CS 6784 Paper Presentation

Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

John Lafferty, Andrew McCallum, Fernando C. N. Pereira

Presenters: Brad Gulko and Stephanie Hyland

February 20, 2014
Main Contribution Summary

- This 2001 paper introduced the Conditional Random Field (CRF).
- Describes efficient representation of field potentials in terms of features.
- Provides two algorithms for finding Maximum Likelihood parameter values.
- Provides some really unconvincing examples...
Main Contribution Summary

... examples are NOT the strongest point of this paper
Talk Structure

- Brad
  - CRF in context
Talk Structure

- Brad
  - CRF in context
  - The Label Bias Problem
Talk Structure

- Brad
  - CRF in context
  - The Label Bias Problem
- Stephanie
  - Parameter Estimation

\[
\rho(y|x) = \frac{1}{Z(x)} \exp \left( \sum_{i \in V_i} \lambda_i g_i (y_i, x) + \sum_{e \in E} \mu_e g_e (y_e, x) \right) \\
\theta = (\lambda_1, \lambda_2, \cdots \mu_1, \mu_2, \cdots) \\
\theta_{t+1} = \theta_t - [Hf(\theta_t)]^{-1} \nabla f(\theta_t)
\]
Talk Structure

- Brad
  - CRF in context
  - The Label Bias Problem
- Stephanie
  - Parameter Estimation
  - Experiments
  - Conclusion

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
CRF in Context

- \( X = \{X_1, X_2, \cdots\} \) be a set of observed RV
- \( Y = \{Y_1, Y_2, \cdots\} \) be a set of label RV
- \( X, Y \) be a set of joint observations of \( X, Y \)

<table>
<thead>
<tr>
<th>Directed</th>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>???</td>
<td>???</td>
</tr>
<tr>
<td>Undirected</td>
<td>???</td>
<td>???</td>
</tr>
</tbody>
</table>
CRF in Context

- $\mathbf{X} = \{X_1, X_2, \cdots \}$ be a set of observed RV
- $\mathbf{Y} = \{Y_1, Y_2, \cdots \}$ be a set of label RV
- $X, Y$ be a set of joint observations of $\mathbf{X}, \mathbf{Y}$

<table>
<thead>
<tr>
<th></th>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directed</td>
<td>HMM</td>
<td>??</td>
</tr>
<tr>
<td>Undirected</td>
<td>MRF</td>
<td>??</td>
</tr>
</tbody>
</table>

HMM, ME-MM, MRF, CRF

In 2001, HMM, ME-MM and MRF were well known, the paper presents the CRF.

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
CRF in Context

- \( X = \{ X_1, X_2, \cdots \} \) be a set of observed RV
- \( Y = \{ Y_1, Y_2, \cdots \} \) be a set of label RV
- \( X, Y \) be a set of joint observations of \( X, Y \)

<table>
<thead>
<tr>
<th>Directed</th>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>ME-MM</td>
<td></td>
</tr>
<tr>
<td>MRF</td>
<td>???</td>
<td></td>
</tr>
</tbody>
</table>

In 2001, HMM, ME-MM and MRF were well known,
CRF in Context

- \( X = \{ X_1, X_2, \cdots \} \) be a set of observed RV
- \( Y = \{ Y_1, Y_2, \cdots \} \) be a set of label RV
- \( X, Y \) be a set of joint observations of \( X, Y \)

<table>
<thead>
<tr>
<th></th>
<th>Generative</th>
<th>Discriminative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Directed</td>
<td>HMM</td>
<td>ME-MM</td>
</tr>
<tr>
<td>Undirected</td>
<td>MRF</td>
<td>CRF</td>
</tr>
</tbody>
</table>

In 2001, HMM, ME-MM and MRF were well known, the paper presents the CRF.

