Constraint-Driven Rank-Based Learning for Information Extraction

By Sameer Singh, Limin Yao, Sebastian Riedel, and Andrew McCallum

Presented by Alan Bauer

The Problem

“Most supervised learning algorithms for undirected graphical models require full inference over the dataset, small subsets of the dataset, or at least a single instance before parameter updates are made. Often this is the main computational bottleneck during training.”

Conditional random fields are one such undirected graphical model. Here is the probability distribution defined by it:

y - assignments to output variables

x - observation

$\mathbf{\Theta}$ - parameters

$$p(\mathbf{y}|\mathbf{x},\Theta) = \frac{1}{Z(\mathbf{x})} \prod_{\Psi_i \in G} e^{\Theta \cdot \mathbf{f}(\mathbf{x}_i,\mathbf{y}_i)}$$

SampleRank

Developed by the authors, it allows updating parameters without the full inference necessary using other learning algorithms.

Every pair of consecutive samples in the MCMC chain is ranked according to two things:

1. An unnormalized conditional probability (model ranking)
2. Ground truth

When the two rankings disagree, the parameters are updated.

Truth Function

$$\mathcal{F}(\mathbf{y}) = -\mathcal{L}(\mathbf{y}, \mathbf{y}_L)$$

y is the possible assignment

y_L is the true assignment

y^a and y^b are consecutive samples

α is the learning rate

$\Delta = \quad \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^a) - \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i^b)$

Θ is updated as follows

$$\Theta \xleftarrow{+} \begin{cases} \alpha\Delta & \text{if } \frac{p(\mathbf{y}^a|\mathbf{x})}{p(\mathbf{y}^b|\mathbf{x})} < 1 \wedge \mathcal{F}(\mathbf{y}^a) > \mathcal{F}(\mathbf{y}^b) \\ -\alpha\Delta & \text{if } \frac{p(\mathbf{y}^a|\mathbf{x})}{p(\mathbf{y}^b|\mathbf{x})} > 1 \wedge \mathcal{F}(\mathbf{y}^a) < \mathcal{F}(\mathbf{y}^b) \\ 0 & \text{otherwise.} \end{cases}$$

Semi-Supervised Rank-Based Learning

The truth function needs to be defined over both labeled and unlabeled data.

$$\mathcal{F} = \mathcal{F}_L \cup \mathcal{F}_U$$

$$y = \mathcal{Y}_L \cup \mathcal{Y}_U$$

Self-Training

$\mathcal{F}_L$ is used as the training set

MAP inference is performed on the unlabeled data, and those predictions $\hat{\mathbf{y}}_U$ are used as ground truth for $\mathcal{F}_U$

So, we have our objective function for the self-training defined as:

$$\mathcal{F}_s(\mathbf{y}) = -\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}_U)$$

Encoding Constraints

Constraints are encoded directly into $\mathcal{F}_U$

Constraint i defined as $\langle p_i, c_i \rangle$

$c_i(\mathbf{y})$ denotes whether assn y satisfies $(+1)$ or violates (-1) i, if i applies

p_i is the constraint strength

$$\mathcal{F}_c(\mathbf{y}) = \sum_i p_i c_i(\mathbf{y})$$

Using $\mathcal{F}_c$

- Every prediction on unlabeled data is ranked only according to the constraints. So the model satisfies those, but is not guaranteed to result in the correct solution.
- To deal with this problem, the ranking function needs to balance the constraints with the current model.
- They provide two ways of doing this.

Self-training with Constraints

$$\begin{aligned}\mathcal{F}_{sc}(\mathbf{y}) &= \mathcal{F}_s(\mathbf{y}) + \lambda_s \mathcal{F}_c(\mathbf{y}) \\ &= -\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}_U) + \lambda_s \sum_i p_i c_i(\mathbf{y})\end{aligned}$$

- This method has two limitations:
- Self-training requires a complete inference step
- The model may have low confidence, but this is not taken into account by self-training.

Model + Constraints

$$
\begin{aligned}
\mathcal{F}_{mc}(\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{x},\Theta) + \log Z(x) + \lambda_m \mathcal{F}_c(\mathbf{y}) \\
&= \sum_{\Psi_i} \Theta \cdot \mathbf{f}(\mathbf{x}_i, \mathbf{y}_i) + \lambda_m \sum_i p_i c_i(\mathbf{y})
\end{aligned}
$$

- This objective function does not require inference and also accounts for model confidence.

