Summarizing Topics: From word lists to phrases

Lauren A. Hannah
Department of Statistics
Columbia University
New York, NY 10012
lah2178@columbia.edu

Hanna Wallach
Microsoft Research
New York, NY 10011
hanna@dirichlet.net

Abstract

We propose a two-stage approach to generate descriptive phrases from the output of a multinomial topic model. First, we propose a Bayesian way to statistically select phrases from a document corpus, using priors associated with LDA [4]. Second, the selected phrases are combined with the topic dictionary to make a list of candidate phrases, which are ranked in terms of topic descriptiveness using a metric based on weighted KL divergence between the topic probabilities implied by the phrase and those implied by the topic model.

1 Introduction

Topic models summarize a set documents in a corpus by giving a weighted association between each document and a set of topics. The topics themselves are a multinomial distribution over a set of words, giving higher probabilities to sets of words that frequently appear together. Topics themselves are usually summarized by the ten or so words with the highest probability under the topic. Topic models are often used for knowledge extraction purposes, such as document clustering and generating candidates for recommender systems. The current topic summary conventions are often problematic for these settings. First, topic summaries are unwieldy and often induce users to generate their own topic names for referencing. Automatically generated descriptive phrases would be useful for applications like recommender systems. Second, automatically generated descriptive phrases can be useful for knowledge discovery, such as highlighting specific but little known terms (like "american heritage river," a specific term used by the EPA to designate a river for special attention). We propose two ideas in this paper for generating phrase-based names from the output of *any* multinomial topic model: (i) statistically defining phrases from a corpus in a Bayesian manner, and (ii) phrase selection using a metric based on weighted KL divergence.

2 Existing Methods

Phrase generation and automatic topic naming are not new ideas. Phrase generation has received attention in the statistical linguistics community since the late 1980's, using frequentist methods like Peason's χ^2 test [7], Gaussian approximations [18], likelihood ratios [9], $t-$tests against the null hypothesis of no difference in mean [6], and mutual information [8]. Most of these methods have significant issues when applied to text mining. Many of the hypothesis testing formulations rely on asymptotic approximations, which are not valid with small sample sizes. Other methods, like mutual information, are biased toward heavily weighting rare events and are difficult to use in a hypothesis testing situation. All proposed methods have been frequentist, which ignores the Bayesian framework underlying most modern topic models [4].

Topic naming has received attention more recently as the popularity of topic models has grown. Many methods find single words that contain information about topic probabilities [10, 2, 5, 3, 21], and cannot be easily extended to settings where multiword phrases are considered for topic names.

Some methods can consider multiword phrases, such as an approach that uses cosine similarity between a phrase and topic word distribution centroid [17], TF-IDF based metrics [20], or a multiphase approach that makes a list of candidate phrases from context, and trims it using topic relevance, marginal relevance, and discrimination [14]. Other methods have included external sources for phrase generation and evaluation, like user-generated names [16], or Wikipedia [13]. We aim to create a statistically principled, stand-alone method that can accommodate both single words and multiword phrases as possible topic labels.

3 Phrase Generation

We use statistical hypothesis testing to determine whether a string of words is actually a phrase, like "white house," or just joined by chance, like "house near." Let $\phi = w_1, \ldots, w_m$ be the set of words in a given n-gram. We deem ϕ to be a phrase if it occurs more frequently than our model would dictate—in this case, if the words are not independent. We start by defining a set of candidate bigrams, and then in future work will add in one word at a time to build full $n-$gram phrases. The first step is to collect all bigrams in the corpus to make the following contingency table:

	$word\ 1\ is\ w_1$	$word\ 1\ is\ not\ w_1$	$Row\ Total$
$word\ 2\ is\ w_2$	a	b	$a+b$
$word\ 2\ is\ not\ w_2$	c	d	$c+d$
$Column\ Total$	$a+c$	$b+d$	n

Previous work has used frequentist ranking [8], or hypothesis testing [7, 18, 6], which rely on asymptotic approximations and are not valid with small sample sizes. When the minimum expected table entry is at least 5, a χ^2 approximation can be used in Pearson's χ^2 Test. However, under an independence assumption the expected values are usually much lower, so we use a Yates' χ^2 Test:

$$\chi^2_{Yates} = \frac{n(|ad - bc| - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}$$

The distribution is a χ^2 with 1 degree of freedom. Word pairs are rejected as phrases if the associated χ^2_{Yates} falls below the $\alpha = 0.999$ quantile using a one-sided χ^2 test, which corresponds to a value of 10.83. However, the Yates corrected χ^2 is too conservative, and can still be inaccurate with the low expected number of observations in some elements of the contingency table.

