A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes

Robert V. Lindsey
University of Colorado, Boulder
robert.lindsey@colorado.edu

William P. Headden III
Two Cassowaries Inc.
headdenw@twocassowaries.com

Michael J. Stipicevic
Google Inc.
stip@google.com

Abstract

Topic models traditionally rely on the bag-of-words assumption. In data mining applications, this often results in end-users being presented with inscrutable lists of *topical unigrams*, single words inferred as representative of their topics. In this article, we present a hierarchical generative probabilistic model of *topical phrases*. The model simultaneously infers the location, length, and topic of phrases within a corpus and relaxes the bag-of-words assumption within phrases by using a hierarchy of Pitman-Yor processes. We use Markov chain Monte Carlo techniques for approximate inference in the model and perform slice sampling to learn its hyperparameters. We show via an experiment on human subjects that our model finds substantially better, more interpretable topical phrases than do competing models.

1 Introduction

Probabilistic topic models have been the focus of intense study in recent years. The archetypal topic model, Latent Dirichlet Allocation (LDA), posits that words within a document are conditionally independent given their topic (Blei et al., 2003). This "bag-of-words" assumption is a common simplification in which word order is ignored, but one which introduces undesirable properties into a model meant to serve as an unsupervised exploratory tool for data analysis.

When an end-user runs a topic model, the output he or she is often interested in is a list of topical unigrams, words probable in a topic (hence, representative of it). In many situations, such as during the use of the topic model for the analysis of a new or ill-understood corpus, these lists can be insufficiently informative. For instance, if a layperson ran LDA on the NIPS corpus, he would likely get a topic whose most prominent words include *policy*, *value*, and *reward*. Seeing these words isolated from their context in a list would not be particularly insightful to the layperson unfamiliar with computer science research. An alternative to LDA which produced richer output like *policy iteration algorithm*, *value function*, and *model-based reinforcement learning* alongside the unigrams would be much more enlightening. Most situations where a topic model is actually useful for data exploration require a model whose output is rich enough to dispel the need for the user's extensive prior knowledge of the data.

Furthermore, lists of topical unigrams are often made only marginally interpretable by virtue of their non-compositionality, the principle that a collocation's meaning typically is not derivable from its constituent words (Schone and Jurafsky, 2001). For example, the meaning of *compact disc* as a music medium comes from neither the unigram *compact* nor the unigram *disc*, but emerges from the bigram as a whole. Moreover, non-compositionality is topic dependent; *compact disc* should be interpreted as a music medium in a music topic, and as a small region bounded by a circle in a mathematical topic. LDA is prone to decompose collocations into different topics and violate the principle of non-compositionality, and its unigram lists are harder to interpret as a result.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 214–222, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

We present an extension of LDA called Phrase-Discovering LDA (PDLDA) that satisfies two desiderata: providing rich, interpretable output and honoring the non-compositionality of collocations. PDLDA is built in the tradition of the "Topical N-Gram" (TNG) model of Wang et al. (2007). TNG is a topic model which satisfies the first desideratum by producing lists of representative, topically cohesive n-grams of the form shown in Figure 1. We diverge from TNG by our addressing the second desideratum, and we do so through a more straightforward and intuitive definition of what constitutes a phrase and its topic. In the furtherance of our goals, we employ a hierarchical method of modeling phrases that uses dependent Pitman-Yor processes to ameliorate overfitting. Pitman-Yor processes have been successfully used in the past in n-gram (Teh, 2006) and LDA-based models (Wallach, 2006) for creating Bayesian language models which exploit word order, and they prove equally useful in this scenario of exploiting both word order and topics.

This article is organized as follows: after describing TNG, we discuss PDLDA and how PDLDA addresses the limitations of TNG. We then provide details of our inference procedures and evaluate our model against competing models on a subset of the TREC AP corpus (Harman, 1992) in an experiment on human subjects which assesses the interpretability of topical n-gram lists. The experiment is premised on the notion that topic models should be evaluated through a real-world task instead of through information-theoretic measures which often negatively correlate with topic quality (Chang et al., 2009).

2 Background: LDA and TNG

LDA represents documents as probabilistic mixtures of latent topics. Each word w in a corpus $\mathbf{w}$ is drawn from a distribution ϕ indexed by a topic z, where z is drawn from a distribution θ indexed by its document d. The formal definition of LDA is

$$\begin{array}{ll} \theta_d \sim \text{Dirichlet}\,(\alpha) & z_i \mid d, \theta \sim \text{Discrete}\,(\theta_d) \\ \phi_z \sim \text{Dirichlet}\,(\beta) & w_i \mid z_i, \phi \sim \text{Discrete}\,(\phi_{z_i}) \end{array}$$

where θ_d is document d's topic distribution, ϕ_z is topic z's distribution over words, z_i is the topic assignment of the ith token, and w_i is the ith word. α and β are hyperparameters to the Dirichlet priors. Here and throughout the article, we use a bold font for vector notation: for example, $\mathbf{z}$ is the vector of all topic assignments, and its ith entry, z_i, corresponds to the topic assignment of the ith token in the corpus.

