Structured Output Learning with Indirect Supervision

Chang,Srikumar,Goldwasser,Roth(ICML2010)

Presented by: Sanjay Subramanian

October 26, 2017

Overview

Example: Object Part Recognition (Source: [1])

Structured Output Learning

Given a car image, where are the body, windows and wheels?

Example: Object Part Recognition (Source: [1])

Structured Output Learning

Given a car image, where are the body, windows and wheels?

Companion Binary Output Problem

Is there a car in this image?

Example: Object Part Recognition (Source: [1])

Structured Output Learning

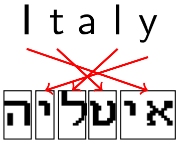
Given a car image, where are the body, windows and wheels?

Companion Binary Output Problem

Is there a car in this image?

- Only a car image can contain car parts in the right position!
- A non-car image cannot have the car parts in the right position

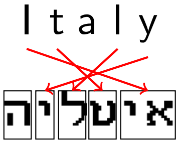
Example: Phonetic Alignment (Source: [1])



Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?

Example: Phonetic Alignment (Source: [1])



Israel

Yes/No

אילינוי

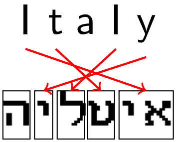
Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?

Companion Binary Output Problem

Are these two NEs a transliteration pair?

Example: Phonetic Alignment (Source: [1])



Israel

Yes/No

אילינוי

Structured Output Learning

Given one English NE and its Hebrew transliteration, tell me what is the phonetic alignment?

Companion Binary Output Problem

Are these two NEs a transliteration pair?

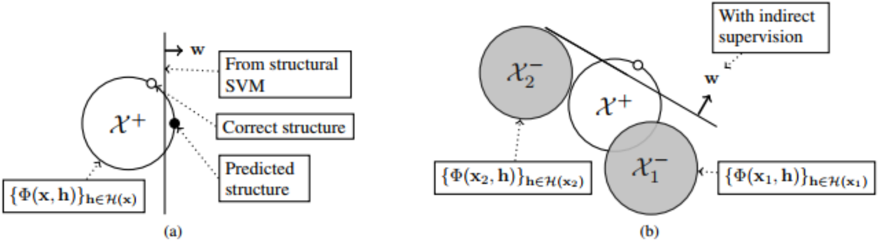
Relationships

- Only a transliteration pair can have good phonetic alignment!
- Non-transliteration pairs cannot have good phonetic alignment!

Model Setup and Assumption

- Let $S = \{(\mathbf{x}_i, \mathbf{h}_i)\}_{i=1}^{l}$ be the direct supervision set.
- Let $B = \{(\mathbf{x}_i, y_i)\}_{i=l+1}^{l+m}$ be the indirect supervision set.
- Let $B^+ = \{(\mathbf{x}_i, y_i) \in B : y_i = 1\}$. Let $B^- = \{(\mathbf{x}_i, y_i) \in B : y_i = -1\}$.
- We want to find $\mathbf{w}$ s.t. $\mathbf{h}_i = \arg\max_{h \in \mathcal{H}(\mathbf{x})} \mathbf{w}^T \Phi(\mathbf{x}_i, \mathbf{h})$.
- Assumption:
  1. $\forall(\mathbf{x}, -1) \in B^-, \forall \mathbf{h} \in \mathcal{H}(\mathbf{x}), \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{h}) \leq 0$
  2. $\forall(\mathbf{x}, +1) \in B^+, \exists \mathbf{h} \in \mathcal{H}(\mathbf{x}), \mathbf{w}^T \Phi(\mathbf{x}, \mathbf{h}) \geq 0$

Visual intuition for assumption



(Source: the paper [2])

Loss function

- Standard structural SVM loss:

$$L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) \equiv \ell\left(\max_{\mathbf{h}}\left[\Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}_i) + \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h})\right]\right)$$

$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$$

Δ: Hamming distance $\quad \ell$: convex, non-decreasing loss function

Loss function

- Standard structural SVM loss:

$$L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) \equiv \ell\left(\max_{\mathbf{h}}\left[\Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}_i) + \mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h})\right]\right)$$

