Graphical Models (IV)

Applications in IR
— Probabilistic Topic Models

Eric Xing
Carnegie Mellon University
June 7, 2007

NLP and Data Mining

We want:
- Semantic-based search
- Infer topics and categorize documents
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- ...
Apoptosis + Medicine

probabilistic generative model

Apoptosis + Medicine
“…probability theory is more fundamentally concerned with the structure of reasoning and causation than with numbers.”

Glenn Shafer and Judea Pearl
*Introduction to Readings in Uncertain Reasoning*, Morgan Kaufmann, 1990
This Talk

- A recap of graphical model
- Two families of probabilistic topics models and approximate inference
  - Bayesian admixture models
  - Random models
- Three applications
  - Topic evolution
  - Machine translation
  - Multimedia inference

Probabilistic Graphical Models

- Graph-theoretic representations of probabilistic distributions

\[
p(X_1, X_2, X_3, X_4, X_5, X_6) = p(X_1) p(X_2|X_1) p(X_3|X_2) p(X_4|X_3) p(X_5|X_4) p(X_6|X_5, X_4)
\]

- Bayesian philosophy

- Modular combination of heterogeneous parts -- divide and conquer
Many modern problems in data mining/NLP can be formulated as probabilistic inference problems

\[ P( \text{query variable} | \text{query data} \& \text{KB}) \]

- Is this text document relevant to my query?
- Which category is this image in?
- What movies would I probably like?
- Create a caption for this image.
- Modeling document collections

General purpose algorithms exist to fully automate such computation

- Computational cost depends on the topology of the network
- Exact inference:
  - The junction tree algorithm
- Approximate inference:
  - Loopy belief propagation, variational inference, Monte Carlo sampling

Two types of GMs

- **Directed edges** give causality relationships (Bayesian Network or Directed Graphical Model):

- **Undirected edges** simply give correlations between variables (Markov Random Field or Undirected Graphical model):
This Talk

- A graphical model primer
- Two families of probabilistic topics models and approximate inference
  - Bayesian admixture models
  - Random models
- Three applications
  - Topic evolution
  - Machine translation
  - Multimedia inference

How to Model Semantic?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning

<table>
<thead>
<tr>
<th>MT</th>
<th>Syntax</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.
Why this is Useful?

- Q: What is it about?
  - A: Mainly MT, with syntax, some learning

<table>
<thead>
<tr>
<th>MT</th>
<th>Syntax</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- Q: give me similar document?
  - Structured way of browsing the collection

- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase-based model achieves a relative improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

Words in Contexts

- “It was a nice shot.”
the opposition Labor Party fared even worse, with a predicted 35 seats, seven less than last election.
Method One:

- **Hierarchical Bayesian Admixture**
Admixture Models

- Objects are bags of elements
- Mixtures are distributions over elements
- Objects have mixing vector \( \theta \)
  - Represents each mixtures’ contributions
- Object is generated as follows:
  - Pick a mixture component from \( \theta \)
  - Pick an element from that component

Topic Models = Admixture Models

Generating a document

- Draw \( \theta \) from the prior
  - For each word \( n \)
    - Draw \( z_n \) from \( \text{multinomial}(\theta) \)
    - Draw \( w_n | z_n, \{\beta_{ik}\} \) from \( \text{multinomial}(\beta_{z_n}) \)
### Prior Comparison

- **Dirichlet (LDA)** (Blei et al. 2003)
  - Conjugate prior means efficient inference
  - Can only capture variations in each topic’s intensity independently

- **Logistic Normal (CTM=LoNTAM)** (Blei & Lafferty 2005, Ahmed & Xing 2006)
  - Capture the intuition that some topics are highly correlated and can rise up in intensity together
  - Not a conjugate prior implies hard inference

### Approximate Inference

(e.g., MF, Jordan et al 1999, GMF, Xing et al 2004)

- **Log Partition Function**
  \[
  \log \left( 1 + \sum_{\gamma=1}^{K} e^{-\gamma} \right)
  \]

