Topic Modeling: Beyond Bag-of-Words

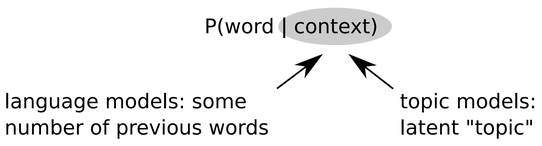
Hanna M. Wallach

University of Cambridge
hmw26@cam.ac.uk

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

Introduction

Generative probabilistic models of text are used in text compression, predictive text entry and information retrieval. These models estimate the probability of a word occurring in a given context. The type of context depends on the type of model:



Here, both types of context are combined to improve model performance. This is accomplished in a single Bayesian framework by combining ideas from two previous models:

Hierarchical Dirichlet Language Model (MacKay & Peto, '95)
Latent Dirichlet Allocation (Blei et al., '03)

Background

A Simple Bigram Language Model

Given a corpus **w**, count

N_w = # of times word w appears in **w**
$N_{v|w}$ = # of times word v follows word w in **w**
N = total # of tokens in **w**

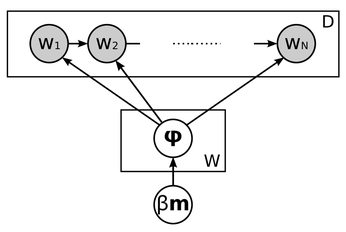
Estimate $f_v = N_v / N$ and $f_{v|w} = N_{v|w} / N_w$ and form the predictive distribution, using, e.g., cross validation to estimate λ:

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

There are many generalizations of this model.

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

A bigram model entirely based on principles of Bayesian inference.



1. For each word w in the vocabulary, draw a distribution over words $\boldsymbol{\varphi}_w$ from a Dirichlet distribution with parameters $\beta\mathbf{m}$
2. For each position i in document d, draw a word w_i from $\boldsymbol{\varphi}_{w_{i-1}}$

Integrate over each $\boldsymbol{\varphi}_w$ and and form the predictive distribution:

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

where

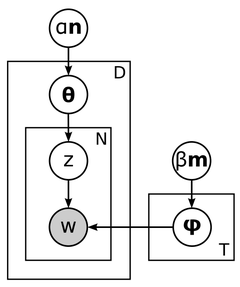
$\lambda_w = \beta / (N_w + \beta)$

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

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

A topic model that treats documents as finite mixtures over underlying latent topics. Topics are inferred from word correlations, independent of word order: "bag-of-words."

Computationally efficient
Not appropriate when word order matters



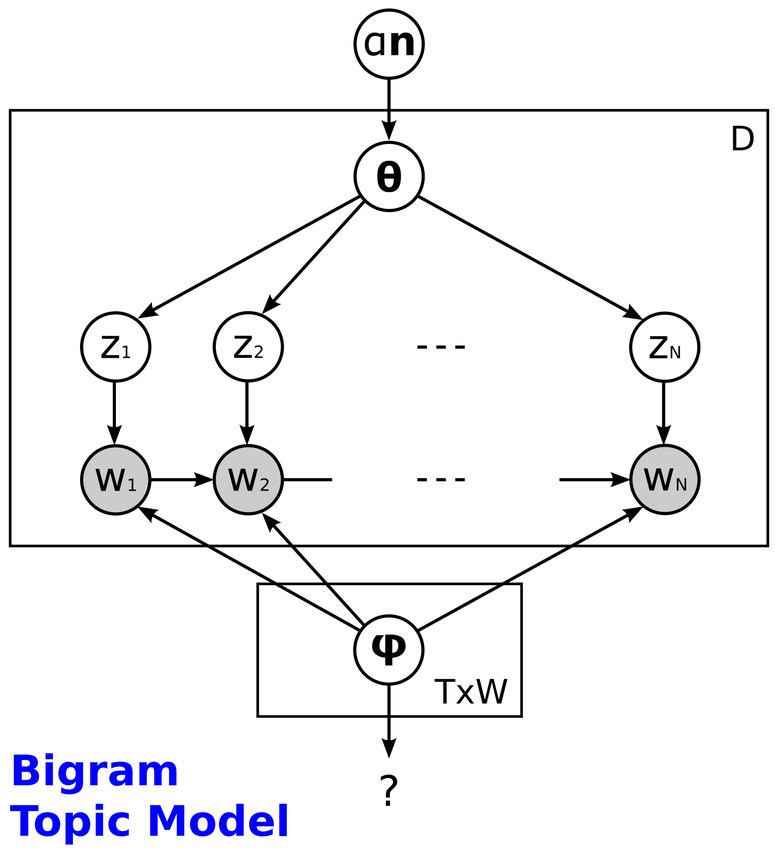
1. For each document d, draw a topic mixture $\boldsymbol{\theta}_d$ from a Dirichlet distribution with parameters $\alpha\mathbf{n}$
2. For each topic t, draw a distribution over words $\boldsymbol{\varphi}_t$ from a Dirichlet distribution with parameters $\beta\mathbf{m}$
3. For each position i in document d
   a) Draw a topic z_i from $\boldsymbol{\theta}_d$
   b) Draw a word w_i from $\boldsymbol{\varphi}_{z_i}$

Integrate over each $\boldsymbol{\varphi}_t$ and $\boldsymbol{\theta}_d$. Let $f_{v|t} = N_{v|t} / N_t$ and $f_{t|d} = N_{t|d} / N_d$. For a single assignment of topics **z** to **w**, the predictive distributions are:

$$P(v \mid t, \mathbf{w}, \mathbf{z}) = \lambda_t m_v + (1 - \lambda_t) f_{v|t}$$
$$P(t \mid d, \mathbf{z}) = \mu_d n_t + (1 - \mu_d) f_{t|d}$$

where

$\lambda_t = \beta / (N_t + \beta)$
$\mu_d = \alpha / (N_d + \alpha)$



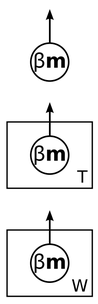
Bigram Topic Model

Combining ideas from the HDLM and LDA leads to a new topic model that moves beyond the bag-of-words assumption. The predictive context is now the immediately preceding word and the current topic.

