Graph-Based Posterior Regularization for Semi-Supervised Structured Prediction

Penn University of Pennsylvania

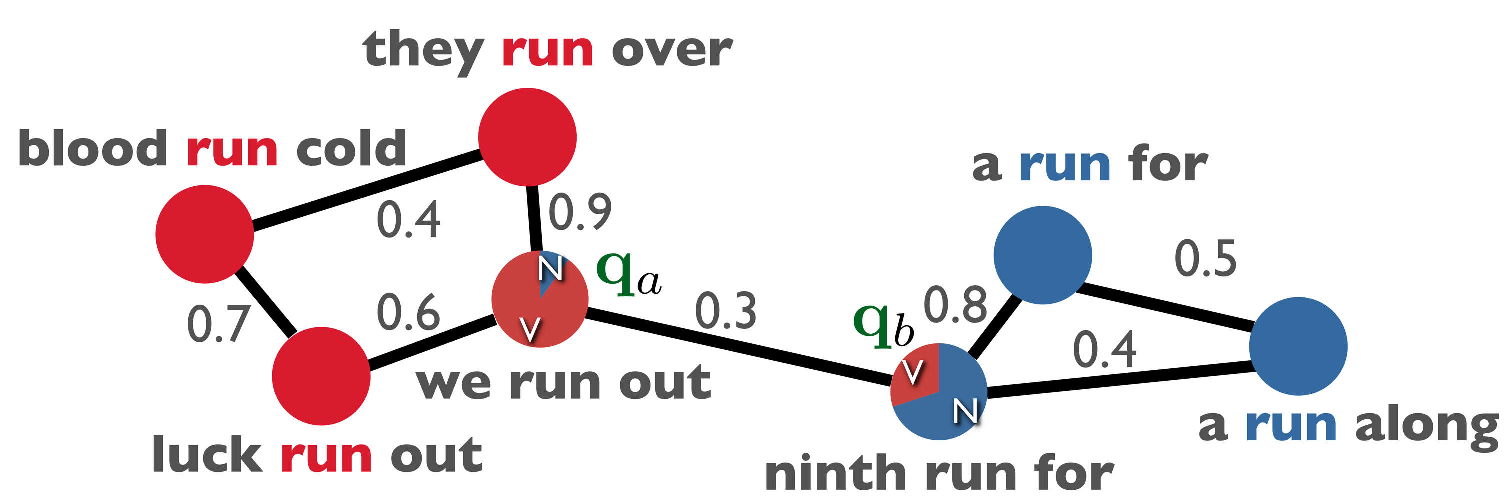
Luheng He Jennifer Gillenwater Ben Taskar

{luhe,jengi}@cis.upenn.edu, taskar@cs.washington.edu

University of Washington

GRAPH-BASED LEARNING

Labels: verb (V), noun (N), etc.

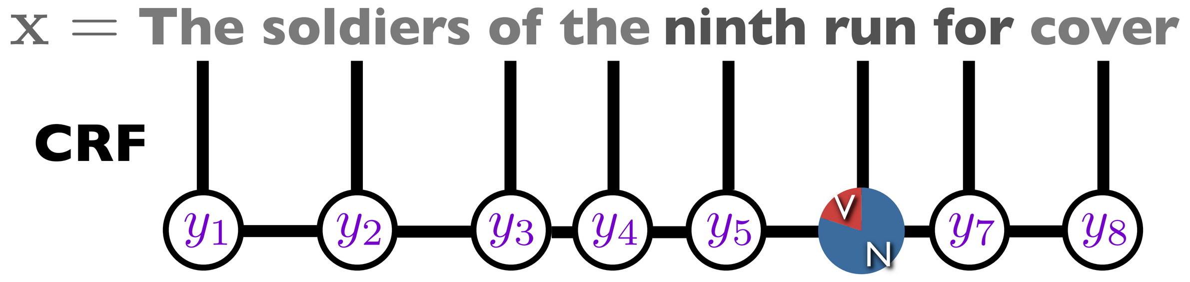


Minimize Laplacian-based objective, summing over all neighbors of unlabeled nodes:

$$\mathrm{Lap}(q) = \sum_{a=1}^{N} \sum_{b=L+1}^{N} w_{ab} ||\mathbf{q}_a - \mathbf{q}_b||_2^2$$

STRUCTURED PREDICTION

$\mathbf{x}$ = **The soldiers of the ninth run for cover**



$$p_\theta(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left[\sum_{t=1}^{T} \theta^\top \mathbf{f}(y_t, y_{t-1}, \mathbf{x})\right]$$

