Generalized Isotonic Conditional Random Fields

Yi Mao
College of Computing
Georgia Institute of Technology
yi.mao@cc.gatech.edu

Guy Lebanon
College of Computing
Georgia Institute of Technology
lebanon@cc.gatech.edu

Abstract

Conditional random fields are one of the most popular structured prediction models. Nevertheless, the problem of incorporating domain knowledge into the model is poorly understood and remains an open issue. We explore a new approach for incorporating a particular form of domain knowledge through generalized isotonic constraints on the model parameters. The resulting approach has a clear probabilistic interpretation and efficient training procedures. We demonstrate the applicability of our framework with an experimental study on sentiment prediction and information extraction tasks.

1 Introduction

The most common technique of estimating a distribution $p_\theta(x)$, $x \in \mathcal{X}, \theta \in \Theta$ based on iid samples $x^{(1)}, \ldots, x^{(n)} \sim p_{\theta_0}$ is to maximize the loglikelihood function $\ell(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(x^{(i)})$ i.e.,

$$\hat{\theta}^{\text{mle}}(x^{(1)}, \ldots, x^{(n)}) = \arg\max_{\theta\in\Theta} \ell(\theta). \tag{1}$$

The maximum likelihood estimator (MLE) $\hat{\theta}^{\text{mle}}$ enjoys many nice theoretical properties. In particular it is strongly consistent i.e. it converges to the true distribution $\hat{\theta}^{\text{mle}}(x^{(1)}, \ldots, x^{(n)}) \to \theta_0$ with probability 1 as $n \to \infty$. It is also asymptotically efficient which indicates that its asymptotic variance is the inverse Fisher information - the lowest possible variance according to the Cramer-Rao lower bound. These theoretical motivations, together with ample experimental evidence have solidified the role of the maximum likelihood estimate as the method of choice in many situations.

In some cases, additional information concerning the domain $\mathcal{X}$ is available which renders some parametric values $N \subset \Theta$ unrealistic. In the presence of this extra information the unrestricted maximum likelihood estimator (1) loses its appeal in favor of the constrained MLE

$$\hat{\theta}^{\text{cmle}}(x^{(1)}, \ldots, x^{(n)}) = \arg\max_{\theta\in\Theta\setminus N} \ell(\theta). \tag{2}$$

The constrained maximum likelihood (2) achieves a lower asymptotic error since its parameteric set is smaller (assuming its underlying assumption $\theta_0 \notin N$ is correct). Though it is defendable on frequentist grounds the constrained MLE is often given a Bayesian interpretation as the maximizer of the posterior under a prior assigning 0 probability to N and a uniform distribution over $\Theta \setminus N$.

The process of obtaining the set $\Theta \setminus N$ may rely on either domain knowledge or auxiliary dataset. In either case it is important to relate the constrained parametric subset $\Theta \setminus N$ to the corresponding set of possible probabilities

$$\mathcal{P}(\Theta \setminus N) = \{p_\theta(x) : \theta \in \Theta \setminus N\}.$$

Identifying p_θ as vectors of probabilities $(p_1, \ldots, p_{|\mathcal{X}|}) \in \mathbb{R}^{|\mathcal{X}|}$ we have that the constrained set of probabilities is a subset of the probability simplex $\mathcal{P}(\Theta \setminus N) \subset \mathbb{P}_\mathcal{X}$ where

$$\mathbb{P}_\mathcal{X} = \left\{ \left(p_1, \ldots, p_{|\mathcal{X}|}\right) : p_i \geq 0 \,, \sum_{i=1}^{|\mathcal{X}|} p_i = 1 \right\}.$$

Above, we assume that the space $\mathcal{X}$ is finite turning the simplex $\mathbb{P}_\mathcal{X}$ of all possible distributions over $\mathcal{X}$ into a subset of a finite dimensional vector space. We maintain this assumption, which is standard in many structured prediction tasks, throughout the paper in order to simplify the notation.

Expressing the constraints as a parametric subset $\Theta \setminus N$ is essential for deriving the constrained MLE estimator (2). Nevertheless, it is important to consider the corresponding subset of probabilities $\mathcal{P}(\Theta \setminus N)$ since it is much more interpretable for a domain expert and easy to test based on auxiliary data. In other words, it is much easier for a domain expert to specify constraints on the probabilities assigned by the model $\mathcal{P}(\Theta \setminus N)$ than constraints on abstract parameters $\Theta \setminus N$. The framework that we propose is thus to first specify the constrained probability set $\mathcal{P}(\Theta \setminus N)$ based on domain knowledge or auxiliary data, and then to convert it to $\Theta \setminus N$ in order to derive effective optimization schemes for the problem (2).

In many cases, the derivation of the set $\Theta \setminus N$ corresponding to $\mathcal{P}(\Theta \setminus N)$ is straightforward. For example in the case of the following simple exponential family model

$$p_\theta(x) = Z^{-1}(\theta) \exp\left(\sum_i \theta_i x_i\right) \quad x, \theta \in \mathbb{R}^d,$$

we have

$$p_\theta(x) > p_\theta(x') \quad \Leftrightarrow \quad \theta^\top (x - x') > 0. \tag{3}$$

In other cases, however, the conversion $\mathcal{P}(\Theta \setminus N) \Rightarrow \Theta \setminus N$ is highly non-trivial. This is also the case with conditional random fields which is the focus of this paper.

We thus consider, in this paper, the following problems in the context of conditional random fields

- Specifying a set of probability constraints $\mathcal{P}(\Theta \setminus N)$ based on domain knowledge or auxiliary data.
- Deriving the equivalent set of parametric constraints $\Theta \setminus N$.
- Deriving efficient algorithms for obtaining the constrained MLE.
- Experimental investigation of the benefit arising from the added constraints in the context of the structured prediction tasks of local sentiment analysis and information extraction.

In the next section we describe in more detail parametric constraints in the context of conditional random fields. Section 3 explores a particularly attractive set of constraints based on generalized isotonic constraints. In Section 4 we describe practical optimization schemes for obtaining the constrained MLE. We describe an experimental study on sentiment prediction and information extraction problems in Sections 6-7 and conclude with related work and discussion in Sections 8-9.

2 Structured Prediction and Conditional Random Fields

Structured prediction is the task of associating a sequence of labels $\mathbf{y} = (y_1, \ldots, y_n), y_i \in \mathcal{Y}$ with a sequence of observed values $\mathbf{x} = (x_1, \ldots, x_n), x_i \in \mathcal{X}$. Two examples are NLP tagging where x_i are words and y_i are morphological or syntactic tags, and image processing where x_i are the pixel brightness values and y_i indicate the segment or object the pixel belongs to.

Conditional random fields (CRF) [10] are parametric families of conditional distributions $p_\theta(\mathbf{y}|\mathbf{x})$ that correspond to joint Markov random fields $p(\mathbf{y},\mathbf{x})$ distributions

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{p(\mathbf{y},\mathbf{x})}{\sum_{\mathbf{y}} p(\mathbf{x},\mathbf{y})} = \frac{\prod_{C\in\mathcal{C}} \phi_C(\mathbf{x}|_C,\mathbf{y}|_C)}{Z(\mathbf{x},\theta)}. \tag{4}$$

Above, $\mathcal{C}$ is the set of cliques in a graph over $\mathcal{X}\times\mathcal{Y}$ and $\mathbf{x}|_C$ and $\mathbf{y}|_C$ are the restriction of $\mathbf{x}$ and $\mathbf{y}$ to variables representing nodes in the clique $C\in\mathcal{C}$. The functions ϕ_C are arbitrary positive-valued functions called clique potentials and $Z(\theta,\mathbf{x})$ represents the conditional normalization term ensuring $\sum_{\mathbf{y}} p_\theta(\mathbf{y}|\mathbf{x}) = 1$ for all $\mathbf{x},\theta$.

It is generally assumed that ϕ_C are exponential functions of features f_C modulated by decay parameters θ_C i.e.

$$\phi_C(\mathbf{x}|_C,\mathbf{y}|_C) = \exp\left(\sum_k \theta_{C,k} f_{C,k}(\mathbf{x}|_C,\mathbf{y}|_C)\right)$$

leading to the parametric family of conditional distributions

$$p_\theta(\mathbf{y}|\mathbf{x}) = Z^{-1}(\mathbf{x},\theta)\exp\left(\sum_{C\in\mathcal{C}}\sum_k \theta_{C,k} f_{C,k}(\mathbf{x}|_C,\mathbf{y}|_C)\right) \quad \theta_{C,k}\in\mathbb{R}. \tag{5}$$

CRF models have been frequently applied to sequence annotation, where $\mathbf{x} = (x_1,\ldots,x_n)$ is a sequence of words and $\mathbf{y} = (y_1,\ldots,y_n)$ is a sequence of labels annotating the words. The standard graphical structure in this case is a chain structure on $y_1,\ldots,y_n$ with noisy observations $\mathbf{x}$ leading to the clique structure $\mathcal{C} = \{\{y_{i-1},y_i\},\{y_i,\mathbf{x}\} : i = 1,\ldots,n\}$ (see Figure 5, left). Note that this graphical structure is more general than the original chain CRF [10] and includes it as a special case.