- **Multivariate Quadratic Approx.**
  - Closed Form Solution for \(\mu^*, \Sigma^*\)
  - Ahmed & Xing

- **Tangent Approx.**
  - Numerical Optimization to fit \(\mu^*, \text{Diag}(\Sigma^*)\)
  - Blei & Lafferty

- **\(\Sigma^*\) is full matrix**
- **\(\Sigma^*\) is assumed to be diagonal**
Variational Inference

\[
\begin{align*}
P(\gamma, z_{1:n} | D) & \quad \text{Approximate the Integral} \\
& \quad \text{Approximate the Posterior} \\
\arg \min_{\mu^*, \Sigma^*, \phi_{1:n}^*} KL(q || p) & \quad \text{Solve} \\
q(\gamma, z_{1:n}) &= q(\gamma | \mu^*, \Sigma^*) \prod q(z_i | \phi_i) \\
\end{align*}
\]

Variational Inference With no Tears

\[
\begin{align*}
P(\gamma, z | D) & \quad \text{Iterate until Convergence} \\
\text{Pretend you know } E[Z_{1:n}] & \quad P(\gamma | E[z_{1:n}], \mu, \Sigma) \\
\text{Now you know } E[\gamma] & \quad P(z_{1:n} | \gamma, w_{1:k}, \beta_{1:k}) \\
\end{align*}
\]

More Formally:

\[
q^*(X_C) = P\left( X_C | \langle S_y \rangle_{\beta_y}, \forall y \in X_{MB} \right)
\]

Message Passing Scheme (GMF)

Equivalent to previous method (Xing et. al.2003)
LoNTAM Variations Inference

- Fully Factored Distribution

\[ q(y, z_{in}) = q(y) \prod q(z_i) \]

- Two clusters: \( \lambda \) and \( Z_{1:n} \)

\[ q^*(X_c) = P\left(X_c \mid \langle S_{\gamma} \rangle_{q_y} : \forall y \in X_{mb}\right) \]

- Fixed Point Equations

\[
q_y^*(\gamma) = P\left(\gamma \mid \langle S_{\gamma} \rangle_{q_y}, \mu, \Sigma\right) \\
q_z^*(z) = P\left(z \mid \langle S_{\gamma} \rangle_{q_y}, \beta_{ik}\right)
\]

Variational \( \gamma \)

\[
q_{\lambda}^*(\gamma) = P\left(\gamma \mid \langle S_{\gamma} \rangle_{q_y}, \mu, \Sigma\right) \\
\propto P(\gamma \mid \mu, \Sigma) P(\langle S_{\gamma} \rangle_{q_y} \mid \gamma)
\]

\[
S_z = m = \left[\sum_n I(z_a = 1), \ldots, \sum_n I(z_a = k)\right]
\]

\[
\propto N(\gamma \mid \mu, \Sigma) \exp\left(\langle m \rangle_{q_y}, \gamma - N \times C(\gamma)\right)
\]

\[
\propto \exp\left(-\frac{1}{2} \gamma' \Sigma^{-1} \gamma + \gamma^{\Sigma^{-1}} m + \langle m \rangle_{q_y}, \gamma - N \times C(\gamma)\right)
\]

\[
C(\gamma) = C(\gamma) + g_{1}^{*} (\gamma - \gamma_{-}) + 0.5 (\lambda - \gamma_{-}) H (\gamma - \gamma_{-})
\]

\[
q_{\lambda}^*(\gamma) = N(\gamma \mid \mu, \Sigma) \\
\mu = \Sigma \{\Sigma^{-1} \mu + NH \gamma + \langle m \rangle - N \gamma\}
\]

Eric Xing
Tangent Approximation

Graph showing tangent approximation

Test on Synthetic Text

Graphs illustrating test on synthetic text
Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

- Varying Num. of Topics
- Varying Voc. Size
- Varying Num. Words Per Document

Comparison: perplexity
Topics and topic graphs

Result on PNAS collection

- PNAS abstracts from 1997-2002
  - 2500 documents
  - Average of 170 words per document
- Fitted 40-topics model using both approaches
- Use low dimensional representation to predict the abstract category
  - Use SVM classifier
  - 85% for training and 15% for testing