TNG extends LDA to model n-grams of arbitrary length in order to create the kind of rich output for text mining discussed in the introduction. It does this by representing a joint distribution $P(\mathbf{z}, \mathbf{c}|\mathbf{w})$ where each c_i is a Boolean variable that signals the start of a new n-gram beginning at the ith token. $\mathbf{c}$ partitions a corpus into consecutive non-overlapping n-grams of various lengths. Formally, TNG differs from LDA by the distributional assumptions

$$\begin{array}{rcl} w_i \mid w_{i-1}, z_i, c_i = 1, \phi & \sim & \text{Discrete}(\phi_{z_i}) \\ w_i \mid w_{i-1}, z_i, c_i = 0, \sigma & \sim & \text{Discrete}(\sigma_{z_i w_{i-1}}) \\ c_i \mid w_{i-1}, z_{i-1}, \pi & \sim & \text{Bernoulli}(\pi_{z_{i-1} w_{i-1}}) \end{array}$$

where the new distributions π_{zw} and σ_{zw} are endowed with conjugate prior distributions: $\pi_{zw} \sim \text{Beta}(\lambda)$ and $\sigma_{zw} \sim \text{Dirichlet}(\delta)$. When $c_i = 0$, word w_i is joined into a topic-specific bigram with w_{i-1}. When $c_i = 1$, w_i is drawn from a topic-specific unigram distribution and is the start of a new n-gram.

An unusual feature of TNG is that words within a *topical n-gram*, a sequence of words delineated by $\mathbf{c}$, do not share the same topic. To compensate for this after running a Gibbs sampler, Wang et al. (2007) analyze each topical n-gram post hoc as if the topic of the final word in the n-gram was the topic assignment of the entire n-gram. Though this design simplifies inference, we perceive it as a shortcoming since the aforementioned principle of non-compositionality supports the intuitive idea that each collocation ought to be drawn from a single topic. Another potential drawback of TNG is that the topic-specific bigram distributions σ_{zw} share no probability mass between each other or with the unigram distributions ϕ_z. Hence, observing a bigram under one topic does not make it more likely under another topic or make its constituent unigrams more probable. To be more concrete, in TNG, observing *space shuttle* under a topic z (or under two topics, one for each word) regrettably does not make *space shuttle* more likely under a topic $z' \neq z$, nor does it make observing *shuttle* more likely under any topic. Smoothing, the sharing of probability mass between

matter	chemical reactions	like charges repel
atoms	atomic number	positively charged nucleus
elements	hydrogen atoms	unlike charges attract
electrons	hydrogen atom	outer energy level
atom	periodic table	reaction takes place
molecules	chemical change	negatively charged electrons
form	physical properties	chemical change takes place
oxygen	chemical reaction	form new substances
hydrogen	water molecules	physical change takes place
particles	sodium chloride	form sodium chloride
element	small amounts	modern atomic theory
solution	positive charge	electrically charged particles
substance	carbon atoms	increasing atomic number
reaction	physical change	second ionization energies
nucleus	chemical properties	higher energy levels

(a) Topic 1

president	supreme court	civil rights act
congress	new york	civil rights movement
vote	democratic party	supreme court ruled
party	vice president	president theodore roosevel
constitution	political parties	second continental congress
state	national government	equal rights amendment
members	executive branch	strong central government
office	civil rights	sherman antitrust act
government	new government	civil rights legislation
states	political party	public opinion polls
elected	andrew jackson	major political parties
representatives	chief justice	congress shall make
senate	federal government	federal district court
house	state legislatures	supreme court decisions
washington	public opinion	american foreign policy

(b) Topic 2

words	main idea	word processing center
word	topic sentence	word processing systems
sentence	english language	word processing equipment
write	following paragraph	speak different languages
writing	words like	use quotation marks
paragraph	quotation marks	single main idea
sentences	direct object	use words like
meaning	word processing	topic sentence states
use	sentence tells	present perfect tense
subject	figurative language	express complete thoughts
language	writing process	word processing software
read	following sentences	use formal english
example	subject matter	standard american english
verb	standard english	collective noun refers
topic	use words	formal standard english

(c) Topic 3

energy	natural resources	nuclear power plants
used	natural gas	nuclear power plant
oil	heat energy	important natural resources
heat	iron ore	electric power plants
coal	carbon dioxide	called fossil fuels
use	potential energy	important natural resource
fuel	solar energy	produce large amounts
produce	light energy	called solar energy
power	fossil fuels	electric light bulb
source	hot water	use electrical energy
light	steam engine	use solar energy
electricity	large amounts	carbon dioxide gas
burn	sun's energy	called potential energy
gas	radiant energy	gas called carbon dioxide
gasoline	nuclear energy	called crude oil