$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i\in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$$

- Structural + Binary loss:

$$L_B(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) \equiv \ell\left(1 - y_i \max_{h\in\mathcal{H}(\mathbf{x})}(\mathbf{w}^T\Phi(\mathbf{x}_i, \mathbf{h}))\right)$$

$$Q(\mathbf{w}) = \min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i\in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i\in B} L_B(\mathbf{x}_i, y_i, \mathbf{w})$$

Δ: Hamming distance ℓ: convex, non-decreasing loss-function

Loss function

- Standard structural SVM loss:

$$L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) \equiv \ell\left(\max_{\mathbf{h}}\left[\Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T \underbrace{(\Phi(\mathbf{x}_i, \mathbf{h}_i) - \Phi(\mathbf{x}_i, \mathbf{h}))}_{\Phi_{\mathbf{h}_i,\mathbf{h}}(\mathbf{x}_i)}\right]\right)$$

$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w})$$

- Structural + Binary loss with normalization:

$$L_B(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) \equiv \ell\left(1 - y_i \max_{h \in \mathcal{H}(\mathbf{x})} (\mathbf{w}^T \underbrace{\frac{\Phi(\mathbf{x}_i, \mathbf{h})}{\kappa(\mathbf{x}_i)}}_{\Phi_B(\mathbf{x},\mathbf{h})})\right)$$

$$Q(\mathbf{w}) = \min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in B} L_B(\mathbf{x}_i, y_i, \mathbf{w})$$

$\boxed{\Delta\text{: Hamming distance} \quad \ell\text{: convex, non-decreasing loss-function}}$

Convex relaxation

$$Q(\mathbf{x}) = \underbrace{\frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} L_S(\mathbf{x}_i, \mathbf{h}_i, \mathbf{w}) + C_2 \sum_{i \in B^-} L_B(\mathbf{x}_i, y_i, \mathbf{w})}_{F(\mathbf{w}) \text{ convex}}$$

$$+ \underbrace{C_2 \sum_{i \in B} \ell \left(1 - \max_{\mathbf{h}} \left(\mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}) \right) \right)}_{G(\mathbf{w}) \text{ no concave/convex guarantee}}$$

We want to approximate $G(\mathbf{w})$ using a function that is convex in $\mathbf{w}$.

Δ: Hamming distance	ℓ: convex, non-decreasing loss-function
L_S: Structural loss	L_B: Binary loss

Convex relaxation

- Iterative approach: when computing $\mathbf{w}_{t+1}$, compute the max using $\mathbf{w}_t$:

$$\mathbf{h}_i^t \equiv \arg\max_{\mathbf{h}} \left(\mathbf{w}_t^T \Phi_B(\mathbf{x}_i, \mathbf{h}) \right)$$

$$\hat{G}(\mathbf{w}, \mathbf{w}_t) = G_t(\mathbf{w}) \equiv \sum_{i \in B} \ell \left(1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t) \right)$$

Δ: Hamming distance	ℓ: convex, non-decreasing loss-function
L_S: Structural loss	L_B: Binary loss

Convex relaxation

- Iterative approach: when computing $\mathbf{w}_{t+1}$, compute the max using $\mathbf{w}_t$:

$$\mathbf{h}_i^t \equiv \arg\max_{\mathbf{h}} \left(\mathbf{w}_t^T \Phi_B(\mathbf{x}_i, \mathbf{h}) \right)$$

$$\hat{G}(\mathbf{w}, \mathbf{w}_t) = G_t(\mathbf{w}) \equiv \sum_{i \in B^+} \ell \left(1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t) \right)$$

- Iteratively compute $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} A(\mathbf{w}, \mathbf{w}_t)$, where