Together with the standard choice of feature functions this leads to the CRF model

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x},\theta)} \cdot \exp\left(\sum_{i=1}^{n}\sum_{\sigma,\tau\in\mathcal{Y}} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y_{i-1},y_i) + \sum_{i=1}^{n}\sum_{\sigma\in\mathcal{Y}}\sum_{k=1}^{l} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y_i,\mathbf{x},i)\right) \tag{6}$$

where $\theta = (\lambda,\mu)$ is the parameter vector and

$$f_{\langle\sigma,\tau\rangle}(y_{i-1},y_i) = 1_{\{y_{i-1}=\sigma\}}1_{\{y_i=\tau\}} \qquad \sigma,\tau\in\mathcal{Y} \tag{7}$$

$$g_{\langle\sigma,A_k\rangle}(y_i,\mathbf{x},i) = 1_{\{y_i=\sigma\}}A_k(\mathbf{x},i) \qquad \sigma\in\mathcal{Y}. \tag{8}$$

The values σ,τ correspond to arbitrary values of labels in $\mathcal{Y}$ and A_k corresponds to binary functions of both observation $\mathbf{x}$ and some position i in the sequence. The choice of A_k is problem dependent. A common practice of choosing $A_k(\mathbf{x},i) = 1_{\{x_i=w_k\}}, k = 1,\ldots,|\mathcal{X}|$ reduces the CRF model to its most traditional form measuring appearances of individual words in a vocabulary. More complex patterns of A_k may consider x_i as well as its neighbors x_{i-1} and x_{i+1} (e.g. $A_k(\mathbf{x},i) = 1_{\{x_i=w,x_{i-1}=w'\}}$ for some $w,w'\in\mathcal{X}$), or consider properties other than word appearance (e.g. $A_k(\mathbf{x},i) = 1_{\{x_i \text{ is capitalized}\}}$). The flexibility in the specification of A_k is the key advantage of CRF over generative sequential models such as hidden Markov models (HMMs). In particular, it enables the modeling of sequences of sentences rather than words as is the case of local sentiment prediction [11].

In the above formulation, we have $|\mathcal{Y}|^2$ feature functions $f_{\langle\sigma,\tau\rangle}$ measuring the transitions between successive label values and $|\mathcal{Y}|\cdot l$ feature functions $\{g_{\langle\sigma,A_k\rangle} : k = 1,\ldots,l,\sigma\in\mathcal{Y}\}$ describing observations $\mathbf{x}$ associated with label σ and function A_k. For the case of an m-order CRF where m is finite, it is possible to write the probabilistic model in the form of (6) by constructing $Y_i = (y_i,\ldots,y_{i+m-1})$, the ordered m-tuple of y_i values. Note, however, that not all transitions between states Y_i and Y_j are allowed for the m-order CRF.

Given a set of iid training samples $D = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) : i = 1, \ldots, m\}$ the parameters $\theta = (\lambda, \mu)$ are typically estimated by maximizing the regularized conditional likelihood

$$\ell(\theta|D) = \frac{1}{m} \sum_{i=1}^{m} \log p_\theta(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) + C\|\theta\|_2^2 \tag{9}$$

which corresponds to the posterior under a Gaussian prior over θ. The maximum likelihood estimation is usually carried out using standard numeric techniques such as iterative scaling, conjugate gradient, or quasi-Newton.

Unlike the situation in Markov random fields (Equation 3), the relationship between parameter and probability constraints in CRF is highly complicated. In particular, constraints over the probability vectors $p_\theta(\mathbf{y}|\mathbf{x}) \in [\alpha - \epsilon, \alpha + \epsilon]$ or $p_\theta(\mathbf{y}|\mathbf{x}) \geq p_\theta(\mathbf{y}'|\mathbf{x}')$ are not easily converted to the corresponding parametric constraints on θ. We explore in the next section several types of probability constraints that are intuitive and interpretable and yet correspond to simple ordering constraints on the parameters θ.

3 Ordered Domain Knowledge and Generalized Isotonic Constraints

In this section, we define a taxonomy of probability ordering constraints for CRF models based on probability ratios. These ordering constraints are intuitive and interpretable, and are easily specified using domain knowledge or auxiliary data. We derive the corresponding parameter constraints which we refer to as generalized isotonic constraints due to the similarity with the parameter constraints in the isotonic regression model [1].

Directly constraining the probability values assigned by the model

$$p_\theta(\mathbf{y}|\mathbf{x}) \in S \tag{10}$$

is impractical due to the large variability in the sequences $\mathbf{x}, \mathbf{y}$. It is difficult to imagine being able to ascertain probabilities of a large set of sequences $\mathbf{x}, \mathbf{y}$.

Another important difficulty in expressing direct probability constraints as in (10) is that it is hard to express domain knowledge in terms of absolute probabilities. Humans are notoriously bad at making statements concerning the probability of observing a particular event.

In this paper, we propose a novel set of probability constraints which eliminates the two difficulties mentioned above, and have a simple corresponding parameter constraints. We resolve the first difficulty by dealing with constraints involving variability in a local region of the graph. For example, in the sentiment prediction task [11] we consider the effect an appearance of a particular word such as `superb` has on the probability of it conveying positive sentiment. We resolve the second difficulty by constraining probabilities ratios involving a text sequence $\mathbf{x}$ and a locally perturbed version of it. As we shall see, such constraints depend only on the perturbed variables and are independent of the values of $\mathbf{x}$ on the remainder of the graph.

Formally, we define the probability constraints in terms of a probability ratio $p_\theta(\mathbf{y}|\mathbf{x})/p_\theta(\mathbf{y}|\mathbf{x}')$ where $\mathbf{x}'$ is identical to $\mathbf{x}$, except on a small graph neighborhood. Thus, instead of specifying the precise probability value, we specify whether the perturbation $\mathbf{x} \mapsto \mathbf{x}'$ increases or decreases the conditional probability of $\mathbf{y}$. Surprisingly, we show that constraining probability ratios corresponds to simple partial order constraints on the parameters or parameter differences.

In the case of linear chain CRF, if we restrict ourselves to perturbations $\mathbf{x} \mapsto \mathbf{x}'$ that modify only the j-component of $\mathbf{x}$ in a simple way, the choices of $\{y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n\}$ and $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ are immaterial making the probability ratio especially easy to assert and interpret.

We start with Proposition 1 below which relates the probability ratio to expectation over the parameters.

Proposition 1. *Let $\mathbf{x}$ be an arbitrary sequence over $\mathcal{X}$ and $\mathbf{x}'$ be identical to $\mathbf{x}$ except that $A_v(\boldsymbol{x}', j) = 1$ whereas $A_v(\boldsymbol{x}, j) = 0$. Then, for a linear chain CRF $p_\theta(\mathbf{y}|\mathbf{x})$ as in (6) we have*

$$\forall \mathbf{y} \qquad \frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x}')} = E_{p_\theta(\mathbf{y}'|\mathbf{x})} \exp\left(\mu_{\langle y'_j, A_v\rangle} - \mu_{\langle y_j, A_v\rangle}\right). \tag{11}$$

Proof.

$$\begin{aligned}
\frac{p_\theta(\mathbf{y}|\mathbf{x})}{p_\theta(\mathbf{y}|\mathbf{x}')} &= \frac{Z(\mathbf{x}',\theta)}{Z(\mathbf{x},\theta)} \frac{\exp\left(\sum_{i,\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y_{i-1},y_i) + \sum_{i,\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y_i,\mathbf{x},i)\right)}{\exp\left(\sum_{i,\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y_{i-1},y_i) + \sum_{i,\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y_i,\mathbf{x}',i)\right)} \\
&= \frac{Z(\mathbf{x}',\theta)}{Z(\mathbf{x},\theta)} \exp\left(-\mu_{\langle y_j,A_v\rangle}\right) \\
&= \exp\left(-\mu_{\langle y_j,A_v\rangle}\right) \cdot \frac{\sum_{\mathbf{y}'} \exp\left(\sum_{i,\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y'_{i-1},y'_i) + \sum_{i,\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y'_i,\mathbf{x}',i)\right)}{\sum_{\mathbf{y}'} \exp\left(\sum_{i,\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y'_{i-1},y'_i) + \sum_{i,\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y'_i,\mathbf{x},i)\right)} \\
&= \exp\left(-\mu_{\langle y_j,A_v\rangle}\right) \frac{\sum_{r\in\mathcal{Y}} \alpha_r(\mathbf{x}) \exp\left(\mu_{\langle r,A_v\rangle}\right)}{\sum_{r\in\mathcal{Y}} \alpha_r(\mathbf{x})} \\
&= \sum_{r\in\mathcal{Y}} \frac{\alpha_r(\mathbf{x})}{\sum_{r'\in\mathcal{Y}} \alpha_{r'}(\mathbf{x})} \exp\left(\mu_{\langle r,A_v\rangle} - \mu_{\langle y_j,A_v\rangle}\right) \\
&= \sum_{\mathbf{y}'} p_\theta(\mathbf{y}'|\mathbf{x}) \exp\left(\mu_{\langle y'_j,A_v\rangle} - \mu_{\langle y_j,A_v\rangle}\right)
\end{aligned}$$

where

$$\alpha_r(\mathbf{x}) = \sum_{\mathbf{y}':y'_j=r} \exp\left(\sum_{i,\sigma,\tau} \lambda_{\langle\sigma,\tau\rangle} f_{\langle\sigma,\tau\rangle}(y'_{i-1},y'_i) + \sum_{i,\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y'_i,\mathbf{x},i)\right).$$

□

Proposition 1 is used below to derive two types of probability ordering constraints and their corresponding parametric constraints.

3.1 One-way Ordering

In one way ordering the probability ratios defined in Proposition 1 are constrained to follow a partial order. This results in a simple ordering between the corresponding parameters.

Proposition 2. *Let $p_\theta(\mathbf{y}|\mathbf{x}), \mathbf{x}, \mathbf{x}'$ be as in Proposition 1. For all label sequences $\mathbf{s}, \mathbf{t}$, we have*

$$\frac{p_\theta(\mathbf{s}|\mathbf{x})}{p_\theta(\mathbf{s}|\mathbf{x}')} \geq \frac{p_\theta(\mathbf{t}|\mathbf{x})}{p_\theta(\mathbf{t}|\mathbf{x}')} \qquad \Longleftrightarrow \qquad \mu_{\langle t_j,A_v\rangle} \geq \mu_{\langle s_j,A_v\rangle}. \tag{12}$$

Proof. By Proposition 1 we have

$$\log \frac{p_\theta(\mathbf{s}|\mathbf{x})}{p_\theta(\mathbf{s}|\mathbf{x}')} - \log \frac{p_\theta(\mathbf{t}|\mathbf{x})}{p_\theta(\mathbf{t}|\mathbf{x}')} = \mu_{\langle t_j,A_v\rangle} - \mu_{\langle s_j,A_v\rangle}.$$

Since $p_\theta(\cdot|\mathbf{x})$, $p_\theta(\cdot|\mathbf{x}')$ are strictly positive, Equation (12) follows. □

Surprisingly, the probability ratio inequality in Proposition 2 is equivalent to an ordering of only two parameters $\mu_{\langle t_j,A_v\rangle} \geq \mu_{\langle s_j,A_v\rangle}$. What makes this remarkable is that only the j-components of the sequences $\mathbf{t}, \mathbf{s}, \mathbf{x}$ matter making the remaining components immaterial. In particular, we can consider $\mathbf{s}, \mathbf{t}$ that are identical except for their j-components. In this case the interpretation of the probability ratio constraint in Proposition 2 is as follows: the perturbation $\mathbf{x} \mapsto \mathbf{x}'$ increases the probability of s_j more than it does the probability of t_j. Since $\mathbf{s}, \mathbf{t}$ and $\mathbf{x}, \mathbf{x}'$ differ in only the j-components such probability ratio constraints are relatively easy to specify and interpret.