(d) Topic 4

water	water vapor	water vapor condenses
air	air pollution	warm air rises
temperature	air pressure	cold air mass
heat	warm air	called water vapor
liquid	cold water	water vapor changes
gas	earth's surface	process takes place
gases	room temperature	warm air mass
hot	boiling point	clean air act
pressure	drinking water	gas called water vapor
atmosphere	atmospheric pressure	dry spell holds
warm	cold war	air pressure inside
cold	high temperatures	sewage treatment plant
surface	liquid water	air pollution laws
oxygen	cold air	high melting points
clouds	warm water	high melting point

(e) Topic 5

china	middle east	2000 years ago
africa	western europe	east india company
india	north africa	eastern united states
europe	mediterranean sea	4000 years ago
people	years ago	southwestern united states
chinese	roman empire	middle atlantic states
asia	far east	northeastern united states
egypt	southeast asia	western united states
world	west africa	southeastern united states
rome	saudi arabia	200 years ago
land	capital letter	middle atlantic region
east	asia minor	indus river valley
trade	united states	western roman empire
countries	capital city	british north america act
empire	centuries ago	coast guard station

(f) Topic 6

Figure 1: Six out of one hundred topics found by our model, PDLDA, on the Touchstone Applied Science Associates (TASA) corpus (Landauer and Dumais, 1997). Each column within a box shows the top fifteen phrases for a topic and is restricted to phrases of a minimum length of one, two, or three words, respectively. The rows are ordered by likelihood.

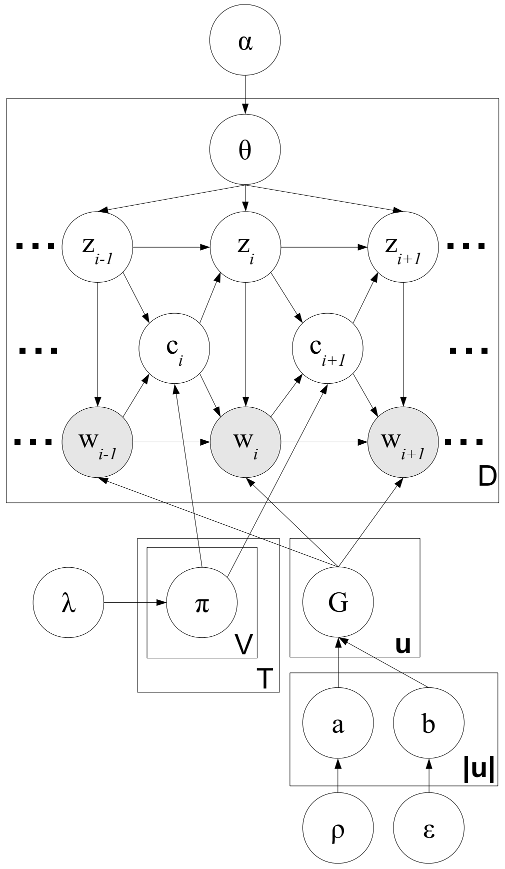


Figure 2: PDLDA drawn in plate notation.

contexts, is desirable so that a model like this does not need to independently infer the probability of every bigram under every topic. The advantages of smoothing are especially pronounced for small corpora or for a large number of topics. In these situations, the observed number of bigrams in a given topic will necessarily be very small and thus not support strong inferences.

3 PDLDA

A more natural definition of a topical phrase, one which meets our second desideratum, is to have each phrase possess a single topic. We adopt this intuitive idea in PDLDA. It can also be understood through the lens of Bayesian changepoint detection. Changepoint detection is used in time series models in which the generative parameters periodically change abruptly (Adams and MacKay, 2007). Viewing a sentence as a time series of words, we posit that the generative parameter, the topic, changes periodically in accordance with the changepoint indicators $\mathbf{c}$. Because there is no restriction on the number of words between changepoints, topical phrases can be arbitrarily long but will always have a single topic drawn from θ_d.