1. For each topic t and word w, draw a distribution over words $\boldsymbol{\varphi}_{w,t}$ from a Dirichlet distribution
2. For each document d, draw a topic mixture $\boldsymbol{\theta}_d$ from a Dirichlet distribution with parameters $\alpha\mathbf{n}$
3. For each position i in document d
   a) Draw a topic z_i from $\boldsymbol{\theta}_d$
   b) Draw a word w_i from $\boldsymbol{\varphi}_{w_{i-1},z_i}$

Prior over $\{\boldsymbol{\varphi}_{w,t}\}$

The prior over $\{\boldsymbol{\varphi}_{w,t}\}$ is coupled, so that learning one $\boldsymbol{\varphi}_{w,t}$ gives information about others. Coupling comes from hyperparameter sharing.



Single: Only one $\beta\mathbf{m}$. Learning about one $\boldsymbol{\varphi}_{w,t}$ will give information about $\boldsymbol{\varphi}_{w',t'}$ for all other w', t' contexts

Per topic: $\beta_t\mathbf{m}_t$ for each topic t. Learning about one $\boldsymbol{\varphi}_{w,t}$ will give information about $\boldsymbol{\varphi}_{w',t'}$ for all other w', t'=t contexts that share this topic.

Per word: $\beta_w\mathbf{m}_w$ for each possible previous word w. Learning about one $\boldsymbol{\varphi}_{w,t}$ will give information about $\boldsymbol{\varphi}_{w',t'}$ for all other w'=w, t' contexts.

Predictive Distributions

Integrate over each $\boldsymbol{\varphi}_{w,t}$ and $\boldsymbol{\theta}_d$. Let $f_{v|w,t} = N_{v|w,t} / N_{w,t}$ and $f_{t|d} = N_{t|d} / N_d$. For a single set of topics **z**, the predictive distribution over words given previous word w and current topic t is (single $\beta\mathbf{m}$)

$$P(v \mid w, t, \mathbf{w}, \mathbf{z}) = \lambda_{w,t} m_v + (1 - \lambda_{w,t}) f_{v|w,t}$$

where

$\lambda_{w,t} = \beta / (N_{w,t} + \beta)$

or ($\beta_t\mathbf{m}_t$ per topic)

$$P(v \mid w, t, \mathbf{w}, \mathbf{z}) = \lambda_{w,t} m_{v|t} + (1 - \lambda_{w,t}) f_{v|w,t}$$

where

$\lambda_{w,t} = \beta_t / (N_{w,t} + \beta_t)$

The predictive distribution over topics is the same as in LDA.

Inference of Hyperparameters

Integrate over each $\boldsymbol{\varphi}_{w,t}$ and $\boldsymbol{\theta}_d$. 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

$$U_{MP} = \arg\max P(\mathbf{w} \mid U) = \arg\max \Sigma_z P(\mathbf{w}, \mathbf{z} \mid U)$$

using a Gibbs EM algorithm. The same algorithm may be used for LDA.

Experiments

Hyperparameters for the HDLM may be inferred using the same optimization algorithm as that used in the M-step of Gibbs EM.

Compare predictive accuracy using information rate of unseen data $\mathbf{w}^*$, measured in bits per word:

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

Lower information rate = better predictive accuracy. Information rate is a direct measure of text compressibility.

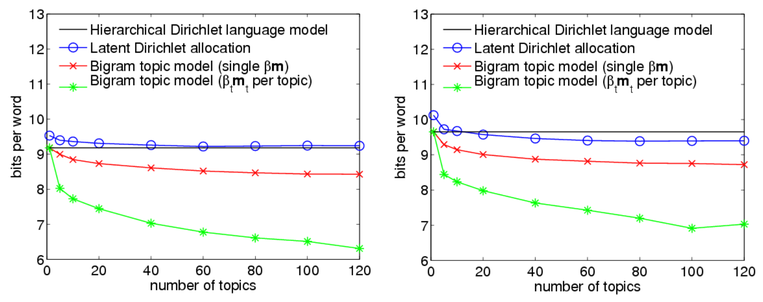
Use Gibbs sampling to draw a single set of topics z for w and multiple sets of topics $\{\mathbf{z}^*_s\}$ for $\mathbf{w}^*$. Then approximate $P(\mathbf{w}^* \mid \mathbf{w})$ by taking the harmonic mean of $\{P(\mathbf{w}^* \mid \mathbf{z}^*_s, \mathbf{w}, \mathbf{z}, U_{MP})\}$.

Corpora

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

For each data set, 100 documents were used to infer the hyperparameters and 50 were used to compare the predictive accuracy.

Information Rate



Left: Psychological Review Abstracts data. Right: 20 Newsgroups data.

Inferred Topics

Content words are in blue. Function words, which are black, were identified by their presence on a standard list of stop words. All three sets of topics were taken from models with 90 topics.

Latent Dirichlet Allocation:

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

Bigram topic model (single $\beta\mathbf{m}$):

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

Bigram topic model ($\beta_t\mathbf{m}_t$ per topic):

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

Findings

1. Combining latent topics and word order improves predictive accuracy.
2. The quality of inferred topics is improved.