Given a set of probability ratio constraints as in Proposition 2, we obtain a partial order on the parameters $\{\mu_{\langle\tau,A_j\rangle} : \tau \in \mathcal{Y}, j = 1,\dots,l\}$ which corresponds to a partial order on the pairs $\{\langle\tau, A_j\rangle : \tau \in \mathcal{Y}, j = 1,\dots,l\}$ i.e.,

$$\langle\tau, A_j\rangle \geq \langle\sigma, A_k\rangle \quad \text{if} \quad \mu_{\langle\tau,A_j\rangle} \geq \mu_{\langle\sigma,A_k\rangle}. \tag{13}$$

In particular fixing a certain A_v we get a partial order on $\mathcal{Y}$ corresponding to the ordering of $\{\mu_{\langle\tau,A_v\rangle} : \tau \in \mathcal{Y}\}$. In the case of sentiment prediction, the elements of $\mathcal{Y}$ correspond to opinions such as very negative, negative, objective, positive, very positive, associated with the standard order. A complete specification of probability ratio constraints would result in a full ordering over $\{\mu_{\langle\tau,A_v\rangle} : \tau \in \mathcal{Y}\}$ for some v. In this case, assuming that A_v measures the presence of word v, we have that if v corresponds to a positive word (e.g. `superb`) we obtain the ordering

$$\mu_{\langle\tau_1,A_v\rangle} \geq \cdots \geq \mu_{\langle\tau_{|\mathcal{Y}|},A_v\rangle} \quad \text{where} \quad \tau_1 \geq \cdots \geq \tau_{|\mathcal{Y}|} \tag{14}$$

and the reverse ordering if v corresponds to a negative word (e.g. `horrible`)

$$\mu_{\langle\tau_1,A_v\rangle} \leq \cdots \leq \mu_{\langle\tau_{|\mathcal{Y}|},A_v\rangle} \quad \text{where} \quad \tau_1 \geq \cdots \geq \tau_{|\mathcal{Y}|}. \tag{15}$$

3.2 Two-way Ordering

Two-way ordering is similar to one-way ordering but, in addition to the activation of a certain feature, it involves the deactivation of a second feature. The following proposition describes the probability constraints more formally and derives the corresponding parameter constraints. The proof is similar to that of Proposition 2 and is omitted.

Proposition 3. *Let* $\mathbf{x}$ *be a sequence over* $\mathcal{X}$ *in which* $A_v(\mathbf{x}, j) = 1$ *and* $A_w(\mathbf{x}, j) = 0$ *and* $\mathbf{x}'$ *be identical to* $\mathbf{x}$ *except that* $A_v(\mathbf{x}', j) = 0$ *and* $A_w(\mathbf{x}', j) = 1$. *Then for a linear chain CRF* $p_\theta(\mathbf{y}|\mathbf{x})$ *as in* (6) *we have that for all* $\mathbf{s}, \mathbf{t}$,

$$\frac{p_\theta(\mathbf{s}|\mathbf{x})}{p_\theta(\mathbf{s}|\mathbf{x}')} \geq \frac{p_\theta(\mathbf{t}|\mathbf{x})}{p_\theta(\mathbf{t}|\mathbf{x}')} \quad \Leftrightarrow \quad \mu_{\langle t_j,A_w\rangle} - \mu_{\langle s_j,A_w\rangle} \geq \mu_{\langle t_j,A_v\rangle} - \mu_{\langle s_j,A_v\rangle}. \tag{16}$$

In a similar way to the one-way ordering, the parameter constraint depends only on the j-components of $\mathbf{s}, \mathbf{t}$ and thus to aid the interpretation we can select $\mathbf{s}, \mathbf{t}$ that are identical except for s_j, t_j. The probability ratio constraint then measures whether perturbing $\mathbf{x} \mapsto \mathbf{x}'$ increases the probability of s_j more than that of t_j. However, in contrast to the one-way ordering the perturbation $\mathbf{x} \mapsto \mathbf{x}'$ involves deactivating the feature A_v and activating A_w. For example in the case of sentiment prediction these features could correspond to the replacement of word v in the j-position with word w.

In contrast to the one-way ordering, a collection of probability ratio constraints in Proposition 3 do not correspond to full or partial ordering on the model parameters. Instead they correspond to a full or partial order on the set of all pairwise differences between the model parameters.

One-way and two-way probability ratio constraints are complimentary in nature and they are likely to be useful in a wide variety of situations. In the case of elicitation from domain knowledge they provide a general framework for asserting statements that are immediately translatable to parameter constraints.

We conclude this section with the following observations regarding possible generalizations of the one-way and two-way constraints

(1) The definition of $f_{\langle\sigma,\tau\rangle}$ in (7) may be extended to a more general form $f_{\langle\sigma,\tau,B_k\rangle}(y_{i-1}, y_i, \mathbf{x}, i) = 1_{\{y_{i-1}=\sigma\}} 1_{\{y_i=\tau\}} B_k(\mathbf{x}, i)$ where B_k are some binary functions of observation $\mathbf{x}$. Without loss of generality, we assume that the set $\{A_k\}$ and $\{B_k\}$ are disjoint. Otherwise, they can be made disjoint by defining a set of new parameters $\lambda_{\sigma,\tau,B_k} \leftarrow \lambda_{\sigma,\tau,B_k} + \mu_{\tau,B_k}$ corresponding to $f_{\sigma,\tau,B_k} \leftarrow f_{\sigma,\tau,B_k} + g_{\tau,B_k}$ for functions that appear in both $\{A_k\}$ and $\{B_k\}$. It is then straightforward to modify Proposition 1 - 3 with respect to parameters $\lambda_{\sigma,\tau,B_k}$.

(2) The simple form of parameter constraints on the right hand side of(12) and (16) results from the fact that only the j-components of the sequences matter in computing the probability ratio (11). For perturbations $\mathbf{x} \mapsto \mathbf{x}'$ involving labels from multiple positions in the sequence, the probability ratio constraints become linear parameter constraints with coefficients 1 or -1. These linear constraints are still considered simple, but they lose the intuitive ordering interpretation and are not the focus of this paper.

4 Algorithms and Optimization

Conceptually, the parameter estimates for generalized isotonic CRF may be found by maximizing the likelihood or posterior subject to a collection of constraints of type (12) or (16). Since the constraints form a convex feasible set, the constrained MLE becomes a convex optimization problem with a unique global optimum. Unfortunately, due to the large number of possible constraints, a direct incorporation of them into a numerical maximizer is a relatively difficult task. We propose a re-parameterization of the CRF model that simplifies the constraints and converts the problem to a substantially easier constrained optimization problem. The re-parameterization, in the case of fully ordered parameter constraints is relatively straightforward. In the more general case of constraints forming a partially ordered set we need the mechanism of Möbius inversions on finite partially ordered sets.

The re-parameterization is based on the partial order on pairs $\{\langle \tau, A_j \rangle : \tau \in \mathcal{Y}, j = 1, \ldots, l\}$ defined in (13). Instead of enforcing the constraints on the original parameters $\mu_{\langle \tau, A_j \rangle}$, we reparameterize the model by introducing a new set of features $\{g^*_{\langle \sigma, A_k \rangle} : \sigma \in \mathcal{Y}, k = 1, \ldots, l\}$ defined as

$$g^*_{\langle \sigma, A_k \rangle}(y_i, x_i) = \sum_{\langle \tau, A_j \rangle : \langle \tau, A_j \rangle \geq \langle \sigma, A_k \rangle} g_{\langle \tau, A_j \rangle}(y_i, x_i) \tag{17}$$

and a new set of corresponding parameters $\mu^*_{\langle \sigma, A_k \rangle}$ satisfying the equality

$$\sum_{\sigma, k} \mu_{\langle \sigma, A_k \rangle} g_{\langle \sigma, A_k \rangle} = \sum_{\sigma, k} \mu^*_{\langle \sigma, A_k \rangle} g^*_{\langle \sigma, A_k \rangle} \tag{18}$$

and leading to the re-parameterized CRF

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \theta)} \cdot \exp\left(\sum_i \sum_{\sigma, \tau} \lambda_{\langle \sigma, \tau \rangle} f_{\langle \sigma, \tau \rangle}(y_{i-1}, y_i) + \sum_i \sum_{\sigma, k} \mu^*_{\langle \sigma, A_k \rangle} g^*_{\langle \sigma, A_k \rangle}(y_i, x_i) \right). \tag{19}$$

Obtaining the maximum likelihood for the reparameterized model (19) instead of the original model has the benefit of converting the complex partial orders in (13) to simple non-negativity constraints $\mu^*_{\langle \sigma, A_k \rangle} \geq 0$ for a subset of the new parameters $\{\mu^*_{\langle \sigma, A_k \rangle} : \sigma \in \mathcal{Y}, k = 1, \ldots, l\}$. As a result, solving the constrained MLE problem on the reparameterized model (19) is substantially simpler to implement and is more efficient computationally. The constrained MLE can be computed in practice using a trivial adaptation of gradient based methods such as conjugate gradient or quasi-Newton.