The full definition of PDLDA is given by

$$\begin{aligned}
w_i \mid \mathbf{u} &\sim \text{Discrete}(G_{\mathbf{u}}) \\
G_{\mathbf{u}} &\sim \text{PYP}(a_{|\mathbf{u}|}, b_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \\
G_{\emptyset} &\sim \text{PYP}(a_0, b_0, H) \\
z_i \mid d, z_{i-1}, \theta_d, c_i &\sim \begin{cases} \delta_{z_{i-1}} & \text{if } c_i = 0 \\ \text{Discrete}\left(\theta_d\right) & \text{if } c_i = 1 \end{cases} \\
c_i \mid w_{i-1}, z_{i-1}, \pi &\sim \text{Bernoulli}\left(\pi_{w_{i-1} z_{i-1}}\right)
\end{aligned}$$

with the prior distriutions over the parameters as

$$\begin{aligned}
\theta_d \sim \text{Dirichlet}\left(\alpha\right) \qquad \pi_{zw} \sim \text{Beta}\left(\lambda\right) \\
a_{|\mathbf{u}|} \sim \text{Beta}\left(\rho\right) \qquad b_{|\mathbf{u}|} \sim \text{Gamma}\left(\epsilon\right)
\end{aligned}$$

Like TNG, PDLDA assumes that the probability of a changepoint c_{i+1} after the ith token depends on the current topic z_i and word w_i. This causes the length of a phrase to depend on its topic and constituent words. The changepoints explicitly model which words tend to start and end phrases in each document. Depending on c_i, z_i is either set deterministically to the preceding topic (when $c_i = 0$) or is drawn anew from θ_d (when $c_i = 1$). In this way, each topical phrase has a single topic drawn from its document's topic distribution. As in TNG, the parameters π_{zw} and θ_d are given conjugate priors parameterized by λ and α.

Let $\mathbf{u}$ be a *context vector* consisting of the phrase topic and the past m words: $\mathbf{u} \triangleq < z_i, w_{i-1}, w_{i-2}, \ldots, w_{i-m} >$. The operator $\pi(\mathbf{u})$ denotes the prefix of $\mathbf{u}$, the vector with the rightmost element of $\mathbf{u}$ removed. $|\mathbf{u}|$ denotes the length of $\mathbf{u}$, and $\emptyset$ represents an empty context. For practical reasons, we pad $\mathbf{u}$ with a special start symbol when the context overlaps a phrase boundary. For example, the first word w_i of a phrase beginning at a position i necessarily has $c_i = 1$; consequently, all the preceding words w_{i-j} in the context vector are treated as start symbols so that w_i is effectively drawn from a topic-specific unigram distribution.

In PDLDA, each token is drawn from a distribution conditioned on its context $\mathbf{u}$. When $m = 1$, this conditioning is analogous to TNG's word distribution. However, in contrast with TNG, the word

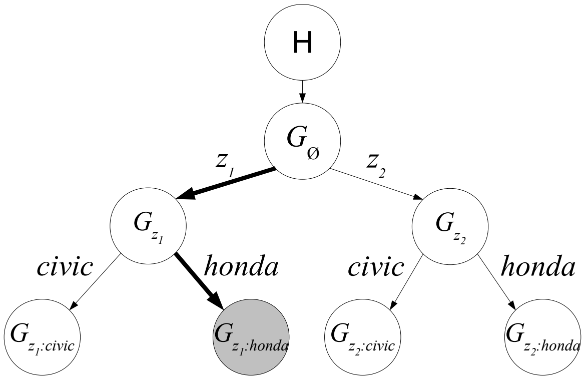


Figure 3: Illustration of the hierarchical Pitman-Yor process for a toy two-word vocabulary $V = \{$*honda*, *civic*$\}$ and two-topic ($T = 2$) model with $m = 1$. Each node G in the tree is a Pitman-Yor process whose base distribution is its parent node, and H is a uniform distribution over V. When, for example, the context is $\mathbf{u} = z_1$: *honda*, the darkened path is followed and the probability of the next word is calculated from the shaded node using Equation 1, which combines predictions from all the nodes along the darkened path.

distributions used are Pitman-Yor processes (PYPs) linked together into a tree structure. This hierarchical construction creates the desired smoothing among different contexts. The next section explains this hierarchical distribution in more detail.

3.1 Hierarchical Pitman-Yor process

Words in PDLDA are emitted from $G_{\mathbf{u}}$, which has a PYP prior (Pitman and Yor, 1997). PYPs are a generalization of the Dirichlet Process, with the addition of a discount parameter $0 \leq a \leq 1$. When considering the distribution of a sequence of words $\mathbf{w}$ drawn *iid* from a PYP-distributed G, one can analytically marginalize G and consider the resulting conditional distribution of $\mathbf{w}$ given its parameters a, b, and base distribution ϕ. This marginal can best be understood by considering the distribution of any $w_i | w_1, \ldots, w_{i-1}, a, b, \phi$, which is characterized by a generative process known as the generalized *Chinese Restaurant Process* (CRP) (Pitman, 2002). In the CRP metaphor, one imagines a restaurant with an unbounded number of tables, where each table has one shared dish (a draw from ϕ) and can seat an unlimited number of customers. The CRP specifies a process by which customers entering the restaurant choose a table to sit at and, consequently, the dish they eat. The first customer to arrive always sits at the first table. Subsequent customers sit at an occupied table k with probability proportional to $c_k - a$ and choose a new unoccupied table with probability proportional to $b + ta$, where c_k is the number of customers seated at table k and t is the number of occupied tables in G. For our language modeling purposes, "customers" are word tokens and "dishes" are word types.

