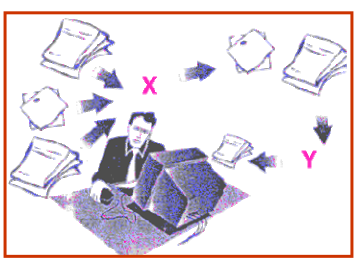
Probabilistic Topic Models and Applications

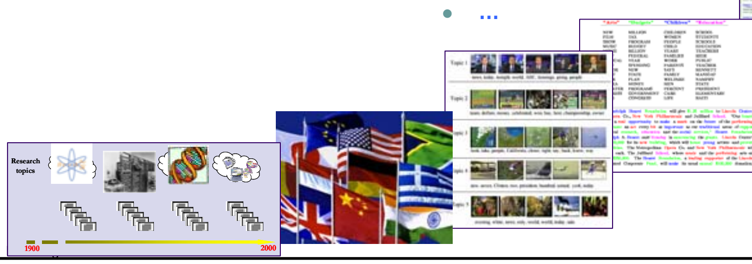
Eric Xing
Machine Learning Department
Language Technology Institute
Computer Science Department
Carnegie Mellon University

NLP and Data Mining

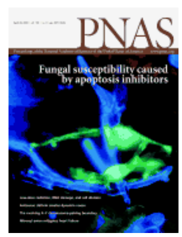


We want:

- Semantic-based search
- infer topics and categorize documents
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- …

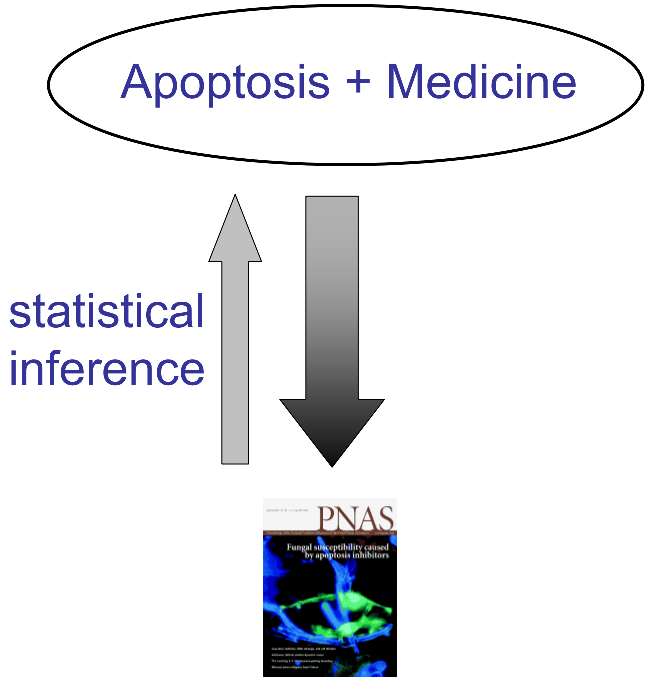


Apoptosis + Medicine



Apoptosis + Medicine

probabilistic generative model

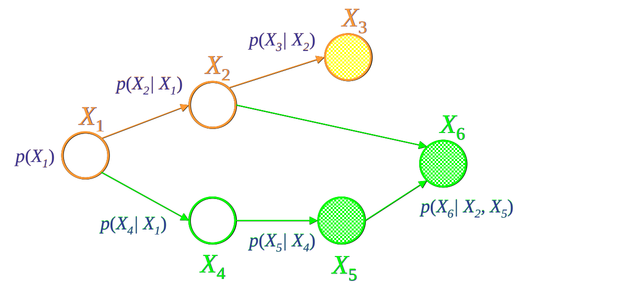


This Talk

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

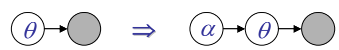
Probabilistic Graphical Models

- Graph-theoretic representations of probabilistic distributions



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

- Bayesian philosophy



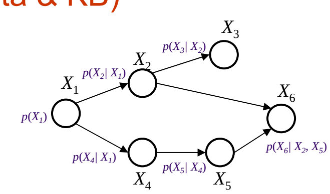
- Modular combination of heterogeneous parts -- **divide and conquer**

Probabilistic Inference

- Many modern problems in data mining/NLP can be formulated as probabilistic inference problems

P(query variable | query data & KB)

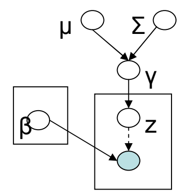
  - Is this text document relevant to my query?
  - Which category is this image in?
  - What movies would I probably like?
  - Create a caption for this image.
  - Modeling document collections



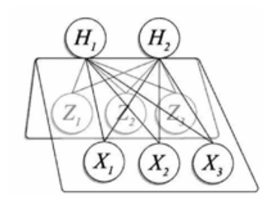
- General purpose algorithms exist to fully automate such computation
  - Computational cost depends on the topology of the network
  - Exact inference:
    - The junction tree algorithm
  - Approximate inference;
    - Loopy belief propagation, variational inference, Monte Carlo sampling

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
  - Image topics
  - Multimedia inference

How to Model Semantic?

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

0.6	0.3	0.1
MT	Syntax	Learning

Mixing Proportion

Source Target SMT Alignment Score BLEU	Parse Tree Noun Phrase Grammar CFG	likelihood EM Hidden Parameters Estimation argMax

Topics

Unigram over vocabulary

Topic Models

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

0.6	0.3	0.1
MT	Syntax	Learning

Mixing Proportion

- 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.

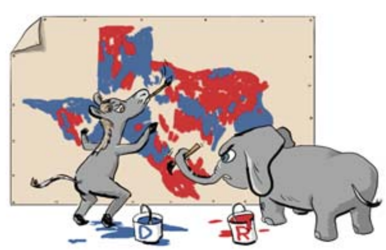
Words in Contexts

- “It was a nice **shot**. ”

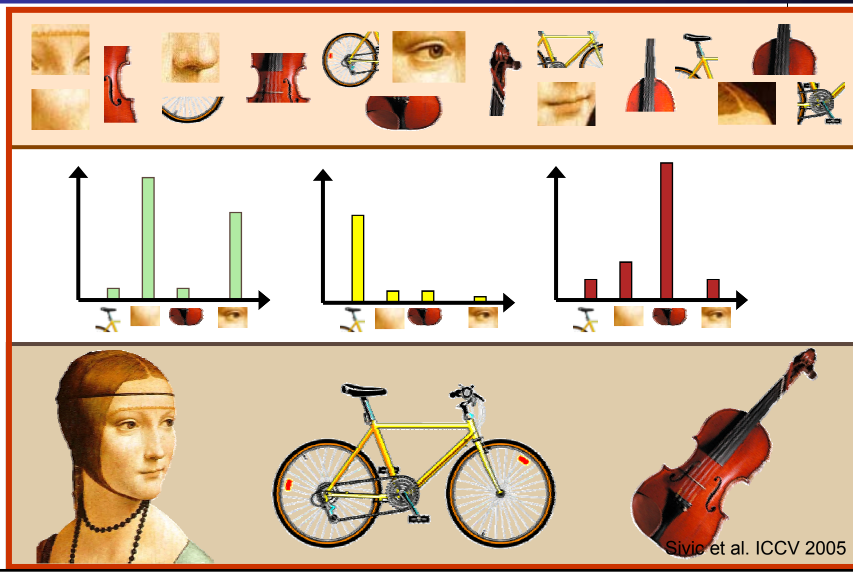


Words in Contexts (con'd)