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Generative vs. Discriminative

- Generative: maximise joint \( P(Y, X) = P(Y | X) P(X) \)
- Discriminative: maximise conditional \( P(Y | X) \)
- When is Discriminative helpful?
  - Tractability requires independence
Generative vs. Discriminative

- **Generative:** maximise joint  \( P(Y, X) = P(Y|X)P(X) \)
- **Discriminative:** maximise conditional  \( P(Y|X) \)
- **When is Discriminative helpful?**
  - Tractability requires independence
  - ...but sometimes there are important correlations in \( X \).
Examples: important correlations

- Long range interactions in human genomics
Examples: important correlations

- Long range interactions in human genomics
- Pronoun definition and binding
Examples: important correlations

- Long range interactions in human genomics
- Pronoun definition and binding
- Context in whole scene image recognition
Examples: important correlations

- Long range interactions in human genomics
- Pronoun definition and binding
- Context in whole scene image recognition
- Recursive structure in language

Generative vs. Discriminative
Directed vs. Undirected

- For a graphical model $G(E, V)$ with joint potential $\Psi(V)$.
- Let $C$ be the set of cliques (fully connected subgroups) in $G$, with $c \in C$ having edges $E_c$ and vertices $V_c$. 

Presenters: Brad Gulko and Stephanie Hyland

CS 6784 Paper Presentation
Directed vs. Undirected

- For a graphical model $G(E, V)$ with joint potential $\Psi(V)$.

- Let $C$ be the set of *cliques* (fully connected subgroups) in $G$, with $c \in C$ having edges $E_c$ and vertices $V_c$.

- Finally, $\text{Dom}(V)$ is the set of all values assumable by the random variables, $V = X \cup Y$.

$$P(V) = \frac{1}{Z} \Psi(V), \quad Z = \sum_{v \in \text{Dom}(V)} \Psi(v)$$

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Directed vs. Undirected, continued

- Compactness requires factorization (Hammersley-Clifford, 1971):

\[
\psi(V) = \prod_{c \in C} \psi_c(V_c)
\]
Directed vs. Undirected, continued

- Compactness requires factorization (Hammersley-Clifford, 1971):
  \[ \psi(V) = \prod_{c \in \mathcal{C}} \psi_c(V_c) \]

- Directed: local Normalization -
  \[ \forall c \in \mathcal{C}, \sum_{v \in \text{Dom}(V_c)} \psi_c(v) = 1 \]
Directed vs. Undirected, continued

- Compactness requires factorization (Hammersley-Clifford, 1971):
  \[ \psi(\mathbf{V}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{V}_c) \]

- Directed: local Normalization - each \( \psi_c \) is a probability.
  \[ \forall c \in \mathcal{C}, \quad \sum_{\mathbf{v} \in \text{Dom}(\mathbf{V}_c)} \psi_c(\mathbf{v}) = 1 \]
Directed vs. Undirected, continued

- Compactness requires factorization (Hammersley-Clifford, 1971):
  \[ \psi(\mathbf{V}) = \prod_{c \in \mathcal{C}} \psi_c(\mathbf{V}_c) \]

- Directed: local Normalization - each \( \psi_c \) is a probability.
  \[ \forall c \in \mathcal{C}, \sum_{\mathbf{v} \in \text{Dom}(\mathbf{V}_c)} \psi_c(\mathbf{v}) = 1 \]

- Undirected: Global Normalization - relaxes this constraint...
  but what does it buy us?
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

\[ Y_1 \in \{1,2\} \quad Y_2 \in \{3,4\} \]

Training Data:
8: \( X = \{RI\} \quad Y = \{13\} \)
2: \( X = \{RO\} \quad Y = \{24\} \)
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

Training Data:
8: $X = \{RI\} \quad Y = \{13\}$
2: $X = \{RO\} \quad Y = \{24\}$

<table>
<thead>
<tr>
<th>$Y_1$</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = R$</td>
<td>0.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\psi(Y_2, X_2, Y_1)$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$X_2 = I$</td>
<td>8</td>
</tr>
<tr>
<td>$X_2 = O$</td>
<td>ε</td>
</tr>
<tr>
<td>$Z = I$</td>
<td>ε</td>
</tr>
<tr>
<td>$Z = O$</td>
<td>ε</td>
</tr>
</tbody>
</table>
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

\[ Y_1 \in \{1,2\} \quad Y_2 \in \{3,4\} \]

