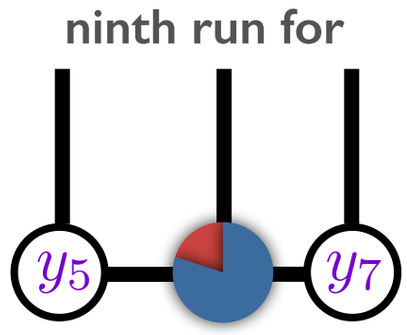
GRAPH-BASED POSTERIOR REGULARIZATION FOR SEMI-SUPERVISED STRUCTURED PREDICTION

Luheng He Jennifer Gillenwater† Ben Taskar

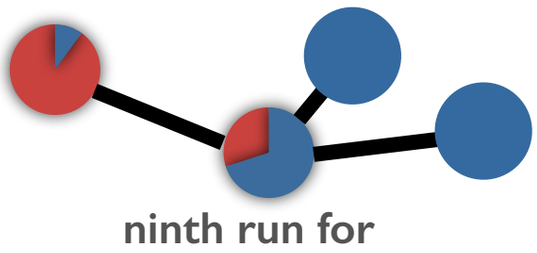
†University of Pennsylvania University of Washington

OVERVIEW

Structured Prediction (CRF)



Graph Propagation



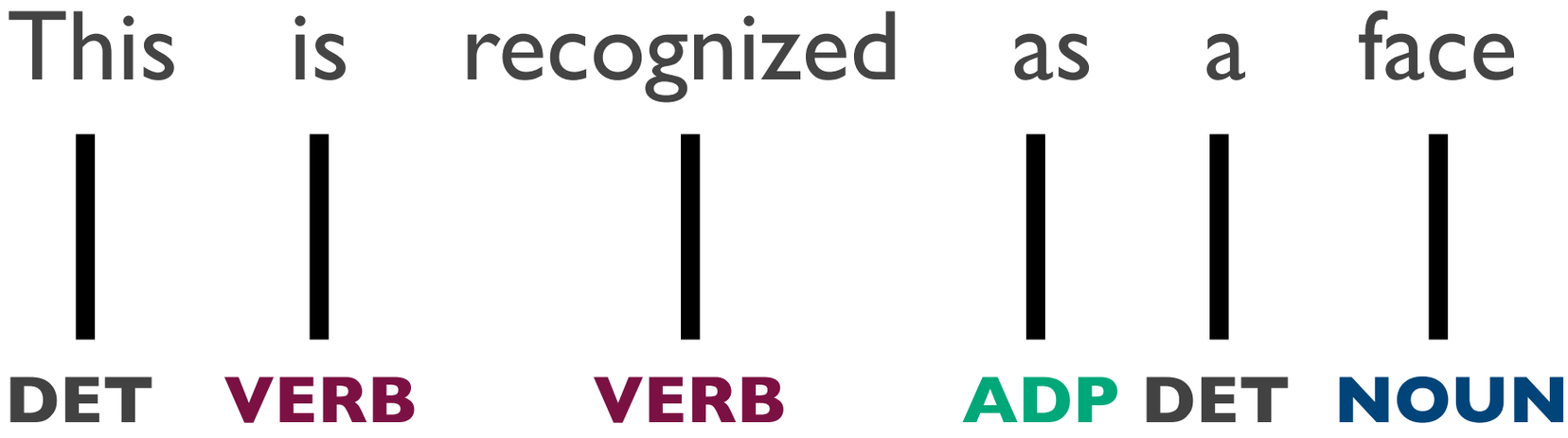
A Joint Objective

$$\mathcal{J}(q, p_\theta)$$

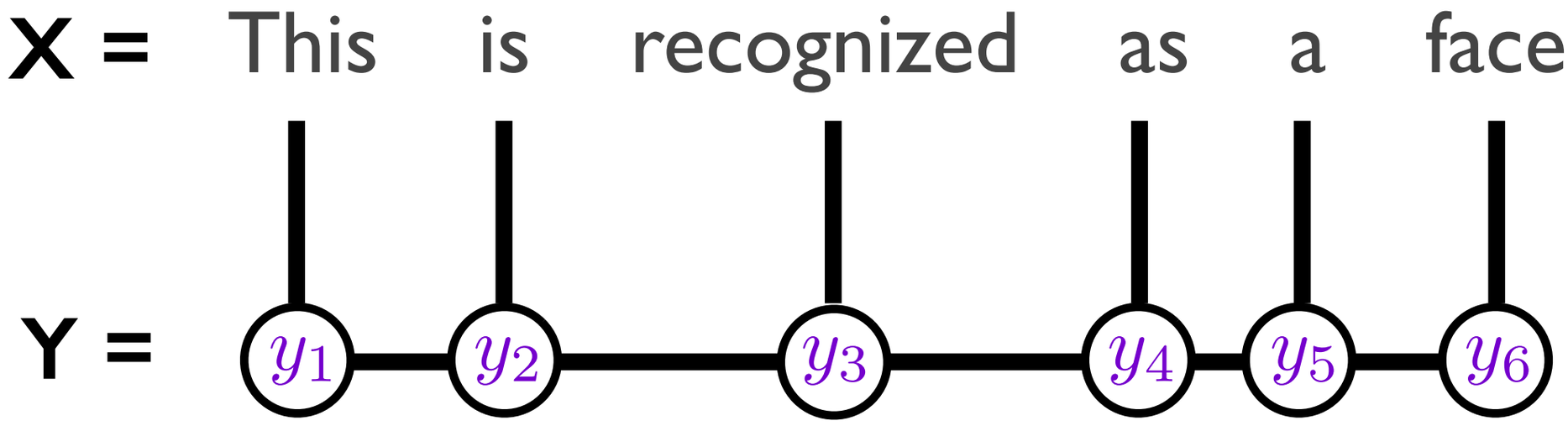
CONDITIONAL RANDOM FIELD

This is recognized as a face

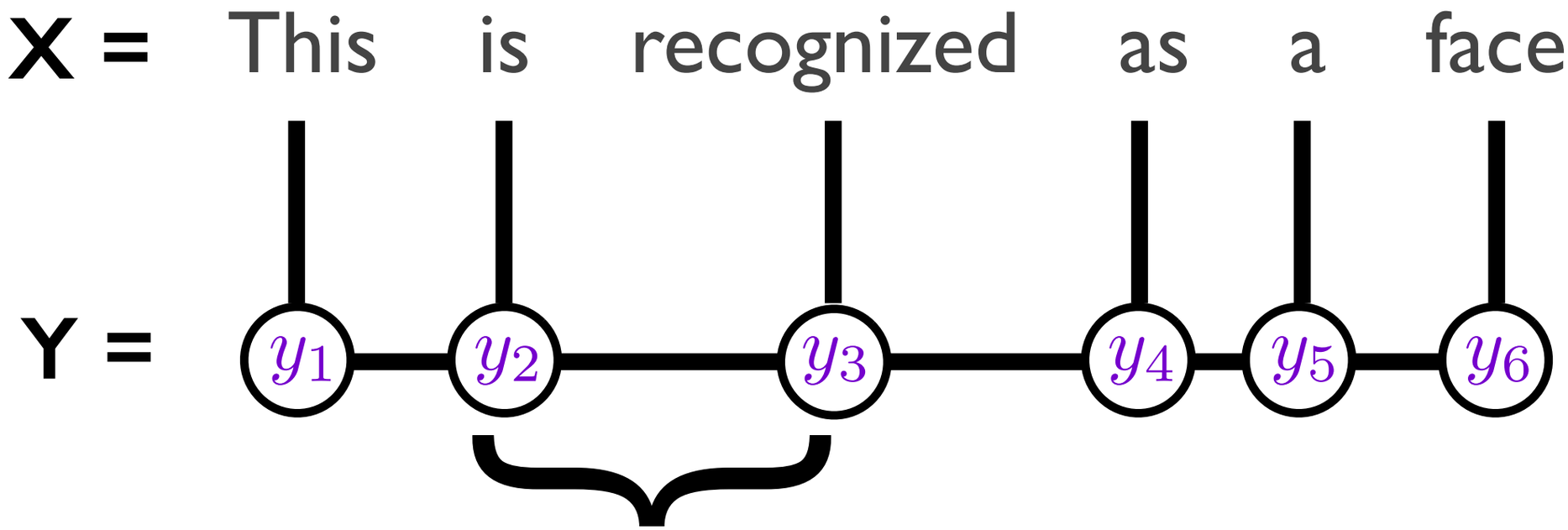
CONDITIONAL RANDOM FIELD



CONDITIONAL RANDOM FIELD

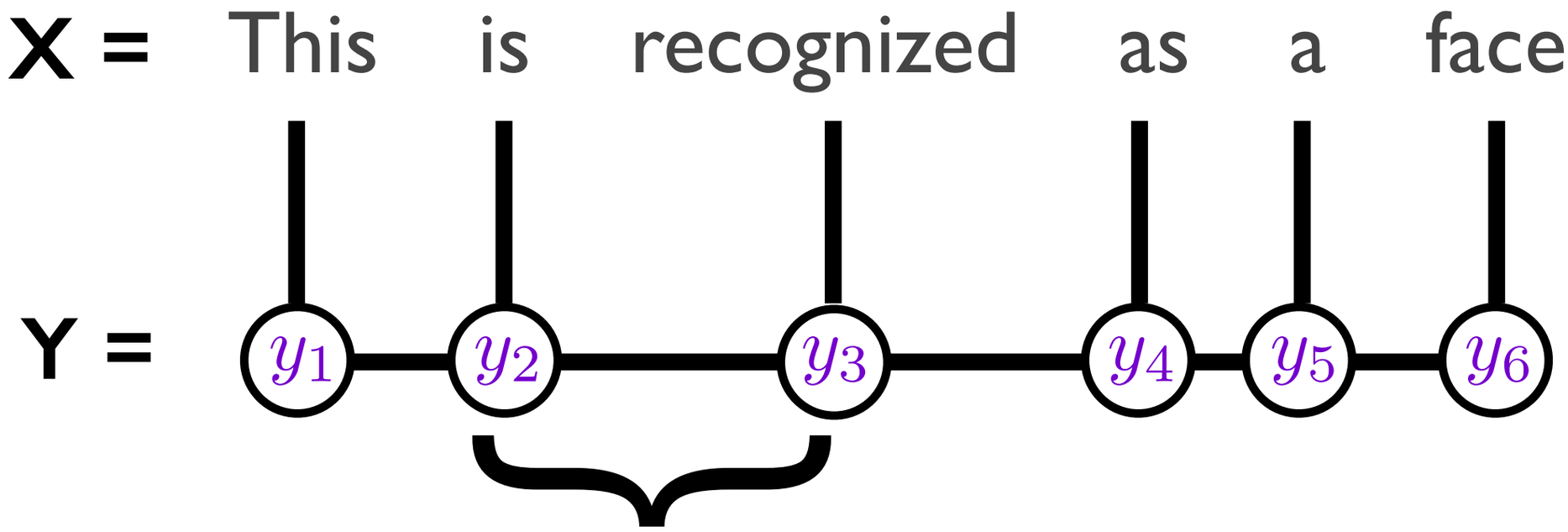


CONDITIONAL RANDOM FIELD



$$\mathbf{f}(\, y_t\,, y_{t-1}, \mathbf{x}\,)$$

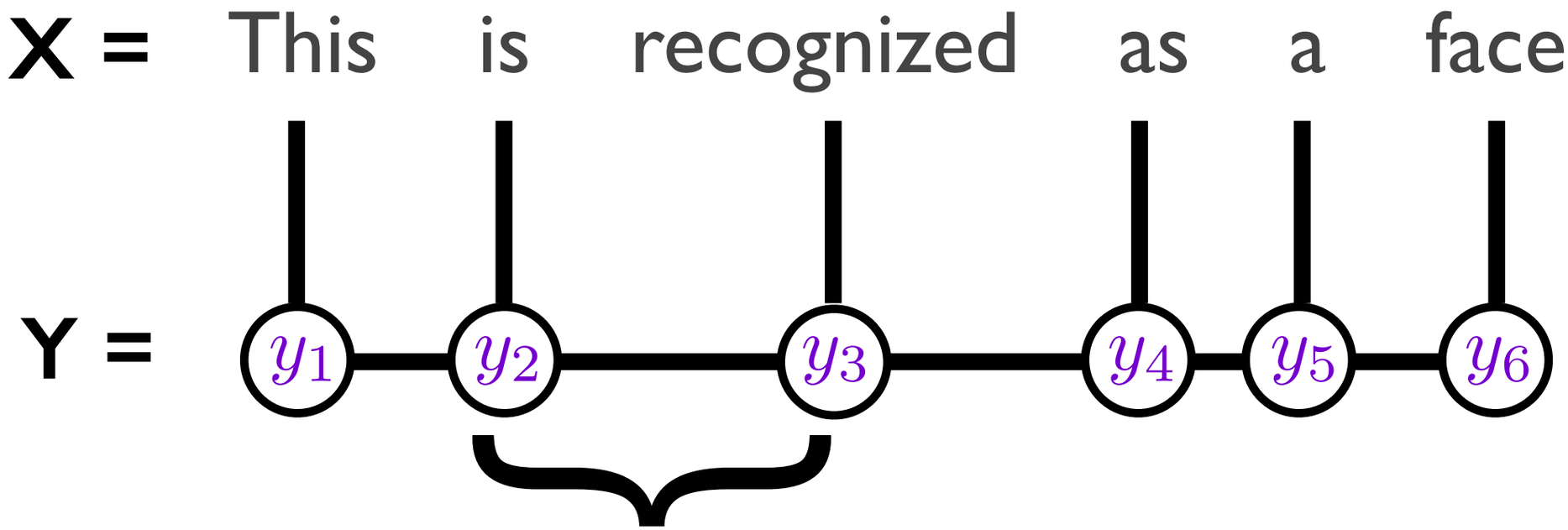