The hierarchical PYP (HPYP) is an intuitive recursive formulation of the PYP in which the base distribution ϕ is itself PYP-distributed. Figure 3 demonstrates this principle as applied to PDLDA. The hierarchy forms a tree structure, where leaves are restaurants corresponding to full contexts and internal nodes correspond to partial contexts. An edge between a parent and child node represents a dependency of the child on the parent, where the base distribution of the child node is its parent. This smooths each context's distribution like the Bayesian n-gram model of Teh (2006), which is a Bayesian version of interpolated Kneser-Ney smoothing (Chen and Goodman, 1998). One ramification of this setup is that if a word occurs in a context $\mathbf{u}$, the sharing makes it more likely in other contexts that have something in common with $\mathbf{u}$, such as a shared topic or word.

The HPYP gives the following probability for a word following the context $\mathbf{u}$ being w:

$$P_{\mathbf{u}}(w \mid \tau, \mathbf{a}, \mathbf{b}) = \frac{c_{\mathbf{u}w\cdot} - a_{|\mathbf{u}|} t_{\mathbf{u}w}}{b_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} + \frac{b_{|\mathbf{u}|} + a_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{b_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} P_{\pi(\mathbf{u})}(w \mid \tau, \mathbf{a}, \mathbf{b}) \qquad (1)$$

where $P_{\pi(\emptyset)}(w | \tau, \mathbf{a}, \mathbf{b}) = G_\emptyset(w)$, $c_{\mathbf{u}w\cdot}$ is the number of customers eating dish w in restaurant $\mathbf{u}$, and $t_{\mathbf{u}w}$ is the number of tables serving w in restaurant $\mathbf{u}$, and τ represents the current seating arrangement. Here and throughout the rest of the paper, we use a dot to indicate marginal counts: e.g., $c_{\mathbf{u}w\cdot} = \sum_k c_{\mathbf{u}wk}$ where $c_{\mathbf{u}wk}$ is the number of customers eating w in $\mathbf{u}$ at table k. The base distribution of $G_\emptyset$ was chosen to be uniform: $H(w) = 1/V$ with V being the vocabulary size. The above equation an interpolation between distributions of context lengths

$|\mathbf{u}|, |\mathbf{u}|-1, \ldots 0$ and realizes the sharing of statistical strength between different contexts.

3.2 Inference

In this section, we describe Markov chain Monte Carlo procedures to sample from $P(\mathbf{z}, \mathbf{c}, \tau | \mathbf{w}, U)$, the posterior distribution over topic assignments $\mathbf{z}$, phrase boundaries $\mathbf{c}$, and seating arrangements τ given an observed corpus $\mathbf{w}$. Let U be shorthand for $\alpha, \lambda, \mathbf{a}, \mathbf{b}$. In order to draw samples from $P(\mathbf{z}, \mathbf{c}, \tau | \mathbf{w}, U)$, we employ a Metropolis-Hastings sampler for approximate inference. The sampler we use is a *collapsed* sampler (Griffiths and Steyvers, 2004), wherein θ, ϕ, and $\mathbf{G}$ are analytically marginalized. Because we marginalize each G, we use the Chinese Restaurant Franchise representation of the hierarchical PYPs (Teh, 2006). However, rather than onerously storing the table assignment of every token in $\mathbf{w}$, we store only the counts of how many tables there are in a restaurant and how many customers are sitting at each table in that restaurant. We refer the inquisitive reader to the appendix of Teh (2006) for further details of this procedure.

Our sampling strategy for a given token i in document d is to jointly propose changes to the changepoint c_i and topic assignment z_i, and then to the seating arrangement τ. Recall that according to the model, if $c_i = 0$, $z_i = z_{i-1}$; otherwise z_i is generated from the topic distribution for document d. Since the topic assignment remains the same until a new changepoint at a position i' is reached, each token w_j for j from position i until $i'-1$ will depend on z_i because for these j, $z_j = z_i$. We call this set of tokens the *phrase suffix* of the ith token and denote it $s(i)$. More formally, let $s(i)$ be the maximal set of continuous indices $j \geq i$ including i such that, if $j \neq i$, $c_j = 0$. That is, $s(i)$ are the indices comprising the remainder of the phrase beginning at position i. In addition, let $x(i)$ indicate the *extended suffix* version of $s(i)$ which includes one additional index: $x(i) \triangleq \{s(i) \cup \{\max(s(i)) + 1\}\}$. In addition to the words in the suffix $s(i)$, the changepoint indicator variables c_j for j in $x(i)$ are also conditioned on z_i. To make these dependencies more explicit, we refer to $\mathbf{z}_{s(i)} \triangleq z_j \; \forall j \in s(i)$, which are constrained by the model to share a topic.

