

# Topic Modeling: Beyond Bag-of-Words

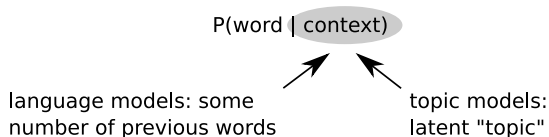
Hanna M. Wallach

University of Cambridge  
hmw26@cam.ac.uk

June 26, 2006

# Generative Probabilistic Models of Text

- ▶ Used in text compression, predictive text entry, information retrieval
- ▶ Estimate probability of a word in a given context:




- ▶ Here, both types of context are combined to improve performance
- ▶ This is done in a single Bayesian framework

# Statistical Language Models

- ▶ Estimate the probability of a word occurring in a given context
- ▶ Context is normally some number of preceding words

... Germany hosted the World ?

what should  
this word be?



- ▶ Used in text compression, predictive text entry, speech recognition
- ▶ There are many different models of this sort

# A Simple Bigram Language Model

- ▶ Given a corpus  $\mathbf{w}$  of  $N$  tokens, count

$N_w = \#$  of times word  $w$  appears in  $\mathbf{w}$

$N_{v|w} = \#$  of times word  $v$  follows word  $w$  in  $\mathbf{w}$

- ▶ Form the predictive distribution:

$$P(v | w, \mathbf{w}) = \lambda f_v + (1 - \lambda) f_{v|w}$$

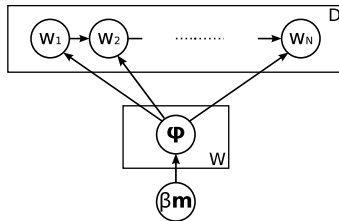
observed marginal  
frequency:  $f_v = N_v / N$

observed conditional  
frequency:  $f_{v|w} = N_{v|w} / N_w$

- ▶ Use, e.g., cross validation to estimate weight  $\lambda$

# Hierarchical Dirichlet Language Model (MacKay & Peto, '95)

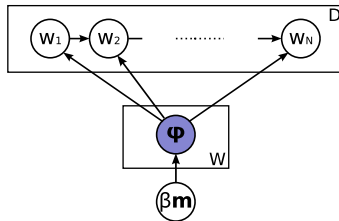
A bigram model based on principles of Bayesian inference:



# Hierarchical Dirichlet Language Model (MacKay & Peto, '95)

A bigram model based on principles of Bayesian inference:

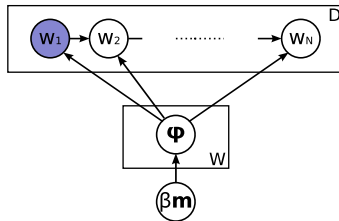
- ▶ For each word  $w$  in the vocabulary, draw a distribution over words  $\phi_w$  from  $\text{Dir}(\phi_w; \beta \mathbf{m})$



# Hierarchical Dirichlet Language Model (MacKay & Peto, '95)

A bigram model based on principles of Bayesian inference:

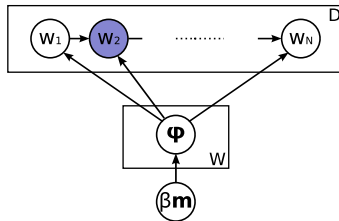
- ▶ For each word  $w$  in the vocabulary, draw a distribution over words  $\phi_w$  from  $\text{Dir}(\phi_w; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ , draw a word  $w_i$  from  $\phi_{w_{i-1}}$



# Hierarchical Dirichlet Language Model (MacKay & Peto, '95)

A bigram model based on principles of Bayesian inference:

- ▶ For each word  $w$  in the vocabulary, draw a distribution over words  $\phi_w$  from  $\text{Dir}(\phi_w; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ , draw a word  $w_i$  from  $\phi_{w_{i-1}}$

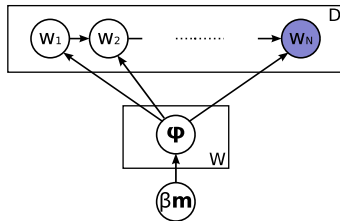




# Hierarchical Dirichlet Language Model (MacKay & Peto, '95)

A bigram model based on principles of Bayesian inference:


- ▶ For each word  $w$  in the vocabulary, draw a distribution over words  $\phi_w$  from  $\text{Dir}(\phi_w; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ , draw a word  $w_i$  from  $\phi_{w_{i-1}}$