Training Data:
8: \( X = \{RI\} \quad Y = \{13\} \)
2: \( X = \{RO\} \quad Y = \{24\} \)

\[
\begin{array}{c|c|c}
\hline
P(Y_1|X_1) & Y_1 & \hline
\hline
 & I & 2 & \\
X_1 \to \& 0.8 & 0.2 & \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c|c}
\hline
\Psi(Y_2,X_2,Y_1) & 3 & 4 & \\
\hline
 Y_1=1 & \hline
X_2=I & 8 & \epsilon & \\
X_2=O & \epsilon & \epsilon & \\
\hline
 Y_1=2 & \hline
X_2=I & \epsilon & \epsilon & \\
X_2=O & \epsilon & 2 & \\
\hline
\end{array}
\]

\[
\begin{array}{c|c|c|c}
\hline
\hline
P(Y_2|X_2,Y_1) & 3 & 4 & \\
\hline
 Y_1=1 & \hline
X_2=I & 1-\epsilon & \epsilon & \\
X_2=O & 0.5 & 0.5 & \\
\hline
 Y_1=2 & \hline
X_2=I & 0.5 & 0.5 & \\
X_2=O & \epsilon & 1-\epsilon & \\
\hline
\end{array}
\]
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

\[ Y_1 \in \{1,2\}, Y_2 \in \{3,4\} \]

Training Data:
\[ 8: X = \{RI\} \quad Y = \{13\} \]
\[ 2: X = \{RO\} \quad Y = \{24\} \]

Viterbi is \( P(Y_1, Y_2 | X) \)
\[ = P(Y_2 | Y_1, X)P(Y_1 | X) \]
\[ = P(Y_1 | X_1)P(Y_2 | Y_1, X_2) \]

Lets try it for \( X = \{RO\} \)

Which labeling wins?
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

Viterbi is $P(Y_1, Y_2 | X)$

\[
P(Y_1, Y_2 | X) = P(Y_2 | Y_1, X)P(Y_1 | X)
\]

or

\[
P(Y_1, Y_2 | X) = P(Y_1 | X)P(Y_2 | Y_1, X_2)
\]

Let's try it for $X = \{RO\}$

But we want

$Y = \{2, 4\}$

What happened?
The Label Bias Problem: Conditional Markov Model (EM-MM)

Toy Problem – fragment of a ME-MM

Local Normalization requires a probability... So...
\[ \frac{\varepsilon}{2\varepsilon} \Rightarrow \frac{1}{2} \]
The Label Bias Problem: Potentials

Toy Problem – fragment of a CRF

\[ \Psi(X, Y_1, Y_2) = \Psi(X_1, Y_1)\Psi(Y_1, Y_2)\Psi(X_2, Y_2) \]

\[ Y_1 \in \{1,2\} \quad Y_2 \in \{3,4\} \]

Training Data:

8: \(X = \{RI\} \quad Y = \{13\}\)

2: \(X = \{RO\} \quad Y = \{24\}\)
The Label Bias Problem: Potentials

Toy Problem – fragment of a CRF

\[ \Psi(X, Y_1, Y_2) = \Psi(X_1, Y_1)\Psi(Y_1, Y_2)\Psi(X_2, Y_2) \]

<table>
<thead>
<tr>
<th>(X)</th>
<th>(Y_1)</th>
<th>(Y_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(R)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>-</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

\( Y_1 \in \{1,2\}, Y_2 \in \{3,4\} \)

Training Data:
1. \(X = \{RI\}, Y = \{13\}\)
2. \(X = \{RO\}, Y = \{24\}\)
The Label Bias Problem: Potentials

Toy Problem – fragment of a CRF

\[ \Psi(X, Y_1, Y_2) = \Psi(X_1, Y_1) \Psi(Y_1, Y_2) \Psi(X_2, Y_2) \]

| \( Y_1, Y_2 \) | \( \Psi(X_1, Y_1) \) | \( \Psi(Y_1, Y_2) \) | \( \Psi(X_2, Y_2) \) | \( \Psi(Y_1, Y_2, X=RO) \) | \( P(Y_1, Y_2|X) \) |
|----------------|----------------|----------------|----------------|----------------|----------------|
| 1,3            | 8              | 8              | 2              |                |                |
| 1,4            | 8              | \( \varepsilon \) | 2              |                |                |
| 2,3            | 2              | \( \varepsilon \) | 2              |                |                |
| 2,4            | 2              | 2              | 2              |                |                |

\( Y_1 \in \{1,2\} \) \( Y_2 \in \{3,4\} \)

Training Data:
- 8: \( X = \{RI\} \) \( Y = \{13\} \)
- 2: \( X = \{RO\} \) \( Y = \{24\} \)

Which labeling wins, now?
The Label Bias Problem: Potentials

Toy Problem – fragment of a CRF

\[ \Psi(X, Y_1, Y_2) = \Psi(X_1, Y_1)\Psi(Y_1, Y_2)\Psi(X_2, Y_2) \]

Training Data:
8: \( X = \{RI\} \) \( Y = \{13\} \)
2: \( X = \{RO\} \) \( Y = \{24\} \)

Because potentials do not have to normalize into probabilities until AFTER aggregation, they don’t suffer from inappropriate conditioning.
Fun fact: We have seen this in class before!