The variables that depend directly on z_i, c_i are $\mathbf{z}_{s(i)}, \mathbf{w}_{s(i)}, \mathbf{c}_{x(i)}$. The proposal distribution first draws from a multinomial over $T+1$ options: one option for $c_i = 0, z_i = z_i - 1$; and one for $c_i = 1$ paired with each possible $z_i = z \in 1 \ldots T$. This is given by

$$P(\mathbf{z}_{s(i)}, c_i \mid \mathbf{z}_{\neg s(i)}, \mathbf{c}_{\neg i}, \tau_{\neg s(i)}, \mathbf{w}, U) \propto$$

$$\prod_{j \in x(i)} \frac{n^{\neg x(j)}_{z_{j-1} w_{j-1} c_j} + \lambda_{c_j}}{n^{\neg x(j)}_{z_{j-1} w_{j-1} \cdot} + \lambda_0 + \lambda_1}$$
$$\prod_{j \in s(i)} P(z_j \mid \mathbf{c}, \mathbf{z}_{\neg s(j)}, U) \qquad P_{\mathbf{u}_j}(\mathbf{w}_j \mid \tau_{\neg s(i)}, U)$$

with

$$P(z_j \mid \mathbf{c}, \mathbf{z}_{\neg s(j)}, U) = \begin{cases} \dfrac{n^{\neg s(j)}_{d z_j} + \alpha}{n^{\neg s(j)}_{d \cdot} + T\alpha} & \text{if } c_j = 1 \\ \delta_{z_j, z_{j-1}} & \text{if } c_j = 0 \end{cases}$$

where $P_{\mathbf{u}_j}(\mathbf{w}_j \mid \tau_{\neg s(i)}, U)$ is given by Equation 1, T is the number of topics, $n^{\neg s(j)}_{dz}$ is the number of phrases in document d that have topic z when $s(j)$'s assignment is excluded, and $n^{\neg s(j)}_{zwc}$ is the number of times a changepoint c has followed a word w with topic z when $s(j)$'s assignments are excluded.

After drawing a proposal for $c_i, \mathbf{z}_{s(i)}$ for token i, the sampler adds a customer eating w_i to a table serving w_i in restaurant $\mathbf{u}_i$. An old table k is selected with probability $\propto \max(0, c_{\mathbf{u}wk} - a_{|\mathbf{u}|})$ and a new table is selected with probability $\propto (b_{|\mathbf{u}_i|} + a_{|\mathbf{u}_i|} t_{\mathbf{u}_i \cdot}) P_{\pi(\mathbf{u})}(w_i)$.

Let $\mathbf{z}'_{s(i)}, c'_i, \tau'_{s(i)}$ denote the proposed change to $\mathbf{z}_{s(i)}, c_i, \tau_{s(i)}$. We accept the proposal with probability $\min(A, 1)$ where

$$A = \frac{\hat{P}(\mathbf{z}'_{s(i)}, c'_i, \tau'_{s(i)}) \; Q(\mathbf{z}_{s(i)}, c_i, \tau_{s(i)})}{\hat{P}(\mathbf{z}_{s(i)}, c_i, \tau_{s(i)}) \; Q(\mathbf{z}'_{s(i)}, c'_i, \tau'_{s(i)})}$$

where Q is the proposal distribution and $\hat{P}$ is the true unnormalized distribution. $\hat{P}$ differs from Q in that the probability of each word w_j and the seating arrangement depends only on $\neg s(j)$, as opposed to the simplification of using $\neg s(i)$. Almost all proposals are accepted; hence, this theoretically motivated Metropolis Hastings correction step makes little difference in practice.

Because the parameters $\mathbf{a}$ and $\mathbf{b}$ have no intuitive interpretation and we lack any strong belief about what they should be, we give them vague priors where $\rho_1 = \rho_2 = 1$ and $\epsilon_1 = 10$, $\epsilon_2 = .1$. We then interleave a slice sampling algorithm (Neal, 2000) between sweeps of the Metropolis-Hastings sampler to learn these parameters. We chose not to do inference on α in order to make the tests of our model against TNG more equitable.