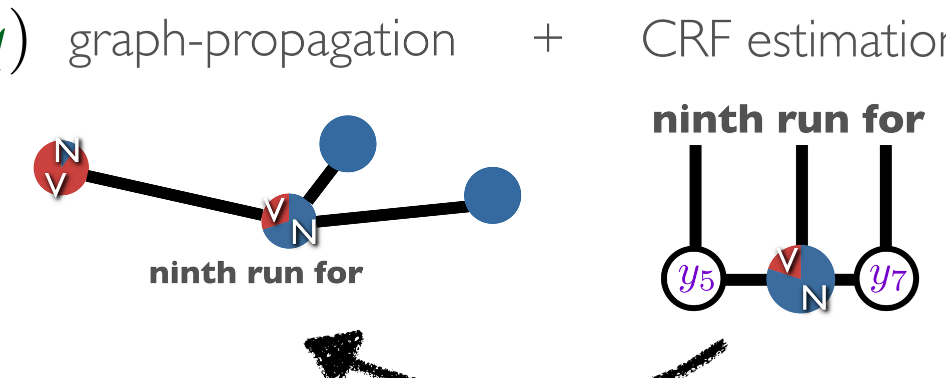
Minimize negative log-likelihood, summing over all labeled sentences:

$$\mathrm{NLik}(p_\theta) = -\sum_{i=1}^{\ell} \log p_\theta(\mathbf{y}^i \mid \mathbf{x}^i)$$

COMBINATION

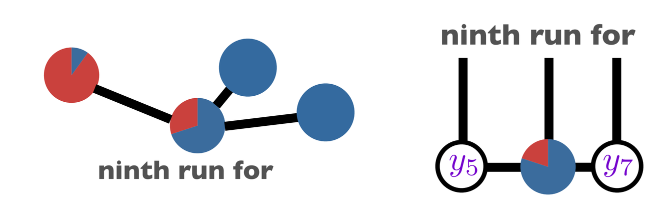
Most closely related work: Subramanya et al. (EMNLP 2010) --- Iterative procedure, marginals of CRF initialize graph-propagation (GP), then GP results provide additional training data for CRF learning.

$\mathrm{Lap}(q)$ graph-propagation + CRF estimation $\mathrm{NLik}(p_\theta)$



This work: retains efficiency of Subramanya et al (EMNLP 2010) while optimizing an extendible, joint objective.

JOINT OBJECTIVE



$$\mathcal{J}(q, p_\theta) = \mathrm{Lap}(q) + \mathrm{NLik}(p_\theta) + \mathrm{KL}(q \parallel p_\theta)$$

Couple the methods via KL divergence.

$(\#\ \text{tags})^8$ values, compactly represented by θ in the case of p

q	p_θ	The	soldiers	of	the	ninth	run	for	cover
7e-5	2e-5	N	N	N	N	N	N	N	N
3e-6	8e-6	N	N	N	N	N	N	N	V
...	...				...				

OPTIMIZATION

p's parameterization makes its update simple:

θ update: $\theta' = \theta - \eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial \theta}$

q has more freedom: $q^i \in \Delta$ of dimension $(\#\ \text{tags})^{(i\text{'s length})}$

standard gradient update: $q_{\mathbf{y}}^i{}' = \mathrm{proj}_\Delta\left(q_{\mathbf{y}}^i - \eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial q_{\mathbf{y}}^i}\right)$

-Problem 1: projection is hard $q_{\mathbf{y}}^i \notin \Delta$

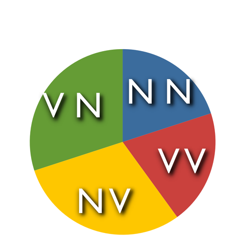
-Problem 2: no compact form $(\#\ \text{tags})^{(i\text{'s length})}$ values

✗ Standard gradient descent on the primal isn't feasible for q

What about optimizing q in the dual? $\mathcal{J}(q, p_\theta) + \gamma\left(\sum_{\mathbf{y}} q_{\mathbf{y}}^i - 1\right)$

Posterior Regularization (PR) of Ganchev et al. (JMLR 2010) uses the dual, and differs from our objective only in the first term.

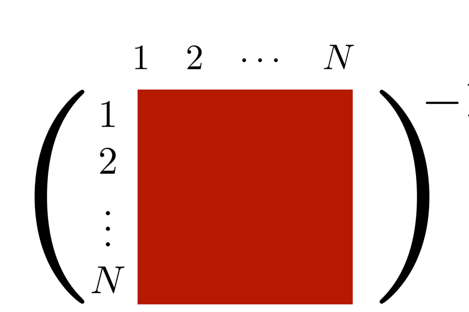
This work: $\mathrm{Lap}(q) \longrightarrow$ Standard PR: $\mathrm{Linear}(m)$



sentence, position, tag t, tag t-1

$$m_{t,k,j}^i = \sum_{\mathbf{y}} \mathbf{1}(y_t = k, y_{t-1} = j) q_{\mathbf{y}}^i \quad \text{} p\text{-factor}$$

$\mathrm{Lap}(m)$ is a quadratic function though, so its dual requires an expensive matrix inverse.