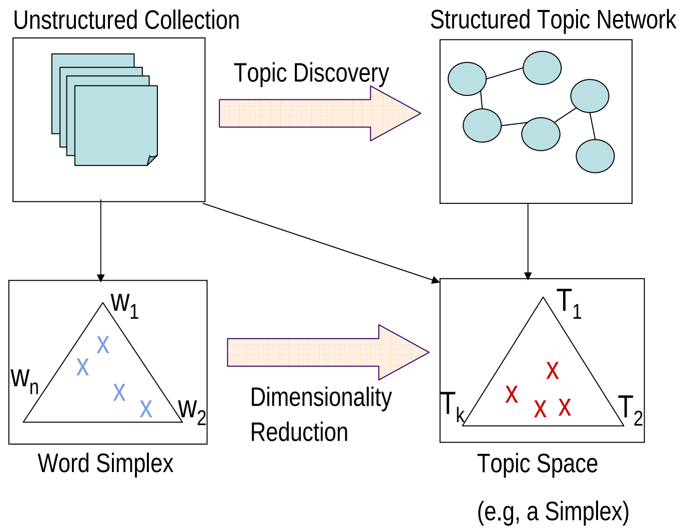
- the opposition Labor **Party** fared even worse, with a predicted 35 **seats**, seven less than last **election**.

"Words" in Contexts (con'd)



Topic Models: The Big Picture



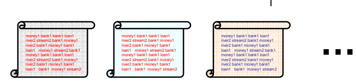
Method One:

- **Hierarchical Bayesian Admixture**

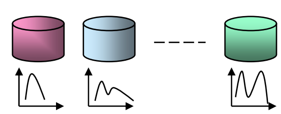
A. Ahmed and E.P. Xing
AISTAT 2007

Admixture Models

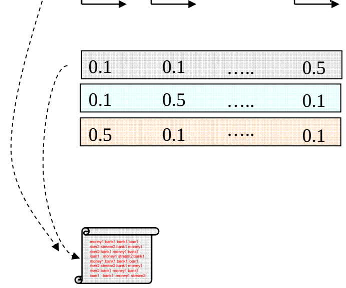
- Objects are bags of elements



- Mixtures are distributions over elements



- Objects have mixing vector θ
  - Represents each mixtures' contributions
- Object is generated as follows:
  - Pick a mixture component from θ
  - Pick an element from that component



Topic Models =Admixture Models

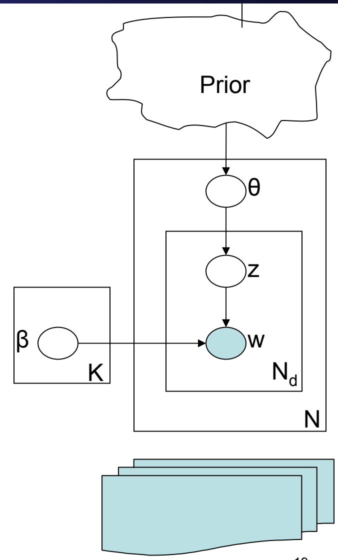
Generating a document

– *Draw* θ from the prior

For each word n

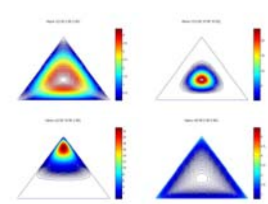
- Draw z_n from $multinomial(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from $multinomial(\beta_{z_n})$

Which prior to use?

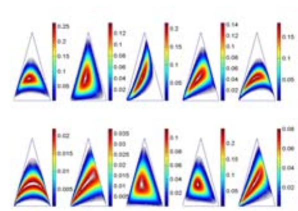


Prior Comparison

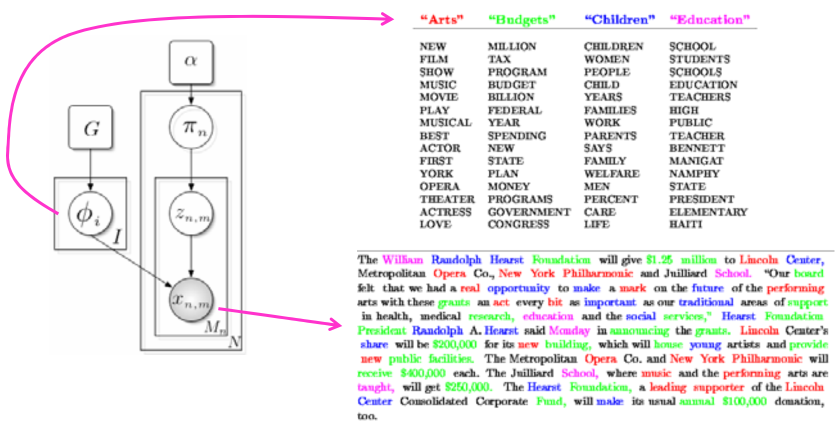
- Dirichlet (LDA) (Blei et al. 2003)
  - Conjugate prior means efficient inference
  - Can only capture variations in each topic's intensity independently



- 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



Inference Tasks

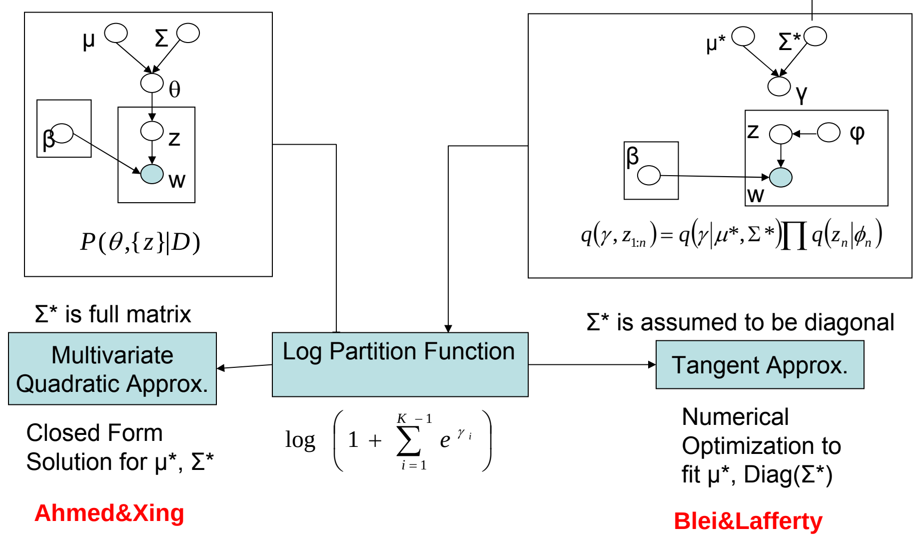


Approximate Inference

- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2nd-order Taylor approximation (Xing)
- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths et al)