The parameters $\mu^*_{\langle \sigma, A_k \rangle}$ may be obtained from the original parameters by convolving $\mu_{\langle \sigma, A_k \rangle}$ with the Möbius function of the partially ordered set (13). The reparameterization (19) is justified by the Möbius inversion theorem which states that $\mu^*_{\langle \sigma, A_k \rangle}$ satisfy

$$\mu_{\langle \sigma, A_k \rangle} = \sum_{\langle \tau, A_j \rangle : \langle \tau, A_j \rangle \leq \langle \sigma, A_k \rangle} \mu^*_{\langle \tau, A_j \rangle}. \tag{20}$$

In the case of two-way ordering, we have ordering on parameter differences rather than the parameters themselves. The mechanism of Möbius inversions can still be applied, but over a transformed feature space

instead of the original feature space. In particular, for (t_j, s_j, A_w, A_v) that satisfy (16), we apply the re-parameterization described in (17) - (19) to the feature functions $\tilde{g}$ defined by

$$\tilde{g}_{\langle t_j, A_v\rangle} = g_{\langle t_j, A_v\rangle} \qquad \tilde{g}_{\langle s_j, A_v\rangle} = g_{\langle s_j, A_v\rangle} + g_{\langle t_j, A_v\rangle}$$
$$\tilde{g}_{\langle t_j, A_w\rangle} = g_{\langle t_j, A_w\rangle} \qquad \tilde{g}_{\langle s_j, A_w\rangle} = g_{\langle s_j, A_w\rangle} + g_{\langle t_j, A_w\rangle}$$

and parameters $\tilde{\mu}$ defined by

$$\tilde{\mu}_{\langle s_j, A_v\rangle} = \mu_{\langle s_j, A_v\rangle} \qquad \tilde{\mu}_{\langle t_j, A_v\rangle} = \mu_{\langle t_j, A_v\rangle} - \mu_{\langle s_j, A_v\rangle}$$
$$\tilde{\mu}_{\langle s_j, A_w\rangle} = \mu_{\langle s_j, A_w\rangle} \qquad \tilde{\mu}_{\langle t_j, A_w\rangle} = \mu_{\langle t_j, A_w\rangle} - \mu_{\langle s_j, A_w\rangle}.$$

More information concerning the Möbius inversion theorem for partially ordered sets may be found in standard textbooks on combinatorics, for example [20].

5 Elicitation of Constraints

There are two ways in which probability constraints such as the ones in Propositions 2 and 3 can be elicited. The first is by eliciting domain knowledge from experts. This is similar to prior elicitation in subjective Bayesian analysis, but has the advantage that the knowledge is specified in terms of probability ratios, rather than model parameters.

The second way to elicit probability constraints is by relying on auxiliary data. The auxiliary data should be related to the domain on which inference is conducted, but does not have to have the same distribution as the training data. Automatic elicitation results in probability ratios satisfying inequalities or more generally having values in some sets. As such, some amount of inconsistency between the auxiliary data and the train and test data is permissible. For example, in sentiment prediction modeling of a particular author, we may have auxiliary data written by another author. In information extraction we may have a secondary corpus from a different source whose label taxonomy is related to the primary dataset.

Inferring probability constraints concerning the full conditionals $p_\theta(\mathbf{y}|\mathbf{x})$ from data is difficult due to the fact that each sequence $\mathbf{x}$ or $\mathbf{y}$ appears only once or a small number of times. The approach below makes some conditional independence assumptions which will simplify the elicitation to the problem of ordering probability ratios of univariate conditional distributions $p(A_v|t_j)/p(A_w|t_j)$.

Proposition 4. *Let $\mathbf{x}, \mathbf{x}'$ be as in Proposition 1 and $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p_\theta(\mathbf{y}|\mathbf{x})$ where $p_\theta(\mathbf{y}|\mathbf{x})$ is a CRF model and $p(x_j|y_j)$ is being modeled by[1] $p\left(\cap_{k\in I} A_k \mid y_j\right)$, $I = \{k \in \{1, \ldots, l\} : x_j \in A_k\}$, satisfying the following conditional independencies*

$$p\left(\bigcap_{k\in I} A_k \mid y_j\right) = \prod_{k\in I} p\left(A_k \mid y_j\right). \tag{21}$$

If the CRF model satisfies (12) *then*

$$p(A_v|t_j) \geq p(A_v|s_j). \tag{22}$$

Proof. We have

$$\begin{aligned} \text{LHS of (12)} \quad &\Rightarrow \quad \frac{p_\theta(\mathbf{s}|\mathbf{x})}{p_\theta(\mathbf{t}|\mathbf{x})} \geq \frac{p_\theta(\mathbf{s}|\mathbf{x}')}{p_\theta(\mathbf{t}|\mathbf{x}')} \quad \Rightarrow \quad \sum_{\mathbf{s}} \frac{p_\theta(\mathbf{s}|\mathbf{x})}{p_\theta(\mathbf{t}|\mathbf{x})} \geq \sum_{\mathbf{s}} \frac{p_\theta(\mathbf{s}|\mathbf{x}')}{p_\theta(\mathbf{t}|\mathbf{x}')} \\ &\Rightarrow \quad \sum_{\mathbf{t}} \frac{p_\theta(\mathbf{t}|\mathbf{x})}{\sum_{\mathbf{s}} p_\theta(\mathbf{s}|\mathbf{x})} \leq \sum_{\mathbf{t}} \frac{p_\theta(\mathbf{t}|\mathbf{x}')}{\sum_{\mathbf{s}} p_\theta(\mathbf{s}|\mathbf{x}')} \Rightarrow \quad \frac{\alpha_{t_j}(\mathbf{x})}{\alpha_{s_j}(\mathbf{x})} \leq \frac{\alpha_{t_j}(\mathbf{x}')}{\alpha_{s_j}(\mathbf{x}')} \end{aligned} \tag{23}$$

[1]We implicitly assume here that sequences $\mathbf{x}$ are identified by their feature signature i.e. the feature functions constitute a 1-1 mapping. In some cases this does not hold and some correction term is necessary.

where the summations are over all label sequences $\mathbf{s}, \mathbf{t}$ having fixed j-components s_j, t_j. See Proposition 1 for a definition of $\alpha_{t_j}, \alpha_{s_j}$.

Due to the conditional independencies expressed in the graphical structure of CRF, the α_r functions satisfy

$$\begin{aligned}\alpha_r(\mathbf{x})/Z(\mathbf{x}) &= \sum_{\mathbf{y}:y_j=r} p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{y}:y_j=r} \frac{p(\mathbf{y},\mathbf{x})}{p(\mathbf{x})} = \sum_{y_{-j}} \frac{p(y_{-j}, y_j = r, \mathbf{x})}{p(\mathbf{x})} \\ &= \frac{p(y_j = r, x_j, x_{-j})}{p(\mathbf{x})} = \frac{p(x_j|y_j = r)p(y_j = r|x_{-j})p(x_{-j})}{p(\mathbf{x})} \end{aligned} \tag{24}$$

where $y_{-j} = \{y_1, \ldots, y_n\} \setminus \{y_j\}$ and $x_{-j} = \{x_1, \ldots, x_n\} \setminus \{x_j\}$.

Substituting (24) in (23) and using the fact that for all $r \in \mathcal{Y}$, $p(y_j = r|x_{-j}) = p(y_j = r|x'_{-j})$ we get

$$\begin{aligned}(23) \quad &\Rightarrow \quad \frac{p(x_j|t_j)}{p(x_j|s_j)} \leq \frac{p(x'_j|t_j)}{p(x'_j|s_j)} \quad \Rightarrow \quad \frac{p(x_j|t_j)}{p(x'_j|t_j)} \leq \frac{p(x_j|s_j)}{p(x'_j|s_j)} \\ &\Rightarrow \quad \frac{p\left(\cap_{k\in I} A_k|t_j\right)}{p\left(\left(\cap_{k\in I} A_k\right) \cap A_v|t_j\right)} \leq \frac{p\left(\cap_{k\in I} A_k|s_j\right)}{p\left(\left(\cap_{k\in I} A_k\right) \cap A_v|s_j\right)} \\ &\Rightarrow \quad p(A_v|t_j) \geq p(A_v|s_j)\end{aligned}$$

where the last implication comes from the conditional independence assumption (21). □

A similar result holds for two-way ordering whose proof is omitted.

Proposition 5. *Under the same conditions as Proposition 4, if the CRF model satisfies* (16) *then*

$$\frac{p(A_v|t_j)}{p(A_w|t_j)} \leq \frac{p(A_v|s_j)}{p(A_w|s_j)} \quad \Rightarrow \quad \frac{p(t_j|A_v)}{p(t_j|A_w)} \leq \frac{p(s_j|A_v)}{p(s_j|A_w)}. \tag{25}$$

Constraints such as (22) or the equivalent (but sometimes easier to estimate)

$$\frac{p(t_j|A_v)}{p(t_j)} \geq \frac{p(s_j|A_v)}{p(s_j)} \tag{26}$$

can be obtained from auxiliary data based on hypothesis tests. More specifically, Equations (25),(26) can be written as $\psi \geq 1$, where ψ is estimated by the odds ratio of a 2×2 continency table obtained from the co-occurrence of a label (s_j or t_j) and a set (A_v or A_w). This can be achieved by a test of independence in a 2×2 table, such as an asymptotic test based on the test statistic $(\log \hat{\psi} - \log \psi)/\text{se}(\log \hat{\psi}) \approx \mathcal{N}(0, 1)$ where se is the standard error [19].

For a large number of constraints, a collection of hypothesis tests can be performed offline. Selecting a certain value as threshold we can order the hypothesis by their p-values and select the ones whose p-values are less than the threshold.

The derivations above are based on the conditional independence assumption (21) which may be too restrictive for arbitrary feature sets. However, in our experiments we found that the constraints identified automatically by hypothesis tests normally overlap with those returned by domain experts. Moreover, even if domain experts are available and human elicitation is taking place, the automatic elicitation described above can substantially reduce human intervention as it can be used to pre-filter a large set of unnecessary features.

6 Sentiment Prediction

Many documents, such as reviews and blogs, are written with the purpose of conveying a particular opinion or sentiment. Other documents may not be written with the purpose of conveying an opinion, but nevertheless

they contain one. Opinions, or sentiments, may be considered in several ways, the simplest of which is varying from positive opinion, through neutral, to negative opinion.