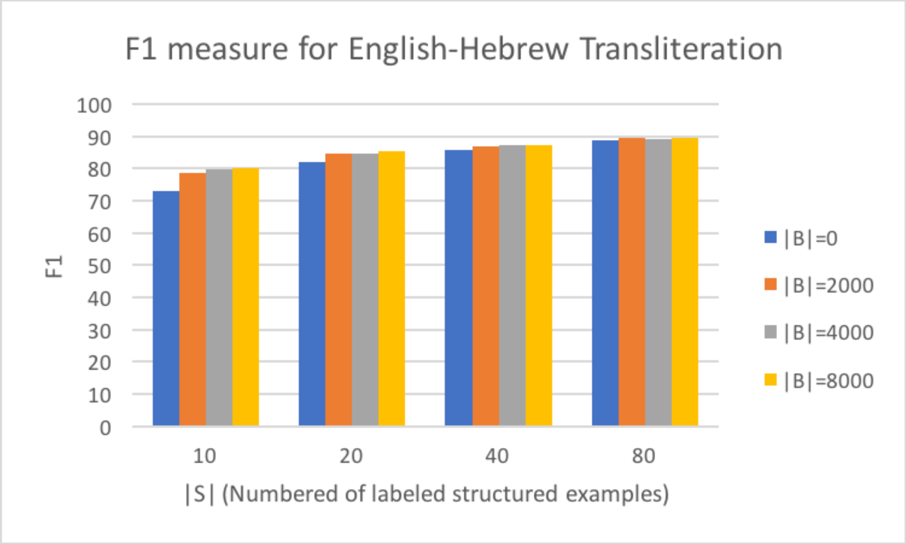
$$A(\mathbf{w}, \mathbf{w}_t) \equiv F(\mathbf{w}) + \hat{G}(\mathbf{w}, \mathbf{w}_t)$$

Theorem

If $\ell(\cdot)$ is convex and non-decreasing, then $Q(\mathbf{w}_{t+1}) \leq Q(\mathbf{w}_t)\ \forall t \geq 0$.

Δ: Hamming distance	ℓ: convex, non-decreasing loss-function
L_S: Structural loss	L_B: Binary loss

Results for Phonetic Transliteration



Formulation for Squared Hinge Loss

$$A(\mathbf{w}, \mathbf{w}_t) = \min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} \ell \left[\Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T \Phi_{\mathbf{h}_i, \mathbf{h}}(\mathbf{x}_i) \right]$$
$$+ C_2 \sum_{i \in B^-} \ell \left[1 + \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}) \right]$$
$$+ C_2 \sum_{i \in B^+} \ell \left(1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t) \right)$$

$$\min_{\mathbf{w}, \xi} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} \xi_i^2 + C_2 \sum_{i \in B} \xi_i^2$$
$$s.t. \ \forall i \in S, \mathbf{h} \in \mathcal{W}_i, \xi_i \geq \Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T \Phi_{\mathbf{h}_i, \mathbf{h}}(\mathbf{x}_i)$$
$$\forall i \in B^-, \mathbf{h} \in \mathcal{V}_i, \xi_i \geq 1 + \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h})$$
$$\forall i \in B^+, \xi_i \geq 1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t)$$

Δ: Hamming distance	$\Phi_{\mathbf{h}_i, \mathbf{h}}(\mathbf{x}_i) = \Phi(\mathbf{x}_i, \mathbf{h}_i) - \Phi(\mathbf{x}_i, \mathbf{h})$
$\mathcal{W}$: "Support vectors" for S	$\mathcal{V}$: "Support vectors" for B

Dual formulation for Squared Hinge Loss

$$\min_{\mathbf{w},\xi} \frac{||\mathbf{w}||^2}{2} + C_1 \sum_{i \in S} \xi_i^2 + C_2 \sum_{i \in B} \xi_i^2$$

$$s.t.\ \forall i \in S, \mathbf{h} \in \mathcal{W}_i, \xi_i \geq \Delta(\mathbf{h}, \mathbf{h}_i) - \mathbf{w}^T \Phi_{\mathbf{h}_i,\mathbf{h}}(\mathbf{x}_i)$$

$$\forall i \in B^-, \mathbf{h} \in \mathcal{V}_i, \xi_i \geq 1 + \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h})$$

$$\forall i \in B^+, \xi_i \geq 1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t)$$