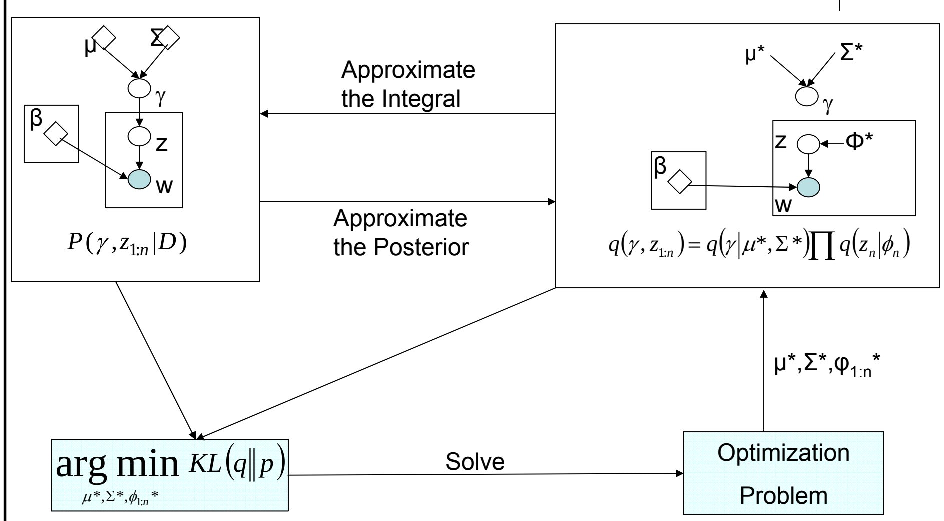
Approximate Inference

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



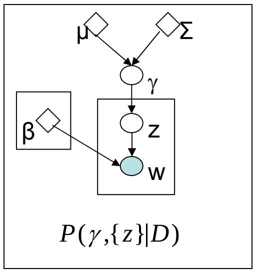
Eric Xing 23

Variational Inference



Eric Xing 24

Variational Inference With no Tears



Iterate until Convergence

- Pretend you know $E[Z_{1:n}]$
  - $P(\gamma|E[z_{1:n}], \mu, \Sigma)$
- Now you know $E[\gamma]$
  - $P(z_{1:n}|E[\gamma], w_{1:n}, \beta_{1:k})$

- More Formally:

$$q^*(X_C) = P\left(X_C \middle| \langle S_Y \rangle_{q_y} : \forall y \in X_{MB} \right)$$

Message Passing Scheme (GMF)

Equivalent to previous method (Xing et. al.2003)

LoNTAM Variations Inference

- Fully Factored Distribution

$$q(\gamma, z_{1:n}) = q(\gamma)\prod q(z_n)$$

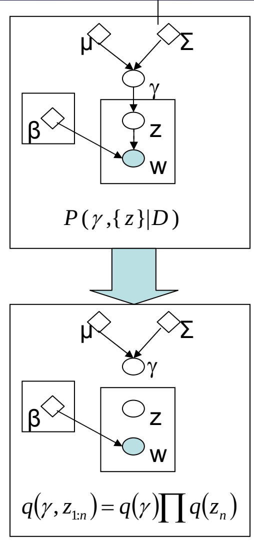
- Two clusters: λ and $Z_{1:n}$

$$q^*(X_C) = P\left(X_C \middle| \langle S_Y \rangle_{q_y} : \forall y \in X_{MB} \right)$$

- Fixed Point Equations

$$q_\gamma^*(\gamma) = P\left(\gamma \middle| \langle S_z \rangle_{q_z}, \mu, \Sigma\right)$$

$$q_z^*(z) = P\left(z \middle| \langle S_\gamma \rangle_{q\gamma}, \beta_{1:k}\right)$$



Variational γ

$$q_\lambda^*(\gamma) = P\left(\gamma \middle| \langle S_z\rangle_{q_z}, \mu, \Sigma\right)$$

$$\propto P(\gamma \mid \mu, \Sigma) P\left(\langle S_z\rangle_{q_z} \middle| \gamma\right)$$

Now what is $\langle S_z\rangle_{q_z}$?

$$S_z = m = \left[\sum_n I(z_n = 1), \dots, \sum_n I(z_n = k)\right]$$

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

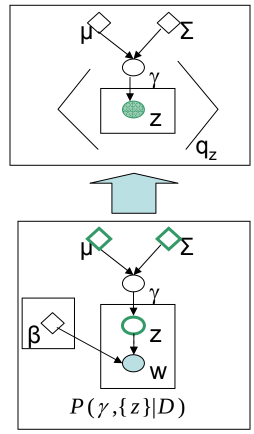
$$\propto \exp\left\{\underbrace{-\frac{1}{2}\gamma'\Sigma^{-1}\gamma + \gamma\Sigma^{-1}\mu + \langle m\rangle_{q_z}\gamma}\; - N \times \underbrace{\boxed{C(\gamma)}}\right\}$$

$$C(\gamma) = C(\gamma_\wedge) + g'_\lambda(\gamma - \gamma_\wedge) + .5(\lambda - \gamma_\wedge)' H (\gamma - \gamma_\wedge)$$

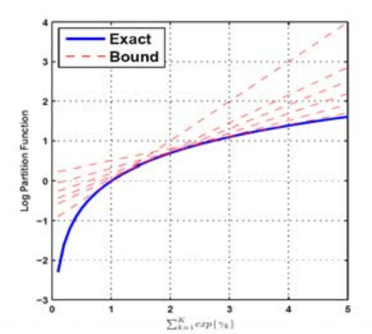
$$q_\lambda^*(\gamma) = N(\mu_\gamma, \Sigma_\gamma)$$

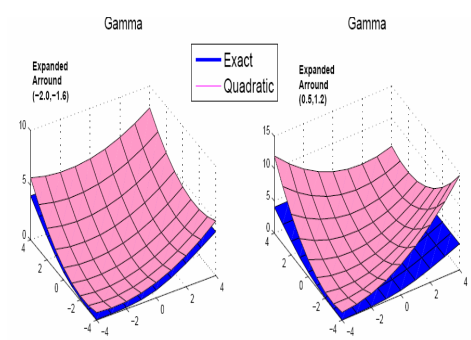
$$\Sigma_\gamma = inv\left(\Sigma^{-1} + NH\right)$$

$$\mu_\gamma = \Sigma_\gamma\left(\Sigma^{-1}\mu + NH\gamma_\wedge + \langle m\rangle - Ng\right)$$

