Constraint-Driven Rank-Based Learning for Information Extraction

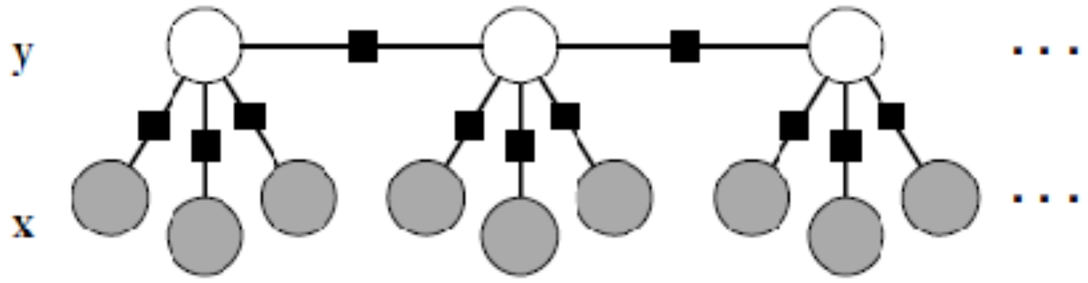
Sameer Singh, Limin Yao, Sebastian Riedel and Andrew McCallum

R222 Presentation by Yuai Liu

Conditional Random Fields (CRF)

- Undirected graphical models (Markov random field) represented as factor graphs.
- Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." *Foundations and Trends® in Machine Learning*

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



Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." *Foundations and Trends® in Machine Learning* 4.4 (2012): p289.
http://homepages.inf.ed.ac.uk/csutton/publications/crftut-fnt.pdf

SampleRank

- A rank-based learning framework.
- Updating parameters within Markov Chain Monte Carlo (MCMC) inference

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

Semi-Supervised Learning

- Generalized Expectation (GE) and Alternating Projections (AP):
  - Encoding preferences by specifying constraints on feature expectations
  - Require expensive full inference
- Constraint-Driven Learning (CODL):
  - Using constraints to guide semi-supervised learning
  - Requires “Top-K” inference step

Semi-Supervised Rank-Based Learning

- Self-training:
  - Maximum a posteriori (MAP) inference is performed on the unlabeled data to estimate ground truth.
  - Self-training objective function: $\mathcal{F}_s(\mathbf{y}) = -\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}_U)$
- Encoding Constraints:
  - A constraint i is specified as: $\langle p_i, c_i \rangle$
  - Objective function: $\mathcal{F}_c(\mathbf{y}) = \sum_i p_i c_i(\mathbf{y})$
  - Constraints may prefer a wrong solution while the model favors the correct solution

Incorporate model predictions

- Combine self-training with constraints:

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

- Two limitations:

1. Requires a complete inference iteration
2. The model might have low confidence in its prediction

Incorporate model predictions

- An objective function incorporates the model score directly:

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

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

Experiments

- Segmenting citations into different fields.
- Cora citation dataset.
- 300 training instances, 100 development instances, 100 test instances.
- Same token-label constraints as Chang et al. (2007)

(a)-Citations
1) Each field must be a consecutive list of words, and can appear at most once in a citation.
2) State transitions must occur on punctuation marks.
3) The citation can only start with author or editor.
4) The words *pp.*, *pages* correspond to *PAGE*.
5) Four digits starting with 20xx and 19xx are *DATE*.
6) Quotations can appear only in titles.
7) The words *note*, *submitted*, *appear* are *NOTE*.
8) The words *CA*, *Australia*, *NY* are *LOCATION*.
9) The words *tech*, *technical* are *TECH_REPORT*.
10) The words *proc*, *journal*, *proceedings*, *ACM* are *JOURNAL* or *BOOKTITLE*.
11) The words *ed*, *editors* correspond to *EDITOR*.

Chang, Ming-Wei, Lev Ratinov, and Dan Roth. "Guiding semi-supervision with constraint-driven learning." *ACL*. 2007. https://pdfs.semanticscholar.org/dc9f/999632bf6d82882cc54e2d2cc2d32eaed932.pdf

Experiments

- $p_i = 1.0,\ \alpha = 1.0,\ \lambda_s = 10,\ \lambda_m = 0.0001.$

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

Conclusion

- Extends rank-based learning to semi-supervised learning
- The integrated models perform competitively compared with SampleRank and CODL.
- The objective function of Model+Cons method does not require inference.
- However, the Model+Cons method did not show higher efficiency than the others in the experiments.
- Did not mention how the parameters (λ_s and λ_m) used in the experiments are tuned.