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.

2. Part Of Speech Tagging.

   He reckons the current account deficit will narrow to only #1.8 billion in September.
   - PRP VBZ DT JJ NN NN MD VB TO RB # CD CD IN NNP
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:

$$\hat{\theta}_{\text{GEN}} = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log p_{y_i} f_{y_i}(x_i; \theta)$$

**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:

\[
\hat{\theta}_{\text{DISC}} = \arg \max_{\theta \in \Theta} \sum_{i=1}^{n} \log \frac{p_{y_i}(x_i; \theta)}{\sum_{k} p_{k} f_{k}(x_i; \theta)}
\]

\[p(y|x)\]

\[x\]

**Figure:** Discriminative Probabilistic Model
Hidden Markov Models - HMMs

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

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

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

**Figure:** Hidden Markov Model
Maximum Entropy Markov Models - MEMMs

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

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

**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
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 $p(Y_v \lor X, Y_w, w \neq v) = p(Y_v \lor X, Y_w, w \sim v)$, 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).

$$p_\theta(y|x) \propto \exp\left( \sum_{e \in E, k} \lambda_k f_k(e, y|e, x) + \sum_{v \in V, k} \mu_k g_k(v, y|v, x) \right).$$
Training - Improved Iterative Scaling - IIS

Algorithm

**IIS Algorithm:**

- **Start with an arbitrary value for each of** $\lambda_k, \mu_k$
- **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):

$$\tilde{E}[f_k] = \sum_{x,y} \tilde{p}(x) p(y|x) \sum_{i=1}^{n+1} f_k(e_i, y|e_i, x) e^{\delta \lambda_k T(x,y)}$$
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
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.
Here we consider Linear-Chain Structured CRFs:

- **Viterbi Decoding by Dynamic Programming.**
- **Shortest Path.**

Each position in the observation, has the matrix:

\[
[M_i(y', y|\mathbf{x})]_{Y \times Y} = e(\Lambda_i(y', y|x)) \quad Y = \{NN, NP, V\}
\]

\[
\begin{pmatrix}
NN & NP & V \\
NN & e^{(\Lambda_i(NN, NN|x))} & e^{(\Lambda_i(NN, NP|x))} & e^{(\Lambda_i(NN, V|x))} \\
NP & e^{(\Lambda_i(NP, NN|x))} & e^{(\Lambda_i(NP, NP|x))} & e^{(\Lambda_i(NP, V|x))} \\
V & e^{(\Lambda_i(V, NN|x))} & e^{(\Lambda_i(V, NP|x))} & e^{(\Lambda_i(V, V|x))}
\end{pmatrix}
\]

- **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

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

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

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

\[ \beta_i(x)^T = \beta_{i+1}(x) M_{i+1}(x). \]
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**: \( g^k(j, x, s) = g'^k(y_j, y_{j-1}, x, t_j, u_j) \)
- **Inference**: \( P(s| x, W) = \frac{1}{Z(x)} e^{W \cdot G(x, s)} \)

![Figure: Semi-Markov Chains](image)
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 (|Y| = 5), and 26 observation values, A-Z (|X| = 26).
- POS tagging experiments on Penn treebank.
Experimental Results

- Graph 1: Comparison of MEMM Error vs. CRF Error.
- Graph 2: Comparison of CRF Error vs. HMM Error.
- Graph 3: Comparison of MEMM Error vs. HMM Error.
- Graph 4: F1 span accuracy across different fractions of available training data.
## Experimental Results

<table>
<thead>
<tr>
<th>CRF/1</th>
<th>baseline F1</th>
<th>+internal dict F1</th>
<th>+external dict F1</th>
<th>+both dictionaries F1</th>
<th>Δbase</th>
<th>Δextern</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>20.8</td>
<td>44.5</td>
<td>69.2</td>
<td>55.2</td>
<td>113.9</td>
<td>232.7</td>
</tr>
<tr>
<td>title</td>
<td>28.5</td>
<td>3.8</td>
<td>38.6</td>
<td>19.9</td>
<td>-86.7</td>
<td>-65.6</td>
</tr>
<tr>
<td>person</td>
<td>67.6</td>
<td>48.0</td>
<td>81.4</td>
<td>64.7</td>
<td>-29.0</td>
<td>-24.7</td>
</tr>
<tr>
<td>city</td>
<td>70.3</td>
<td>60.0</td>
<td>80.4</td>
<td>69.8</td>
<td>-14.7</td>
<td>-15.1</td>
</tr>
<tr>
<td>company</td>
<td>51.4</td>
<td>16.5</td>
<td>55.3</td>
<td>15.6</td>
<td>-67.9</td>
<td>-77.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRF/4</th>
<th>baseline F1</th>
<th>+internal dict F1</th>
<th>+external dict F1</th>
<th>+both dictionaries F1</th>
<th>Δbase</th>
<th>Δextern</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>15.0</td>
<td>25.4</td>
<td>46.8</td>
<td>43.1</td>
<td>69.3</td>
<td>212.0</td>
</tr>
<tr>
<td>title</td>
<td>23.7</td>
<td>7.9</td>
<td>36.4</td>
<td>14.6</td>
<td>-66.7</td>
<td>-92.0</td>
</tr>
<tr>
<td>person</td>
<td>70.9</td>
<td>64.5</td>
<td>82.5</td>
<td>74.8</td>
<td>-9.0</td>
<td>-10.9</td>
</tr>
<tr>
<td>city</td>
<td>73.2</td>
<td>70.6</td>
<td>80.8</td>
<td>76.3</td>
<td>-3.6</td>
<td>-6.1</td>
</tr>
<tr>
<td>company</td>
<td>54.8</td>
<td>20.6</td>
<td>61.2</td>
<td>25.1</td>
<td>-62.4</td>
<td>-65.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>semi-CRF</th>
<th>baseline F1</th>
<th>+internal dict F1</th>
<th>+external dict F1</th>
<th>+both dictionaries F1</th>
<th>Δbase</th>
<th>Δextern</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>25.6</td>
<td>35.5</td>
<td>62.7</td>
<td>65.2</td>
<td>38.7</td>
<td>144.9</td>
</tr>
<tr>
<td>title</td>
<td>33.8</td>
<td>37.5</td>
<td>41.1</td>
<td>40.2</td>
<td>10.9</td>
<td>21.5</td>
</tr>
<tr>
<td>person</td>
<td>72.2</td>
<td>74.8</td>
<td>82.8</td>
<td>83.7</td>
<td>3.6</td>
<td>14.7</td>
</tr>
<tr>
<td>city</td>
<td>75.9</td>
<td>75.3</td>
<td>84.0</td>
<td>83.6</td>
<td>-0.8</td>
<td>10.7</td>
</tr>
<tr>
<td>company</td>
<td>60.2</td>
<td>59.7</td>
<td>60.9</td>
<td>60.9</td>
<td>-0.8</td>
<td>1.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>model</th>
<th>error</th>
<th>oov error</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>5.69%</td>
<td>45.99%</td>
</tr>
<tr>
<td>MEMM</td>
<td>6.37%</td>
<td>54.61%</td>
</tr>
<tr>
<td>CRF</td>
<td>5.55%</td>
<td>48.05%</td>
</tr>
<tr>
<td>MEMM+</td>
<td>4.81%</td>
<td>26.99%</td>
</tr>
<tr>
<td>CRF+</td>
<td>4.27%</td>
<td>23.76%</td>
</tr>
</tbody>
</table>

+ 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

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?