CONDITIONAL RANDOM FIELD



Conditional Distribution

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

CONDITIONAL RANDOM FIELD



$$\mathbf{f}(\, y_t\,, y_{t-1}, \mathbf{x}\,)$$

Conditional Distribution

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

CRF objective

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

GRAPH PROPAGATION

This is recognized as a face .

VERB

The painting is considered as a work of genius .

?

Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning

GRAPH PROPAGATION

This is recognized as a face .

VERB

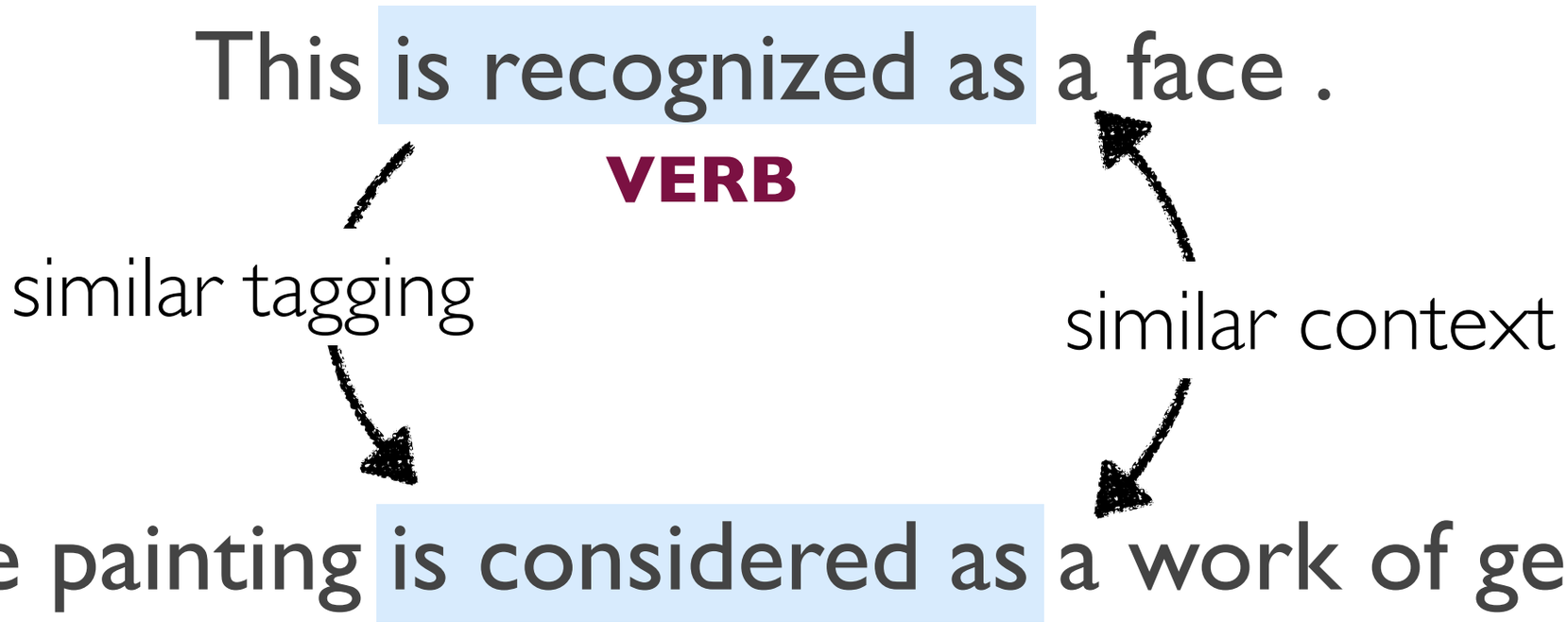
similar context

The painting is considered as a work of genius .

?

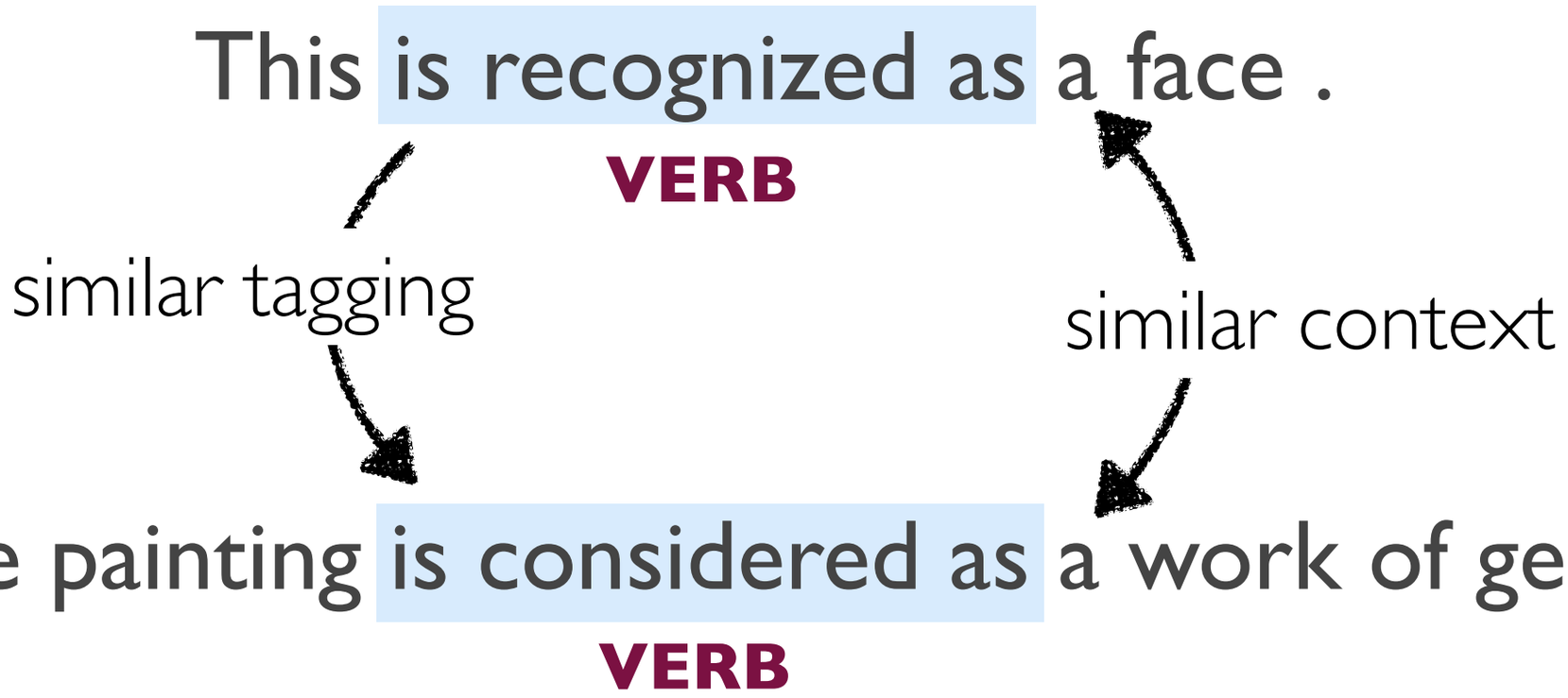
Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning

GRAPH PROPAGATION



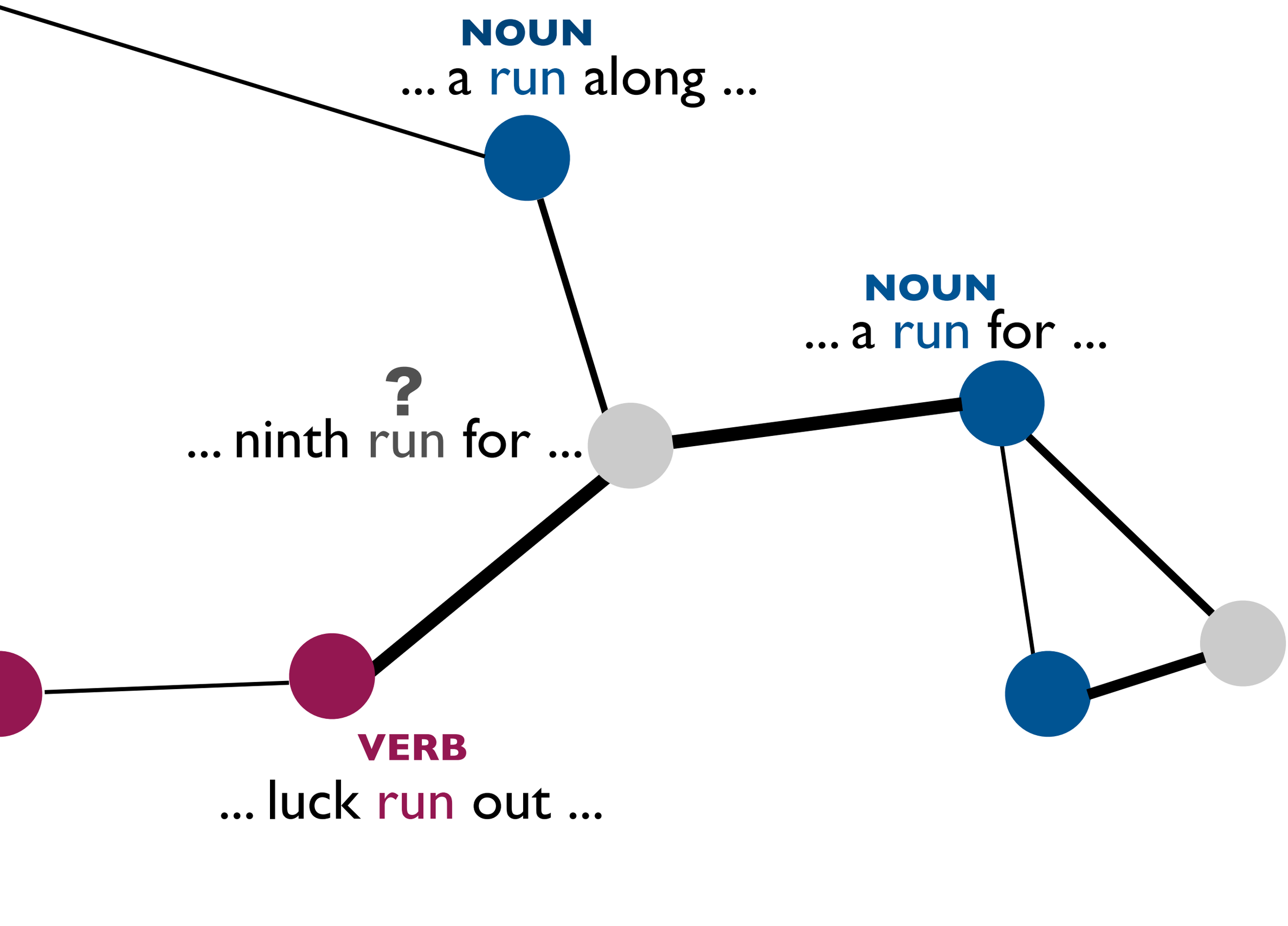
Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning

GRAPH PROPAGATION

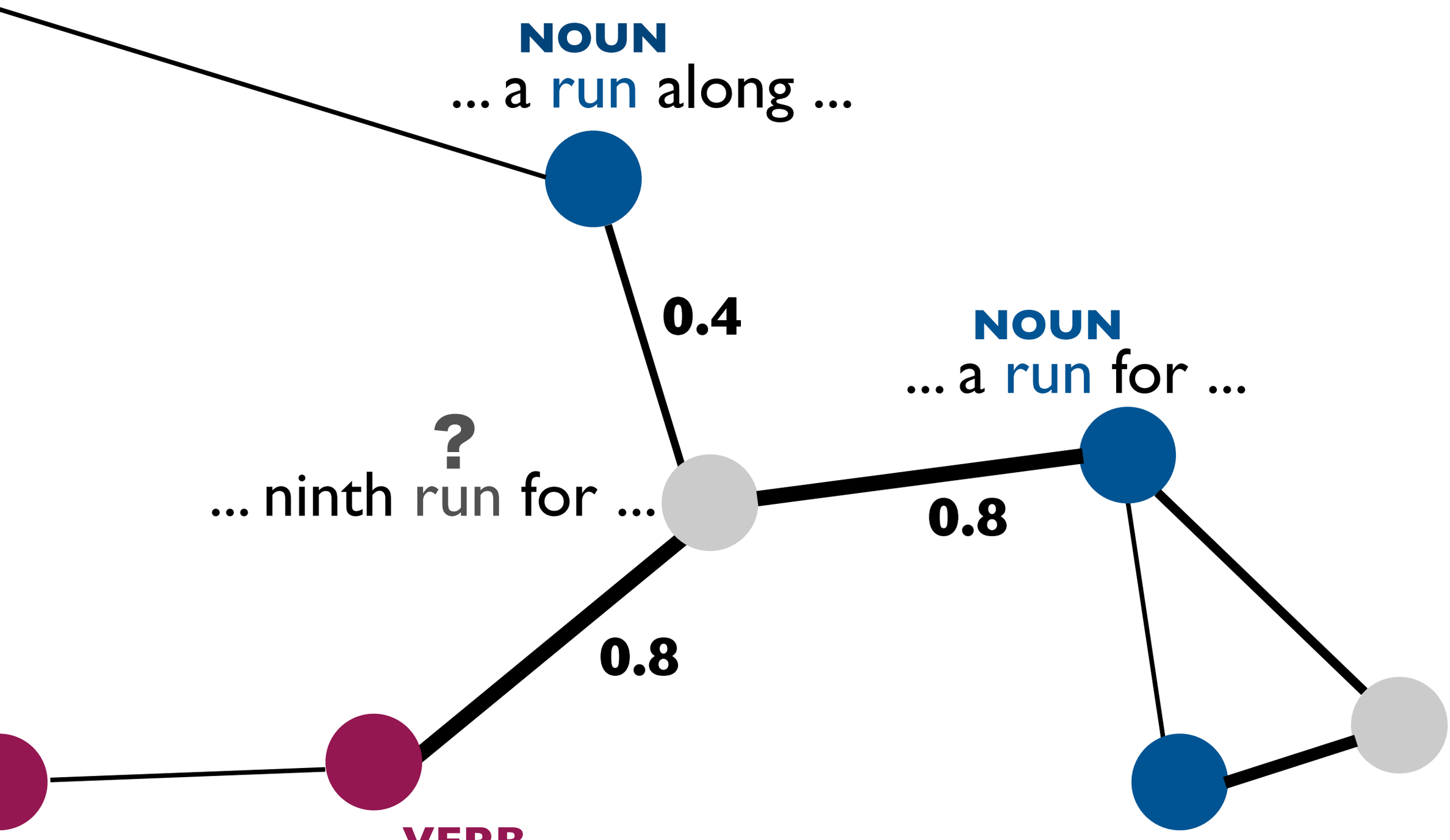


Zhu et al. (ICML 2013): Graph-based Semi-supervised Learning

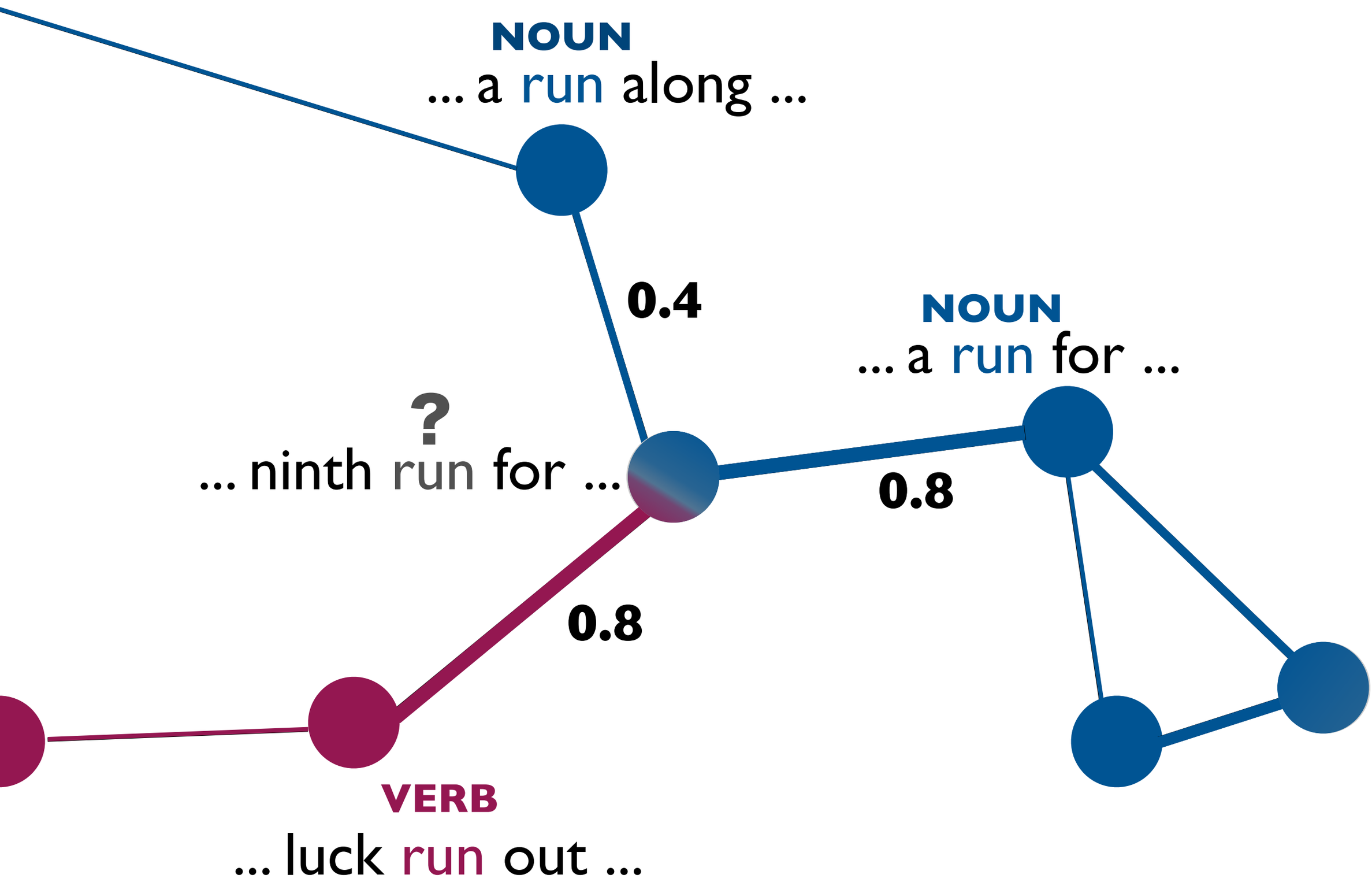
GRAPH LAPLACIAN REGULARIZER



GRAPH LAPLACIAN REGULARIZER



GRAPH LAPLACIAN REGULARIZER



GRAPH LAPLACIAN REGULARIZER

NOUN

... a run along ...