### Classification Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Doc</th>
<th>BL</th>
<th>AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics</td>
<td>21</td>
<td>61.9</td>
<td>61.9</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>86</td>
<td>65.1</td>
<td>77.9</td>
</tr>
<tr>
<td>Immunology</td>
<td>24</td>
<td>70.8</td>
<td>66.6</td>
</tr>
<tr>
<td>Biophysics</td>
<td>15</td>
<td>53.3</td>
<td>66.6</td>
</tr>
<tr>
<td>Total</td>
<td>146</td>
<td>64.3</td>
<td>72.6</td>
</tr>
</tbody>
</table>
Method Two:

- Layered Boltzmann machines

The Harmonium

hidden units

visible units

Boltzmann machines:

\[
p(x,h \mid \theta) = \exp \left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{i,j} \phi_{i,j}(x_i, h_j) - A(\theta) \right\}
\]
Properties of Harmoniums

- Factors are marginally dependent.
- Factors are conditionally independent given observations on the visible nodes.
\[ P(\ell | w) = \prod_i P(\ell_i | w) \]
- Iterative Gibbs sampling.
- Learning with contrastive divergence

A Binomial Word-count Model

\[ h_j = 3: \text{topic } j \text{ has strength } 3 \]
\[ h_j \in \mathbb{R}, \quad \langle h_j \rangle = \sum_i W_{i,j} x_i \]
\[ x_i = n: \text{word } i \text{ has count } n \]
\[ x_i \in \mathbb{I} \]

\[
p(h | x) = \prod_j \text{Normal}_{h_j} \left[ \sum_i W_{i,j} \bar{x}_i, 1 \right]
\]
\[
p(x | h) = \prod_i \text{Bi}_{x_i} \left[ N, \frac{\exp(\alpha_i + \sum_j W_{i,j} h_j)}{1 + \exp(\alpha_i + \sum_j W_{i,j} h_j)} \right]
\]
\[ \Rightarrow p(x) \propto \exp \left\{ \sum \alpha_i x_i - \log \Gamma(x_i) - \log \Gamma(N - x_i) + \frac{1}{2} \sum_j \left\{ \langle x_j \rangle - \sum_i W_{i,j} x_i \right\}^2 \right\} \]

Bi\(_N\)[\(N, p\)] = \(C_N^p (1-p)^{N-p} = C_N^p \left(\frac{N}{N+1}\right)^p (1-p)^{N-p} \)

Let \(p = \frac{\exp(\alpha_i + \sum_j W_{i,j} h_j)}{1 + \exp(\alpha_i + \sum_j W_{i,j} h_j)}\),
\[ \text{Bi}_i[N, p] = C_N^p \left(\frac{N}{N+1}\right)^p \]
\[ \propto C_N^p \exp \left\{ \left(\alpha_i + \sum_j W_{i,j} h_j\right) x_i + A_i \right\} \]
Reduce to softmax when \(N=1\)!
The Computational Trade-off

**Undirected model**: Learning is hard, inference is easy.

**Directed Model**: Learning is "easier", inference is hard.

Example: Document Retrieval.

Retrieval is based on comparing (posterior) topic distributions of documents.
- **directed models**: inference is slow. Learning is relatively "easy".
- **undirected model**: inference is fast. Learning is slow but can be done offline.