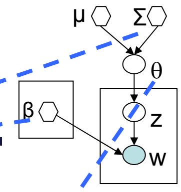


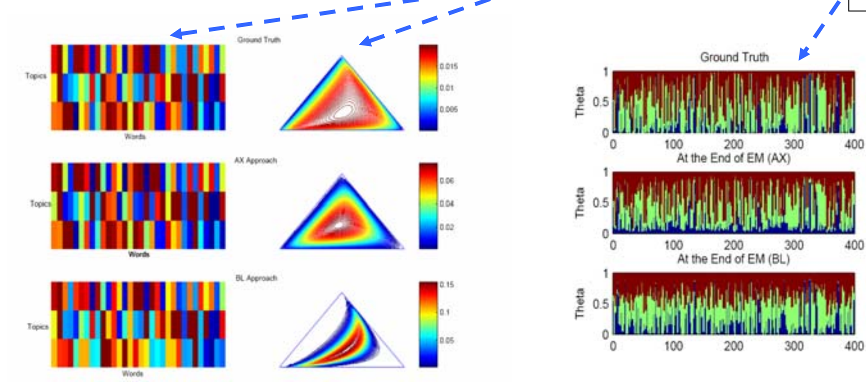
Tangent Approximation





Test on Synthetic Text

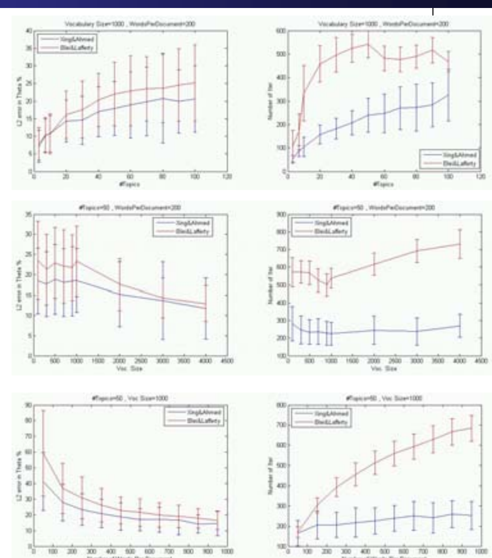




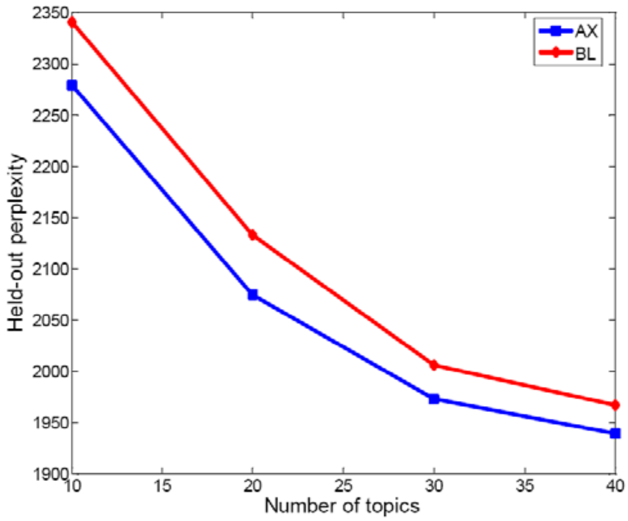
Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

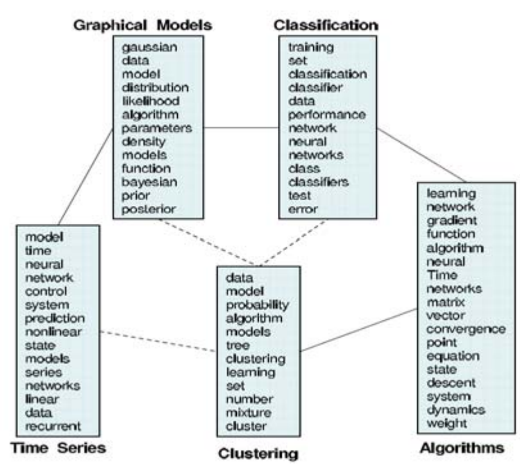
- Varying Num. of Topics
- Varying Voc. Size
- Varying Num. Words Per Document



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

Category	Doc	BL	AX
Genetics	21	61.9	61.9
Biochemistry	86	65.1	77.9
Immunology	24	70.8	66.6
Biophysics	15	53.3	66.6
Total	146	64.3	72.6

Method Two:

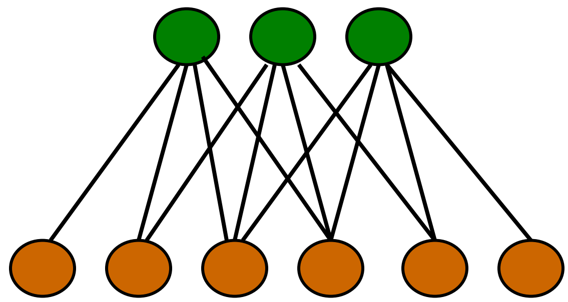
- **Layered Boltzmann machines**

E.P. Xing, R. Yan and A. G. Hauptmann, UAI 2006

The Harmonium

hidden units

visible units

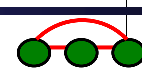


Boltzmann machines:

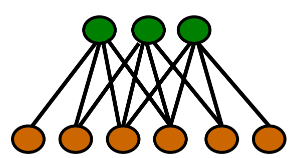
$$p(x,h\,|\,\theta)=\exp\Big\{\ \sum_i \theta_i\phi_i(x_i)+\sum_j \theta_j\phi_j(h_j)+\sum_{i,j}\theta_{i,j}\phi_{i,j}(x_i,h_j)-A(\boldsymbol{\theta})\ \Big\}$$

Properties of Harmoniums

- Factors are marginally *dependent*.

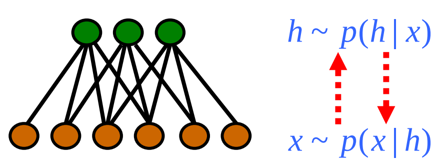


- Factors are conditionally *independent* given observations on the visible nodes.

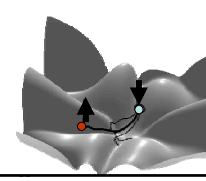


$$P(\boldsymbol{\ell}\,|\,\mathbf{w})=\prod_i P(\ell_i\,|\,\mathbf{w})$$

- Iterative Gibbs sampling.

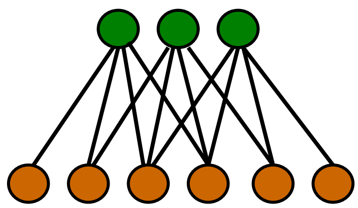


- Learning with contrastive divergence



A Binomial Word-count Model

topics



words counts

$h_j = 3$: *topic j has strength 3*

$h_j \in \mathbf{R}$, $\quad \langle h_j \rangle = \Sigma_i W_{i,j} x_i$

$x_i =$ n: *word i has count n*

$x_i \in \mathbf{I}$

$$\mathrm{Bi}_{x_i}[N, p] = C_{x_i}^N p^{x_i}(1-p)^{N-x_i} = C_{x_i}^N \left(\tfrac{p}{1-p}\right)^{x_i}(1-p)^N$$

Let $p = \frac{\exp(\alpha_j + \Sigma_j W_{ij} h_j)}{1+\exp(\alpha_j + \Sigma_j W_{ij} h_j)}$,

$$\mathrm{Bi}_{x_i}[N, p] = C_{x_i}^N \frac{\left(\exp(\alpha_i + \Sigma_j W_{ij} h_j)\right)^{x_i}}{\left(1+\exp(\alpha_i + \Sigma_j W_{ij} h_j)\right)^N}$$

$$\propto C_{x_i}^N \exp\left\{\left(\alpha_i + \Sigma_j W_{ij} h_j\right) x_i + A_i\right\}$$

Reduce to softmax when N=1 !

$$p(\mathbf{h} \mid \mathbf{x}) = \prod_j \mathrm{Normal}_{h_j}\left[\ \sum_i \vec{W}_{ij}\, \vec{x}_i, 1\ \right]$$

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_i \mathrm{Bi}_{x_i}\left[\ N, \frac{\exp(\alpha_j + \Sigma_j W_{ij} h_j)}{1+\exp(\alpha_j + \Sigma_j W_{ij} h_j)}\ \right]$$

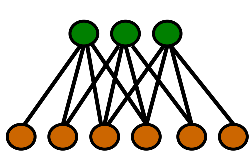
$$\Rightarrow\ p(\mathbf{x}) \propto \exp\left\{\left(\Sigma_i\, \alpha_i x_i - \log\Gamma(x_i) - \log\Gamma(N - x_i)\right) + \tfrac{1}{2}\, \Sigma_j \left(\Sigma_i W_{i,j} x_i\right)^2\right\}$$

The Computational Trade-off

<u>**Undirected model**</u>: Learning is hard, inference is easy.

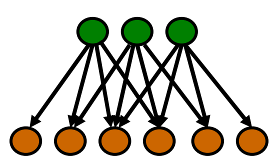
<u>**Directed Model**</u>: Learning is "easier", inference is hard.

Example: Document Retrieval.