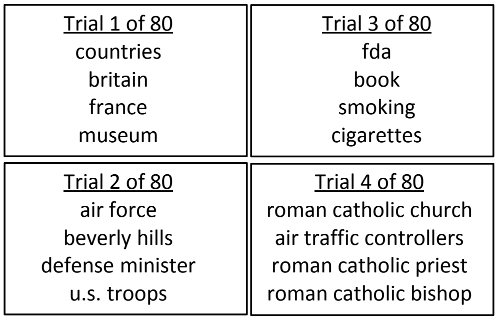
4 Related Work

An integral part of modeling topical phrases is the relaxation of the bag-of-words assumption in LDA. There are many models that make this relaxation. Among them, Griffiths and Steyvers (2005) present a model in which words are generated either conditioned on a topic or conditioned on the previous word in a bigram, but not both. They use this to model human performance on a word-association task. Wallach (2006) experiments with incorporating LDA into a bigram language model. Her model uses a hierarchical Dirichlet to share parameters across bigrams in a topic in a manner similar to our use of PYPs, but it lacks a notion of the topic being shared between the words in an n-gram. The Hidden Topic Markov Model (HTMM) (Gruber et al., 2007) assumes that all words in a sentence have the same topic, and consecutive sentences are likely to have the same topic. By dropping the independence assumption among topics, HTMM is able to achieve lower perplexity scores than LDA at minimal additional computational costs. These models are unconcerned with topical n-grams and thus do not model phrases.

Johnson (2010) presents an Adaptor Grammar model of topical phrases. Adaptor Grammars are a framework for specifying nonparametric Bayesian models over context-free grammars in which certain subtrees are "memoized" or remembered for reuse. In Johnson's model, subtrees corresponding to common phrases for a topic are memoized, resulting in a model in which each topic is associated with a distribution over whole phrases. While it is a theoretically elegant method for finding topical phrases, for large corpora we found inference to be impractically slow.

5 Phrase Intrusion Experiment

Perplexity is the typical information theoretic measure of language model quality used in lieu of extrinsic measures, which are more difficult and costly to run. However, it is well known that perplexity scores may negatively correlate with actual quality as assessed by humans (Chang et al., 2009). With that fact in mind, we expanded the methodology of Chang et al. (2009) to create a "phrase intrusion" task that quantitatively compares the quality of the topical n-gram lists produced by our model against those of other models.

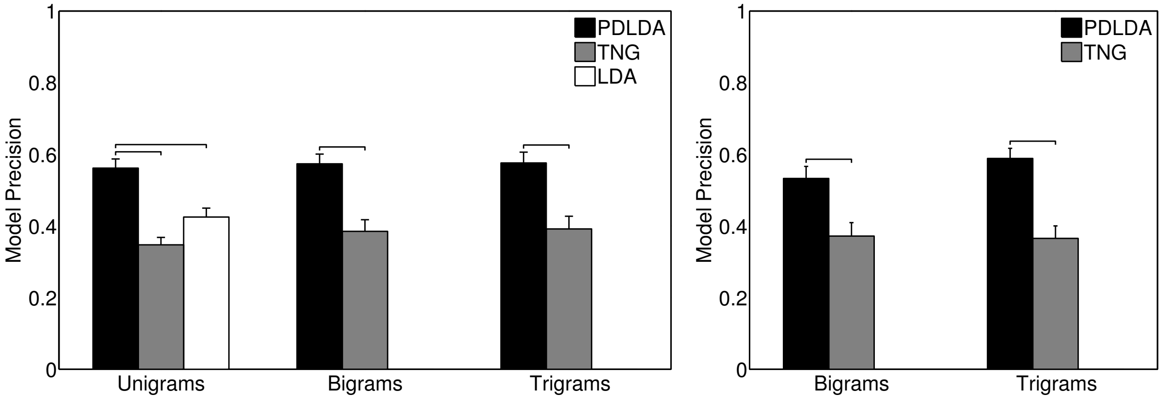


Trial 1 of 80 countries britain france museum	Trial 3 of 80 fda book smoking cigarettes
Trial 2 of 80 air force beverly hills defense minister u.s. troops	Trial 4 of 80 roman catholic church air traffic controllers roman catholic priest roman catholic bishop

Figure 4: Experimental setup of the phrase intrusion experiment in which subjects must click on the n-gram that does not belong.

Each of 48 subjects underwent 80 trials of a web-based experiment on Amazon Mechanical Turk, a reliable (Paolacci et al., 2010) and increasingly common venue for conducting online experiments. In each trial, a subject is presented with a randomly ordered list of four n-grams (cf. Figure 4). Each subject's task is to select the *intruder phrase*, a spurious n-gram not belonging with the others in the list. If, other than the intruder, the items in the list are all on the same topic, then subjects can easily identify the intruder because the list is semantically cohesive and makes sense. If the list is incohesive and has no discernible topic, subjects must guess arbitrarily and performance is at random.

To construct each trial's list, we chose two topics z and z' ($z \neq z'$), then selected the three most probable n-grams from z and the intruder phrase, an n-gram probable in z' and improbable in z. This design ensures that the intruder is not identifiable due solely to its being rare. Interspersed among the phrase intrusion trials were several simple screening trials intended to affirm that subjects possessed a minimal level of attentiveness and reading comprehension. For example, one such screening trial presented subjects with the list *banana*, *apple*, *television*, *orange*. Subjects who got any of these trials



(a) Word repetition allowed within a list.