# HDLM: Predictive Distribution

- ▶ Integrate out each  $\phi_w$
- ▶ Predictive probability of word  $v$  following word  $w$  is

$$P(v | w, \mathbf{w}) = \lambda_w m_v + (1 - \lambda_w) f_{v|w}$$

  $m_v$  has taken on the role of the marginal statistic  $f_v$  from the simple bigram language model

- ▶ Weight per context:  $\lambda_w = \frac{\beta}{N_w + \beta}$
- ▶ Bayesian version of the simple bigram language model

# Statistical Topic Models

- ▶ Documents are modeled as finite mixture of topics
- ▶ The topic mixture provides an explicit representation of a document

... Germany hosted the World

- ▶ Each word is generated by a single topic
- ▶ Used in information retrieval, classification, collaborative filtering

# Statistical Topic Models

- ▶ Documents are modeled as finite mixture of topics
- ▶ The topic mixture provides an explicit representation of a document

— countries  
— facial hair  
— other



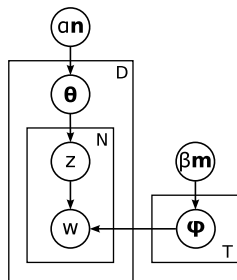
... Germany hosted the World Beard and Mustache Championships <sup>[1]</sup>

- ▶ Each word is generated by a single topic
- ▶ Used in information retrieval, classification, collaborative filtering

[1] <http://www.worldbeardchampionships.com>

# Latent Dirichlet Allocation (Blei et al., '03)

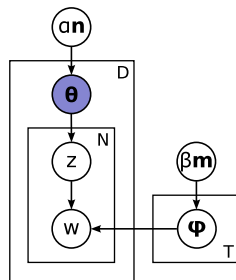
Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”



# Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”

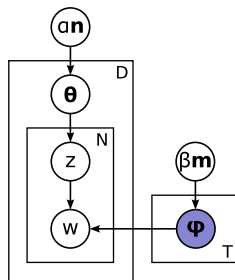
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$



# Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”

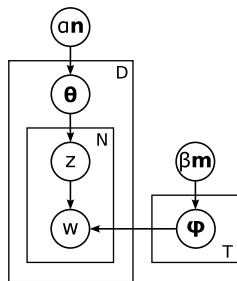
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each topic  $t$ , draw a distribution over words  $\phi_t$  from  $\text{Dir}(\phi_t; \beta \mathbf{m})$



# Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”

- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each topic  $t$ , draw a distribution over words  $\phi_t$  from  $\text{Dir}(\phi_t; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ :

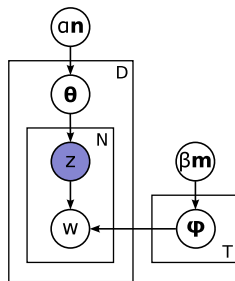




# Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”

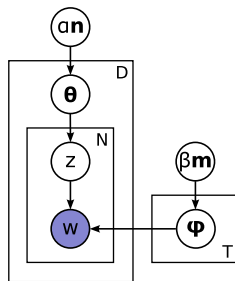
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each topic  $t$ , draw a distribution over words  $\phi_t$  from  $\text{Dir}(\phi_t; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$



# Latent Dirichlet Allocation (Blei et al., '03)

Models documents as mixtures of latent topics. Topics inferred from word correlations, independent of word order: “bag-of-words”

- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each topic  $t$ , draw a distribution over words  $\phi_t$  from  $\text{Dir}(\phi_t; \beta \mathbf{m})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{z_i}$



# Combining Word Order and Topic

- ▶ Each type of model has something to offer the other
- ▶ Context-based language models can be improved by topics:

 ... Germany hosted the World 

- ▶ Topic models can be improved by notion of word order:

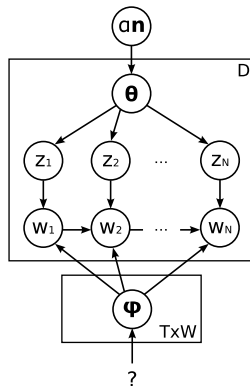
... the department chair ...

which topic?



# Bigram Topic Model

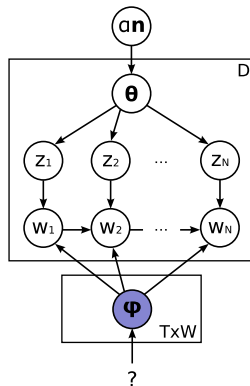
Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

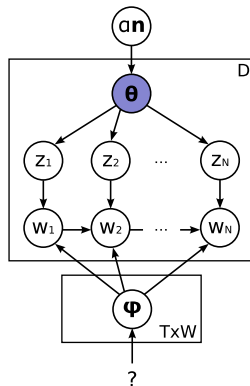
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