Prob (**NOUN** | ninth run for) = 0.6
Prob (**VERB** | ninth run for) = 0.4
Prob (ADV | ninth run for) = 0
Prob (DET | ninth run for) = 0
…

VERB

... luck run out ...

GRAPH LAPLACIAN REGULARIZER

NOUN

... a run along ...

$$\text{Prob}(\text{tag} \mid \text{ninth run for}) =$$

$$\arg\min_{m} \quad 0.4 \times \| m - \text{Prob}(\text{tag} \mid \text{a run along}) \|_2^2$$

$$+0.8 \times \| m - \text{Prob}(\text{tag} \mid \text{a run for}) \|_2^2$$

$$+0.8 \times \| m - \text{Prob}(\text{tag} \mid \text{luck run out}) \|_2^2$$

VERB

... luck run out ...

GRAPH LAPLACIAN REGULARIZER

NOUN

... a run along ...

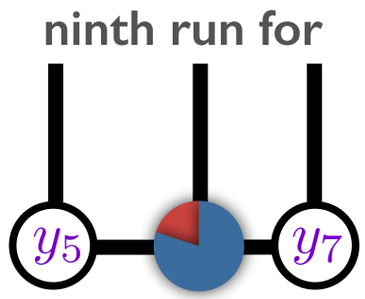
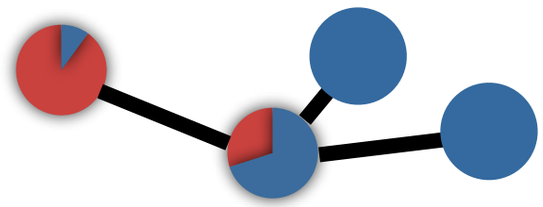
$$\min_{\mathbf{m}} \quad \mathrm{Lap}(\mathbf{m}) = \sum_{a \in \mathrm{Unlab}} \sum_{b \in \mathrm{Neighbors}(a)} \sum_{k \in \mathrm{Tags}} w_{ab}(m_{a,k} - m_{b,k})^2$$

$m_{a,k}$: The proportion of time trigram **a** has tag **k**

VERB

... luck run out ...

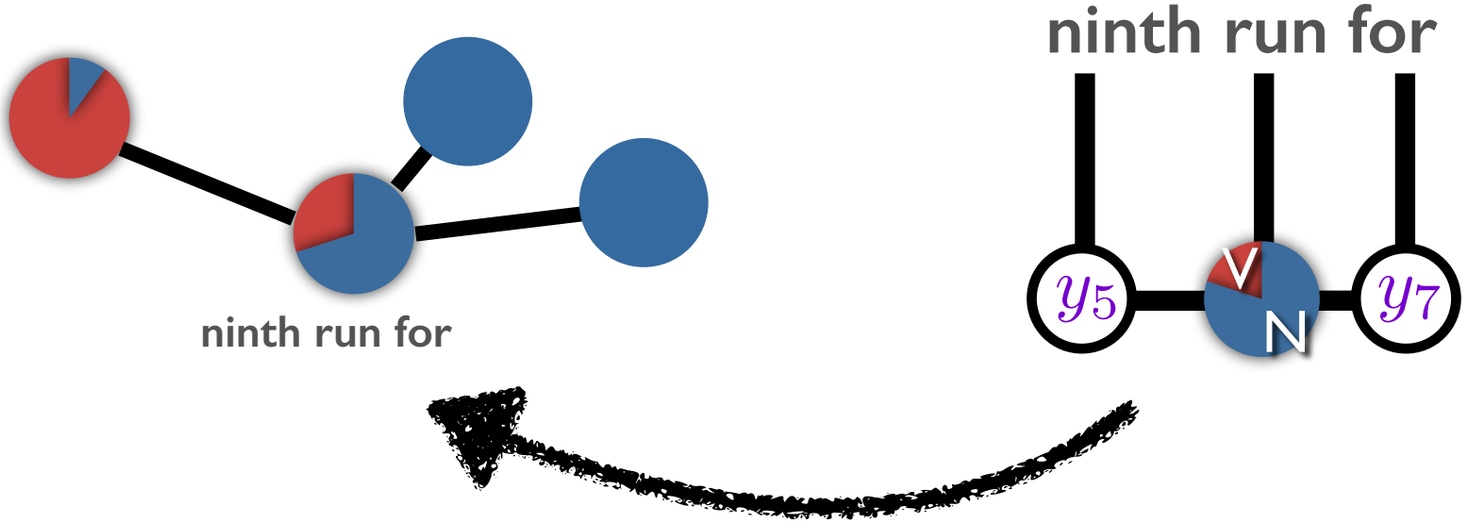
COMBINING THE TWO

	ninth run for  CRF estimation	 ninth run for graph-propagation
Data	**labeled**	**labeled + unlabeled**
Model	$p(\text{tags} \mid \text{sentence}; \theta)$	$m(\text{tag} \mid \text{trigram})$

PRIOR WORK

Subramanya et al. (EMNLP 2010)

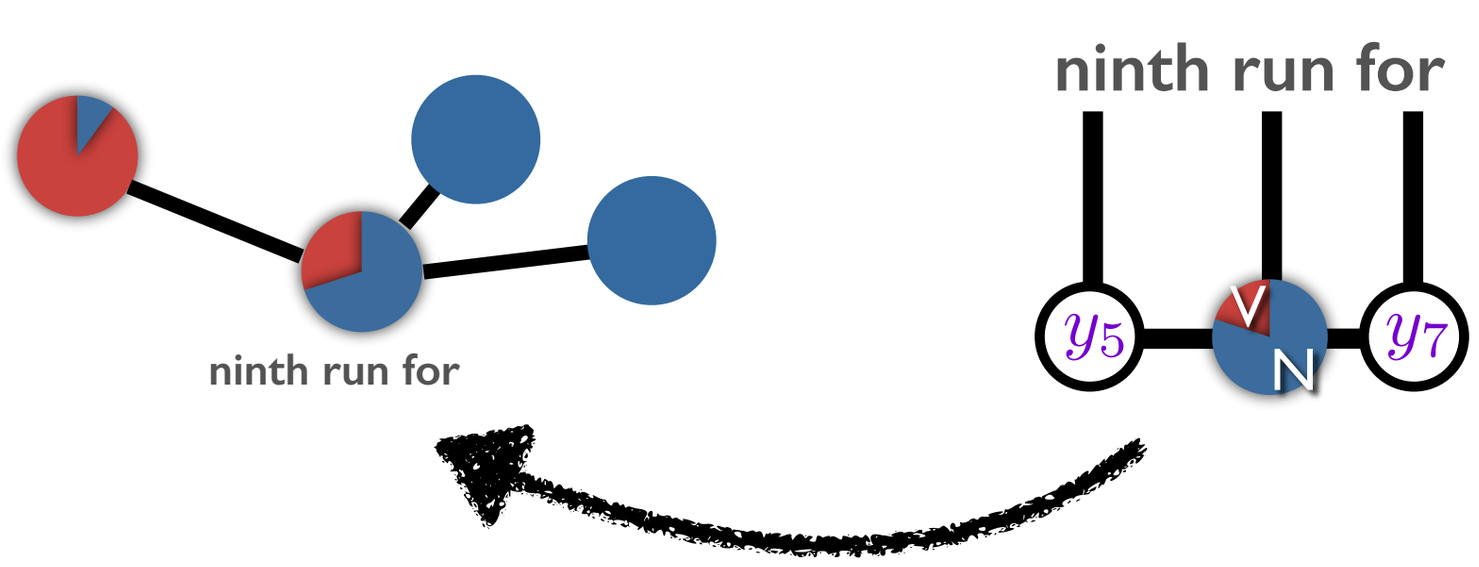
graph-propagation + CRF estimation



PRIOR WORK

Subramanya et al. (EMNLP 2010)

graph-propagation + CRF estimation



Our work: retains efficiency while optimizing an extendible, joint objective.

HOW TO COMBINE?

introduce auxiliary variables q

HOW TO COMBINE?

introduce auxiliary variables q

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

$$q_y^i = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

HOW TO COMBINE?

introduce auxiliary variables q

1. Normalized

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

$$q_y^i = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

HOW TO COMBINE?

introduce auxiliary variables q

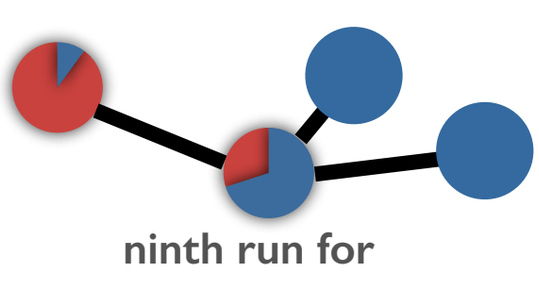
1. Normalized

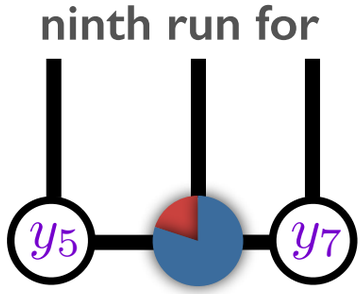
2. Decomposed into local factors

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

$$q_y^i = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

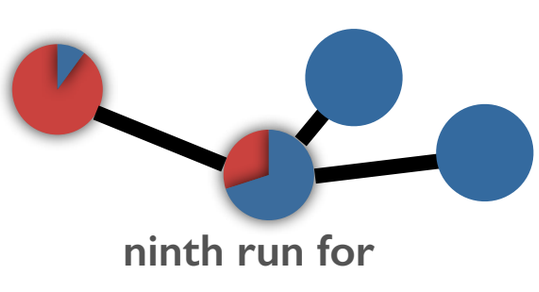
JOINT OBJECTIVE

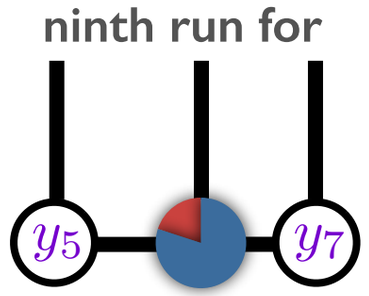




$$\mathcal{J}(q, p_\theta) \;=\; \mathrm{Lap}(q) + \;\mathrm{NLik}(p_\theta) \;+\; \mathrm{KL}(q \,\|\, p_\theta)$$