Testing for independence can also be viewed in a Bayesian setting, as the underlying topic models have a Bayesian structure. We assume that word pairs can be generated from one of two models: a model where each word in a pair is drawn independently from a Bernoulli distribution, and a model where the pair of words together is drawn from a multinomial. A number of different Bayes factors have been derived for testing independence in contingency tables by using different prior formulations. These have included Dirichlet priors on the multinomial parameters [11, 12], and Gaussian priors on coefficients of a linear model that describes the log odds of collocations [19, 15, 1]. In all situations, the independent model is nested within the dependent model. Unlike traditional χ^2 tests, Bayes factors do not rely on asymptotic approximations that are inherent in χ^2 approximations. This makes Bayes factors especially favorable for this setting, where expected cell values can be close to 0.

We choose to use Bayes factors with a Dirichlet prior for the multinomial parameters as this is a common prior for popular topic models like LDA [4]. Bayes factors testing contingency table row/column independence under a Dirichlet prior have been studied by [12]. They proposed the following model. To enhance model tractability, the counts in each part of the contingency table are modeled as independent Poisson random variables conditioned on the total table count with mean parameters $\boldsymbol{\lambda} = (\lambda_a, \lambda_b, \lambda_c, \lambda_d)$; let $\bar{\lambda} = \lambda_a + \lambda_b + \lambda_c + \lambda_d$. These can be used to generate multinomial probabilities, $\boldsymbol{\pi} = (\pi_a, \pi_b, \pi_c, \pi_d)$, with $\pi_i = \lambda_i / \bar{\lambda}$.

Under the alternative model with dependent rows and columns, a Dirichlet prior is placed on the multinomial parameters and a gamma prior is placed on the total count:

$$\boldsymbol{\pi} \sim \mathrm{Dir}_4(\alpha_a, \alpha_b, \alpha_c, \alpha_d), \qquad \bar{\lambda} \sim \Gamma(\bar{\alpha}, \beta),$$

where $\bar{\alpha} = \alpha_a + \alpha_b + \alpha_c + \alpha_d$. Under the null model, row and column probabilities are modeled independently. Both are given independent Dirichlet (beta) priors, while the count is also given a gamma prior:

$$\boldsymbol{\pi}_c \sim \mathrm{Dir}_2(\alpha_a + \alpha_c - 1, \alpha_b + \alpha_d - 1),\ \boldsymbol{\pi}_r \sim \mathrm{Dir}_2(\alpha_a + \alpha_b - 1, \alpha_c + \alpha_d - 1),\ \bar{\lambda} \sim \Gamma(\bar{\alpha} - 1, \beta).$$

Here $\boldsymbol{\pi}_c$ is a vector of column probabilities and $\boldsymbol{\pi}_r$ is a vector of row probabilities.

Bayes factors can be computed for different sets of information; we consider Bayes factors when our data is the observed counts, (a, b, c, d), conditioned on the total count, which removes the dependency on the Gamma scaling parameter, β. The factor is given in Equation (4.4) of [12],

$$\begin{aligned} B_{01}(a,b,c,d \,|\, n) = & \left[\frac{\Gamma(a+b+\alpha_a+\alpha_b-1)\Gamma(c+d+\alpha_c+\alpha_d-1)\Gamma(\bar{\alpha}-2)}{\Gamma(n+\bar{\alpha}-2)\Gamma(\alpha_a+\alpha_b-1)\Gamma(\alpha_c+\alpha_d-1)}\right] \\ & \times \left[\frac{\Gamma(a+c+\alpha_a+\alpha_c-1)\Gamma(b+d+\alpha_b+\alpha_d-1)\Gamma(\bar{\alpha}-2)}{\Gamma(n+\bar{\alpha}-2)\Gamma(\alpha_a+\alpha_c-1)\Gamma(\alpha_b+\alpha_d-1)}\right] \\ & \times \left[\frac{\Gamma(\alpha_a)\Gamma(\alpha_b)\Gamma(\alpha_c)\Gamma(\alpha_d)\Gamma(n+\bar{\alpha})}{\Gamma(\bar{\alpha})\Gamma(a+\alpha_a)\Gamma(b+\alpha_b)\Gamma(c+\alpha_c)\Gamma(d+\alpha_d)}\right]. \end{aligned}$$

We used a flat Dirichlet prior, with $\boldsymbol{\alpha} = \mathbf{1}$. The threshold was set at $1/10$, meaning the odds ratio for all selected phrases is greater that or equal to 10.

