the modeling of long-range dependencies (Poon and Domingos, 2008; Wick et al., 2008; Wellner et al., 2004).

The goal of this work is to enable constraint-based training for complex models, where exact inference is intractable, using MCMC methods. We compare two training algorithms. First, we propose an MCMC-based approximation to *Generalized Expectation* (GE; Mann and McCallum (2007)) training, including the incorporation of a temperature term that increases accuracy. Second, we explore the use of *Constraint-Driven SampleRank* (CDSR; Singh et al. (2010)). In Section 4, we present sequence labeling experiments that demonstrate that MCMC GE closely approximates exact GE, and that GE can substantially outperform CDSR. Then, in Section 5, we use these approaches for lightly supervised training of an entity disambiguation model that contains exponential-domain variables and pairwise factors, yielding high accuracy on a standard data set with just two simple constraints. This is the first application of constraint-based supervision to loopy, undirected graphical models (although CDSR uses approximate inference, it was previously applied only to linear chains in Singh et al. (2010)).

2 Background

In this section, we give a brief summary of the background material and introduce notation.

2.1 Discriminative Log-Linear Graphical Models

Given observed (input) variables $\mathbf{x}$, the probability distribution over the output variables $\mathbf{y}$ of interest can be modeled using a probabilistic graphical model. In particular, an undirected graphical model (Kschischang et al., 2001) consists of factors Ψ_a that assign scores to assignments of subsets of input and output variables $(\mathbf{x}_a, \mathbf{y}_a)$. In log-linear models, the score for each factor Ψ_a is the dot product of a factor-specific feature vector f_a and a global set of parameters $\boldsymbol{\theta}$. Given these features and parameters, the probability of an assignment of the variables is:

$$\begin{aligned} p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) &= \frac{1}{Z(\mathbf{x})} \exp \sum_a \Psi_a(\mathbf{y}_a, \mathbf{x}_a) \\ &= \frac{1}{Z(\mathbf{x})} \exp \sum_a \boldsymbol{\theta} \cdot f_a(\mathbf{y}_a, \mathbf{x}_a) = \frac{1}{Z(\mathbf{x})} \exp \boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{y}, \mathbf{x}) \end{aligned} \quad (1)$$

where $Z(\mathbf{x})$ is the input specific normalization constant.

Supervised estimation of the parameters $\boldsymbol{\theta}$ relies on labeled training data $\mathcal{D}_l = \{\hat{\mathbf{y}}_i, \mathbf{x}_i\}_n$. Specifically, parameters are typically estimated to maximize the model log-likelihood of the labeled data,

$$\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}_l) = \sum_{\mathcal{D}_l} \log p(\hat{\mathbf{y}}_i|\mathbf{x}_i;\boldsymbol{\theta}) - \lambda||\boldsymbol{\theta}||^2. \quad (2)$$

Log-linear models have been applied to many problems in natural language processing and information extraction. NLP tasks such as chunking, part of speech tagging, and named-entity recognition can be cast as sequence labeling problems, and exact inference in a sequential log-linear model is efficient (Sutton and McCallum, 2011). Further, these models provide considerable flexibility in designing features.

2.2 Constraint-based Learning

Obtaining labeled data is often much more difficult than obtaining unlabeled data for a number of reasons. First, labeling can be prohibitively expensive and time-consuming for some tasks, as in dependency parsing, where a complex structure needs to be identified for every sentence, or entity resolution, where the true partitioning needs to be chosen amongst *exponential* possible clusterings. Second,

a large number of languages of interest have a significant lack of resources. We would still like to be able to learn the parameters of log-linear models in such contexts, when labeled data is unavailable.