$$L(\mathbf{w}, \xi, \alpha) = \frac{||\mathbf{w}||^2}{2} - \sum_{i \in S} \sum_{\mathbf{h}_{i,j} \in \mathcal{W}_i} \alpha_{i,j} \left[\xi_i - \Delta(\mathbf{h}_{i,j}, \mathbf{h}_i) + \mathbf{w}^T \Phi_{\mathbf{h}_i,\mathbf{h}_{i,j}}(\mathbf{x}_i)\right]$$

$$+ C_1 \sum_{i \in S} \xi_i^2 - \sum_{i \in B^-} \sum_{\mathbf{h}_{i,j} \in \mathcal{V}_i^-} \alpha_{i,j} \left[\xi_i - 1 - \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_{i,j})\right]$$

$$+ C_2 \sum_{i \in B} \xi_i^2 - \sum_{i \in B^+} \alpha_i \left[\xi_i - 1 + \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t)\right]$$

$\Phi_{\mathbf{h}_i,\mathbf{h}}(\mathbf{x}_i) = \Phi(\mathbf{x}_i, \mathbf{h}_i) - \Phi(\mathbf{x}_i, \mathbf{h})$	$\mathcal{W}$: "Support vectors" for S
$\mathcal{V}$: "Support vectors" for B	$\alpha_{i,j}$: Dual variables for each primal constraint

Dual formulation for Squared Hinge Loss

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \implies \mathbf{w} = \sum_{i \in S} \sum_{j:\mathbf{h}_{i,j} \in \mathcal{W}_i} \alpha_{i,j} \Phi_{\mathbf{h}_i, \mathbf{h}_{i,j}}(\mathbf{x}_i)$$

$$- \sum_{i \in B^-} \sum_{j:\mathbf{h}_{i,j} \in \mathcal{V}_i} \alpha_{i,j} \Phi_B(\mathbf{x}_i, \mathbf{h}_{i,j}) + \sum_{i \in B^+} \alpha_i \Phi_B(\mathbf{x}_i, \mathbf{h}_i^t)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \implies \xi_i = \frac{1}{2C_1} \sum_{j:\mathbf{h}_{i,j} \in \mathcal{W}_i} \alpha_{i,j} \text{ if } i \in S$$

$$\xi_i = \frac{1}{2C_2} \sum_{j:\mathbf{h}_{i,j} \in \mathcal{V}_i} \alpha_{i,j} \text{ if } i \in B$$

Dual Update Rule

Substitute for $\mathbf{w}$ and ξ in Lagrangian, fix i, j, and derive following update rule for $\alpha_{i,j}$. Iteratively update the $\alpha_{i,j}$'s and $\mathbf{w}$ until convergence.

Case 1: $i \in S$:

$$\alpha^*_{i,j} = \max\left(0, \alpha_{i,j} + \frac{\Delta(\mathbf{h}_i, \mathbf{h}_{i,j}) - \mathbf{w}^T \Phi_{\mathbf{h}_i,\mathbf{h}_{i,j}}(\mathbf{x}_i) - \frac{\sum_j \alpha_{i,j}}{2C_1}}{||\Phi_{\mathbf{h}_i,\mathbf{h}_{i,j}}(\mathbf{x}_i)||^2 + \frac{1}{2C_1}}\right)$$

Case 2: $i \in B$ ($z_i = 1$ if $i \in B^-$ and $z_i = 0$ if $i \in B^+$)

$$\alpha^*_{i,j} = \max\left(0, \alpha_{i,j} + \frac{1 - z_i \mathbf{w}^T \Phi_B(\mathbf{x}_i, \mathbf{h}_{i,j}) - \frac{\sum_j \alpha_{i,j}}{2C_2}}{||\Phi_B(\mathbf{x}_i, \mathbf{h}_{i,j})||^2 + \frac{1}{2C_2}}\right)$$

References

[1] Chang, Ming-Wei, et al. Structured Output Learning with Indirect Supervision. http://cogcomp.org/papers/Chang11.pdf.
[2] Chang, Ming-Wei, et al. Structured Output Learning with Indirect Supervision. International Conference on Machine Learning (ICML), 2010.