JOINT OBJECTIVE

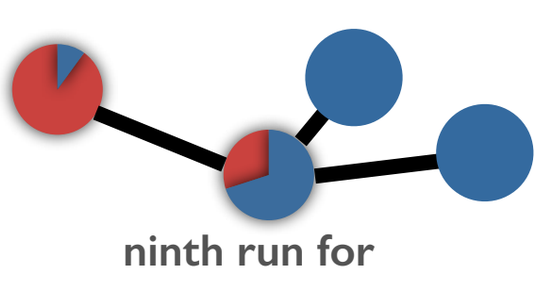


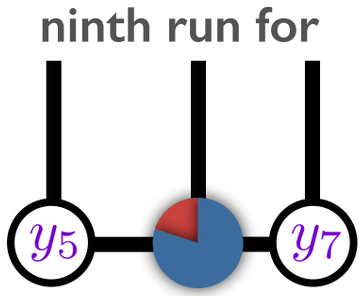


$$\mathcal{J}(q, p_\theta) = \text{Lap}(q) + \text{NLik}(p_\theta) + \text{KL}(q \,\|\, p_\theta)$$

$$\text{Lap}(q) = \sum_{a \in \text{Unlab}} \sum_{b \in \text{Neighbors}(a)} \sum_{k \in \text{Tags}} w_{ab}(m_{a,k}(q) - m_{b,k}(q))^2$$

JOINT OBJECTIVE

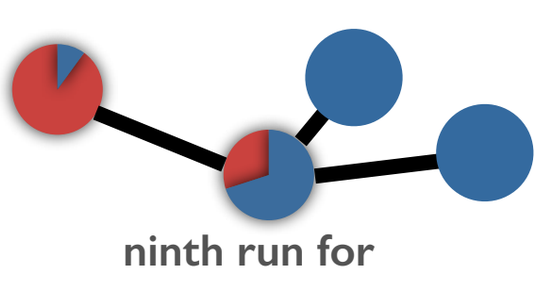


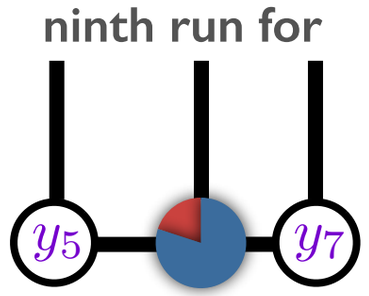


$$\mathcal{J}(q, p_\theta) \;=\; \mathrm{Lap}(q) + \;\mathrm{NLik}(p_\theta) \;+\; \mathrm{KL}(q \,\|\, p_\theta)$$

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

JOINT OBJECTIVE





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

$$\text{KL}(q \parallel p_\theta) = \sum_{i=1}^{n} \sum_{\mathbf{y}} q^i_{\mathbf{y}} \log \frac{q^i_{\mathbf{y}}}{p_\theta(\mathbf{y} \mid \mathbf{x}^i)}$$

HOW TO OPTIMIZE?

$$\min_{q,\theta} \mathcal{J}(q, p_\theta)$$

HOW TO OPTIMIZE?

$$\min_{q,\theta} \mathcal{J}(q, p_\theta)$$

Δ unconstrained

HOW TO OPTIMIZE?

$$\min_{q,\theta} \mathcal{J}(q, p_\theta)$$

Δ ↗ q,θ ↖ unconstrained

p update:

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

HOW TO OPTIMIZE?

$$\min_{q,\theta} \mathcal{J}(q, p_\theta)$$

Δ ↗ ↖ unconstrained

p update:

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

q update:

projection is hard $\quad \sum_{\mathbf{y}} q^i_{\mathbf{y}} = 1$

no compact form $\quad (\#\ \text{tags})^{(i\text{'s length})}\ \text{values}$

UPDATE Q

UPDATE Q

q can be represented by local factors **r**

$$q^i_{\mathbf{y}} = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

UPDATE Q

q can be represented by local factors **r**

$$q^i_{\mathbf{y}} = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

doing an additive gradient update

$$q^{i\,\prime}_{\mathbf{y}} = q^i_{\mathbf{y}} - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_{\mathbf{y}}}$$

UPDATE Q

q can be represented by local factors **r**

$$q^i_{\mathbf{y}} = \frac{1}{Z_q(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right]$$

doing an additive gradient update

$$q^{i\,\prime}_{\mathbf{y}} = q^i_{\mathbf{y}} - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_{\mathbf{y}}}$$

q' cannot be written as product of local factors!

EXPONENTIATED GRADIENT

multiplicative gradient update:

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

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

EXPONENTIATED GRADIENT

multiplicative gradient update:

$$q_{\mathbf{y}}^{i\,\prime} = \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right] \exp\left[-\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_{\mathbf{y}}^i}\right]$$

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

EXPONENTIATED GRADIENT

multiplicative gradient update:

$$q_{\mathbf{y}}^{i\,\prime} = \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right] \exp\left[-\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_{\mathbf{y}}^i}\right]$$

decompose into local factors

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

EXPONENTIATED GRADIENT

multiplicative gradient update:

$$q_{\mathbf{y}}^{i\,\prime} = \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right] \exp\left[-\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q_{\mathbf{y}}^i}\right]$$

decompose into local factors

$$= \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r'_{i,t}(y_t, y_{t-1})\right]$$

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

EXPONENTIATED GRADIENT

multiplicative gradient update:

$$q^{i\,\prime}_{\mathbf{y}} = \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r_{i,t}(y_t, y_{t-1})\right] \exp\left[-\eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial q^i_{\mathbf{y}}}\right]$$

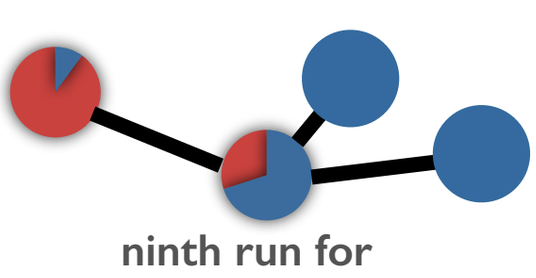
decompose into local factors

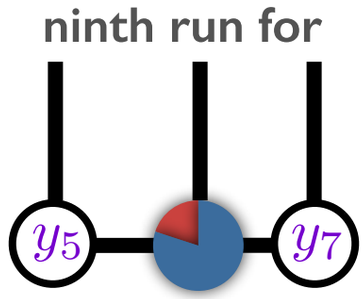
$$= \frac{1}{Z_{q'}(\mathbf{x}^i)} \exp\left[\sum_{t=1}^{T} r'_{i,t}(y_t, y_{t-1})\right]$$

only updating $(\#\text{tags})^2 \times (i\text{'s length})$ variables!

Collins et al. (JMLR 2008): Exponentiated gradient for CRFs

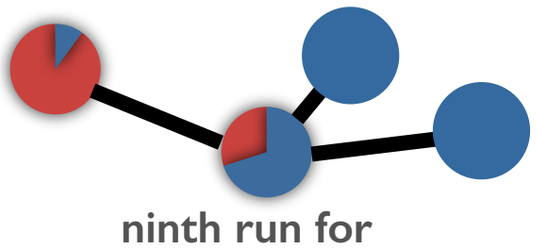
SUMMARY

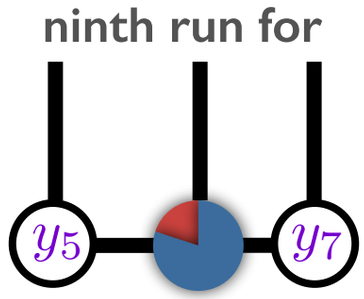




$$\mathcal{J}(q, p_\theta) \;=\; \mathrm{Lap}(q) \;+\; \mathrm{NLik}(p_\theta) \;+\; \mathrm{KL}(q \,\|\, p_\theta)$$

SUMMARY





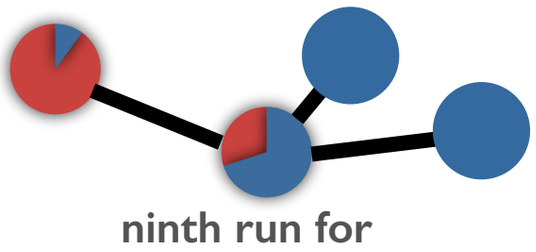
$$\mathcal{J}(q, p_\theta) \;=\; \mathrm{Lap}(q) \;+\; \mathrm{NLik}(p_\theta) \;+\; \mathrm{KL}(q \,\|\, p_\theta)$$

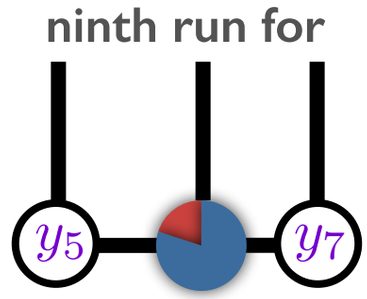
M-step: $\theta' = \theta - \eta \frac{\partial \mathcal{J}(q, p_\theta)}{\partial \theta}$

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

(update each $r_{i,t}$ in practice)

SUMMARY





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

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

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

(update each $r_{i,t}$ in practice)