Constraint-based learning is a recent area of research in semi-supervised learning that addresses this problem (Mann and McCallum, 2008; Bellare et al., 2009; Ganchev et al., 2010). While labeling data is difficult, it is often much easier to specify *auxiliary* information about the labels. This auxiliary information can be represented as targets for the model expectations of a few *constraint features*. As an example, consider the task of named entity recognition for a new language, for which labeling would involve reading every sentence and specifying the person/location/organization token spans in that sentence. Instead, it would be easier to specify that all instances of a particular name (say "John") should be labeled as a person with a high probability (say 80%). Formally, this corresponds to a constraint feature $\phi = [[y = \texttt{PER}, x = \text{"John"}]]$ and a target $\tilde{\phi} = 0.8$. The model expectation of ϕ over the unlabeled data is then encouraged to match the target $\tilde{\phi}$, i.e. $\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}[\phi] \approx 0.8$. These constraint features and accompanying expectation targets are a very powerful and intuitive way to encode the supervision, and can save significant labeling effort. Note that the set of constraint features $\boldsymbol{\phi}$ may be disjoint from the model features $\mathbf{f}$, but often in practice the constraints are specified over a subset of the model features.

While previous work on constraint-based learning has proposed the use of variational inference in *Posterior Regularization* (Naseem et al., 2010), and the use of MCMC to leverage non-Markov constraints for sequence labeling (Bellare et al., 2009), in this paper we are interested in MCMC-based methods and complex, loopy models.

2.3 Generalized Expectations

A number of different approaches for learning with expectation constraints have been proposed. *Generalized Expectation (GE) Criteria* has been shown to be effective for estimating parameters of discriminative log-linear models without labeled data (Mann and McCallum, 2007; Druck et al., 2009). Although this approach can easily be combined with labeled data, in this work we will focus on the task of learning parameters without any labeled data. Further, we focus on the following "squared error" GE objective over the unlabeled data $\mathcal{D}_u$:

$$\max_{\boldsymbol{\theta}} \mathcal{O}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} - \sum_{\mathbf{x} \in \mathcal{D}_u} ||\tilde{\boldsymbol{\phi}} - \mathrm{E}_{p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}[\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})]||_2^2. \quad (3)$$

We use numerical optimization to maximize $\mathcal{O}(\boldsymbol{\theta})$. The gradient of $\mathcal{O}(\boldsymbol{\theta})$ is given by

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{O}(\boldsymbol{\theta}) = 2 \sum_{\mathbf{x} \in \mathcal{D}_u} (\tilde{\boldsymbol{\phi}} - \mathrm{E}_{p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}[\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})])^T \mathrm{COV}_{\boldsymbol{\theta}}(\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}), \mathbf{f}(\mathbf{x}, \mathbf{y})) \quad (4)$$

where $\mathrm{COV}_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \mathbf{f})$ is the covariance between constraint and model features. The covariance and the expectation terms in the gradient can only be computed exactly if the model features and the constraint features form a tree-shaped graph. Although this is true for many sequential models, many NLP tasks such as dependency parsing, coreference resolution, and relation extraction are better represented as models with long-range dependencies. Intractability of exact inference, and hence the computation of the covariance and expectation terms, restricts the utility of GE as a learning algorithm for these tasks. In this work we will study the extension of GE that can be applied to arbitrary graphical models.

3 Approximate Constraint-based Learning

In this section, we explore approaches that enable supervision from constraints for models for which exact inference is intractable.

3.1 MCMC for GE Training

In this paper, we approximate GE training using Markov chain Monte Carlo (MCMC) methods (Andrieu et al., 2003). In particular, once we obtain our set of samples from MCMC, $S = \{s : s \sim p(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})\}$, we use them to approximate expectations.

$$\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})}[\boldsymbol{\phi}(\mathbf{x}, \mathbf{y})] \approx \frac{1}{|S|} \sum_{\mathbf{y} \in S} \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \quad (5)$$