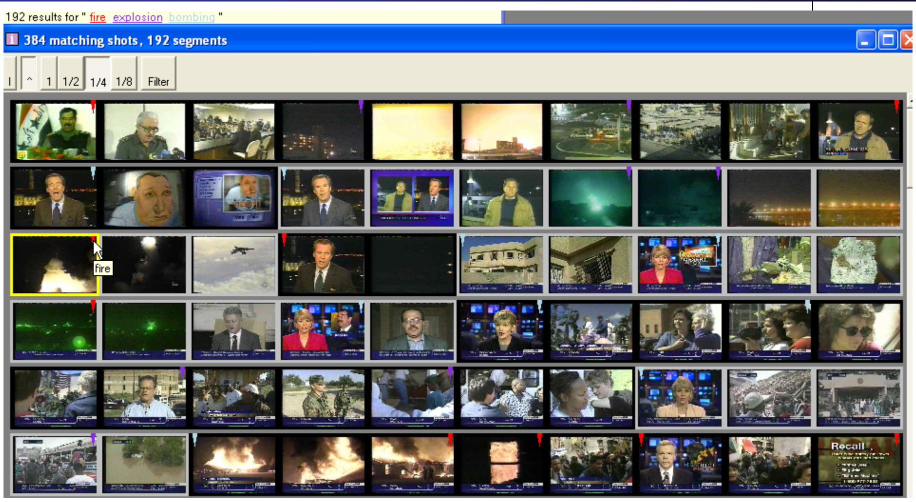
topics

words



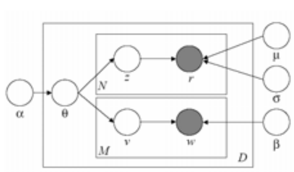
Retrieval is based on comparing (posterior) topic distributions of documents.
- <u>directed models</u>: inference is slow. Learning is relatively "easy".
- <u>undirected model</u>: inference is fast. Learning is slow but can be done offline.

Multi-Source Data

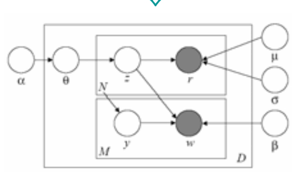


TRECVID 2004 Example Images

Inter-Source Associations

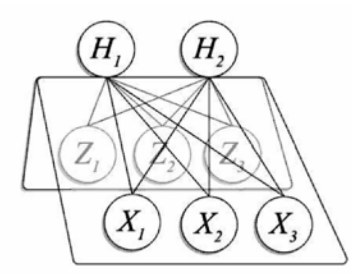


GM LDA ⇓



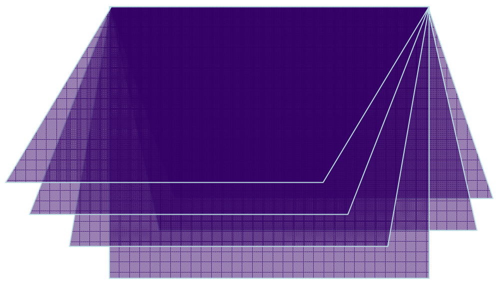
Co-LDA

DWH



Z and *X* are marginally dependent (same as GM-LDA)

Multi-wing Harmoniums



Learning and Inference

- Maximal likelihood learning based on gradient ascent.

$$\delta\theta_i \propto \langle f_i(x_i) \rangle_{\text{data}} - \langle f_i(x_i) \rangle_p$$

  - gradient computation requires model distribution $p(.)$
  - $p(.)$ is intractable
- Contrastive Divergence
  - approximate $p(.)$ with Gibbs sampling
- Variational approximation
  - GMF approximation

$$q(\mathbf{x},\mathbf{z},\mathbf{h}) = \prod_i q(x_i \mid \nu_i) \prod_k q(z_k \mid \mu_k, \sigma_k) \prod_j q(h_j \mid \gamma_i)$$

Inter-source Inference

- GMF approximation to DWH

$$q(\mathbf{x},\mathbf{z},\mathbf{h}) = \prod_i q(x_i \mid N,\nu_i)\prod_k q(z_k \mid \mu_k,\sigma_k)\prod_j q(h_j \mid \gamma_j)$$

  - Expected mean value of topic strength:

$$\gamma_j = \Sigma_i W_{i,j}\nu_i + \Sigma_k U_{k,j}\mu_k$$

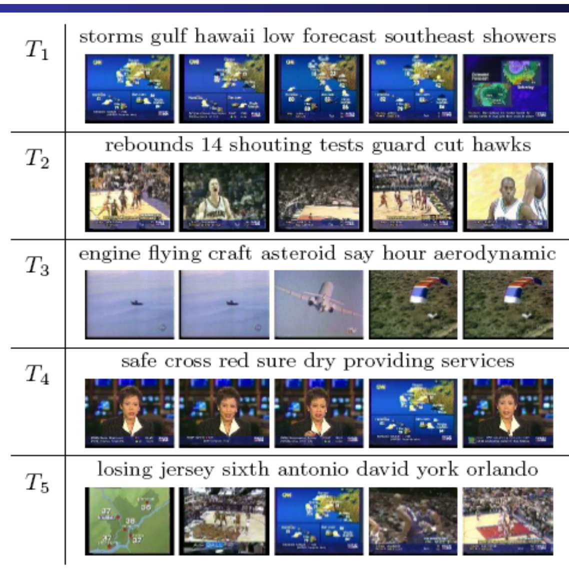
  - Expected mean value of image-feature :

$$\mu_k = \sigma_k^2\left(\beta_k + \Sigma_j U_{k,j}\gamma_j\right)$$

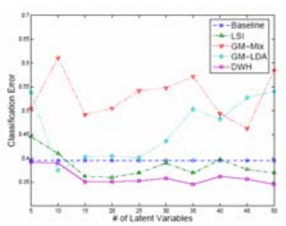
  - Expected mean count

$$N\nu_i = N\frac{\exp(\alpha_j + \Sigma_j W_{ij}\gamma_j)}{1+\exp(\alpha_j + \Sigma_j W_{ij}\gamma_j)}$$

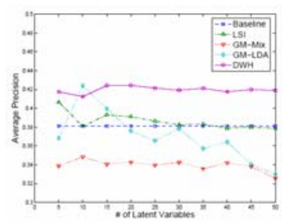
Examples of Latent Topics



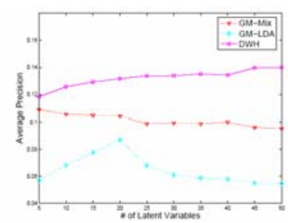
Performance



Classification



Retrieval

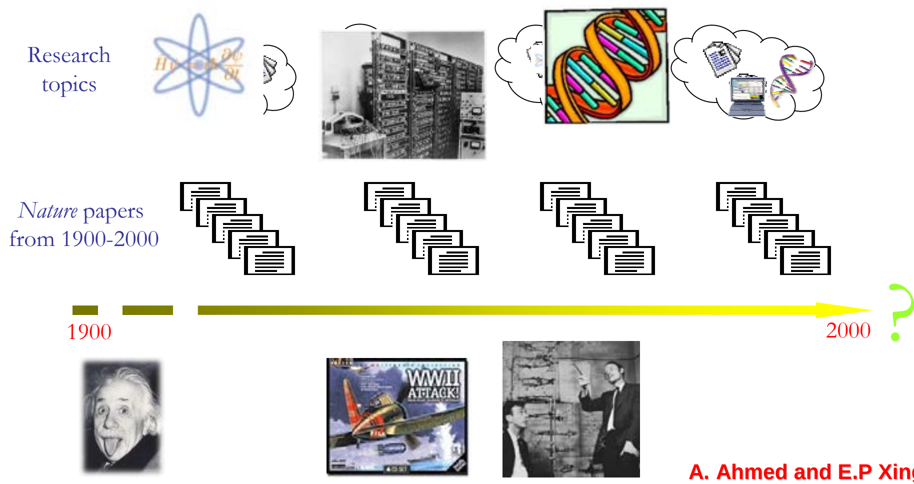


Annotation

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
  - Image topics
  - Multimedia inference

Application 1: topic evolution?