Theorem:
Converges to a local optimum of
$\mathcal{J}(q, p_\theta)$

EXPERIMENT SETTING

10 Languages (CoNLL-X and CoNLL-2007)

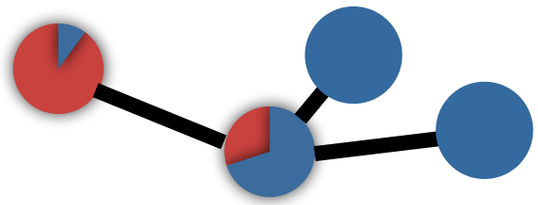
100 Randomly sampled labeled sentences

Averaged over 10 sampling runs

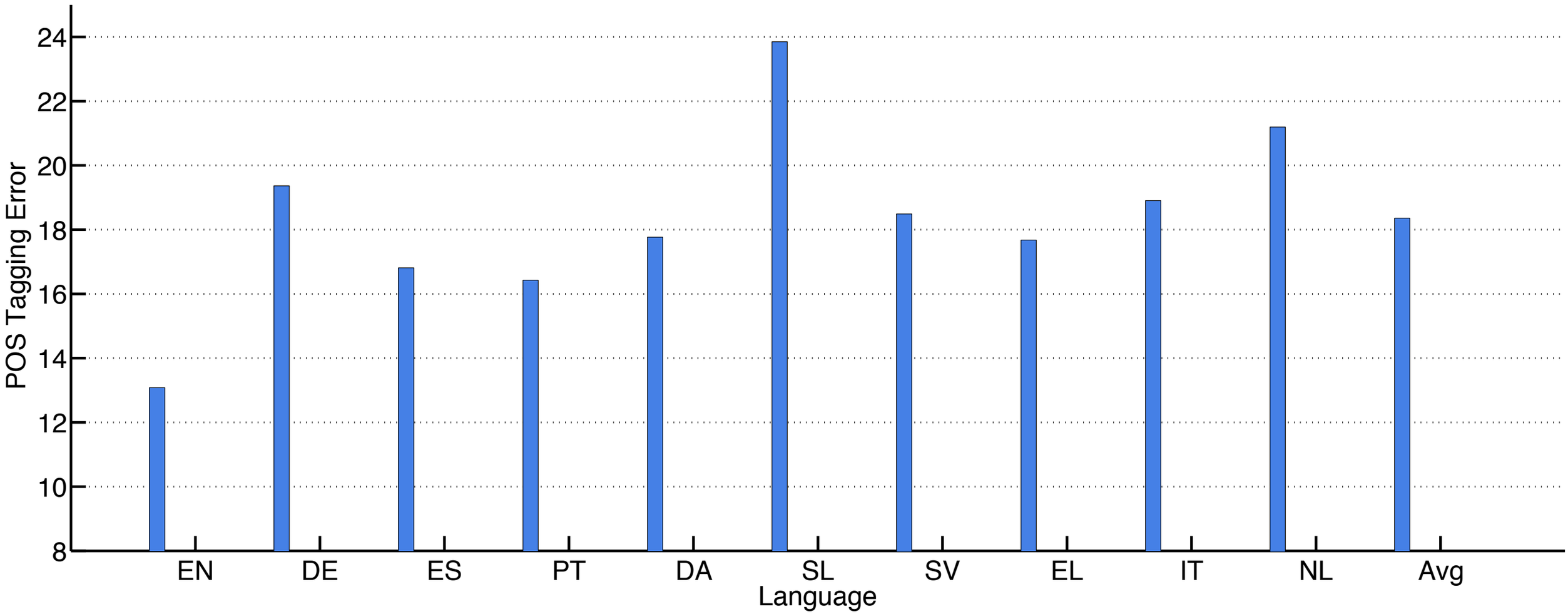
Universal POS Tags (Petrov et al. 2011)

Second Order CRF Model $\mathbf{f}(y_t, y_{t-1}, y_{t-2}, \mathbf{x})$

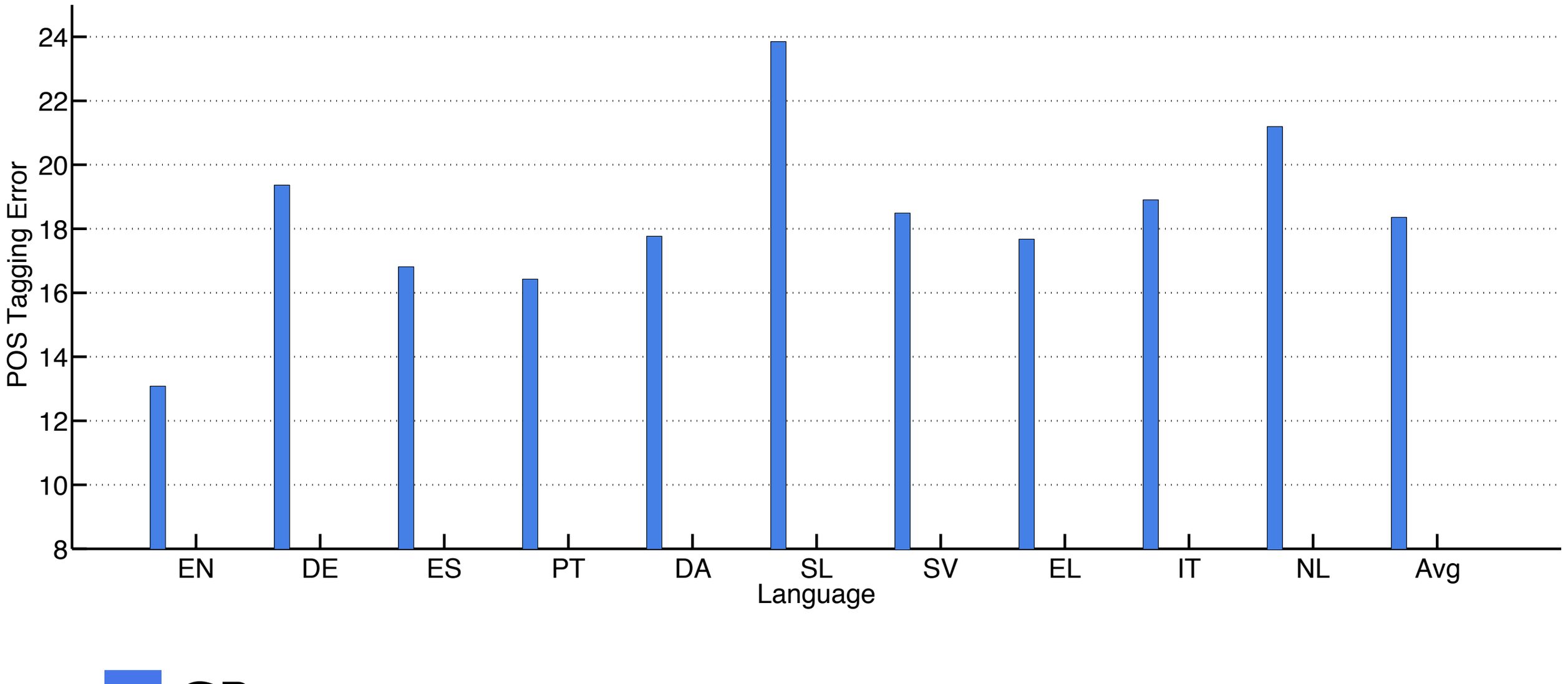
graph propagation



ninth run for

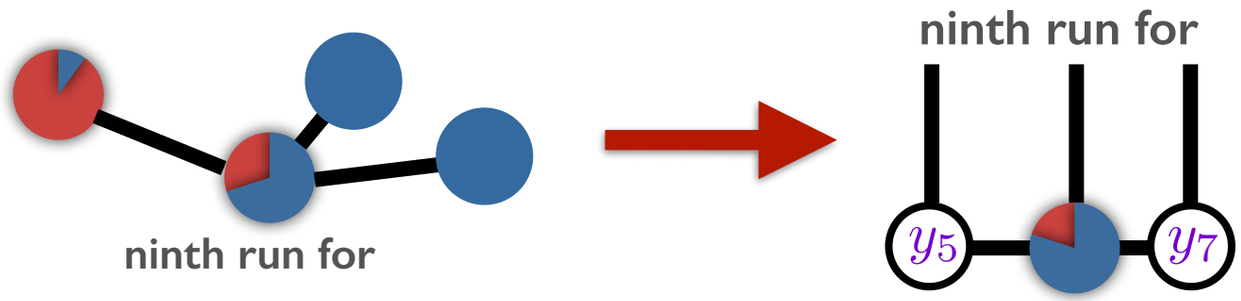


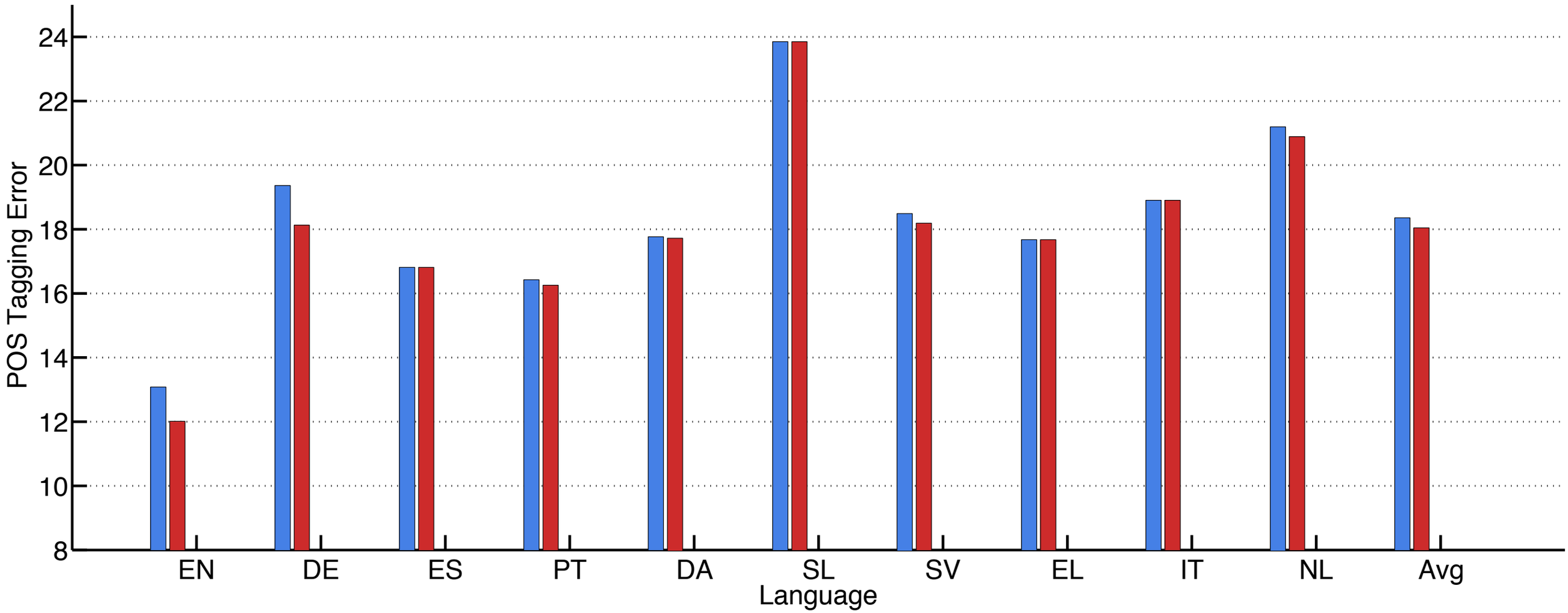
GP


POS Tagging Error
24
22
20
18
16
14
12
10
8
EN
DE
ES
PT
DA
SL
SV
EL
IT
NL
Avg
Language