Similarly, we approximate the covariance:

$$\mathrm{COV}_{\boldsymbol{\theta}}(\boldsymbol{\phi}, \mathbf{f}) \approx \frac{1}{|S|} \sum_{\mathbf{y} \in S} \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \mathbf{f}(\mathbf{x}, \mathbf{y})^{\mathrm{T}} - \frac{1}{|S|} \sum_{\mathbf{y} \in S} \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) \frac{1}{|S|} \sum_{\mathbf{y} \in S} \mathbf{f}(\mathbf{x}, \mathbf{y})^{\mathrm{T}}. \quad (6)$$

By the law of large numbers, these estimates converge almost surely to the exact values as $|S| \to \infty$. In fact, it can be shown using the Hoeffding Inequality (Hoeffding, 1963) that the probability of the estimated mean in Equation 5 deviating from the true mean is exponentially small in $|S|$.

Though we do not explore this setting here, the proposed approach is also applicable when exact inference in the model is tractable, but the constraint features do not decompose in the same way as the model features.

3.1.1 Temperature

Mann and McCallum (2007) note that a label constraint with target distribution $[0.6, 0.4]$ can be satisfied by a model that assigns a label distribution of $[0.6, 0.4]$ to every output variable. As these low-entropy solutions are undesirable, the model probabilities may be modified with a *temperature*.

$$p_T(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) \propto \exp(\frac{1}{T}\boldsymbol{\theta} \cdot \mathbf{f}(\mathbf{x}, \mathbf{y}))$$

If $T < 1$, the model distribution becomes more peaked. As $T \to 0$, p_T approaches a distribution with all mass concentrated on the maximum probability output.

A temperature is also often used in MCMC to focus sampling effort around the mode of the distribution (Andrieu et al., 2003). Consequently, using a temperature $T < 1$ in MCMC GE could increase accuracy and may reduce the number of samples required to obtain an accurate approximation.

3.2 Constraint-Driven SampleRank (CDSR)

Singh et al. (2010) introduce an approach that incorporates supervision in form of constraints to the SampleRank algorithm. SampleRank (Wick et al., 2011) is a supervised training method that performs updates to the parameters *during* inference. The algorithm ensures that the model ranking of pairs of assignments is consistent with an objective function $\mathcal{F} : \mathcal{Y} \to \mathcal{R}$ defined using the labeled data. This approach is well-suited for MCMC as the pairs are generated during sampling.

CDSR defines a custom objective function that uses the constraints to define the ranking of assignments over unlabeled data. In particular, for a given configuration $\mathbf{y}, \mathbf{x}$, the objective is the sum of the targets of the constraint features that appear in $\mathbf{y}, \mathbf{x}$, i.e. $\mathcal{F}(\mathbf{y}, \mathbf{x}) = \tilde{\boldsymbol{\phi}} \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{y})$. Since CDSR only examines pairs of samples, it does not match the target expectations; instead the probabilities act as a proxy for constraint *strength*. Effectively, this results in the mode of the target distribution acting as a *hard* constraint. If multiple constraints apply, the constraints with higher probabilities dominate. Note that this is preferable to directly enforcing the constraints at test time; CDSR will propagate the supervision to learn parameters for model features that do not have a target.

4 Comparison on Sequence Labeling

The use of approximate inference in GE has not been studied. Therefore, in this section we investigate how the accuracy of the expectation estimates affects the accuracy of the trained model. We also compare exact and approximate GE to CDSR.

We use linear chain conditional random fields (CRFs; (Lafferty et al., 2001)) for these experiments to enable tractable comparison with exact GE. Note that we do not necessarily expect the approximate GE method to be faster in this setting, as both exact inference and sampling can be performed in linear time (in the number of variables) for linear chains. When inference is NP-hard (such as for loopy models like the entity resolution model in Section 5), the sampling approach enables constraint-based training where it would otherwise be infeasible.

