Probabilistic Models for Sequence Labeling

Besnik Fetahu

June 9, 2011

- Background & Motivation
  - Problem introduction
  - Generative vs. Discriminative models
  - Existing approaches for sequence labeling
  - Label bias problems
  - Factor graphs
- Conditional Random Fields
  - Parameter estimation
  - Inference
- Semi-Markov CRFs
- Experimental Results

Segmentation and Labeling Problem

- Probabilistic nature of the problem.
- Dependence on previous, and future labels.
- Generalization.
- Data Sparsity.
- Ambiguity.
- Combinatorial explosions.

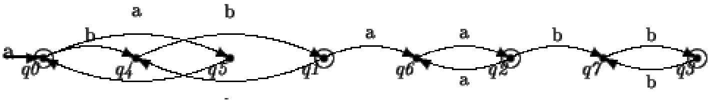


Figure: Dependence Labeling Problem.

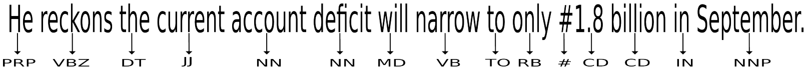
Problems that are considered more often on this field

1. Named Entity Recognition.

   Jim bought 300 shares of ACME Corp. in 2006.

   - Persons: Jim
   - Quantities: 300
   - Companies: ACME Corp.
   - Dates: 2006.

2. Part Of Speech Tagging.



Generative vs. Discriminative Probabilistic Approaches

Probabilistic Generative Models

- Model joint distribution
- Build models for each label
- Minimum variance
- Biased parameter estimation
- Aim: Find $p(y|x)$
- Maximize Likelihood:
  $\hat{\theta}_{GEN} = arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_{y_i} f_{y_i}(x_i; \theta)$

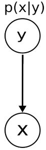


Figure: Generative Probabilistic Model

Generative vs. Discriminative Probabilistic Approaches

Probabilistic Discriminative Models

- Model conditional distribution
- Best classification performance
- Minimize classification loss
- Parameters that influence only the conditional distribution
- Maximize the logistic regression:

$$\hat{\theta}_{DISC} = arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log \frac{p_{y_i} f_{y_i}(x_i;\theta)}{\sum_k p_k f_k(x_i;\theta)}$$

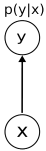


Figure: Discriminative Probabilistic Model

Hidden Markov Models - HMMs

- Generative model
- Consider many combinations
- Observation, depends directly at a state, in some time.
- Evaluate:

$$p(y,x) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t)$$

$$p(y,x) = \frac{1}{Z}\exp\{\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_t)\}$$

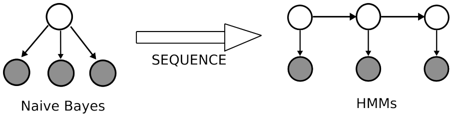


Figure: Hidden Markov Model

Maximum Entropy Markov Models - MEMMs

- Discriminative model
- Exponential model for each state-observation transition

$$p(y'|y,x) = \frac{1}{Z(y,x)} \exp\{\sum_{k=1}^{K} \lambda_k f_k(y,y',x)\}$$

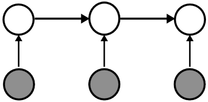


Figure: Maximum Entropy Markov Model

Problems with previous approaches

Label Ambiguity Reasons:

1. Local model construction
2. Competing states against each other
3. Non-Discriminatory state transitions

Proposed Approaches:

1. Delay branching of state transitions
2. Start with a fully connected graph

Disadvantages of these approaches

1. Discretization can lead to combinatorial explosions.
2. Exclude prior knowledge.

Label Bias Problem

Problems:

1. State transitioning
2. Both paths equally probable

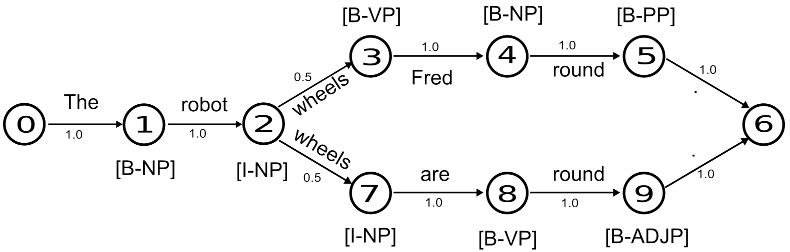


Figure: Label Bias Problem

From Directed Graphs to Undirected Graphs

Generative models represented as directed graphs

1. Outputs precede inputs.
2. Describe how outputs generate inputs.

Discriminative models as factor graphs

1. Define factors, as the dependence of features with observations.
2. Arbitrary number of features i.e. *Capital Letters, Noun, ...*
   $\Psi_k(y, x_k) = p(x_k|y)$

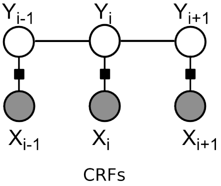


Figure: Factor Graph

Modeling of CRFs

Definition

Let $G = (V, E)$ be a graph such that $Y = (Y_v)_{v \in V}$ so that Y is indexed by the vertices of G. Then (X, Y) is a conditional random field in case, when conditioned on X, the random variable Y_v obeys the Markov property with respect to the graph $p(Y_v \vee X, Y_w, w \neq v) = p(Y_v \vee X, Y_w, w \sim v)$, where $w \sim v$ means that w and v are neighbors in G.

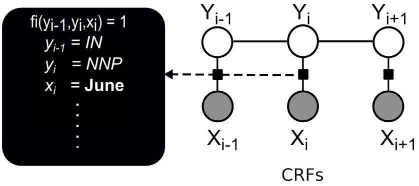


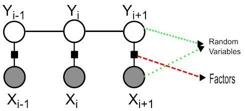
Figure: Factor Graph

Modeling of CRFs

Properties of CRFs:

- Condition globally on observation X.
- Similar to bipartite graphs: Two sets of random variables as vertices, with factorized edges.
- Normalize probabilities, of labels y given observation x, by the product of potential functions.
- Fixed set of features.
- Usually more expensive than HMMs (Arbitrary dependencies on observation sequence).

$$p_\theta(\mathbf{y}|\mathbf{x}) \propto exp(\sum_{e \in E, k} \lambda_k f_k(e, \mathbf{y}|_e, x) + \sum_{v \in V, k} \mu_k g_k(v, \mathbf{y}|_v, x)).$$



Training - Improved Iterative Scaling - IIS

Algorithm

IIS Algorithm:

- *Start with an arbitrary value for each of* λ_k, μ_k
- *Repeat until convergence:*
  - *Solve:* $\tilde{E}[f_k], \tilde{E}[g_k]$
  - *Set:*
    $\lambda_k \leftarrow \lambda_k + \delta\lambda_k$
    $\mu_k \leftarrow \mu_k + \delta\mu_k$

Properties of IIS:

- Global optimum.
- Slow convergence.

Objective for maximization (for edge features, similar for vertex features):

$$\tilde{E}[f_k] = \sum_{x,y} \tilde{p}(x)p(y|x) \sum_{i=1}^{n+1} f_k(e_i, y|_{e_i}, x)e^{\delta\lambda_k T(x,y)}$$

Training - Stochastic Gradient Ascent - SDG

Consider Stochastic Gradient Ascent (difference from descent that is for minimization)

- Increase the log likelihood.
- One example at a time.
- Most features of an example are 0 (skip), complexity $O(nfp)$.
- Change parameters once.
- Works good on sparse environments.

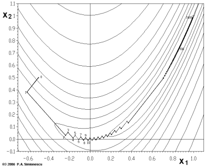


Figure: Stochastic Gradient Ascent

Training - State of the Art: Limited-memory BFGS

Properties of L-BFGS:

- Second Order Derivatives.
- Build Approximations to the Hessian Matrix.
- Quick Convergence.
- Great performance for unconstrained problems.

Approximations made in the gradient steps:

- $\delta^k = B^k G(\theta^k)\ B^{(k)} y^{(k)=\delta^{(k-1)}}$
  Where matrix B, represents the approximated inverse Hessian Matrix.