4 Phrase Selection

Words or phrases that contain a lot of information about the topic should be: (i) Precise, as the word or phrase should be identify the topic with little ambiguity; and (ii) Recognizable, as the word or phrase should be common enough that somebody with some familiarity with documents containing it has a reasonable probability to recognize the word or phrase. Precision can be viewed as the ability of a word or phrase to point to a given topic, but little more than that topic. Mathematically, we say that a word or phrase ϕ has high precision for topic t if it greatly changes the KL divergence between the topic distribution given ϕ (in a random variable sense) from the unconditional topic distribution. This should eliminate from consideration high probability words or phrases that are common over a set of topics. However, this sort of metric is often skewed toward very rare but highly topic specific words and phrases. Recognizability is highly correlated with the commonness of a word or phrase; the more it is used in a set of documents, the higher the likelihood that the word is well known to a relatively large group of people. Mathematically, we say that a word is common if $p(\phi)$, the probability of the word or phrase in the entire corpus, is high.

Let ϕ be a word w or a bigram $w_1 w_2$. A metric that balances precision and recognizability is the expected KL divergence of the topic distribution given ϕ, $p(t|\phi)$, from the unconditional topic distribution, $p(t)$, generated by the topic model:

$$Q(\phi, t) = p(\phi)\left[\sum_{s=t,\sim t} p(s|\phi)\log\frac{p(s|\phi)}{p(s)}\right] + p(\sim\phi)\left[\sum_{s=t,\sim t} p(s|\sim\phi)\log\frac{p(s|\sim\phi)}{p(s)}\right], \quad (1)$$

$$\begin{aligned} p(\phi) &= \frac{\#\,\phi:\ \text{all terms in same topic}}{\#\,\mathrm{n-grams}:\ \text{all terms in same topic}} \\ p(t|\phi) &= \frac{\#\,\phi:\ \text{all terms in topic } t}{\#\,\phi:\ \text{all terms in same topic}} \\ p(t|\sim\phi) &= \frac{\#\,\mathrm{n-grams\ excluding}\ \phi:\ \text{all terms in topic } t}{\#\,\mathrm{n-grams\ excluding}\ \phi:\ \text{all terms in same topic}}. \end{aligned}$$

Here n-gram is defined by the length of ϕ (unigram or bigram). Note that even seeing a bigram can change the distribution over topics, regardless of the bigram content. The first part of (1) is similar to the saliency metric of [5], although the latter is over the entire topic distribution rather than a single topic. This weights the KL divergence of the topics given that ϕ has been seen from the unconditional distribution with the probability of ϕ. The second part of the right hand side weights the KL divergence of the topic distribution given that ϕ is absent from the unconditional distribution by the probability that ϕ is absent. The second term should always be close to 0 for bigrams and unigrams, so the first term dominates $Q(\phi, t)$. Since $Q(\phi, t)$ is not dependent on the length of a phrase and simply measures the change in topic distributions given ϕ, it can be used to compare phrases of differing lengths.

5 Results

We applied our methods to two corpora, Federal Reserve Minutes from (ADD DOCUMENTATION HERE), and emails from the Clinton Library (MORE DOX). LDA topic models were fit using Mallet, which uses Gibbs sampling inference. Outputs consist of each word labeled by document, position, and topic. Both χ^2 and Bayes factor hypothesis tests were run on the corpora; the resulting candidate phrase lists were used to generate descriptive phrases. We compare the top 5 selected candidate phrases in Table 1. The Bayes factor test tends to give higher scores to phrases which occur often, while the χ^2 test often givens high scores to phrases that occur only a handful of times; this is due to the influence of the prior in the Bayesian model. However, the lists of selected words are quite similar between the methods. Selected phrases are given in Tables 2 and 3. Selected phrases can include ligature errors, such as "certi cation" and "signi cantly", common phrases, like "bully pulpit", or uncommon phrases not included in the top 10 single words, like "tri-party repo." In the latter situations, high information phrases may direct the user to new lines of inquiry.

Table 1: Candidate phrase generation for Federal Reserve Minutes.

χ^2			Bayes Factor		
Phrase	Count	Value	Phrase	Count	Log Value
st. louis	28	557381	funds rate	2008	-8620
moral hazard	67	539486	monetary policy	1227	-5688
san francisco	21	533437	basis points	939	-5171
ad hoc	9	513574	fed funds	709	-3437
pros cons	16	502282	inflation expectations	1176	-3351

Table 2: Topic phrases for Federal Reserve Minutes.