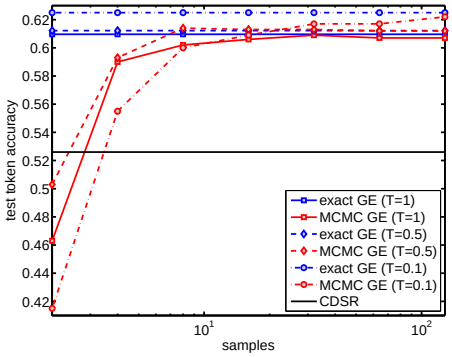
We use *Gibbs sampling* to sample from $p(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})$. To ensure the quality of samples, we have a burn-in period of 100 iterations, and do 10 iterations between each sample. We evaluate on the *apartment* listings data set, with features and processing as in Druck et al. (2009), and *CoNLL03* named entity recognition, with features and processing as in Ratinov and Roth (2009). We automatically select 40 feature constraints from labeled data (that have high mutual information with the true labels), and coarsen the target probabilities to be one of $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$, similar to Druck et al. (2009). For example, *apartments* constraints include `lease`$\rightarrow\{$*rent* : 1.0$\}$ and `bedroom`$\rightarrow\{$*features* : 0.4, *size* : 0.6$\}$[1]. We do not use any of the labels for training.

Figure 1 compares MCMC GE, with an increasing number of samples per sequence, to exact GE and CDSR (at convergence) on a held-out test set. Interestingly, we observe that a relatively small number of samples (~ 10) is sufficient to obtain an accurate model with MCMC GE training. With more samples, the approach quickly obtains similar accuracy to exact GE.

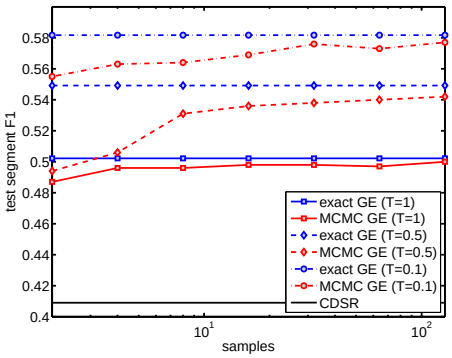
In both sets of experiments, we find that using a temperature $T < 1$ improves accuracy. However, using $T < 1$ in MCMC GE reduces the rate of convergence to exact GE, likely as a consequence of the known difficulty that MCMC methods have in transitioning between modes. This motivates the investigation of *annealing* procedures in which T varies over time, though initial attempts failed to yield improvements.

The accuracy obtained by CDSR at convergence is substantially lower than GE. Both of the tasks in our evaluation use constraints that have *soft* targets, i.e. many of the constraint features are weak indicators of the label, and the expectation

[1] In other words, if we encounter the word "bedroom" in an advertisement, it is likely to be talking about *features* or *size*, with the given probabilities, where *size* indicates that the text describes the number of rooms.



(a) Apartment Listings Segmentation



(b) CoNLL03 Named Entity Recognition

Figure 1: **Sequence Labeling:** Comparison of Exact GE with CDSR and MCMC GE as the temperature and the number of MCMC samples is varied.

targets need to match precisely for learning an accurate model. CDSR treats the constraints instead as hard constraints, and is unable to match the targets.

5 Lightly Supervised Entity Resolution

Entity resolution, an important component of the information extraction pipeline, is the task of partitioning mentions (or records) so that mentions that refer to the same entity are in the same partition. This task forms the basis of a number of problems such as within-document coreference, citation matching, entity disambiguation, document clustering, and so on. Labeling data for this task is incredibly difficult and time consuming because it involves either labeling every pair of men-

If the string match is $\geq 80\%$, then the citations match 85% of the time.
If the string match is $\leq 20\%$, then the citations do not match 99% of the time.

Table 1: Constraints used for citation resolution.

tions (quadratic), or identifying sets of mentions that are coreferent (exponential). Further, deciding whether two mentions are coreferent or not can require further exploration, such as conducting research on the web. In this section we perform entity resolution without labeled data using a discriminative model and two simple constraints. In particular, from the set of citations in the Cora[2] dataset, we resolve the citation strings that refer to the same paper.