- Graphical model $G(E, V)$ with joint potential $\Psi(V)$, $C$ the set of cliques in $G$ with $c \in C$ having edges $E_c$ and vertices $V_c$

$$P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c)$$
Fun fact: We have seen this in class before!

- Graphical model $G(E, V)$ with joint potential $\Psi(V)$, $C$ the set of cliques in $G$ with $c \in C$ having edges $E_c$ and vertices $V_c$

$$P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c)$$

- **M$^3$ nets**: cliques are pairs, and all conditioned on observed $x$:

$$P(y|x) \propto \prod_{(i,j) \in E} \psi_{ij}(y_i, y_j, x)$$
Fun fact: We have seen this in class before!

- Graphical model $G(E, V)$ with joint potential $\Psi(V)$, $C$ the set of cliques in $G$ with $c \in C$ having edges $E_c$ and vertices $V_c$

$$P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c)$$

- $M^3$ nets: cliques are pairs, and all conditioned on observed $x$:

$$P(y|x) \propto \prod_{(i,j) \in E} \psi_{ij}(y_i, y_j, x)$$

- AMN: cliques are pairs of nodes and singletons:

$$P_\phi(y) = \frac{1}{Z} \prod_i^{N} \phi_i(y_i) \prod_{i,j \in E} \phi_{i,j}(y_i, y_j)$$
Where do Parameters Come From?

CRF’s are part of the same general class, \( P(\mathbf{V}) \propto \Psi(\mathbf{V}) = \prod_{c \in \mathcal{C}} \Psi_c(\mathbf{V}_c) \)
Where do Parameters Come From?

CRF’s are part of the same general class, $P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c)$

For trees, cliques are pairs of vertices sharing an edge $(y|e)$, and single vertices $(y|v)$:

$$
\Psi(V) = \prod_{e \in Edge} \Psi_e(y|e) \prod_{v \in V} \Psi_v(y|v)
$$
Where do Parameters Come From?

CRF’s are part of the same general class, \( P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c) \)

For trees, cliques are pairs of vertices sharing an edge \( (y|_e) \), and single vertices \( (y|_v) \):

\[
\Psi(V) = \prod_{e \in \text{Edge}} \Psi_e(y|_e) \prod_{v \in V} \Psi_v(y|_v)
\]

And because this is a conditional network with \( V = X \cup Y \)

\[
\Psi(Y|X) = \prod_{e \in \text{Edge}} \Psi_e(y|_e, x) \prod_{v \in V} \Psi_v(y|_v, x)
\]
Where do Parameters Come From?

CRF’s are part of the same general class, \( P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c) \)

For trees, cliques are pairs of vertices sharing an edge \((y|_e)\), and single vertices \((y|_v)\):

\[
\Psi(V) = \prod_{e \in Edge} \Psi_e(y|_e) \prod_{v \in V} \Psi_v(y|_v)
\]

And because this is a conditional network with \( V = X \cup Y \)

\[
\Psi(Y|X) = \prod_{e \in Edge} \Psi_e(y|_e,x) \prod_{v \in V} \Psi_v(y|_v,x)
\]

An exponential identity gives us

\[
\Psi(Y|X) = \exp \left( \sum_{e \in E} \log \left( \Psi_e(y|_e,x) \right) + \sum_{v \in V} \log \left( \Psi_v(y|_v,x) \right) \right)
\]
Where do Parameters Come From?