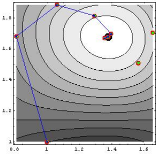


Figure: Quasi-Newton Line Search.

Inference with CRFs

Here we consider Linear-Chain Structured CRFs:

- Viterbi Decoding by Dynamic Programming.
- Shortest Path.
- Each position in the observation, has the matrix:$[M_i(y', y|x)]_{|\mathcal{Y}x\mathcal{Y}|} = e(\Lambda_i(y', y|x))$ $\mathcal{Y} = \{NN, NP, V\}$

$$
\begin{array}{c c}
 & \begin{array}{ccc} \text{NN} & \text{NP} & \text{V} \end{array} \\
\begin{array}{c} \text{NN} \\ \text{NP} \\ \text{V} \end{array} &
\begin{pmatrix}
e^{(\Lambda_i(NN,NN|x))} & e^{(\Lambda_i(NN,NP|x))} & e^{(\Lambda_i(NN,V|x))} \\
e^{(\Lambda_i(NP,NN|x))} & e^{(\Lambda_i(NP,NP|x))} & e^{(\Lambda_i(NP,V|x))} \\
e^{(\Lambda_i(V,NN|x))} & e^{(\Lambda_i(V,NP|x))} & e^{(\Lambda_i(V,V|x))}
\end{pmatrix}
\end{array}
$$

- Where:
$\Lambda_i(y', y|x) = \sum_k \lambda_k f_k(e_i, Y|_{e_i} = (y', y), x) + \sum_k \mu_k g_k(v_i, Y|_{v_i} = y, x))$

Inference with CRFs

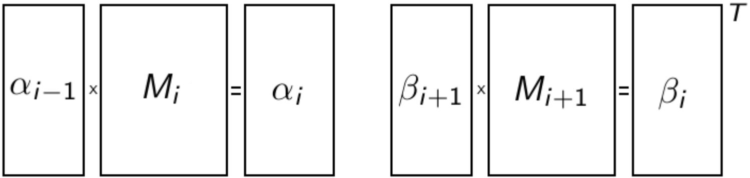


Figure: Forward-Backward Inference Calculation.

$$\alpha_0(y|x) = \begin{cases} 1 & : \text{if y} = \text{start} \\ 0 & : \text{otherwise} \end{cases}$$

$$\alpha_i(x) = \alpha_{i-1}(x) M_i(x).$$

$$\beta_{n+1}(y|x) = \begin{cases} 1 & : \text{if y} = \text{stop} \\ 0 & : \text{otherwise} \end{cases}$$

$$\beta_i(x)^T = \beta_{i+1}(x) M_{i+1}(x).$$

Semi-Markov Conditional Random Fields

Semi-Markov Models:

- Semi-Markov Chains.
- Persist states for time d.
- Segment observations.
- Features built on the segmented observation.
- Faster Inference than order-L CRFs.

Observation Segmentation

- **I went skiing with Fernando Pereira in British Colombia**
- $s = \langle (1,1,O),(2,2,O),(3,3,O),(4,4,O),(5,6,I),(7,7,O),(8,9,I) \rangle$
- $y = \langle O,O,O,O,I,I,O,I,I \rangle$

Semi-Markov Conditional Random Fields

Modeling of Semi-Markov CRFs:

- *Segment:* $s_j = \langle t_j, u_j, y_j \rangle$
- Segment Feature Functions: $g^k(j, x, s) = g'^k(y_j, y_{j-1}, x, t_j, u_j\rangle$
- Inference: $P(s|x, W) = \frac{1}{Z(x)} e^{W \cdot G(x,s)}$