Topic	LDA Output	Descriptive Phrases	KL Values
0	issue, terms, inflation, problem, important, expectations, fact, policy, situation, markets	issue, terms, problem, important, situation	0.00091, 0.00057, 0.00044, 0.00029, 0.00028
1	inflation, objective, price, stability, goal, committee, target, numerical, percent, explicit	price stability, objective, inflation objective, dual mandate, numerical objective	0.0044, 0.0032, 0.0029, 0.0028, 0.0021
2	liquidity, institutions, financial, markets, market, lending, problem, facilities, chairman, institution	moral hazard, unusual exigent, exigent circumstances, institutions, liquidity	0.0022, 0.0009, 0.0008, 0.0008, 0.0008
4	market trading futures hedge morning money contracts fund early stock	hedge fund, market, eurodollar futures, trading, money market	0.0007, 0.0006, 0.0004, 0.0004, 0.0004
15	rate, funds, reserves, target, federal, reserve, interest, balance, sheet, option	funds rate, excess reserves, federal funds, balance sheet, interest rates	0.0040, 0.0032, 0.0029, 0.0028, 0.0026
22	u.s., growth, foreign, dollar, exports, prices, oil, economies, forecast, countries	current account, industrial countries, account deficit, net exports, foreign economies	0.0028, 0.0028, 0.0026, 0.0024, 0.0023
28	ceo, price, percent, growth, terms, chairman, reports, texas, largest, economy	ceo, texas instruments, texas, dallas, burlington northern	0.0009, 0.0004, 0.0004, 0.0004, 0.0003
32	important, chairman, issues, point, agree, staff, forward, comments, discussion, issue	exit strategy, issues raised, ad hoc, federal reserve, pros cons	0.0033, 0.0031, 0.0030, 0.0030, 0.0030
35	auction, collateral, facility, taf, term, primary, banks, discount, window, stigma	discount window, auction, auction facility, collateral, taf	0.0035, 0.0018, 0.0017, 0.0015, 0.0014
45	capital, firms, risk, lehman, bank, pdcf, banks, management, regulatory, primary	bear stearns, tri-party repo, morgan stanley, stress testing, lehman	0.0011, 0.0008, 0.0005, 0.0005, 0.0005
49	rate, funds, basis, policy, today, inflation, market, points, point, move	funds rate, 25 basis, basis points, fed funds, 50 basis	0.0077, 0.0056, 0.0054, 0.0052, 0.0040

References

[1] James H. Albert. Bayesian testing and estimation of association in a two-way contingency table. *Journal of the American Statistical Association*, 92(438):685–693, 1997.

[2] Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori. A new document clustering algorithm for topic discovering and labeling. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 161–168. Springer, 2008.

[3] Jonathan Bischof and Edoardo M Airoldi. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 201–208, 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Table 3: Topic phrases for Clinton Library Emails.