CRF’s are part of the same general class, $P(V) \propto \Psi(V) = \prod_{c \in C} \Psi_c(V_c)$

For trees, cliques are pairs of vertices sharing an edge $(y|_e)$, and single vertices $(y|_v)$:

$$
\Psi(V) = \prod_{e \in Edge} \Psi_e(y|_e) \prod_{v \in V} \Psi_v(y|_v)
$$

And because this is a conditional network with $V = X \cup Y$

$$
\Psi(Y|X) = \prod_{e \in Edge} \Psi_e(y|_e, x) \prod_{v \in V} \Psi_v(y|_v, x)
$$

An exponential identity gives us

$$
\Psi(Y|X) = \exp \left( \sum_{e \in E} \log \left( \Psi_e(y|_e, x) \right) + \sum_{v \in V} \log \left( \Psi_v(y|_v, x) \right) \right)
$$

Potentials can be ANY positive values... like linear combinations of arbitrary features

$$
\Psi(Y|X) = \exp \left( \sum_{e \in E, k \in K} \lambda_k f_k(e, y|_e, x) + \sum_{v \in V, k' \in K'} \mu_{k'} g_{k'}(v, y|_v, x) \right)
$$
Improved iterative scaling

- Want to maximize log-likelihood with respect to parameters

\[ \theta = (\lambda_1, \lambda_2, \cdots; \mu_1, \mu_2, \cdots) \]
Improved iterative scaling

- Want to maximize log-likelihood with respect to parameters

\[ \theta = (\lambda_1, \lambda_2, \cdots; \mu_1, \mu_2, \cdots) \]


\(^1\)Della Pietra *et al.* (1997)

Presenters: Brad Gulko and Stephanie Hyland

CS 6784 Paper Presentation
Improved iterative scaling

- Want to maximize log-likelihood with respect to parameters
  \[ \theta = (\lambda_1, \lambda_2, \cdots; \mu_1, \mu_2, \cdots) \]
- Method: Improved Iterative Scaling\(^1\): Extension of Generalised Iterative Scaling (Darroch and Ratcliff 1972).
- **Improved** because features need not sum to constant.

\(^1\)Della Pietra et al. (1997)

Presenters: Brad Gulko and Stephanie Hyland

CS 6784 Paper Presentation
Improved iterative scaling

- Want to maximize log-likelihood with respect to parameters
  \[ \theta = (\lambda_1, \lambda_2, \cdots ; \mu_1, \mu_2, \cdots) \]
- *Improved* because features need not sum to constant.
- Idea: new set of parameters
  \[ \theta' = \theta + \delta\theta = (\lambda_1 + \delta\lambda_1, \cdots ; \mu_1 + \delta\mu_1 \cdots) \] which will not decrease objective function. Iteratively apply!

---

\(^1\)Della Pietra et al. (1997)
Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Improved iterative scaling

- Want to maximize log-likelihood with respect to parameters

\[ \theta = (\lambda_1, \lambda_2, \cdots ; \mu_1, \mu_2, \cdots) \]

- Method: Improved Iterative Scaling\(^1\): Extension of Generalised Iterative Scaling (Darroch and Ratcliff 1972).

- *Improved* because features need not sum to constant.

- Idea: new set of parameters

\[ \theta' = \theta + \delta \theta = (\lambda_1 + \delta \lambda_1, \cdots ; \mu_1 + \delta \mu_1 \cdots) \] which will not decrease objective function. Iteratively apply!

- Problem: slow, and nobody uses this any more.

\(^1\)Della Pietra et al. (1997)

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Modern CRF training - L-BFGS

- Generally use L-BFGS² algorithm.

²Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Modern CRF training - L-BFGS

- Generally use L-BFGS\(^2\) algorithm.
- Approximates Newton’s method. Optimise multivariate function \(f(\theta)\) through updates

\[
\theta_{t+1} = \theta_t - [Hf(\theta_t)]^{-1} \nabla f(\theta_t)
\]

\(^2\)Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm
Modern CRF training - L-BFGS

- Generally use L-BFGS\(^2\) algorithm.
- Approximates Newton’s method. Optimise multivariate function \(f(\theta)\) through updates

\[
\theta_{t+1} = \theta_t - [Hf(\theta_t)]^{-1} \nabla f(\theta_t)
\]

- Quasi-Newtonian: approximates Hessian \(Hf(\theta)\).

\(^2\)Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm
Modern CRF training - L-BFGS

- Generally use L-BFGS\(^2\) algorithm.
- Approximates Newton’s method. Optimise multivariate function \(f(\theta)\) through updates

\[
\theta_{t+1} = \theta_t - [Hf(\theta_t)]^{-1} \nabla f(\theta_t)
\]

- Quasi-Newtonian: approximates Hessian \(Hf(\theta)\).
- Limited-memory: doesn’t store full (approximate) Hessian.