We distinguish between the tasks of global sentiment prediction and local sentiment prediction. Global sentiment prediction is the task of predicting the sentiment of the document based on the word sequence. Local sentiment prediction [11] is the task of predicting a sequence of sentiments $\mathbf{y} = (y_1, \ldots, y_n), y_i \in \mathcal{Y}$ based on a sequence of sentences $\mathbf{x} = (x_1, \ldots, x_n)$. In this case, each sentiment measures the local sentiment of the sentence x_i in the document.

Previous research on sentiment prediction has generally focused on predicting the sentiment of the entire document. A commonly used application is the task of predicting the number of stars assigned to a movie, based on a review text. Typically, the problem is considered as standard multiclass classification or regression using the bag of words representation.

In addition to the sentiment of the entire document, which we call global sentiment, we define the concept of local sentiment as the sentiment associated with a particular part of the text. It is reasonable to assume that the global sentiment of a document is a function of the local sentiment and that estimating the local sentiment is a key step in predicting the global sentiment. Moreover, the concept of local sentiment is useful in a wide range of text analysis applications including document summarization and visualization.

Formally, we view local sentiment as a function on the sentences in a document taking values in a finite partially ordered set, or a poset, $(\mathcal{Y}, \leq)$. To determine the local sentiment at a particular word, it is necessary to take context into account. For example, due to context the local sentiment at each of the following words `this is a horrible product` is low. Since sentences are natural components for segmenting document semantics, we view local sentiment as a piecewise constant function on sentences. Occasionally we encounter a sentence that violates this rule and conveys opposing sentiments in two different parts. In this situation we break the sentence into two parts and consider them as two sentences. We therefore formalize the problem as predicting a sequence of sentiments $\mathbf{y} = (y_1, \ldots, y_n), y_i \in \mathcal{Y}$ based on a sequence of sentences $\mathbf{x} = (x_1, \ldots, x_n)$ where we consider each sentence as a bag of words $x_i = \{w_{i1}, \ldots, w_{il_i}\}$.

We examine the performance of the CRF model in the local sentiment task and the benefit arising from incorporating parameter constraints through auxiliary data and domain knowledge. The CRF is based on Equation (6) with the feature functions $A_k(\mathbf{x}, i) = 1_{\{w_k \in x_i\}}$ that measure the appearance of vocabulary words in each sentence. The dataset that we use contains 249 movie reviews, randomly selected from the Cornell sentence polarity dataset v1.0[2], all written by the same author. The local sentiment labeling was performed manually by the author by associating with each sentence one of the following sentiment values $\mathcal{Y} = \{-2, -1, 0, 1, 2\}$ where 2 corresponds to highly praised, 1 corresponds to something good, 0 corresponds to objective description, -1 corresponds to something that needs improvement, and -2 corresponds to strong aversion.

6.1 Sentence Level Sentiment Prediction

Figure 1 displays the testing accuracy of local sentiment prediction both as a function of the training data size and as a function of the number of constrained words averaged over 40 train-test splits. In all cases, limited memory BFGS was used to train the CRF. The constraints were enforced using the barrier method. The objective function was the regularized MLE with a Gaussian prior on θ with variance 10.

The dataset presents one particular difficulty where more than 75% of the sentences are labeled objective (or 0). As a result, the prediction accuracy for objective sentences is over-emphasized. To correct for this fact, we report our results by averaging the test-set performance for each individual label. Note that since there are 5 labels, random guessing yields a baseline of 0.2 accuracy.

As described in Section 5, for one-way ordering, we obtained 500 words from an auxiliary data set that received the smallest p values in a test of (26) to set the constraints (12). The auxiliary data set is the additional 201 movie reviews from a second author described in 6.3. Table 1 displays the top 15 positive and negative words.

[2]Available at `http://www.cs.cornell.edu/People/pabo/movie-review-data`

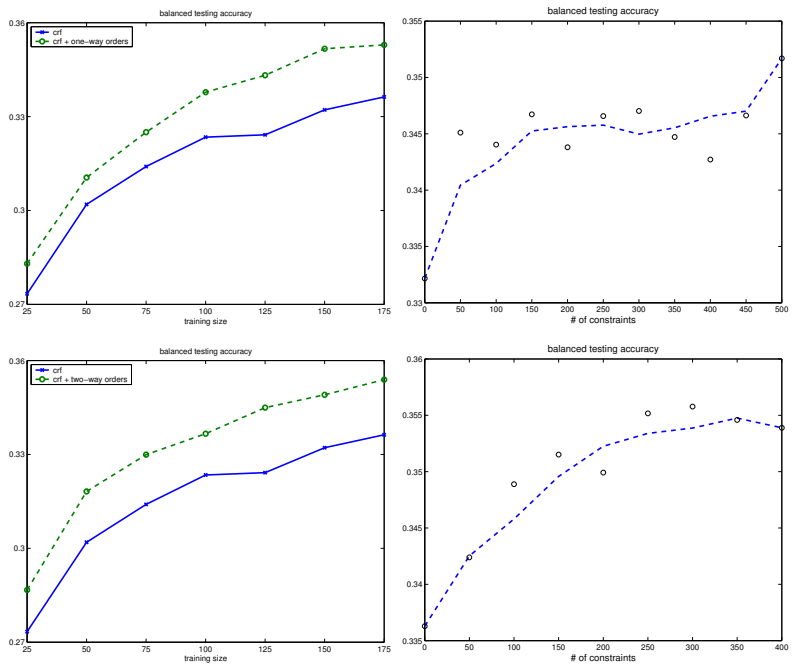

Figure 1: Balanced test accuracy for local sentiment prediction both as a function of training size (left column) and as a function of number of constrained words (right column). 500 words that received the smallest p values in a test of (26) are subject to one-way ordering (top row). 400 pairs of words that received the smallest p values in a test of (25) are subject to two-way ordering (bottom row). Blue lines in the right column are obtained by smoothing the data (represented by black circles). In this case, the training size is fixed to be 150.

Similarly, we may apply a test of (25) on the auxiliary data set to get pairs of words for setting constraints of (16). Figure 2 shows a portion of the graph by connecting a pair of ordered words with a line and drawing the higher ordered words above. A total of 400 pairs of words are selected for two-way ordering constraints in Figure 1 (bottom left).

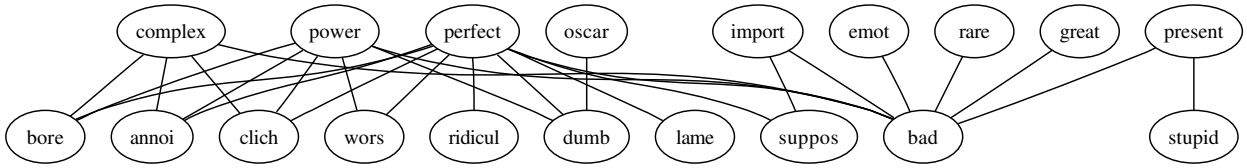

Figure 2: Ordering of stemmed words with respect to the positive sentiment. The words with higher order are drawn at the top.

The results in Figure 1 indicate that by incorporating either one-way or two-way ordering information, the generalized isotonic CRF perform consistently better than regular CRF. The advantage of incorporating sequential information in sentiment prediction has already been demonstrated in [11] and we therefore omit results comparing CRF and isotonic CRF with non-sequential models such as naive Bayes or SVM here.

We note that the information provided by one-way and two-way ordering are somewhat overlapping. For example, setting the words `great` and `bad` for one-way ordering automatically implies that the word pair

great	perfect	power	love	complex
import	emot	present	fascin	rare
oscar	true	simpl	polit	beauti
bad	suppos	bore	stupid	wors
dumb	minut	tediou	annoi	wrong
bland	ridicul	worst	lifeless	lame

Table 1: Lists of first 15 positive or negative stemmed words with the smallest p values.

(`great`, `bad`) satisfies the two-way ordering. We therefore avoid considering generalized isotonic CRF with mixed constraint types.

6.2 Global Sentiment Prediction

We also evaluated the contribution of the local sentiment analysis in helping to predict the global sentiment of documents. The sentence-based definition of sentiment flow is problematic when we want to fit a model that uses sentiment flows from multiple documents. Different documents have different number of sentences and it is not clear how to compare them or how to build a model from a collection of discrete flows of different lengths. We therefore convert the sentence-based flow to a smooth length-normalized flow that can meaningfully relate to other flows.

In order to account for different lengths, we consider the sentiment flow as a function $h : [0, 1] \to \mathcal{Y} \subset \mathbb{R}$ that is piecewise constant on the intervals $[0, l), [l, 2l), \ldots, [(k-1)l, 1]$ where k is the number of sentences in the document and $l = 1/k$. Each of the intervals represents a sentence and the function value on it is its sentiment.

To create a more robust representation we smooth out the discontinuous function by convolving it with a smoothing kernel. The resulting sentiment flow is a smooth curve $f : [0, 1] \to \mathbb{R}$ that can be easily related or compared to similar sentiment flows of other documents (see Figure 3 for an example). We can then define natural distances between two flows, for example the L_p distance

$$d_p(f_1, f_2) = \left(\int_0^1 |f_1(r) - f_2(r)|^p \, dr \right)^{1/p} \tag{27}$$

for use in a distance based classifier that predicts the global sentiment.

We compared a nearest neighbor classifier for the global sentiment, where the representation varied from bag of words to smoothed length-normalized local sentiment representation (with and without objective sentences). The smoothing kernel was a bounded Gaussian density (truncated and renormalized) with $\sigma^2 = 0.2$. Figure 3 displays discrete and smoothed local sentiment labels, and the smoothed sentiment flow predicted by isotonic CRF.

Figure 4 and Table 2 display test-set accuracy of global sentiment prediction as a function of the train set size. The distance in the nearest neighbor classifier was either L_1 or L_2 for the bag of words representation or their continuous version (27) for the smoothed sentiment curve representation. The results indicate that the classification performance of the local sentiment representation is better than the bag of words representation. In accordance with the conclusion of [15], removing objective sentences (that correspond to sentiment 0) increased the local sentiment analysis performance by 20.7%. We can thus conclude that for the purpose of global sentiment prediction, the local sentiment flow of the non-objective sentences holds most of the relevant information.