5.1 Model

We represent the prediction task as a discriminative log-linear graphical model. The model consists of *set-valued* entity variables that take subsets of mentions as their values. Conversely, mentions are random variables that are defined over their entity assignments. The factors are defined over all pairs of mentions, resulting in a fully-connected model for which exact inference is intractable. The features for these factors represent the string similarity between the mention strings. We use the *similar title* and *venue* features as computed in Poon and Domingos (2007), as well as standard features based on the count and proportion of token matches in the citation strings.

5.2 Setup

As supervision, we select two simple constraints, displayed in Table 1. They encode the intuition that "two citations that have a high string overlap usually refer to the same paper", and that "two citations that have very lower string overlap almost never refer to the same paper". Target probabilities are estimated using a combination of domain knowledge and a cursory examination of the mentions.

We compare a number of approaches that learn using these constraints. First, our **Hard** baseline deterministically enforces the constraints from Table 1. For example, a pair of citations that have $\geq 80\%$ string match are required to be coreferent. There is no learning in this baseline. The MaxEnt (**ME**) approach estimates weights for the constraint features using maximum entropy estimation, with

[2] Available at http://alchemy.cs.washington.edu/

gradients computed using MCMC samples. Specifically, the objective for this approach is to maximize the entropy of model predictions while minimizing the L_2^2 difference between constraint feature and target expectations. Hence, it can only learn the parameters that correspond to the constraint features. In contrast, **GE** and **CDSR** additionally learn parameters for unconstrained model features. For GE, this transfer occurs via the covariance term discussed in Section 3.1.

Sampling is used to compute the gradients for ME and GE. We use 500 Metropolis-Hastings samples in each iteration. When using GE, the parameters are initialized using the final parameters from ME. For evaluation, we compute the pairwise evaluation metric by treating entity resolution as a binary classification task defined over pairs of mentions. Since this metric can often be misleading for practical problems, we also compute the B^3 metric (Bagga and Baldwin, 1998) that is widely used for entity disambiguation.

5.3 Results

The results of our experiments are shown in Table 2. Results with the Hard baseline demonstrate that using the constraints as simple rules during inference is not sufficient to obtain accurate predictions. Although ME learns parameters using the constraints, it performs poorly as well because it only learns parameters for the two constraint features. GE substantially outperforms the Hard and ME baselines in all metrics. The difference in performance between ME and GE can be attributed to the propagation of information to non-constraint features. CDSR also outperforms the ME and Hard baselines, and gives comparable results to GE (slightly better results on fold 1).

It is somewhat unexpected that GE and CDSR provide comparable results. Because CDSR does not observe the model's expectations, it cannot match the targets; instead it uses the modes of the target distribution as a hard constraint over individual assignments. This approximation has a substantial adverse effect with high entropy constraints (Section 4). However, because our constraints for entity resolution have targets close to $0/1$, the approximation that CDSR introduces does not have a significant adverse impact. We expect GE to outperform CDSR on tasks where the target expectations are less peaked, or where matching the targets precisely is required for high accuracy.

The fact that GE and CDSR yield high accuracy on an important task with very little supervision demonstrates the potential and importance of constraint-based training for complex models.