  λ controls the relative importance of the constraints. High values mean SampleRank will not violate any constraints.

Experiment

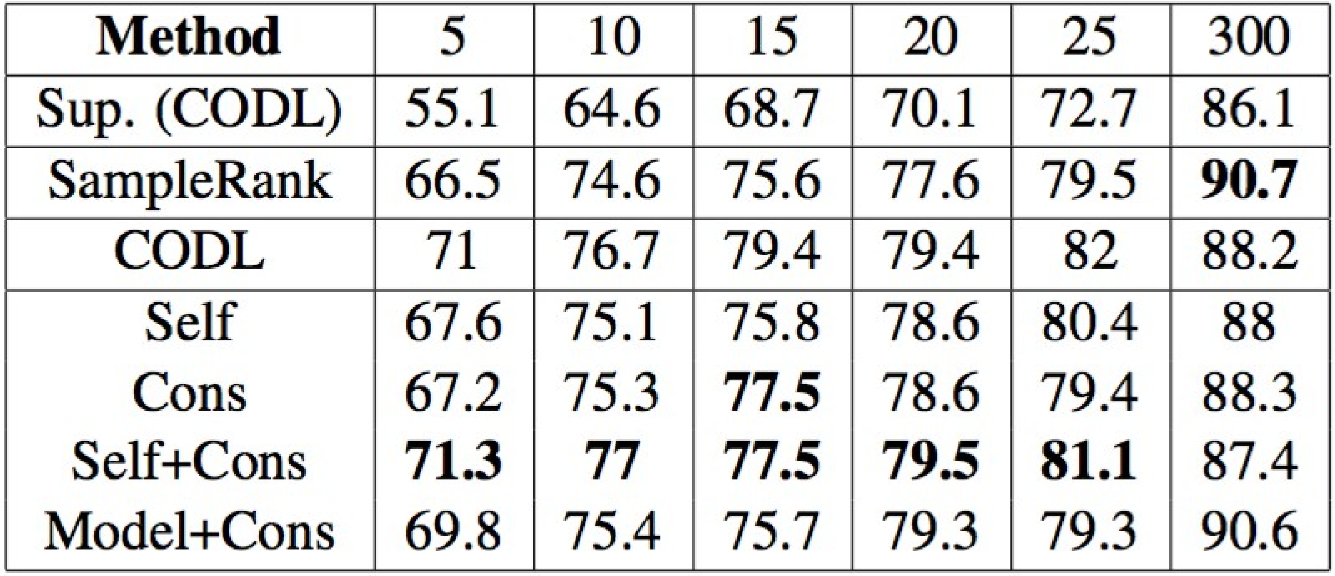
- Segment citations into different fields, e.g. “author”, “title”, with citations from the Cora citation dataset
- Compare the four objective functions plus supervised results from SampleRank and from the Constraint-Driven Learning Algorithm (CODL), plus CODL semi-supervised results.
- Settings used:

$$p_i = 1.0, \alpha = 1.0, \lambda_s = 10, \text{ and } \lambda_m = 0.0001$$

Experiment

- 300 instances training data
  - Varying the amount of labeled data
- 100 instances development
- 100 test data

Results



Method	5	10	15	20	25	300
Sup. (CODL)	55.1	64.6	68.7	70.1	72.7	86.1
SampleRank	66.5	74.6	75.6	77.6	79.5	**90.7**
CODL	71	76.7	79.4	79.4	82	88.2
Self	67.6	75.1	75.8	78.6	80.4	88
Cons	67.2	75.3	**77.5**	78.6	79.4	88.3
Self+Cons	**71.3**	**77**	**77.5**	**79.5**	**81.1**	87.4
Model+Cons	69.8	75.4	75.7	79.3	79.3	90.6

Table 1: **Tokenwise Accuracy:** for different methods as we vary the size of the labeled data

Time:
Self-training – 90 minutes
Self+Cons and Model+Cons – 100 minutes
Cons – 30 minutes
*CODL times not reported

Conclusion

- Integrating the two paradigms of semi-supervised learning retains the efficiency of parameter updates within inference while using unlabeled data.
- They think as datasets get larger and more complex, this will become a more useful technique.