Topic	LDA Output	Descriptive Phrases	KL Values
0	law, requirements, federal, laws, legislation, provision, provide, including, agencies, required	certi cation, signi cantly, current law, enacted law, adversely affect	0.0025, 0.0025, 0.0024, 0.0024, 0.0024
6	africa, african, south, opportunity, program, including, president, policy, case, american	africa, south africa, communique icag, approach policy, planning program	0.0004, 0.0002, 0.0001, 0.0001, 0.0001
21	reform, election, president, statement, meet, change, union, speech, major, pulpit	reform, election, dramatic reform, bully pulpit, conference statement	0.00010, 0.00005, 0.00005, 0.00005, 0.00004
34	america, children, american, americans, give, today, country, families, challenge, working	common ground, american dream, america challenge, common sense, families communities	0.0035, 0.0035, 0.0035, 0.0035, 0.0035
49	act, scoring, budget, pay, legislative, omb, direct, subject, omnibus, iad	scoring, omnibus budget, direct spending, reconciliation act, budget reconciliation	0.0002, 0.0002, 0.0002, 0.0002, 0.0002
55	congress, reform, congressional, limits, term, amendment, president, republicans, press, cut	term limits, congress, lobby reform, constitutional amendment, gift ban	0.0005, 0.0003, 0.0003, 0.0003, 0.0003
65	burma, regime, issue, nigeria, proposal, effort, congress, development, narcotics, program	burma, burma heroin, source opium, largest source, personally involved	0.0001, 0.0001, 0.0001, 0.0001, 0.0001
124	service, smoking, law, opinion, question, tobacco, nicotine, jack, disease, misconduct	jack thompson, nicotine dependence, service, willful misconduct, smoking	0.0003, 0.0002, 0.0002, 0.0002, 0.0001
209	bank, initiative, im, usg, aids, export, hiv, aid, drugs, nancing	im bank, hiv aids, bank, aids initiative, tied aid	0.0005, 0.0003, 0.0003, 0.0003, 0.0003
247	members, public, vote, campaign, senators, list, president, plan, votes, states	senators targeted, republican senators, intergovernmental affairs, dnc intergovernmental, swing votes	0.0001, 0.0001, 0.0001, 0.0001, 0.0001
257	french, paris, privacy, magaziner, domain, issue, management, data, france, views	digital signatures, hears french, thomas marten, french views, domain management	0.0003, 0.0002, 0.0002, 0.0002, 0.0002
262	water, environmental, clean, environment, forest, act, land, national, protection, lands	water, clean water, endangered species, water act, drinking water	0.0006, 0.0006, 0.0004, 0.0003, 0.0003
282	consumer, financial, cards, fraud, card, treasury, sarah, sec, consumers, loan	sarah rosen, consumer, loan checks, debit cards, cards	0.0003, 0.0003, 0.0002, 0.0002, 0.0002
285	government, federal, cut, programs, deficit, vice, tax, president, job, works	government, reinventing government, deficit reduction, federal workforce, government works	0.0005, 0.0003, 0.0003, 0.0003, 0.0003
286	president, thing, money, put, picture, lot, things, sentence, section, stuff	president, president yeah, blah blah, thing, paint picture	0.0006, 0.0003, 0.0003, 0.0003, 0.0003
376	workers, wage, minimum, jobs, companies, businesses, business, job, pay, economic	minimum wage, raise minimum, workers, corporate citizenship, increase minimum	0.0009, 0.0005, 0.0005, 0.0004, 0.0004
377	immigration, illegal, legal, policy, subject, governor, wilson, border, asylum, issue	immigration, illegal immigration, border patrol, rahm emanuel, immigration policy	0.0003, 0.0002, 0.0001, 0.0001, 0.0001
381	crime, police, bill, assault, criminals, streets, guns, weapons, gun, community	police, crime, assault weapons, brady bill, assault	0.0007, 0.0007, 0.0003, 0.0003, 0.0002
416	insurance, premium, individual, rate, coverage, market, premiums, effect, group, insurers	insurance, insurance market, groups individuals, rate increases, preexisting condition	0.0002, 0.0002, 0.0002, 0.0001, 0.0001
424	programs, support, federal, program, funding, million, initiative, assistance, provide, increase	microenterprise development, technical assistance, microenterprise programs, block grant, cdfi fund	0.0012, 0.0011, 0.0011, 0.0011, 0.0011
468	referendum, national, reform, vote, states, campaign, congress, tv, voted, public	national referendum, pac donations, broadcasting industry, free tv, dole ross	0.0003, 0.0003, 0.0003, 0.0002, 0.0002
469	china economy global million economic leaders competition understand percent create	china, china leaders, global competition, class industries, plagued corruption	0.0003, 0.0001, 0.0001, 0.0001, 0.0001

[5] Jason Chuang, Christopher D Manning, and Jeffrey Heer. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77. ACM, 2012.

[6] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploit- ing On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum, Hillsdale, NJ, 1991.

[7] Kenneth W Church and William A Gale. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62, 1991.

[8] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[9] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74, 1993.

[10] Filippo Geraci, Marco Pellegrini, Marco Maggini, and Fabrizio Sebastiani. Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution. In *String Processing and Information Retrieval*, pages 25–36. Springer, 2006.

[11] I. J. Good. A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society: Series B*, 29(3):399–431, 1967.

[12] Erdogan Gunel and James Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, 1974.

[13] Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1536–1545. Association for Computational Linguistics, 2011.

[14] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM, 2007.

[15] A. E. Raftery. A note on Bayes factors for log-linear contingency table models with vague prior information. *ournal of the Royal Statistical Society: Series B*, 48(2):249–250, 1986.

[16] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[17] Marian-Andrei Rizoiu and Julien Velcin. Topic extraction for ontology learning. *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, pages 38–61, 2011.

[18] Frank Smadja. Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177, 1993.

[19] D. J. Spiegelhalter and A. F. M. Smith. Bayes factors for log-linear contingency table models with vague prior information. *ournal of the Royal Statistical Society: Series B*, 44(3):377–387, 1982.

[20] Pucktada Treeratpituk and Jamie Callan. Automatically labeling hierarchical clusters. In *Proceedings of the 2006 international conference on Digital government research*, pages 167–176. Digital Government Society of North America, 2006.

[21] Yuen-Hsien Tseng. Generic title labeling for clustered documents. *Expert Systems with Applications*, 37(3):2247–2254, 2010.