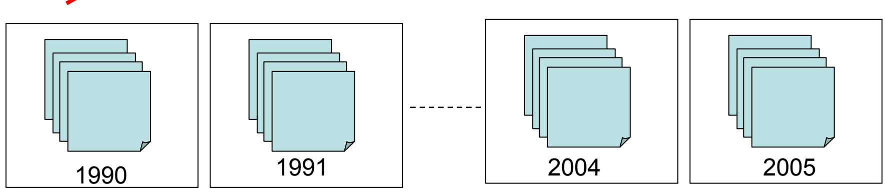


A. Ahmed and E.P Xing, Submitted 2007

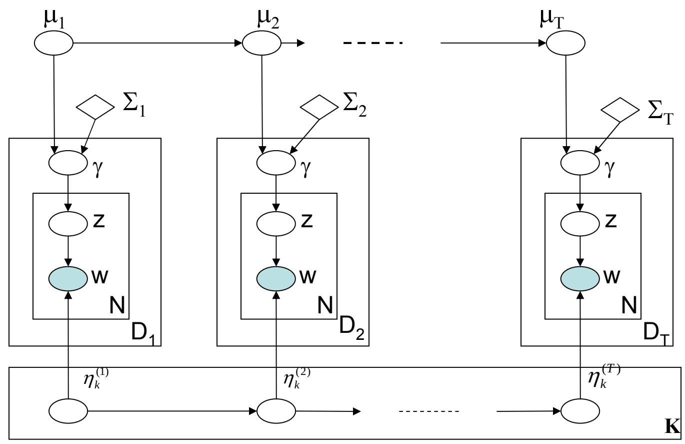
How to Model Topic Evolution

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

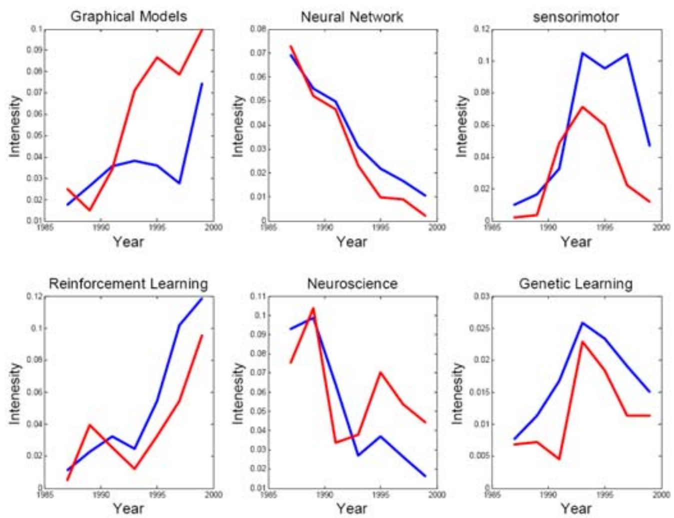
The Dynamic Correlated Topic model



The Dynamic CTM



Topic Trends

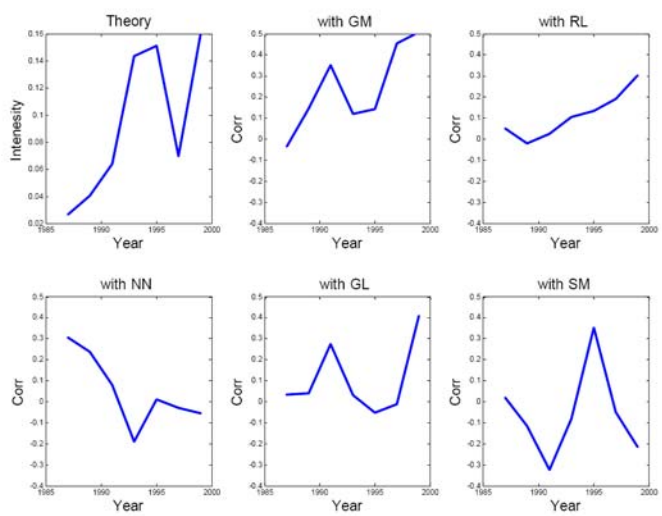


Topic Words over Time

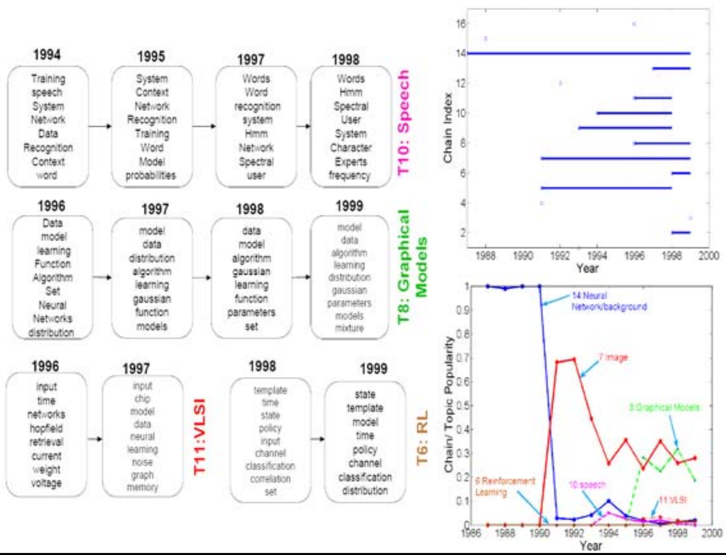
Leg Insect Controller Sensory sweeping → learning Sweeping state prioritized time control → learning State neural control reinforcement policy → Policy learning State reinforcement Optimal action Reward markov nonlinear — RL

brain information neurons fibers potential model synaptic → model activity network neurons cell phase firing → Model time neurons Spike input Cells head information Rate synaptic temporal → Synergy synaptic Pruning neurons adaptation Spike neuron degradation Irregular hebbian — Neuroscience