6.3 Measuring the rate of sentiment change

Thus far, we have ignored the dependency of the labeling model $p_\theta(\mathbf{y}|\mathbf{x})$ on the author, denoted here by the variable a. We now turn to account for different sentiment-authoring styles by incorporating this variable

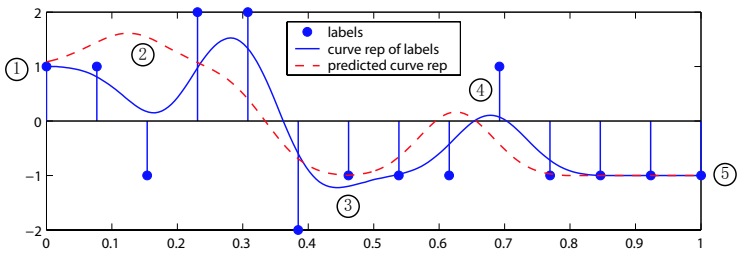

Figure 3: Sentiment flow and its smoothed curve representation. The blue circles indicate the labeled sentiment of each sentence. The blue solid curve and red dashed curve are smoothed representations of the labeled and predicted sentiment flows. Only non-objective labels are kept in generating the two curves. The numberings correspond to sentences displayed in Section 6.4.

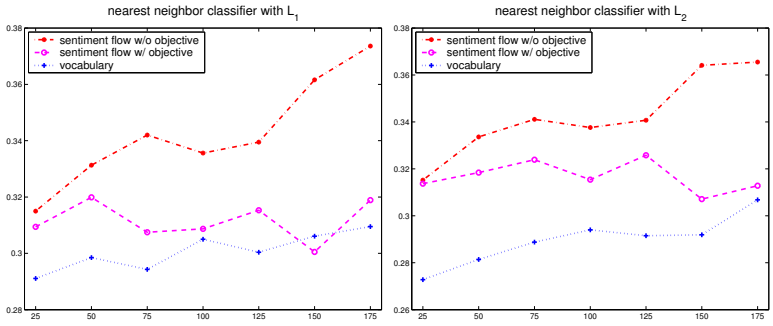

Figure 4: Accuracy of global sentiment prediction (4-class labeling) as a function of train set size.

into the model. The word emissions $y_i \rightarrow w_i$ in the CRF structure are not expected to vary much across different authors. The sentiment transitions $y_{i-1} \rightarrow y_i$, on the other hand, typically vary across different authors as a consequence of their individual styles. For example, the review of an author who sticks to a list of self-ranked evaluation criteria is prone to strong sentiment variations. In contrast, the review of an author who likes to enumerate pros before he gets to cons (or vice versa) is likely to exhibit more local homogeneity in sentiment.

Accounting for author-specific sentiment transition style leads to the graphical model in Figure 5. The corresponding author-dependent CRF model

$$p_\theta(\mathbf{y}|\mathbf{x}, a) = \frac{1}{Z(\mathbf{x}, a)} \exp\left(\sum_{i,a'}\sum_{\sigma,\tau}\left(\lambda_{\langle\sigma,\tau\rangle} + \lambda_{\langle\sigma,\tau,a'\rangle}\right) f_{\langle\sigma,\tau,a'\rangle}(y_{i-1}, y_i, a) + \sum_i \sum_{\sigma,k} \mu_{\langle\sigma,A_k\rangle} g_{\langle\sigma,A_k\rangle}(y_i, \mathbf{x}, i)\right)$$

uses features $f_{\langle\sigma,\tau,a'\rangle}(y_{i-1}, y_i, a) = f_{\langle\sigma,\tau\rangle}(y_{i-1}, y_i)\delta_{a,a'}$ and transition parameters that are author-dependent $\lambda_{\langle\sigma,\tau,a\rangle}$ as well as author-independent $\lambda_{\langle\sigma,\tau\rangle}$. Setting $\lambda_{\langle\sigma,\tau,a\rangle} = 0$ reduces the model to the standard CRF model. The author-independent parameters $\lambda_{\langle\sigma,\tau\rangle}$ allow parameter sharing across multiple authors in case the training data is too scarce for proper estimation of $\lambda_{\langle\sigma,\tau,a\rangle}$. For simplicity, the above ideas are described

	L_1		L_2	
vocabulary	0.3095		0.3068	
sentiment flow with objective sentences	0.3189	3.0%	0.3128	1.95%
sentiment flow without objective sentences	0.3736	20.7%	0.3655	19.1%

Table 2: Accuracy results and relative improvement when training size equals 175.

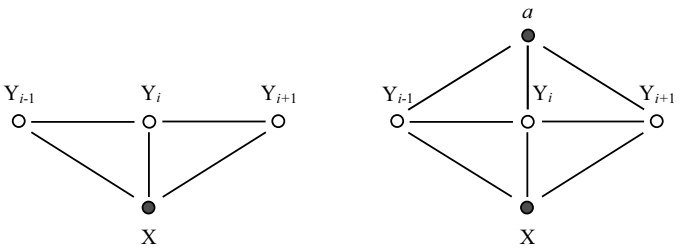

Figure 5: Graphical models corresponding to CRF (left) and author-dependent CRF (right).

in the context of non-isotonic CRF. However, it is straightforward to combine author-specific models with generalized isotonic restrictions.

We examine the rate of sentiment change as a characterization of the author's writing style using the isotonic author-dependent model. We assume that the CRF process is a discrete sampling of a corresponding continuous time Markov jump process. A consequence of this assumption is that the time T the author stays in sentiment σ before leaving is modeled by the exponential distribution $p_\sigma(T > t) = e^{-q_\sigma(t-1)}, t > 1$. Here, we assume $T > 1$ and q_σ is interpreted as the rate of change of the sentiment $\sigma \in \mathcal{Y}$: the larger the value, the more likely the author will switch to other sentiments in the near future.

To estimate the rate of change q_σ of an author we need to compute $p_\sigma(T > t)$ based on the marginal probabilities $p(\mathbf{s}|a)$ of sentiment sequences $\mathbf{s}$ of length l. The probability $p(\mathbf{s}|a)$ may be approximated by

$$\begin{aligned} p(\mathbf{s}|a) &= \sum_{\mathbf{x}} p(\mathbf{x}|a) p_\theta(\mathbf{s}|\mathbf{x}, a) \\ &\approx \sum_{\mathbf{x}} \frac{\tilde{p}'(\mathbf{x}|a)}{n-l+1} \times \left(\sum_i \frac{\alpha_i(s_1|\mathbf{x}, a) \prod_{j=i+1}^{i+(l-1)} M_j(s_{j-i}, s_{j-i+1}|\mathbf{x}, a) \beta_{i+(l-1)}(s_l|\mathbf{x}, a)}{Z(\mathbf{x}, a)} \right) \end{aligned} \tag{28}$$

where $\tilde{p}'$ is the empirical probability function $\tilde{p}'(\mathbf{x}|a) = \frac{1}{|C|} \sum_{\mathbf{x}' \in C} \delta_{\mathbf{x},\mathbf{x}'}$ for the set C of documents written by author a of length no less than l. α, M, β are the forward, transition and backward probabilities analogous to the dynamic programming method in [10].

Using the model $p(\mathbf{s}|a)$ we can compute $p_\sigma(T > t)$ for different authors at integer values of t which would lead to the quantity q_σ associated with each author. However, since (28) is based on an approximation, the calculated values of $p_\sigma(T > t)$ will be noisy resulting in slightly different values of q_σ for different time points t and cross validation iterations. A linear regression fit for q_σ based on the approximated values of $p_\sigma(T > t)$ for two authors using 10-fold cross validation is displayed in Figure 6. The data was the 249 movie reviews from the previous experiments written by one author, and additional 201 movie reviews from a second author. Interestingly, the author associated with the red dashed line has a consistent lower q_σ value in all those figures, and thus is considered as more "static" and less prone to quick sentiment variations.

6.4 Text Summarization

We demonstrate the potential usage of sentiment flow for text summarization with a very simple example. The text below shows the result of summarizing the movie review in Figure 3 by keeping only sentences associated with the start, the end, the top, and the bottom of the predicted sentiment curve. The number before each sentence relates to the circled number in Figure 3.

1 What makes this film mesmerizing, is not the plot, but the virtuoso performance of Lucy Berliner (Ally Sheedy), as a wily photographer, retired from her professional duties for the last ten years and living with a has-been German actress, Greta (Clarkson). 2 The less interesting story line involves the ambitions of an attractive, baby-faced assistant editor at the magazine, Syd (Radha Mitchell), who lives with a boyfriend (Mann) in an emotionally chilling relationship. 3 We just lost interest in the characters, the film began to look like a commercial for a magazine that wouldn't stop and get to the main article. 4 Which left the film only somewhat satisfying;

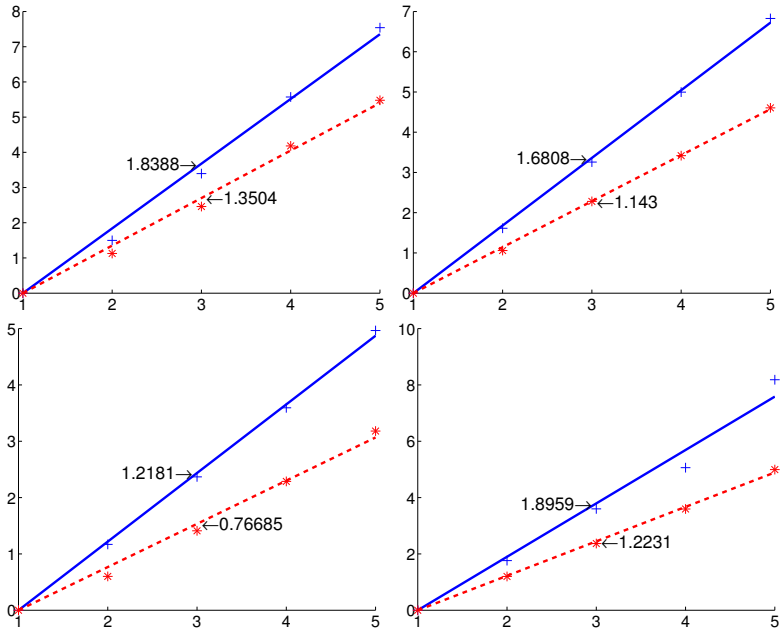

Figure 6: Linear regression fit for q_σ, $\sigma = 2, 1, -1, -2$ (left to right) based on approximated values of $p_\sigma(T > t)$ for two different authors. X-axis: time t; Y-axis: negative log-probability of $T > t$.

acclaim	activ	admir	ador	aesthet
aliv	allure	amaz	amus	appeal
appreci	apt	artfully	artifice	astonish
attract	authent	awe	award	
abruptly	absurd	adolesc	ambigu	annoy
arrog	awkward	arti		

Table 3: Stemmed words starting with 'a' that are chosen manually to conveying positive or negative sentiment.

it did create a proper atmosphere for us to view these lost characters, and it did have something to say about how their lives are being emotionally torn apart. 5 It would have been wiser to develop more depth for the main characters and show them to be more than the superficial beings they seemed to be on screen.