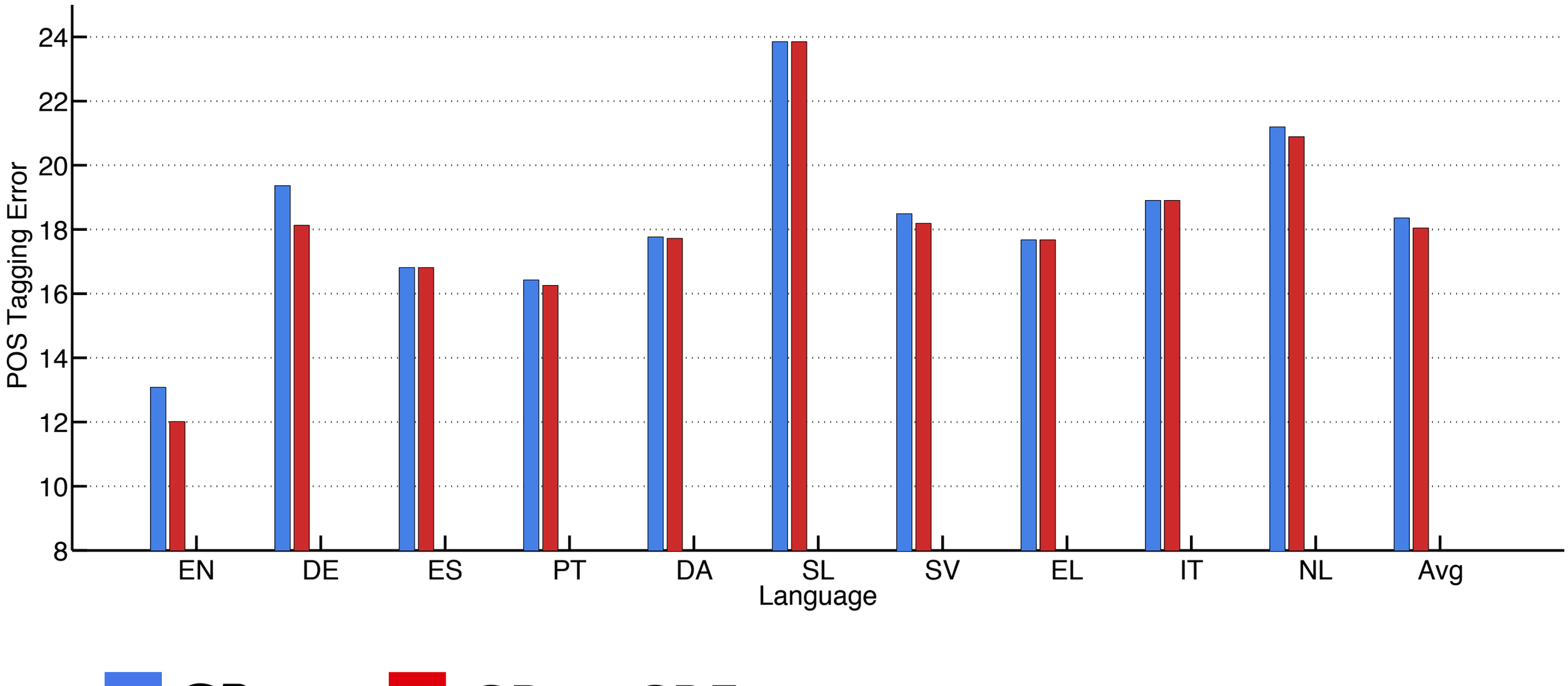
GP

GP → CRF



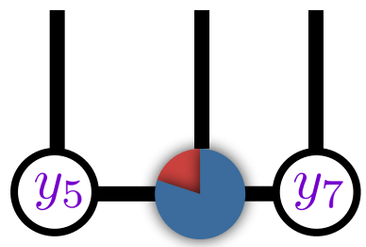


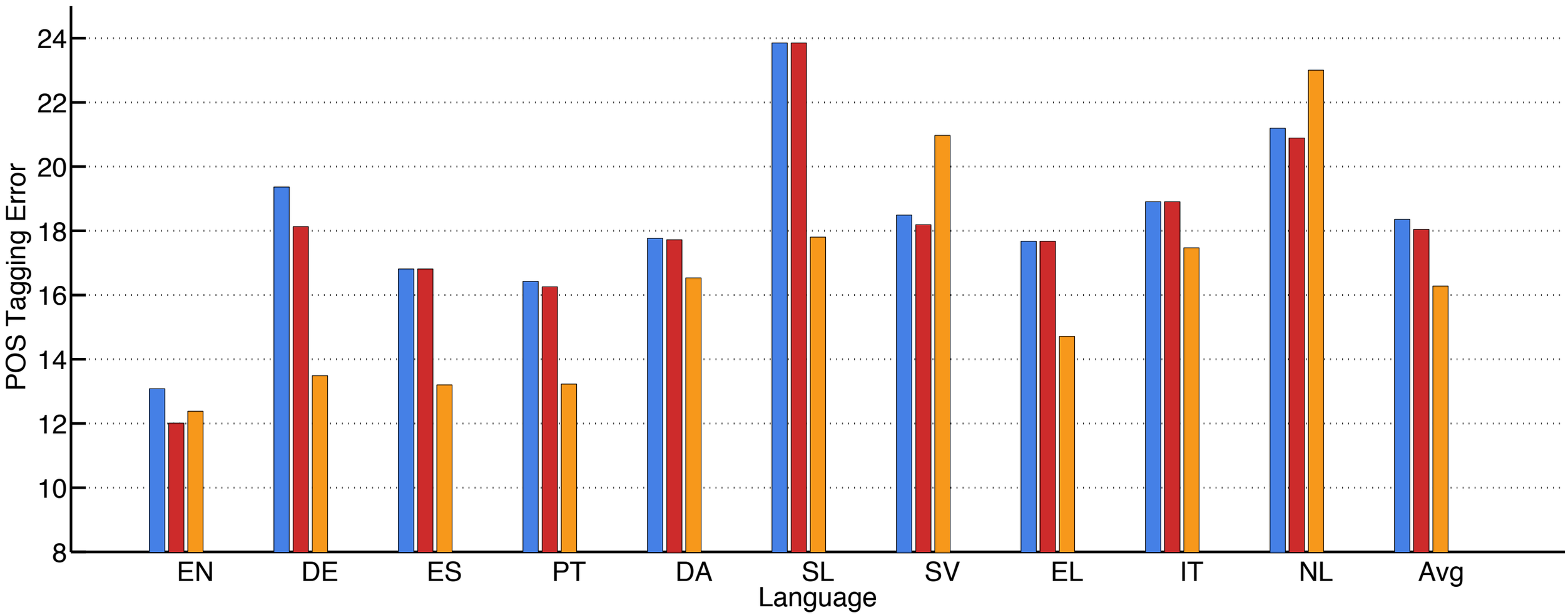
GP GP → CRF


POS Tagging Error
24
22
20
18
16
14
12
10
8
EN
DE
ES
PT
DA
SL
SV
EL
IT
NL
Avg
Language
GP
GP → CRF

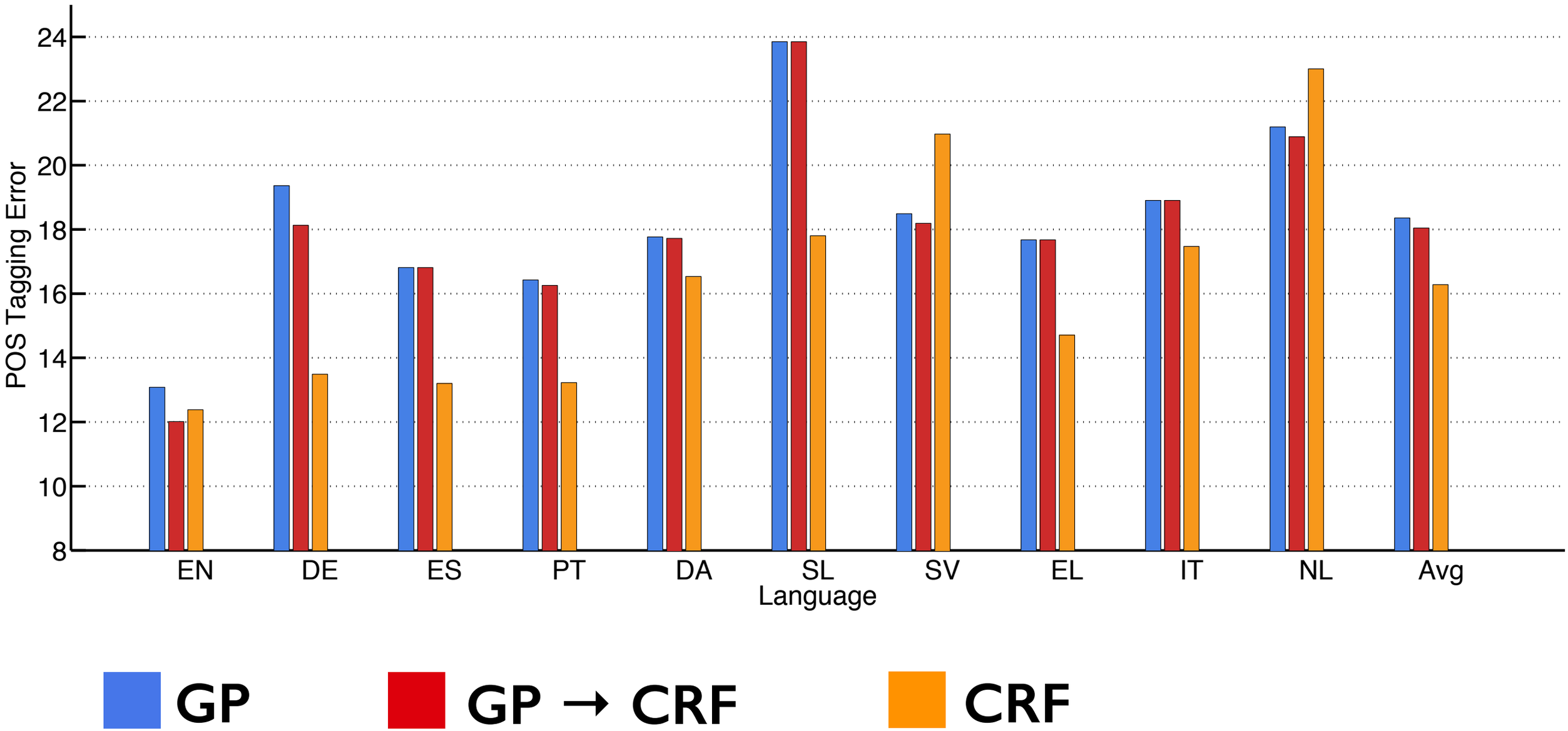
CRF

ninth run for

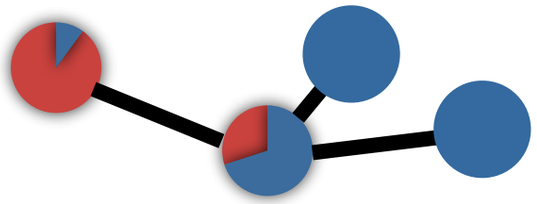
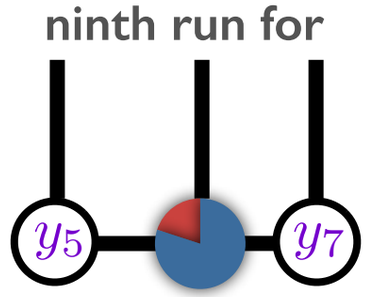