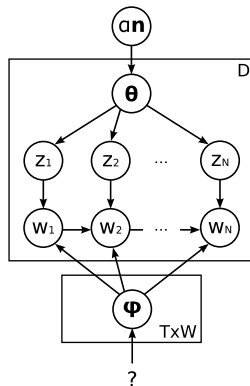
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

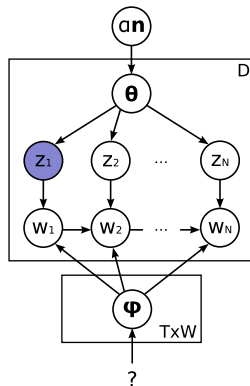
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$

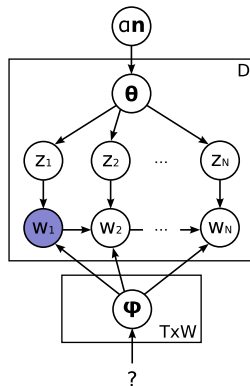




# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

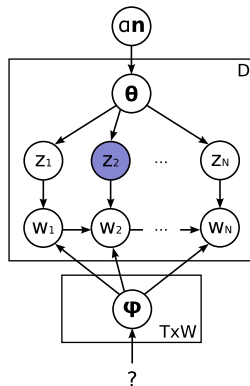
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{w_{i-1}, z_i}$



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

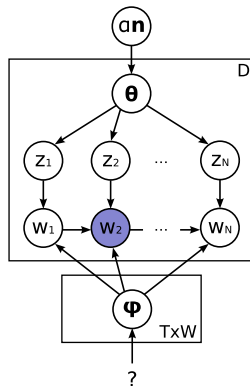
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{w_{i-1}, z_i}$



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

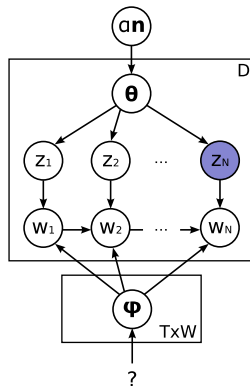
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{w_{i-1}, z_i}$



# Bigram Topic Model

Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

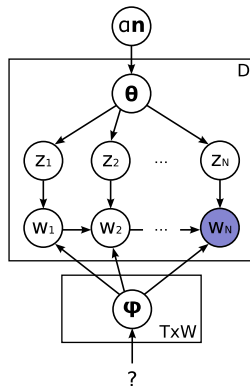
- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{w_{i-1}, z_i}$



# Bigram Topic Model

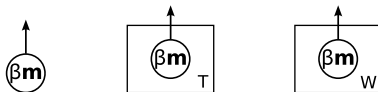
Combining ideas from HDLM and LDA gives a new topic model that moves beyond the bag-of-words assumption

- ▶ For each topic  $t$  and word  $w$ , draw a distribution over words  $\phi_{w,t}$  from a Dirichlet prior
- ▶ For each document  $d$ , draw a topic mixture  $\theta_d$  from  $\text{Dir}(\theta_d; \alpha \mathbf{n})$
- ▶ For each position  $i$  in document  $d$ :
  - ▶ Draw a topic  $z_i$  from  $\theta_d$
  - ▶ Draw a word  $w_i$  from  $\phi_{w_{i-1}, z_i}$



## Prior over $\{\phi_{w,t}\}$

- ▶ Prior over  $\{\phi_{w,t}\}$  must be “coupled” so that learning about one  $\phi_{w,t}$  gives information about others
- ▶ Coupling comes from hyperparameter sharing
- ▶ Several ways of doing this:
  - ▶ Single: Only one  $\beta\mathbf{m}$
  - ▶ Per topic:  $\beta_t\mathbf{m}_t$  for each topic  $t$
  - ▶ Per word:  $\beta_w\mathbf{m}_w$  for each possible previous word  $w$



# Inference of Hyperparameters

- ▶ Integrate over  $\phi_{w,t}$  and  $\theta_d$
- ▶ Let  $U = \{\alpha\mathbf{n}, \beta\mathbf{m}\}$  or  $U = \{\alpha\mathbf{n}, \{\beta_t\mathbf{m}_t\}\}$
- ▶ Assume uniform hyperpriors over all hyperparameters
- ▶ Find the maximum of the evidence

$$U^{\text{MP}} = \arg \max_U \sum_{\mathbf{z}} P(\mathbf{w}, \mathbf{z} | U)$$

using a Gibbs EM algorithm

# Comparing Predictive Accuracy

- ▶ Information rate of unseen test data  $\mathbf{w}^*$  in bits per word:

$$R = -\frac{\log_2 P(\mathbf{w}^*|\mathbf{w})}{N^*}$$

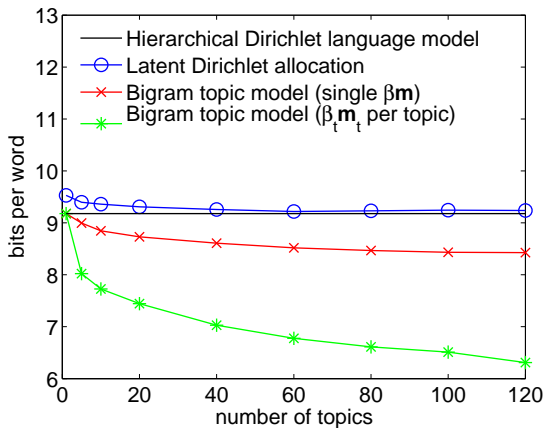
- ▶ Lower information rate = better predictive accuracy
- ▶ Direct measure of text compressibility
- ▶ Use Gibbs sampling to approximate  $P(\mathbf{w}^*|\mathbf{w})$



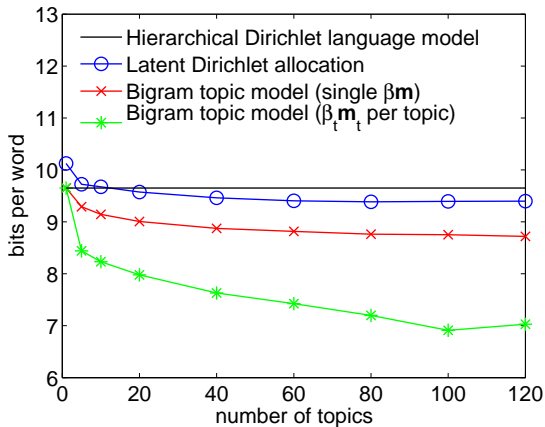
# Data Sets

- ▶ 150 abstracts from Psychological Review
  - ▶ Vocabulary size: 1,374 words
  - ▶ 13,414 tokens in training data, 6,521 in test data
- ▶ 150 postings from 20 Newsgroups data set
  - ▶ Vocabulary size: 2,281 words
  - ▶ 27,478 tokens in training data, 13,579 in test data

# Information Rate: Psychological Review



# Information Rate: 20 Newsgroups



# Inferred Topics: Latent Dirichlet Allocation

the <i>[number]</i> in to <b>espn</b> <b>hockey</b> a this as <b>run</b>	i is <b>satan</b> the which and of <b>metaphorical</b> <b>evil</b> there	that <b>proteins</b> the of to i if <i>[number]</i> you <b>fact</b>	<b>easter</b> <b>ishtar</b> a the have with but <b>english</b> and is
---	---	--	--

# Inferred Topics: Bigram Topic Model (Single $\beta\mathbf{m}$ )

to	the	the	the
<b>party</b>	<b>god</b>	and	a
<b>arab</b>	is	between	to
not	<b>belief</b>	<b>warrior</b>	i
<b>power</b>	<b>believe</b>	<b>enemy</b>	of
any	<b>use</b>	<b>battlefield</b>	<i>[number]</i>
i	there	a	is
is	<b>strong</b>	of	in
this	<b>make</b>	there	and
<b>things</b>	i	<b>way</b>	it

## Inferred topics: Bigram Topic Model ( $\beta_t \mathbf{m}_t$ per topic)

<b>party</b>	<b>god</b>	<i>[number]</i>	the
<b>arab</b>	<b>believe</b>	the	to
<b>power</b>	about	<b>tower</b>	a
as	<b>atheism</b>	<b>clock</b>	and
<b>arabs</b>	<b>gods</b>	a	of
<b>political</b>	before	<b>power</b>	i
are	see	<b>motherboard</b>	is
<b>rolling</b>	<b>atheist</b>	<b>mhz</b>	<i>[number]</i>
<b>london</b>	most	<b>socket</b>	it
<b>security</b>	<b>shafts</b>	<b>plastic</b>	that

# Findings and Future Work

## Findings:

- ▶ Combining latent topics and word order improves predictive accuracy
- ▶ The quality of inferred topics is improved

## Future work:

- ▶ Per word:  $\beta_w \mathbf{m}_w$  for each possible previous word  $w$
- ▶ Other model structures
- ▶ Evaluation on larger corpora

# Questions?

hmw26@cam.ac.uk

<http://www.inference.phy.cam.ac.uk/hmw26/>

Work conducted in part at the University of Pennsylvania, supported by NSF ITR grant EIA-0205456 and DARPA contract NBCHD030010.