\(^2\)Limited-Memory Broyden-Fletcher-Goldfarb-Shanno Algorithm

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Label bias

- Generate data with noisy HMM.
- 4-state system (not counting ‘initial state’), transitions:
  - 1 \Rightarrow 2 \Rightarrow 3
  - 4 \Rightarrow 5 \Rightarrow 3
- Emissions: highly biased!
  - \( P(X = Y’s\ preferred\ value\ |\ Y) = 29/32 \)
  - \( P(X = other\ |\ Y) = 1/32 \)
Label bias

- Generate data with noisy HMM.
- 4-state system (not counting ‘initial state’), transitions:
  - 1 $\Rightarrow$ 2 $\Rightarrow$ 3
  - 4 $\Rightarrow$ 5 $\Rightarrow$ 3
- Emissions: highly biased!
  - $P(X = Y’s$ preferred value $| Y) = 29/32$
  - $P(X = other | Y) = 1/32$
- Preferred values: 1 $\rightarrow$ ‘r’, 4 $\rightarrow$ ‘r’, 2 $\rightarrow$ ‘i’, 5 $\rightarrow$ ‘o’, 3 $\rightarrow$ ‘b’.
- Result: CRF error 4.6%, MEMM error 42%.

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Mixed-order sources

- Generate data with mixed-order HMM:
  - Transitions: \((1 - \alpha)p_1(y_i|y_{i-1}) + \alpha p_2(y_i|y_{i-1}, y_{i-2})\)
  - Emissions: \((1 - \alpha)p_1(x_i|y_i) + \alpha p_2(x_i|y_i, x_{i-1})\)

- Five labels, 26 observation values.

- Training/testing: 1000 sequences of length 25.

- CRF trained with Algorithm S (modified IIS). MEMM trained with iterative scaling.

- Viterbi to label test set.

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Mixed-order sources: results

- Squares: $\alpha < 0.5$. 

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Mixed-order sources: results

- Squares: $\alpha < 0.5$.
- CRF sort of wins?

Presenters: Brad Gulko and Stephanie Hyland

CS 6784 Paper Presentation
Part of Speech Tagging

- Penn Treebank: 45 syntactic tags, label each word in sentence.
- Train first-order HMM, MEMM, CRF.
Part of Speech Tagging

- Penn Treebank: 45 syntactic tags, label each word in sentence.

- Train first-order HMM, MEMM, CRF.

<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

Presenters: Brad Gulko and Stephanie Hyland

CS 6784 Paper Presentation
Part of Speech Tagging

- Penn Treebank: 45 syntactic tags, label each word in sentence.

- Train first-order HMM, MEMM, CRF.

<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>

- Spelling features exploit conditional framework.

- Examples: starts with number/upper case?, contains hyphen, has suffix?

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Skip-chain CRF

- Example: skip-chain CRF\(^3\).

Skip-chain CRF

- Example: skip-chain CRF\(^3\).
- Has long-range features!


Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Skip-chain CRF

- Example: skip-chain CRF$^3$.
- Has long-range features!
- Basic idea: extend linear-chain CRF by joining some distant observations with ‘skip edges’.

---

Skip-chain CRF

- Example: skip-chain CRF\(^3\).
- Has long-range features!
- Basic idea: extend linear-chain CRF by joining some distant observations with ‘skip edges’.
- Connect multiple mentions of entity across whole document.

Example: Note: Squares denote factors (e.g. potential functions).

Skip-chain CRF

X5  Mr  Y5
X6  Smith  Y6
X7  is  Y7
X8  here.  Y8
X9  Smith  Y9
X10  says

Question: Ignoring the skip edges (in blue), what potentials does $Y_i$ appear in?

Answer: $(Y_i, Y_{i-1}, X_i), (Y_{i+1}, Y_i, X_{i+1})$.
Example: Note: Squares denote factors (e.g. potential functions).

Question: Ignoring the skip edges (in blue), what potentials does $Y_i$ appear in?

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Example: Note: Squares denote factors (e.g. potential functions).

Question: Ignoring the skip edges (in blue), what potentials does $Y_i$ appear in? 
Answer: $\psi(Y_i, Y_{i-1}, X_i), \psi(Y_{i+1}, Y_i, X_{i+1})$.