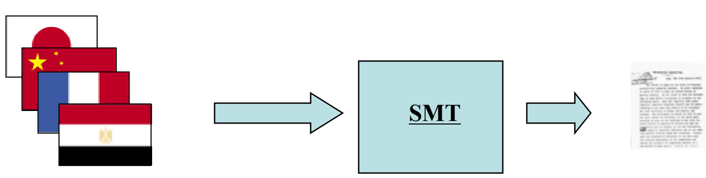
Topic Correlations Over Time



Birth-Death of Topics via Dynamic Dirichlet Process

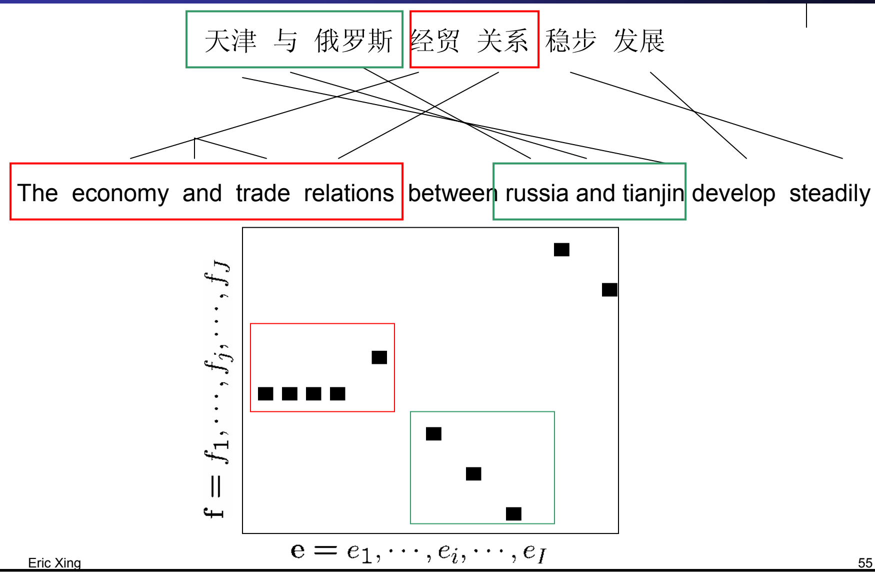


Application 2: Machine translation

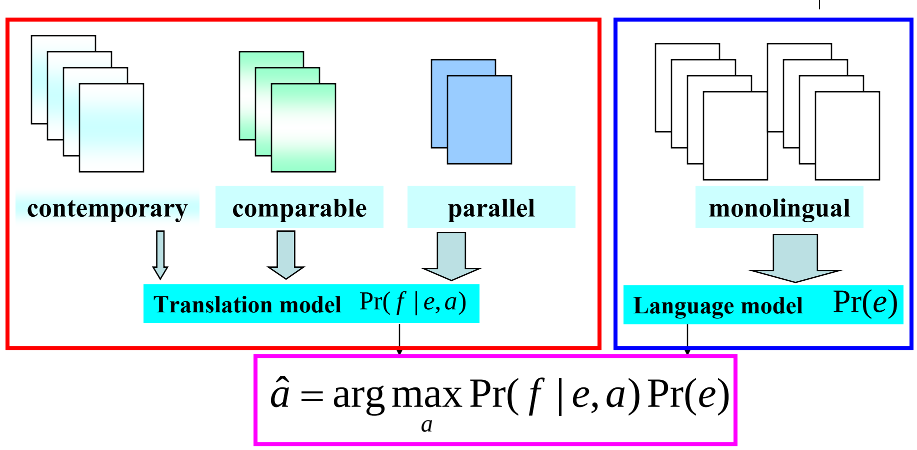


B. Zhao and E.P Xing, ACL 2006, NIPS 2007

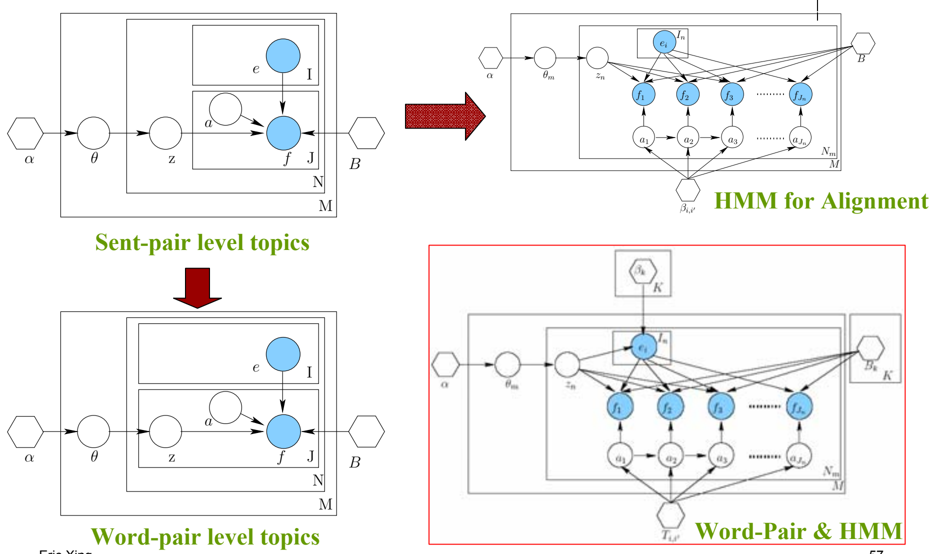
Word Alignment



The Statistical Formulation



An upgrade path for BiTAMs



Experiments

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

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

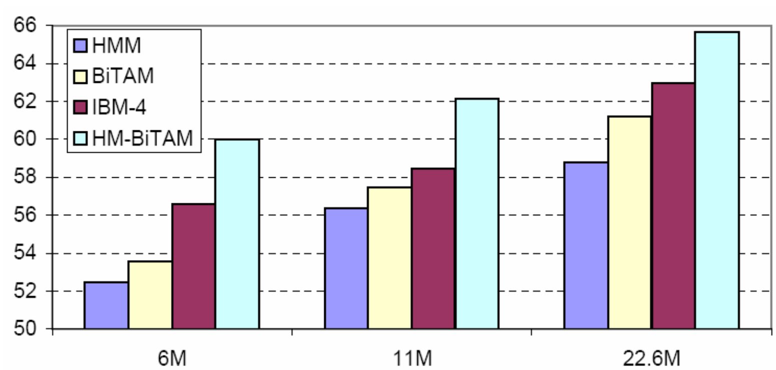
- Word Alignment Accuracy & Translation Quality
  - F-measure
  - BLEU

Topics

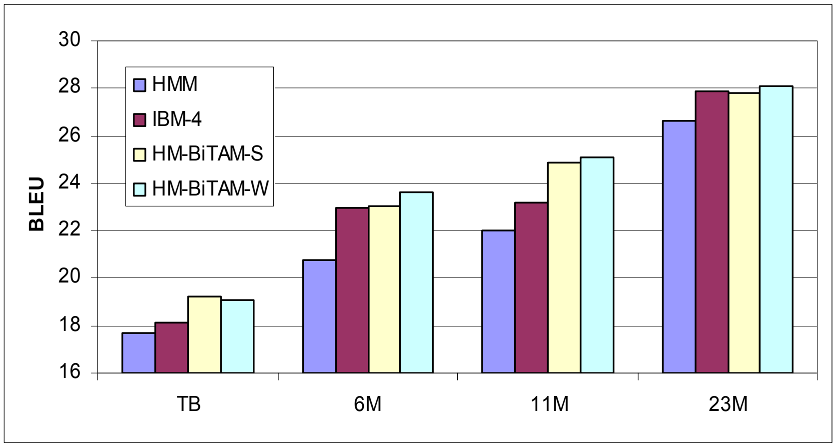
T1	Teams, sports, disabled, games members, people, cause, water, national, handicapped
T2	Shenzhen, singapore, hongkong, stock, national, investment, yuan, options, million, dollar
T3	Chongqing, company, takeover, shenzhen, tianjin, city, national, government, project, companies
T4	Hongkong, trade, export,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

T1	人, 残疾, 体育, 事业, 水, 世界, 区, 新华社, 队员, 记者
T2	深圳, 深, 新, 元, 有, 股, 香港, 国有, 外资, 新华社
T3	国家, 重庆, 市, 区, 厂, 天津, 政府, 项目, 国, 深圳
T4	香港, 贸易, 出口, 外资, 合作, 今年, 项目, 利用, 新, 技术
T5	住房, 房, 九江, 建设, 澳门, 元, 职工, 目前, 国家, 占, 省
T6	公司, 天然气, 两, 国, 美国, 记者, 关系, 俄, 法, 重庆