Comparison of model semantics

- **LSI**: $\bar{x} = W^* \tilde{d}$
- **LDA**: $p(X) \leftarrow Z \leftarrow \tilde{\theta}$
- **Harmonium**: $p(X) \leftarrow W^T \tilde{\theta}$
Multi-Source Data

TRECVID 2004 Example Images

Inter-Source Associations

Z and X are marginally dependent (same as GM-LDA)
Multi-wing Harmoniums

Learning and Inference

- Maximal likelihood learning based on gradient ascent.
  \[ \delta \theta \propto \{ f_i(x_i) \}_{\text{data}} - \{ f_i(x_i) \}_p \]

  - Gradient computation requires model distribution \( p(\cdot) \)
  - \( p(\cdot) \) is intractable

- Contrastive Divergence
  - Approximate \( p(\cdot) \) with Gibbs sampling

- Variational approximation
  - GMF approximation
  \[ q(x, z, h) = \prod_i q(x_i | v_i) \prod_k q(z_k | \mu_k, \sigma_k) \prod_j q(h_j | y_j) \]
Inter-source Inference

- GMF approximation to DWH

\[ q(x, z, h) = \prod_i q(x_i | N, \nu) \prod_k q(z_k | \mu_k, \sigma_k) \prod_j q(h_j | \gamma_j) \]

- Expected mean value of topic strength:

\[ \gamma_j = \sum_i W_{i,j} \nu_i + \sum_k U_{k,j} \mu_k \]

- Expected mean value of image-feature:

\[ \mu_k = \sigma^2_k \left( \beta_k + \sum_j U_{k,j} \gamma_j \right) \]

- Expected mean count

\[ NV_i = N \frac{\exp(\gamma_j + \sum_j W_{i,j} \nu_j)}{1 + \exp(\gamma_j + \sum_j W_{i,j} \nu_j)} \]

Examples of Latent Topics

| \(T_1\) | storms gulf hawaii low forecast southeast showers |
| \(T_2\) | rebounds 14 shouting tests guard cut hawks |
| \(T_3\) | engine flying craft asteroid say hour aerodynamic |
| \(T_4\) | safe cross red sure dry providing services |
| \(T_5\) | losing jersey sixth antonio david york orlando |
This Talk

- A graphical model primer
- Two families of probabilistic topics models and approximate inference
  - Bayesian admixture models
  - Random models
- Three applications
  - Learning topic graphs and topic evolution
  - Machine translation
  - Multimedia inference
Application 1: How to model topic correlation?

(a) CTM  (b) PAM  (c) sCTM

And topic evolution?
Sparse Correlated Model (SCTM)

NIPS: Example Network
**NIPS: Held-out Perplexity**

![Graphs showing topic trends and perplexity](image)

**How to Model Topic Evolution**

- Topic Trends
- Topic Keywords
- Topic correlations
- Number of topics

The Dynamic Correlated Topic model

![Diagrams showing topic evolution over years](image)
The Dynamic CTM

Topic Trends
Topic Words over Time

Topic Correlations Over Time
Application 2: Machine translation

B. Zhao and E.P Xing, ACL 2006

Word Alignment

The economy and trade relations between Russia and Tianjin develop steadily.
The Statistical Formulation

\[ \hat{a} = \arg \max_a \Pr(f \mid e, a) \Pr(e) \]

BiTAM Model-1

- Graphical Model (a language to encode dependencies)

\[
p(F \mid A, E, \alpha, B) = \int_\theta p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(f_n \mid a_n, e_n, B_n) d\theta
\]
**An upgrade path for BiTAMs**

Sent-pair level topics

Word-pair level topics

HMM for Alignment

Word-Pair & HMM

**Experiments**

- **Training data**
  - Small: Treebank 316 doc-pairs (133K English words)
  - Large: FBIS-Beijing, Sinorama, XinHuaNews, (15M English words).

<table>
<thead>
<tr>
<th>Train</th>
<th>#Doc.</th>
<th>#Sent.</th>
<th>#Tokens English</th>
<th>#Tokens Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treebank</td>
<td>316</td>
<td>4172</td>
<td>133K</td>
<td>105K</td>
</tr>
<tr>
<td>FBIS.BJ</td>
<td>6,111</td>
<td>105K</td>
<td>4.18M</td>
<td>3.54M</td>
</tr>
<tr>
<td>Sinorama</td>
<td>2,373</td>
<td>103K</td>
<td>3.81M</td>
<td>3.60M</td>
</tr>
<tr>
<td>XinHua</td>
<td>19,140</td>
<td>115K</td>
<td>3.85M</td>
<td>3.93M</td>
</tr>
<tr>
<td>FOUO</td>
<td>15,478</td>
<td>368K</td>
<td>13.14M</td>
<td>11.93M</td>
</tr>
<tr>
<td>Test</td>
<td>95</td>
<td>627</td>
<td>25,500</td>
<td>19,726</td>
</tr>
</tbody>
</table>