6.5 Elicitation of Constraints from Domain Experts

In all previous experiments, the probability ordering constraints are obtained by testing hypotheses such as (26) or (25) on the auxiliary data set. We now demonstrate that we may achieve similar or even better results by eliciting constraints from domain experts.

During the experiment, one of the authors was presented with the vocabulary of the sentiment data set, and was asked to pick a subset of words from it which they thought would indicate either positive or negative sentiment. A total of 402 words were picked, and a subset of them starting with 'a' are listed in Table 3.

This set of words are then used to define one-way ordering constraints for CRF corresponding to a full ordering on the labels $\mathcal{Y}$. Figure 7 shows the test-set performance as a function of training size averaged over 40 cross validations. Compared with Figure 1 (top left), applying domain knowledge directly achieves similar, or even higher accuracy. This demonstrates the flexibility of our framework in the sense that domain knowledge may come from multiple sources, including domain experts and auxiliary data sets.

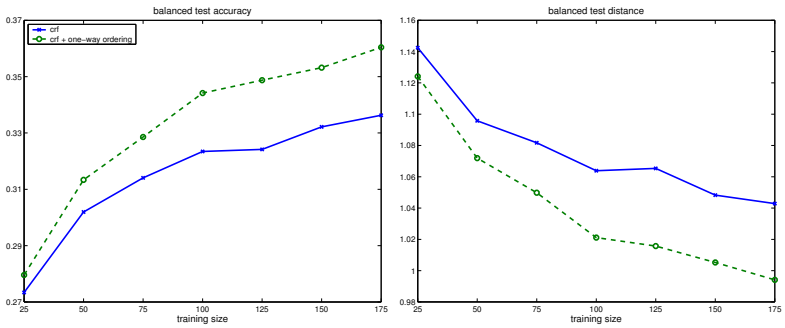

Figure 7: Balanced test-set accuracy (left) and distance of predicted sentiment from true sentiment (right) as a function of training size average over 40 cross validations. One-way ordering constraints are elicited from a domain expert without the use of auxiliary data set.

7 Information Extraction

The idea of generalized isotonic CRF can also be applied to information extraction in natural language processing. In contrast to the case of local sentiment prediction, the set of labels $\mathcal{Y}$ in information extraction is categorical and there is no natural order on it. The sequences $\mathbf{x}$ corresponds to a sentence or a document with x_i being vocabulary words.

We use a CRF model (6) with a set of features $A = \{A_v(\mathbf{x}, i) = 1_{\{x_i=v\}}\}$ that measure the appearance of word v at the current position. We consider isotonic constraints that define a partial order on the $\mu_{\langle\sigma,A_v\rangle}$ as follows. For each word v, we determine the most likely tag $\sigma \in \mathcal{Y}$ and if deemed significant we enforce

$$\mu_{\langle\sigma,A_v\rangle} \geq \mu_{\langle\tau,A_v\rangle} \quad \forall\tau \in \mathcal{Y}, \tau \neq \sigma. \tag{29}$$

We conducted our experiments on the advertisements data for apartment rentals[3] which contains 302 documents labeled with 12 fields, including `size`, `rent`, `restrictions`, etc. During each iteration, 100 documents are randomly selected for testing and the remaining documents are used for training. As previously noted, we use limited memory BFGS for L_2 regularized likelihood with the barrier method enforcing constraints.

As before, one of the authors was presented with the vocabulary of the advertisements data, and was asked to pick a subset of words from it which he thought would be indicative of some field. As a part of the elicitation, he was allowed to observe a few labeled documents (≤ 5) from the data set before the actual selection of words. Table 4 lists the picked words and the field column gives the highest ranked label σ for each word v on the right.

We also use features that model the local context, including

$$B = \{B_v^-(\mathbf{x}, i) = 1_{\{x_{i-1}=v\}}, B_v^+(\mathbf{x}, i) = 1_{\{x_{i+1}=v\}}\}$$

which consider words appearing before and after the current position, and

$$C = \{C_{u,v}^-(\mathbf{x}, i) = 1_{\{x_{i-1}=u\}}1_{\{x_i=v\}}, C_{u,v}^+(\mathbf{x}, i) = 1_{\{x_i=u\}}1_{\{x_{i+1}=v\}}\}$$

which consider bigrams containing the current word. Table 5 lists a set of bigrams that are deemed indicative of some field.

Table 6 and 7 display the prediction accuracy (which equals micro-averaged $F_{1.0}$) and macro-averaged $F_{1.0}$ for test data subject to one-way ordering induced by Table 4 and 5. The results are averaged over 20 cross-validation iterations. In all cases, generalized isotonic CRFs consistently outperform the CRF.

[3]Available at: http://nlp.stanford.edu/∼grenager/data/unsupie.tgz

Field	Words
contact	*EMAIL* *PHONE* *TIME* today monday tuesday wednesday friday sat saturday sunday weekend(s) am pm appointment visit reply contact email fax tel schedule questions information details interested @
size	ft feet sq sqft
neighborhood	airport restaurant(s) safeway school(s) shop(s) shopping store(s) station(s) theater(s) transit transportation freeway(s) grocery hwy(s) highway(s) expressway near nearby close mall park banks churches bars cafes
rent	*MONEY* term(s) yearly yr lease(s) contract deposit year month
available	immediately available june july aug august
restrictions	smoke smoker(s) smoking pet(s) cat(s) dog(s) preferred
address	ave avenue blvd
features	backyard balcony(-ies) basement dishwasher(s) dryer(s) furniture fridge garage(s) jacuzzi kitchen(s) kitchenette laundry lndry lobby oven(s) parking pool(s) refrig refrigerator(s) sauna(s) sink(s) spa storage stove(s) swimming tub(s) washer(s)
photos	image(s) photo(s) picture(s)
utilities	utility utilities utils electricity pays
roommates	roommate student

Table 4: Words selected for one-way ordering in generalized isotonic CRF. The label on the left is determined to be the most likely label corresponding to the words on the right. Words between two asterisks, e.g. *EMAIL*, represent tokens that match the given regular expressions. Words with parentheses denote a group of similar words, e.g. image(s) is used to represent both image and images.

Field	Words
size	single-family *NUMBER*-story *NUMBER*-bedroom(s) one-bath *NUMBER*-bath(s) one-bathroom *NUMBER*-bathroom one-bedroom two-bedroom(s) *NUMBER*-br square-feet sq-feet sq-ft
neighborhood	walking-distance easy-access convenient-to close-to access-to mile-from distance-to block(s)-to away-from located-near block(s)-from block(s)-away minutes-to(away,from)
features	lots-of plenty-of living-room dining-room gas-stove street-parking
contact	open-house set-up stop-by
rent	*NUMBER*-month application-fee security-deposit per-month /-month /-mo a(one,first,last)-month
address	located-at
restrictions	at-least may-be

Table 5: Bigrams selected for one-way ordering in generalized isotonic CRF. The label on the left is determined to be the most likely label corresponding to the bigrams on the right.

N	accuracy		$F_{1.0}$		accuracy		$F_{1.0}$	
	CRF	iso-CRF	CRF	iso-CRF	CRF	iso-CRF	CRF	iso-CRF
10	0.5765	**0.5862***	0.2804	**0.3264**	0.5942	**0.6255**	0.3153	**0.3923**
15	0.6265	**0.6578**	0.3479	**0.4002**	0.6294	**0.6614**	0.3703	**0.4503**
20	0.6354	**0.6750**	0.3760	**0.4433**	0.6553	**0.6931**	0.4110	**0.5090**
25	0.6760	**0.6968**	0.4257	**0.4687**	0.6712	**0.7100**	0.4412	**0.5320**
50	0.7062	**0.7491**	0.5064	**0.5734**	0.7187	**0.7409**	0.5226	**0.5818**
75	0.7533	**0.7658**	0.5716	**0.6038**	0.7391	**0.7528**	0.5594	**0.6061**
100	0.7696	**0.7814**	0.5992	**0.6287**	0.7514	**0.7628**	0.5857	**0.6256**
200	0.7910	**0.8012***	0.6348	**0.6691**	0.7810	**0.7859**	0.6294	**0.6540**

Table 6: Labeling accuracy and macro-averaged $F_{1.0}$ for various training size N. Models are trained using the set of features A (left) as well as $A \cup B$ (right) subject to one-way ordering induced by Table 4. An asterisk (*) indicates that the difference is not statistically significant according to the paired t test at the 0.05 level.

	accuracy		$F_{1.0}$	
	CRF	iso-CRF	CRF	iso-CRF
10	0.5760	**0.5902**	0.2745	**0.2954***
15	0.6146	**0.6322**	0.3310	**0.3560**
20	0.6439	**0.6508**	0.3685	**0.3880**
25	0.6610	**0.6883**	0.4043	**0.4495**
50	0.7190	**0.7370**	0.5043	**0.5503**
75	0.7428	**0.7576**	0.5488	**0.5902**
100	0.7615	**0.7727**	0.5796	**0.6122**
200	0.7921	**0.7999**	0.6405	**0.6667**

Table 7: Labeling accuracy and macro-averaged $F_{1.0}$ for various training size N. Models are trained using the set of features $A \cup B \cup C$ subject to one-way ordering induced by both Table 4 and 5. We omit the results for one-way ordering induced by Table 4 only, which are almost identical to those reported for iso-CRF. An asterisk (*) indicates that the difference is not statistically significant according to the paired t test at the 0.05 level.