Method (Fold)	**Pairwise Metric** P / R	F1	**B^3 Metric** P / R	F1
Hard (0)	100.0 / 64.3	78.3	100.0 / 61.2	75.9
ME (0)	83.5 / 66.1	73.7	76.5 / 63.8	69.6
CDSR (0)	95.8 / 98.9	**97.4**	95.5 / 98.1	96.8
GE (0)	94.8 / 100.0	97.3	94.3 / 100.0	**97.1**
Hard (1)	99.1 / 56.7	72.2	99.0 / 62.1	76.3
ME (1)	80.8 / 60.7	69.3	81.6 / 67.7	74.0
CDSR (1)	99.2 / 91.1	**94.9**	98.1 / 92.4	**95.1**
GE (1)	87.5 / 99.4	93.0	88.9 / 98.7	93.5
Hard (2)	97.1 / 60.1	74.4	99.2 / 60.5	75.1
ME (2)	67.1 / 60.9	63.8	68.1 / 61.8	64.8
CDSR (2)	88.9 / 91.9	**90.4**	91.7 / 92.9	92.3
GE (2)	82.5 / 99.2	90.1	86.8 / 98.8	**92.4**

Table 2: **Entity Resolution:** on the Cora data, using constraints to learn the model.

Conclusion and Future Work

We studied two methods for constraint-based training of complex models. We approximated the expectations and covariances required for GE training using MCMC. Although MCMC methods are well known, the use of MCMC methods in constraint-based training has not been studied. We conducted sequence labeling experiments that demonstrated that a small number of samples is sufficient to approximate exact GE, and considerably outperform CDSR, a fast and approximate training approach. We also applied MCMC GE and CDSR to lightly supervised entity resolution, obtaining accurate models with only two simple, intuitive constraints. This novel application of constraint-based training demonstrates the potential and importance of the approaches.

Directions for future work include investigating efficient sampling schemes and exploring similar approximations for Posterior Regularization (Ganchev et al., 2010).

Acknowledgements

This work was supported in part by the Center for Intelligent Information Retrieval, in part by UPenn NSF medium IIS-0803847 and the University of Massachusetts gratefully acknowledges the support of Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Andrieu, C., de Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50:5–43.

Bagga, A. and Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Intl Conf on Comp Linguistics*.

Bellare, K., Druck, G., and McCallum, A. (2009). Alternating projections for learning with expectation constraints. In *Uncertainty in Artificial Intelligence (UAI)*, pages 35–42.

Chang, M., Ratinov, L., and Roth, D. (2007). Guiding semi-supervision with constraint-driven learning. In *Association for Computational Linguistics (ACL)*.

Druck, G., Settles, B., and McCallum, A. (2009). Active learning by labeling features. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ganchev, K., Graça, J., Gillenwater, J., and Taskar, B. (2010). Posterior regularization for structured latent variable models. *JMLR*, 11:2001–2049.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Kschischang, F. R., Frey, B. J., and Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

Mann, G. and McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. In *International Conference on Machine Learning (ICML)*.

Mann, G. and McCallum, A. (2008). Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Association for Computational Linguistics (ACL)*, pages 870–878.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1234–1244.

Poon, H. and Domingos, P. (2007). Joint inference in information extraction. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 913–918.

Poon, H. and Domingos, P. (2008). Joint unsupervised coreference resolution with markov logic. In *Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP '08, pages 650–659, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *CoNLL*, pages 147–155.

Singh, S., Yao, L., Riedel, S., and McCallum, A. (2010). Constraint-driven rank-based learning for information extraction. In *North American Chapter of the Association for Computational Linguistics and Human Language Technologies (NAACL HLT)*, pages 729–732.

Sutton, C. and McCallum, A. (2011). An introduction to conditional random fields. *Foundations and Trends in Machine Learning*. To appear.

Wellner, B., McCallum, A., Peng, F., and Hay, M. (2004). An integrated, conditional model of information extraction and coreference with application to citation matching. In *Uncertainty in Artificial Intelligence (UAI)*, pages 593–601.

Wick, M., Rohanimanesh, K., Bellare, K., Culotta, A., and McCallum, A. (2011). Samplerank: training factor graphs with atomic gradients. In *International Conference on Machine Learning (ICML)*.

Wick, M. L., Rohanimanesh, K., Schultz, K., and McCallum, A. (2008). A unified approach for schema matching, coreference and canonicalization. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, KDD '08, pages 722–730, New York, NY, USA. ACM.