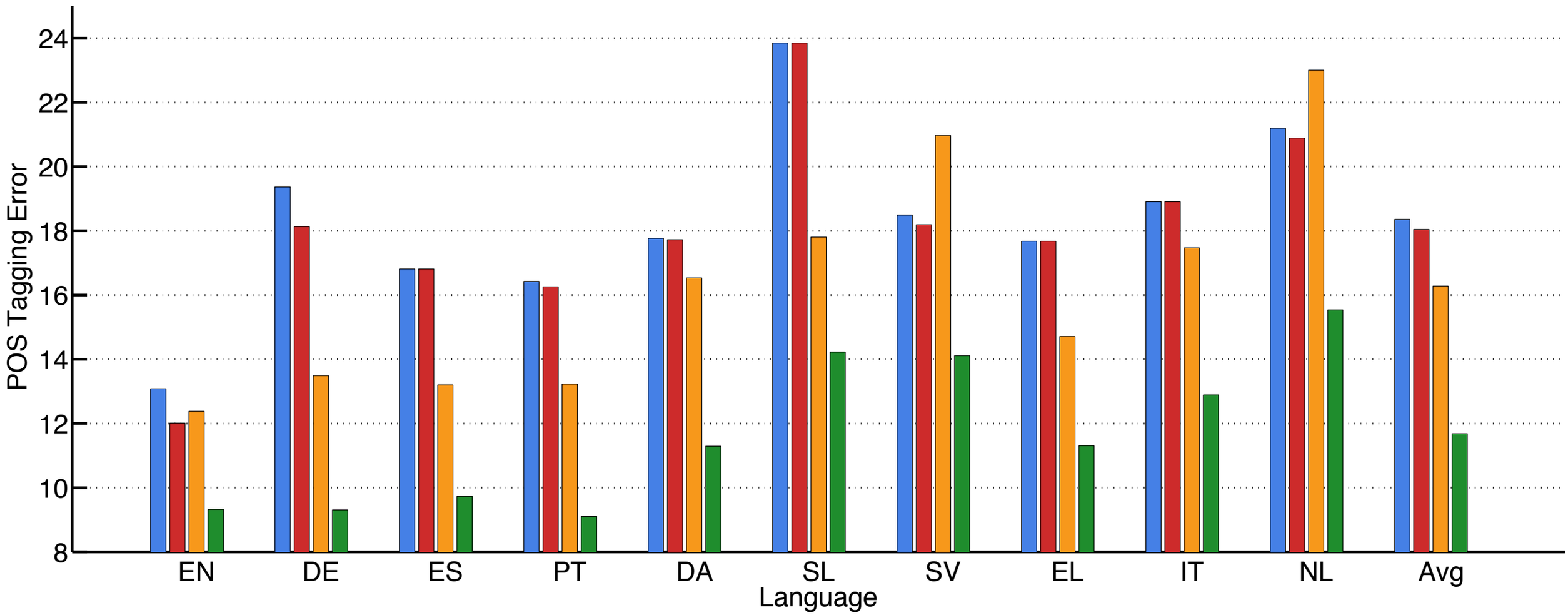
GP **GP → CRF** **CRF**


POS Tagging Error
24
22
20
18
16
14
12
10
8
EN
DE
ES
PT
DA
SL
SV
EL
IT
NL
Avg
Language
GP
GP → CRF
CRF

J

$$\mathrm{KL}\Big($$  ninth run for $$\Big\|$$ ninth run for 

$$\Big)$$



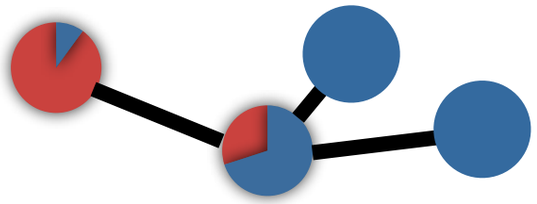
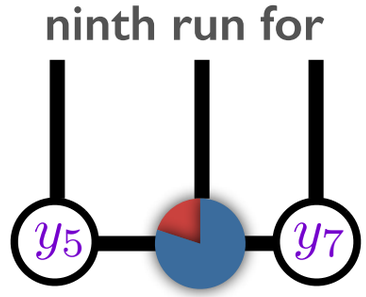
GP

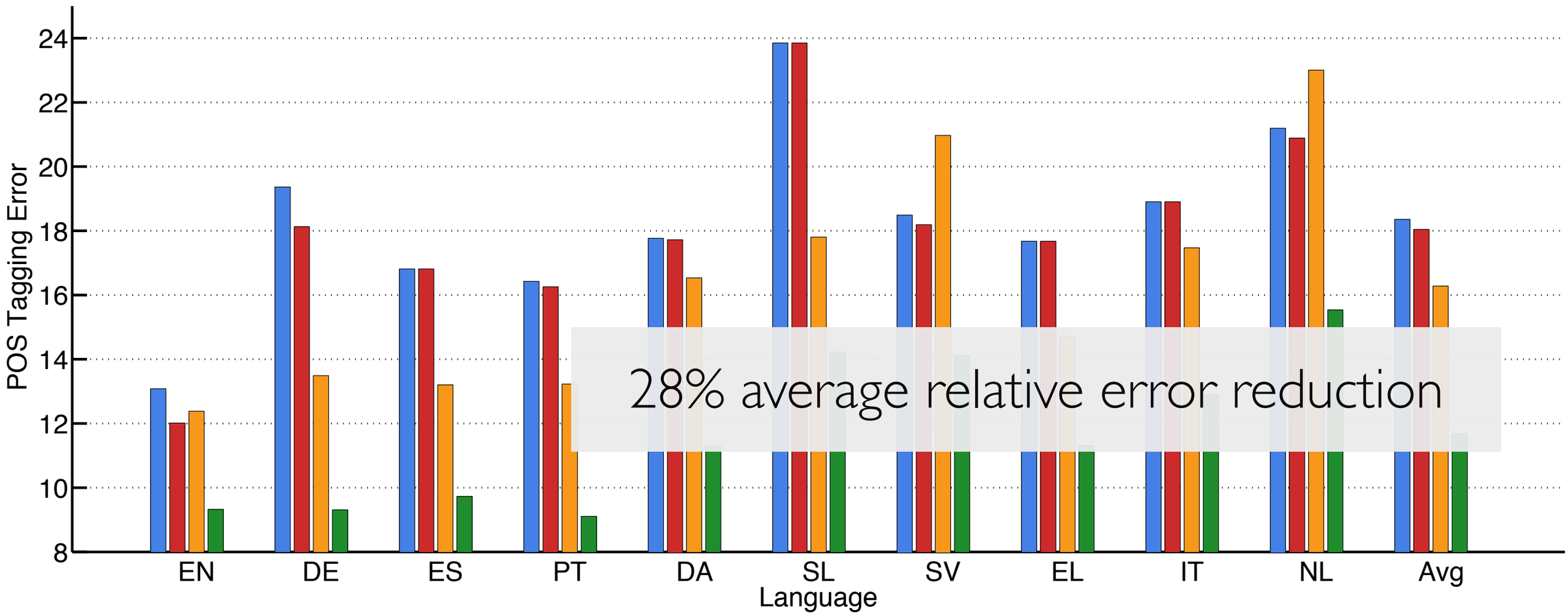
GP → CRF

CRF

J

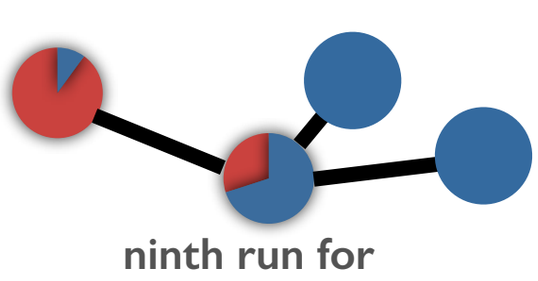
J

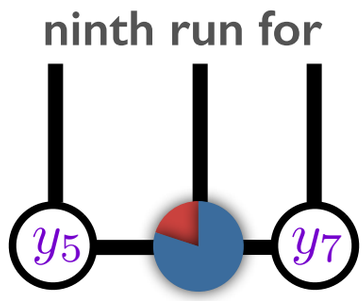
$$\mathrm{KL}\Big($$  ninth run for $$\Big\|$$ ninth run for  $$\Big)$$



GP GP → CRF CRF J

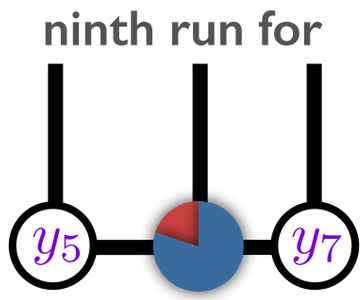
CONCLUSION





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

CONCLUSION

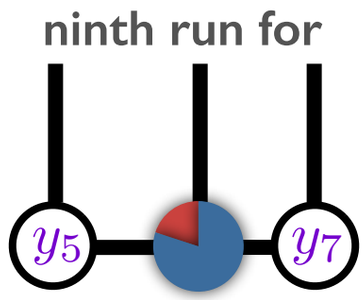


Posterior Constraint

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

any convex, differentiable regularizer

CONCLUSION



Posterior Constraint

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

any convex, differentiable regularizer

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