8 Related Work

Sequentially modeling the data is a key step in many applications, such as part-of-speech tagging, information extraction, and protein secondary structure prediction. Hidden Markov models (HMM) [18], maximum entropy Markov models (MEMM) [12] and conditional random fields (CRF) [10] are three of the most popular sequential models up to date.

HMM models the joint probability of the observation sequence and the label sequence. It is a generative model that makes a strong independence assumption about observations to ensure the tractability of the inference. This assumption is often inappropriate for real applications, where we believe that the representation should consist of many overlapping features. MEMM remove the assumption by modeling the conditional probability of the next state given the current state and the current observation. Since they use per-state exponential models, MEMM potentially suffer from the label bias problem. CRF combine the advantages of two previous models by introducing a single exponential model for the joint probability of the entire label sequence given the observation sequence, and reports superior experimental results in the areas mentioned above.

Given a set of iid training samples, the parameters of CRF are typically estimated by maximizing the regularized conditional likelihood defined in Equation 9. Other popular approaches of learning a CRF model include maximum margin Markov networks [21] where the model is trained discriminatively using a margin-based optimization problem, and Searn [9], an algorithm that decomposes a structured prediction problem into a set of classification problems solved by standard classification methods. The generalized perceptron proposed by Collins [3] is another widely used model for NLP tasks and is closely related to the CRF.

Another contributing factor to our work is isotonic regression [1], which is an important method in statistical inference with monotonicity constraints. It can be traced back to the problem of maximizing the likelihood of univariate normal distributions subject to an ordered restriction on the means. The term *isotonic* is interpreted as order-preserving: for a finite set $S = \{1, \cdots, n\}$ on which a full order $\leq$ is defined, a real vector $(\beta_1, \cdots, \beta_n)$ is isotonic if $i, j \in S$, $i \leq j$ imply $\beta_i \leq \beta_j$. Given real vector $(x_1, \cdots, x_n)$ with weights $(w_1, \cdots, w_n)$, the isotonic regression takes the form of a weighted least square fitting which minimizes $\sum_{i=1}^{n} w_i(x_i - \beta_i)^2$ subject to the constraint that $(\beta_1, \ldots, \beta_n)$ is isotonic.

Various extensions have been proposed for isotonic regression. Some of them consider relationships other than a full order. Examples include the tree order $\beta_1 \leq \beta_2, \ldots, \beta_1 \leq \beta_n$, and the umbrella order $\beta_1 \leq \ldots \leq \beta_i \geq \ldots \geq \beta_n$ for some fixed i. Most similar to our framework is the ordering constraint proposed in [8] for normal means from a two-way layout experiment

$$\beta_{i+1,j+1} - \beta_{i+1,j} - \beta_{i,j+1} + \beta_{i,j} \geq 0 \quad i = 1, \ldots, m-1,\ j = 1, \ldots, n-1$$

which states that the differences $\beta_{i'j} - \beta_{ij}$ grow as the level j increases for any $i' > i$.

Sentiment prediction was first formulated as a binary classification problem to answer questions such as: "What is the review's polarity, positive or negative?" Pang et al. [17] demonstrated the difficulties in sentiment prediction using solely the empirical rules (a subset of adjectives), which motivates the use of statistical learning techniques. The task was then refined to allow multiple sentiment levels, facilitating the use of standard text categorization techniques [16].

Various statistical learning techniques have been suggested for sentiment prediction, treating the data either as categorial (naive Bayes, maximum entropy and support vector machine [17, 16]) or as ordinal (support vector regression and metric labeling [16]). Although most methods report over 90% accuracy on text categorization, their performance degrades drastically when applied to sentiment prediction.

Indeed, sentiment prediction is a much harder task than topic classification tasks such as Reuters or WebKB. It is different from traditional text categorization: (1) in contrast to the categorical nature of topics, sentiments are ordinal variables; (2) several contradicting opinions might co-exist, which interact with each other to produce the global document sentiment; (3) context plays a vital role in determining the sentiment. In view of this, Mao and Lebanon [11] suggest to model local sentiment flow in documents rather than predicting the sentiment of the entire document directly. The idea is further exploited in [13] where the sentiments of text at varying levels of granularity are jointly classified.

In a statistical framework, the expert's knowledge has to be in probabilistic form for it to be used. However, unless the expert is a statistician, or is very familiar with statistical concepts, efforts have to be made to formulate the expert's knowledge and beliefs in probabilistic terms. This is done through elicitation [5, 14] in the statistical literature. Psychological literature suggests that people are prone to certain heuristics and biases in the way they respond to situations involving uncertainty. As a result, elicitation is conducted in a principled way where stages involving eliciting summaries, fitting a distribution and testing adequacy may repeat several times before a faithful elicitation is reached. The usefulness of elicitation has been demonstrated in statistical literature where most work concentrates on eliciting univariate probability distributions. Multivariate elicitation is largely unexplored due to the complexity of formulating variable interactions.

Most work of incorporating prior knowledge into structured prediction models is done in the context of part-of-speech tagging or information extraction. Chang et al. [2] specify the prior knowledge for two information extraction tasks [12, 6] as a set of constraints to be satisfied by label-observation pairs. Haghighi and Klein [7] define prototype to be some canonical examples (e.g. words) of each target label (e.g. part-of-speech). Their method is similar to our one-way ordering for information extraction in the sense that they activate prototypes, in addition to the observed word, at each sequence position. However, neither approach offers a probabilistic interpretation. Druck et al. [4] apply the generalized expectation criteria for learning a sliding window multinomial logistic regression for name entity recognition. The prior knowledge is in the form of a probability distribution over labels conditioned on some feature f. Such prior knowledge is hard to specify in the case of a structured prediction model since the sample space size scales exponentially with the sequence length.

9 Discussion

Regularized maximum likelihood estimation is one of the most popular estimation techniques in statistical learning. A natural way to incorporate domain knowledge into this framework is through the use of an informative or subjective prior. Assuming the prior is uniform over an admissible area the maximum posterior estimate becomes the constrained version of the maximum likelihood.

An informative prior or frequentist constraints are usually specified on the parameter space Θ. Unfortunately, it is highly non-trivial to obtain a statistical interpretation of the informative prior in terms of the underlying probabilities. This is especially true for conditional random fields which is perhaps the most popular model for structured prediction.

We argue that domain knowledge, whether elicited from a domain expert or from auxiliary data, is best specified directly in terms of probability constraints. Such constraints have a clear interpretation in terms of probability of certain events. We define several types of probability constraints that lead directly to simple parameter constraints thereby facilitating their use as a subjective prior in the statistical learning process. Moreover, the probability constraints can be described in terms of simple queries corresponding to the increase of the probability of a label t_j as a result of a local perturbation of the input sequence $\mathbf{x} \mapsto \mathbf{x}'$. The increase in probability is then compared to the increase in probability of another label s_j. Since it incorporates relative judgement corresponding to an ordering of probability ratios, it is more likely to be accurately elicited than specific probability values.

We present a general framework for incorporating several types of constraints into a simple informative prior consisting of partial ordering constraints on the model parameters. The framework applies to a wide range of applications and leads to efficient computational procedure for solving the constrained regularized maximum likelihood. We demonstrate its applicability to the problems of local sentiment analysis and predicting syntactic and morphological tags in natural language processing.

Our experiments indicate that incorporating the constraints leads to a consistent improvement in prediction accuracy over the regularized CRF model which is considered the state-of-the-art for sentiment prediction and information extraction. In our experiments we study both elicitation from a domain expert and from auxiliary data. In the latter case, we develop an effective mechanism for automatically deriving constraints based on hypothesis testing.

The developed framework applies directly to CRF but could be modified to other structured prediction models such as max-margin discriminative networks. With some simple modifications it applies also to other conditional models such as multinomial logistic regression and in general other forms of conditional graphical models.

References

[1] R. E. Barlow, D.J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions (the theory and application of isotonic regression)*. John Wiley and Sons, Inc., 1972.

[2] M. Chang, L. Ratinov, and D. Roth. Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[3] Michael Collins. Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002.

[4] G. Druck, G. Mann, and A. McCallum. Leveraging existing resources using generalized expectation criteria. In *NIPS Workshop on Learning Problem Design*, 2006.

[5] P. H. Garthwaite, J. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100:680–701, 2005.

[6] T. Grenager, D. Klein, and C. D. Manning. Unsupervised learning of field segmentation models for information extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[7] A. Haghighi and D. Klein. Prototype-driven learning for sequence models. In *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

[8] C. Hirotsu. Ordered alternatives for interaction effects. *Biometrika*, 65(3):561–570, 1978.

[9] H. Daumé III, J. Langford, and D. Marcu. Search-based structured prediction. Submitted to Machine Learning Journal, 2006.

[10] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning*, 2001.

[11] Y. Mao and G. Lebanon. Isotonic conditional random fields and local sentiment flow. In *Advances in Neural Information Processing Systems 19*, pages 961–968, 2007.

[12] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. of the International Conference on Machine Learning*, 2000.

[13] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[14] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, 2006.

[15] B. Pang and L. Lee. A sentimental eduction: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the Association of Computational Linguistics*, 2004.

[16] B. Pang and L. Lee. Seeing stars: Exploiting class relationship for sentiment categorization with respect to rating scales. In *Proc. of the Association of Computational Linguistics*, 2005.

[17] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 2002.

[18] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[19] M. J. Silvapulle and P. K. Sen. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley, 2004.

[20] R. P. Stanley. *Enumerative Combinatorics*, volume 1. Cambridge University Press, 2000.

[21] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In *Advances in Neural Information Processing Systems*, 2004.