- **Word Alignment Accuracy & Translation Quality**
  - F-measure
  - BLEU
## Topics

<table>
<thead>
<tr>
<th>T1</th>
<th>Teams, sports, disabled, games members, people, cause, water, national, handicapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>Shenzhen, singapore, hongkong, stock, national, investment, yuan, options, million, dollar</td>
</tr>
<tr>
<td>T3</td>
<td>Chongqing, company, takeover, shenzhen, tianjin, city, national, government, project, companies</td>
</tr>
<tr>
<td>T4</td>
<td>Hongkong, trade, export, import, foreign, tech., high, 1998, year, technology</td>
</tr>
<tr>
<td>T5</td>
<td>House, construction, government, employee, living, provinces, macau, anhui, yuan</td>
</tr>
<tr>
<td>T6</td>
<td>Gas, company, energy, usa, russia, france, chongqing, resource, china, economy, oil</td>
</tr>
</tbody>
</table>

## Comparison

![Graph comparing negative log-likelihood of HMI-STAN vs IBM Model-4 and HMI with forward-backward EM](image-url)
HM-BiTAM versus others

Translation Evaluations
### Application 3: video representation/classification

- **Video**: a complex, multi-modal data type for representation and classification
  - Image, text (closed-captions, speech transcript), audio

- **Goal**: classify video segments called **video shots** into semantic categories

---

<table>
<thead>
<tr>
<th>Systems</th>
<th>1-gram</th>
<th>2-gram</th>
<th>3-gram</th>
<th>4-gram</th>
<th>BLEU@4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hero Sys.</td>
<td>73.92</td>
<td>40.57</td>
<td>23.21</td>
<td>13.84</td>
<td>30.70</td>
</tr>
<tr>
<td>Gale Sys.</td>
<td>75.63</td>
<td>42.71</td>
<td>25.00</td>
<td>14.30</td>
<td>32.78</td>
</tr>
<tr>
<td>HM-BiTAM</td>
<td>76.77</td>
<td>42.99</td>
<td>25.42</td>
<td>14.04</td>
<td>33.19</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>76.10</td>
<td>43.85</td>
<td>26.70</td>
<td>15.73</td>
<td>34.17</td>
</tr>
</tbody>
</table>
Harmoniums for Multi-modal Data

- Dual-wing harmoniums (DWH) [Xing et al. 05]
  - modeling bi-modal data: captioned images, video
  - learning hidden topics from two "wings" of observed features

\begin{align*}
H_j & \cdots & H_k \\
X_j & \cdots & X_N \\
Z_j & \cdots & Z_M \\
\text{text features} & & \text{image features}
\end{align*}

Mixture-of-Harmoniums (MoH)

- A family of category-specific dual-wing harmoniums

\begin{align*}
\gamma \\
H_j & \cdots & H_k \\
X_j & \cdots & X_N \\
\text{category 1} \\
H_j & \cdots & H_k \\
X_j & \cdots & X_N \\
\text{category 2} \\
H_j & \cdots & H_k \\
X_j & \cdots & X_N \\
\text{category T}
\end{align*}

- classification by finding the "best-fitting" harmonium
Hierarchical Harmonium (HH)

- Incorporate category labels as a layer of hidden nodes on top of latent topic nodes

Semantic Topics by FoH

- Revealing “sub-topics” of each category
- Co-clusters of both text and image features
Semantic Topics by HH

- Reveal the "common topics" of all the data

Inter-category relationship
Classification Accuracy

- Harmonium models outperform directed models (e.g., LDA)

![Classification Accuracy Chart]

Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive

- There are many ways of implementing topic models
  - Directed
  - Undirected

- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC

- Many applications
  - ...