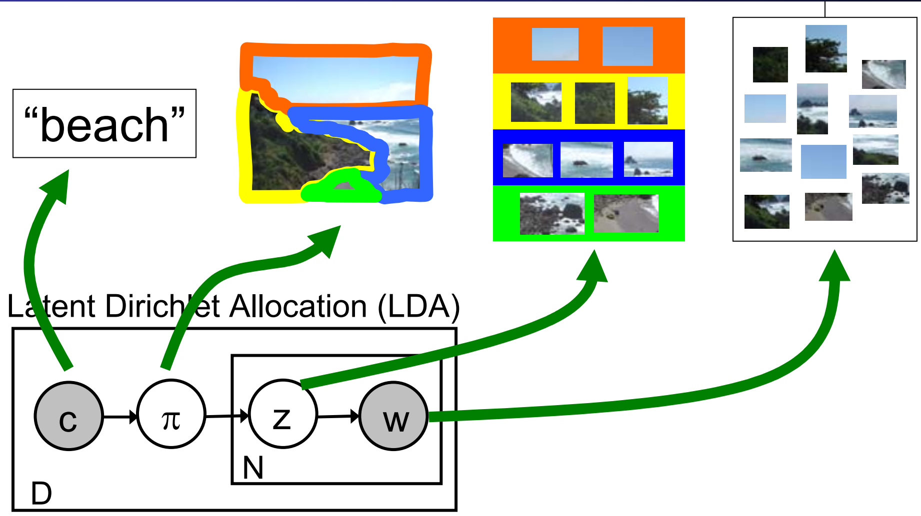
HM-BiTAM versus others



Translation Evaluations

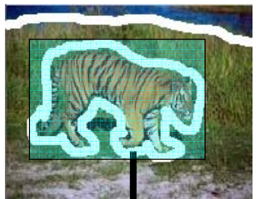


Application 3: Topic Models for Images



Fei-Fei et al. ICCV 2005

Image Representation



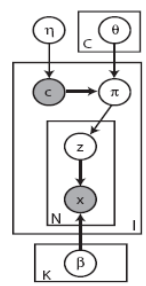
cat, grass, tiger, water

$$[r_{11} \cdots r_{1d}] \;,\; [w_1 \cdots w_{|V|}]$$

representation vector (real, 1 per image segment) : **annotation vector** (binary, same for each segment)

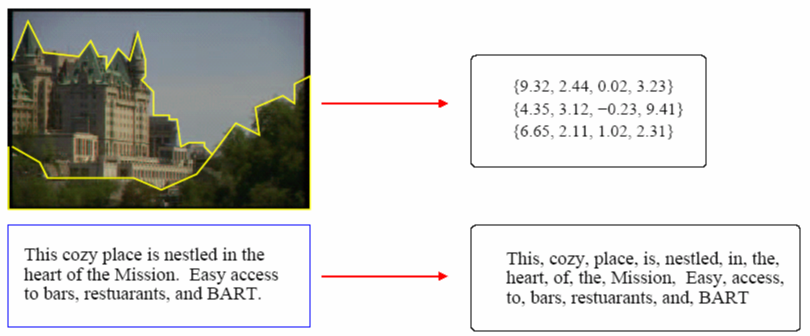
$$[r_{n1} \cdots r_{nd}] \;,\; [w_1 \cdots w_{|V|}]$$

To Generate an Image …



$$\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\pi}, c|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\beta}) &= p(c|\boldsymbol{\eta})p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) \cdot \\
&\quad \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi})p(x_n|z_n, \boldsymbol{\beta}) \\
p(c|\boldsymbol{\eta}) &= \mathrm{Mult}(c|\boldsymbol{\eta}) \\
p(\boldsymbol{\pi}|c, \boldsymbol{\theta}) &= \prod_{j=1}^{C} Dir(\boldsymbol{\pi}|\boldsymbol{\theta}_{j\cdot})^{\delta(c,j)} \\
p(z_n|\boldsymbol{\pi}) &= \mathrm{Mult}(z_n|\boldsymbol{\pi}) \\
p(x_n|z_n, \boldsymbol{\beta}) &= \prod_{k=1}^{K} p(x_n|\boldsymbol{\beta}_{k\cdot})^{\delta(z_n^k, 1)}
\end{aligned}$$

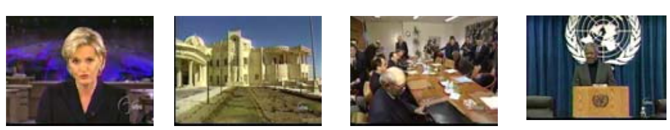
Annotated images



- Forsyth et. al. (2001): images as documents where region-specific feature vectors are like visual words.
- A captioned image can be thought of as annotated data: two documents, one of which describes the other.

Application 4: 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

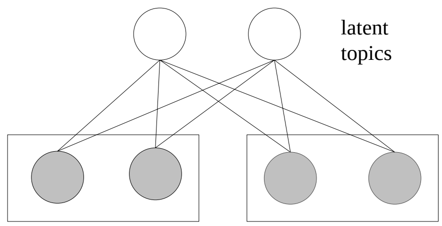


anchor building meeting speech

J. Yang, Y. Liu, E. P. Xing and A. Hauptmann, SDM 2007, **BEST PAPER Award**

Harmoniums for Multi-modal Data

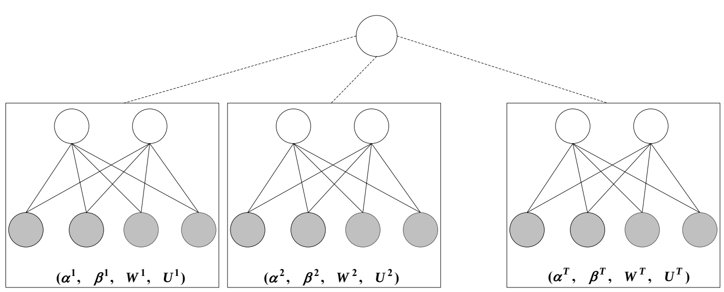
- Dual-wing harmoniums (DWH) [Xing et al. 05]
  - modeling bi-modal data: captioned images, video
  - learning hidden topics from two ``wings'' of observed features



Eric Xing 67

Mixture-of-Harmoniums (MoH)

- A family of category-specific dual-wing harmoniums



  - classification by finding the "best-fitting" harmonium

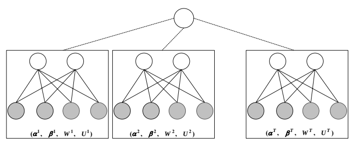
Eric Xing 68

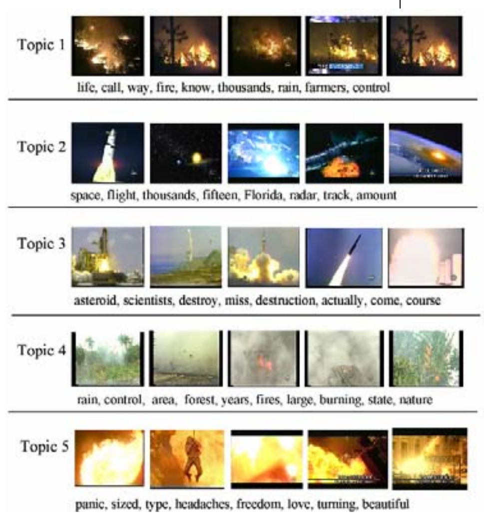
$H_1 \quad \cdots$

$\cdots \quad X_N$

Semantic Topics by FoH

- Revealing "sub-topics'' of each category
- Co-clusters of both text and image features





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
  - …
  - Word-sense disambiguation
  - Word-net
  - Network inference

$H_1 \cdots H_K$

$X_1 \cdots X_N \quad Z_1 \cdots Z_M$

category T