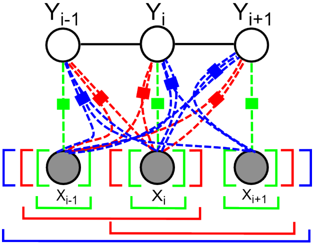


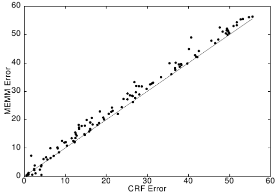
Figure: Semi-Markov Chains

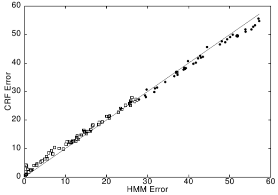
Experimental Results

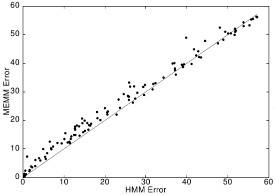
Experimental setup:

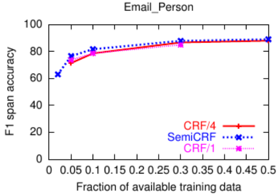
- Label bias verification.
- Synthetic data, generated by randomly chosen HMMs.
  - Transition probabilities are:
    $p_\alpha(y_i|y_{i-1}, y_{i-2}) = \alpha p_2(y_i|y_{i-1}, y_{i-2}) + (1-\alpha)p_1(y_i|y_{i-1})$
  - Emission probabilities:
    $p_\alpha(x_i|y_i, x_{i-1}) = \alpha p_2(x_i|y_i, x_{i-1}) + (1-\alpha)p_1(x_i|y_i)$
- Mixture of first-order and second-order models.
- Five labels, a-e ($|\mathcal{Y}| = 5$), and 26 observation values, A-Z ($|\mathcal{X}| = 26$).
- POS tagging experiments on Penn treebank.

Experimental Results









Experimental Results

	baseline	+internal dict		+external dict		+both dictionaries		
	F1	F1	Δbase	F1	Δbase	F1	Δbase	Δextern
CRF/1								
state	20.8	**44.5**	113.9	**69.2**	232.7	55.2	165.4	-67.3
title	28.5	3.8	-86.7	38.6	35.4	19.9	-30.2	-65.6
person	67.6	48.0	-29.0	81.4	20.4	64.7	-4.3	-24.7
city	70.3	60.0	-14.7	80.4	14.4	69.8	-0.7	-15.1
company	51.4	16.5	-67.9	55.3	7.6	15.6	-69.6	-77.2
CRF/4								
state	15.0	25.4	69.3	46.8	212.0	43.1	187.3	-24.7
title	23.7	7.9	-66.7	36.4	53.6	14.6	-38.4	-92.0
person	70.9	64.5	-9.0	82.5	16.4	74.8	5.5	-10.9
city	73.2	70.6	-3.6	80.8	10.4	76.3	4.2	-6.1
company	54.8	20.6	-62.4	**61.2**	11.7	25.1	-54.2	-65.9
semi-CRF								
state	**25.6**	35.5	38.7	62.7	144.9	**65.2**	154.7	9.8
title	**33.8**	**37.5**	10.9	**41.1**	21.5	**40.2**	18.9	-2.5
person	**72.2**	**74.8**	3.6	**82.8**	14.7	**83.7**	15.9	1.2
city	**75.9**	**75.3**	-0.8	**84.0**	10.7	**83.6**	10.1	-0.5
company	**60.2**	**59.7**	-0.8	60.9	1.2	**60.9**	1.2	0.0

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM$^+$	4.81%	26.99%
CRF$^+$	4.27%	23.76%

$^+$Using spelling features

Conclusion

- Generative vs. Discriminative models.
- Arbitrary number of features.
- Global Modeling vs. Local modeling.
- Convex optimization problem.
- Different solutions to parameter estimation.
- Factor Graphs vs. Directed graphs.
- Semi-Markov CRFs.

Bibliography

1. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data (Lafferty et al.)
2. Semi-Markov Conditional Random Fields for Information Extraction (Sarawagi S. and Cohen W).
3. The Trade-Off Between Generative and Discriminative Classifiers (Bouchard G. and Triggs B.)
4. Log Linear Models and Conditional Random Fields - Tutorial (Elkan Ch.)
5. Conditional Random Fields: An Introduction (Wallach H.)

Thank you!

Questions?