Presenters: Brad Gulko and Stephanie Hyland
CS 6784 Paper Presentation
Skip-chain CRF

Skip-chain task

- Data: 485 email announcements for seminars at CMU.
Skip-chain task

- Data: 485 email announcements for seminars at CMU.
- Task: identify start time, end time, location, speaker.
Skip-chain task

- Data: 485 email announcements for seminars at CMU.
- Task: identify start time, end time, location, speaker.
- Linear chain CRF with skip edges between identical capitalised words.
Skip-chain task

- Data: 485 email announcements for seminars at CMU.
- Task: identify start time, end time, location, speaker.
- Linear chain CRF with skip edges between identical capitalised words.
- Other word-specific features e.g. ‘appears in list of first names’, ‘upper case’, ‘appears to be part of time/date’ (by regex), etc.
## Skip-chain results

<table>
<thead>
<tr>
<th>System</th>
<th>stime</th>
<th>etime</th>
<th>location</th>
<th>speaker</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIEN Peshkin and Pfeffer [2003]</td>
<td>96.0</td>
<td>98.8</td>
<td>87.1</td>
<td>76.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Linear-chain CRF</td>
<td>97.5</td>
<td>97.5</td>
<td>88.3</td>
<td>77.3</td>
<td>90.2</td>
</tr>
<tr>
<td>Skip-chain CRF</td>
<td>96.7</td>
<td>97.2</td>
<td>88.1</td>
<td>80.4</td>
<td>90.6</td>
</tr>
</tbody>
</table>

- Values are F1 scores.
Skip-chain results

<table>
<thead>
<tr>
<th>System</th>
<th>stime</th>
<th>etime</th>
<th>location</th>
<th>speaker</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIEN Peshkin and Pfeffer [2003]</td>
<td>96.0</td>
<td><strong>98.8</strong></td>
<td>87.1</td>
<td>76.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Linear-chain CRF</td>
<td><strong>97.5</strong></td>
<td>97.5</td>
<td><strong>88.3</strong></td>
<td>77.3</td>
<td>90.2</td>
</tr>
<tr>
<td>Skip-chain CRF</td>
<td>96.7</td>
<td>97.2</td>
<td>88.1</td>
<td><strong>80.4</strong></td>
<td><strong>90.6</strong></td>
</tr>
</tbody>
</table>

- Values are F1 scores.
- Repeated occurrences of speaker improve skip-chain performance.
Skip-chain results

<table>
<thead>
<tr>
<th>System</th>
<th>stime</th>
<th>etime</th>
<th>location</th>
<th>speaker</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIEN Peshkin and Pfeffer [2003]</td>
<td>96.0</td>
<td>98.8</td>
<td>87.1</td>
<td>76.9</td>
<td>89.7</td>
</tr>
<tr>
<td>Linear-chain CRF</td>
<td>97.5</td>
<td>97.5</td>
<td>88.3</td>
<td>77.3</td>
<td>90.2</td>
</tr>
<tr>
<td>Skip-chain CRF</td>
<td>96.7</td>
<td>97.2</td>
<td>88.1</td>
<td>80.4</td>
<td>90.6</td>
</tr>
</tbody>
</table>

- Values are F1 scores.
- Repeated occurrences of speaker improve skip-chain performance.
- Tokens are *consistently* classified by skip-chain. Linear-chain is inconsistent on 30.2 speakers, skip-chain: 4.8.
Summary

- CRFs combine *discriminative* (e.g. MEMM) and *undirected* (e.g. MRF) properties to solve problems:
Summary

- CRFs combine *discriminative* (e.g. MEMM) and *undirected* (e.g. MRF) properties to solve problems:
  - Global normalisation avoids label bias.
Summary

- CRFs combine *discriminative* (e.g. MEMM) and *undirected* (e.g. MRF) properties to solve problems:
  - Global normalisation avoids label bias.
  - Conditioning on observations avoids modelling complex dependencies.
Summary

- CRFs combine *discriminative* (e.g. MEMM) and *undirected* (e.g. MRF) properties to solve problems:
  - Global normalisation avoids label bias.
  - Conditioning on observations avoids modelling complex dependencies.
  - Enables use of features using global structure.
Summary

- CRFs combine *discriminative* (e.g. MEMM) and *undirected* (e.g. MRF) properties to solve problems:
  - Global normalisation avoids label bias.
  - Conditioning on observations avoids modelling complex dependencies.
  - Enables use of features using global structure.

- Examples in paper strangely insubstantial, but CRFs are widely and successfully used.