(b) Word repetition not allowed.

Figure 5: An across-subject measure of the ability to detect intruders as a function of n-gram size and model. Excluding trials with repeated words does not qualitatively affect the results.

wrong were excluded from our analyses.

Each subject was presented with trials constructed from the output of PDLDA and TNG for unigrams, bigrams, and trigrams. For unigrams, we also tested the output of the original smoothed LDA (Blei et al., 2003). The experiment was conducted twice for a 2,246-document subset of the TREC AP corpus (Blei et al., 2003; Harman, 1992): the first time proceeded as described above, but the second time did not allow word repetition within a topic's list. The topical phrases found by TNG and PDLDA often revolve around a central n-gram, with other words pre- or post- appended to it. In this intrusion experiment, any n-gram not containing the central word or phrase may be trivially identifiable, regardless of its relevance to the topic. For example, the intruder in Trial 4 of Figure 4 is easily identifiable even if a subject does not understand English. This second experiment was designed to test whether our conclusions hinge on word repetition.

We used the MALLET toolbox (McCallum, 2002) for the implementations of LDA and TNG. Each model was run with 100 topics for 5,000 iterations. We set $m = 2$, $\alpha = .01$, $\beta = .01$, $\lambda = 1$, $\pi_1 = \pi_2 = 1$, $\rho_1 = 10$, and $\rho_2 = .1$. For all models, we treated certain punctuation as the start of a phrase by setting $c_j = 1$ for all tokens j immediately following periods, commas, semicolons, and exclamation and question marks. To reduce runtime, we removed stopwords occuring in the MALLET toolbox's stopword list. Because TNG and LDA had trouble with single character words not in the stoplist, we manually removed them before the experiment. Any token immediately following a removed word was treated as if it were the start of a phrase.

As in Chang et al. (2009), performance is measured via model precision, the fraction of subjects agreeing with the model. It is defined as $\text{MP}_k^{m,n} = \sum_s \mathbb{1}(i_{k,s}^{m,n} = \omega_{k,s}^{m,n})/S$ where $\omega_{k,s}^{m,n}$ is the index of the intruding n-gram for subject s among the words generated from the kth topic of model m, $i_{k,s}^{m,n}$ is the intruder selected by s, and S is the number of subjects. The model precisions are shown in Figure 5. PDLDA achieves the highest precision in all conditions. Model precision is low in all models, which is a reflection of how challenging the task is on a small corpus laden with proper nouns and low-frequency words. Figure 5b demonstrates that the outcome of the experiment does not depend strongly on whether the topical n-gram lists have repeated words.

6 Conclusion

We presented a topic model which simultaneously segments a corpus into phrases of varying lengths and assigns topics to them. The topical phrases found by PDLDA are much richer sources of information than the topical unigrams typically produced in topic modeling. As evidenced by the phrase-intrusion experiment, the topical n-gram lists that PDLDA finds are much more interpretable than those found by TNG.

The formalism of Bayesian changepoint detection arose naturally from the intuitive assumption that the topic of a sequence of tokens changes periodically, and that the tokens in between changepoints comprise a phrase. This formalism provides a principled way to discover phrases within the LDA framework. We presented a model embodying these principles and showed how to incorporate dependent Pitman-Yor processes into it.

Acknowledgements

The first author is supported by an NSF Graduate Research Fellowship. The first and second authors began this project while working at J.D. Power & Associates. We are indebted to Michael Mozer, Matt Wilder, and Nicolas Nicolov for their advice.

References

Ryan Prescott Adams and David J.C. MacKay. 2007. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing Technology, Harvard University.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press.

Thomas L. Griffiths, Joshua B. Tenenbaum, and Mark Steyvers. 2007. Topics in semantic representation. *Psychological Review*, 114:211–244.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic Markov models. *Journal of Machine Learning Research - Proceedings Track*, 2:163–170.

Donna Harman. 1992. Overview of the first text retrieval conference (trec–1). In *Proceedings of the first Text REtrieval Conference (TREC–1)*, Washington DC, USA.

Mark Johnson. 2010. PCFGs, Topic Models, Adaptor Grammars and Learning Topical Collocations and the Structure of Proper Names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July. Association for Computational Linguistics.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211 – 240.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Radford Neal. 2000. Slice sampling. *Annals of Statistics*, 31:705–767.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

J. Pitman and M. Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

J. Pitman. 2002. Combinatorial stochastic processes. Technical Report 621, Department of Statistics, University of California at Berkeley.

Patrick Schone and Daniel Jurafsky. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 100–108.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 985–992, Morristown, NJ, USA. Association for Computational Linguistics.

Hanna M. Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984.

Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining*.