EXPONENTIATED GRADIENT

Alternative type of gradient update makes "projection" efficient:

$$q_{\mathbf{y}}^i{}' = \frac{1}{Z_q(\mathbf{x}^i)} q_{\mathbf{y}}^i \exp\left[-\eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial q_{\mathbf{y}}^i}\right]$$

$$\exp\left[-\eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial q_{\mathbf{y}}^i}\right] =$$

$$\exp\left[-\eta \sum_{t=1}^{T} \frac{\partial \mathrm{Lap}(m_{\mathbf{y}}^i)}{\partial m_{t,y_t,y_{t-1}}^i} + \eta\left(\log p_\theta(\mathbf{y} \mid \mathbf{x}^i) - \log q_{\mathbf{y}}^i - 1\right)\right]$$

$$= \underbrace{\exp\left[-\eta \sum_{t=1}^{T} \frac{\partial \mathrm{Lap}(m_{\mathbf{y}}^i)}{\partial m_{t,y_t,y_{t-1}}^i}\right] p_\theta(\mathbf{y} \mid \mathbf{x}^i)^\eta (q_{\mathbf{y}}^i)^{-\eta} e}_{\text{product of p-factors}}, \quad q = p_\theta$$

$\mathrm{proj}_\Delta \longrightarrow Z_q(\mathbf{x}^i)$, computable via forward-backward

EXTENSION

$\mathrm{Lap}(q) \longrightarrow$ any convex, differentiable $g(m)$

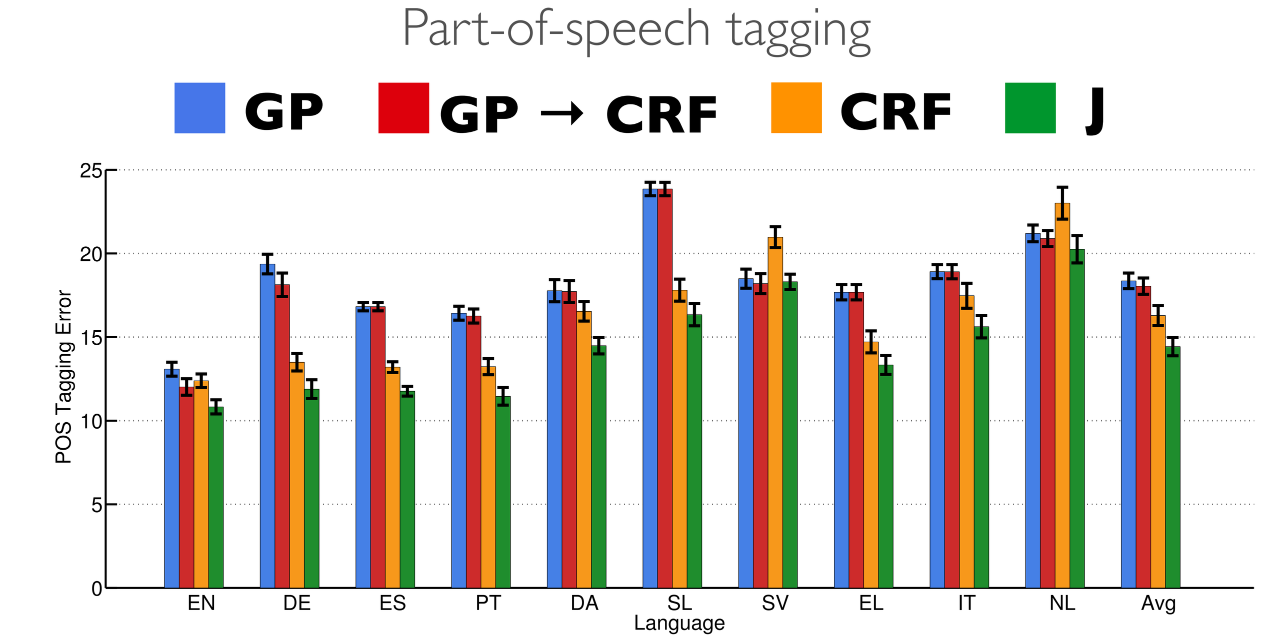
Theorem: The EM-like optimization procedure below converges to a local optimum of the joint objective

M-step: $\theta' = \theta - \eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial \theta}$

E-step: $q_{\mathbf{y}}^i{}' = \frac{1}{Z_q(\mathbf{x}^i)} q_{\mathbf{y}}^i \exp\left[-\eta \frac{\partial \mathcal{J}(q,p_\theta)}{\partial q_{\mathbf{y}}^i}\right]$

EXPERIMENTS

Part-of-speech tagging



Handwriting recognition

	GP	GP → CRF	CRF	J
Mean	17.57	15.07	9.82	**4.89**
StdDev	0.30	0.35	0.48	**0.42**

Code: https://code.google.com/p/pr-graph/