# Summarizing Topics: From Word Lists to Phrases

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

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

## Abstract

In this paper, we present a two-stage approach to generating descriptive phrases from the output of a statistical topic model, such as LDA [4]. First, we propose a Bayesian method for selecting statistically significant phrases from a corpus of documents, using inferred parameter values from LDA. Second, the selected phrases are combined with the topic assignments to make a list of candidate phrases for each topic. These phrases then are ranked in terms of descriptiveness using a metric based on the weighted KL divergence between topic probabilities implied by the phrase and those implied by inferred parameter values from LDA.

## 1 Introduction

Statistical topic models summarize a set of documents, or corpus, by providing a weighted association between each document and a set of topics. Each topic is characterized by a categorical distribution over some shared vocabulary that assigns higher probabilities to sets of words that tend to occur together. Since these categorical distributions are tend to be extremely high dimensional (on the order of tens or hundreds of thousands), each topic is usually summarized by the ten or so words with the highest probabilities under that topic. Since topic models are often used for exploratory analysis of corpora that are far too large for a single human to read, their output needs to be meaningful to humans. Unfortunately, current topic summary conventions (i.e., lists of five to ten words) are at best insufficient and can even be misleading. First, word lists are unwieldy and, as evidenced by the results sections of numerous papers using topic models in the social sciences, induce users to generate their own topic names for referencing. We argue that automatically generated descriptive phrases would make these users lives easier. Second, since topic models are often used for exploratory purposes, automatically generated descriptive phrases would enhance the exploration process itself by highlighting specific but little-known terms (like "american heritage river," a specific term used by the EPA to designate a river for special attention). In this paper, we outline a two-stage approach to generating descriptive phrases from the inferred parameters of any LDA-based statistical topic model. This approach involves 1) identifying statistically significant phrases in a Bayesian manner and 2) selecting phrases using a metric based on KL divergence.

## 2 Existing Methods

Phrase generation and automatic topic naming are not new ideas. Phrase generation first received attention in the natural language processing community in the late 1980s, via frequentist methods like Pearson's $\chi^2$ test [7], Gaussian approximations [18], likelihood ratios [9], $t-$tests against the null hypothesis of no difference in mean [6], and mutual information [8]. Unfortunately, most of these methods have significant issues when applied to text. 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. Moreover, all proposed methods have been frequentist, ignoring the

Bayesian framework underlying most modern topic models [4]. Topic naming has received increasing attention as the popularity of topic models has grown. Many methods find single words that convey information about topic probabilities [10, 2, 5, 3, 21], and cannot be easily extended handle multiword phrases. Some methods can accommodate multiword phrases, such as an approach that uses the cosine similarity between a phrase and topic's centroid [17], TF-IDF-based metrics [20], or an two-stage approach that makes a list of candidate phrases from context and then trims it using topic relevance, marginal relevance, and discrimination [14]. Other methods have included external sources for phrase generation and evaluation [16, 13]. Here, we present a statistically principled, stand-alone method that can seamlessly accommodate both single words and multiword phrases.

## 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 $\psi = w_1, \dots, w_n$ be the sequence of words in a given $n$-gram. We deem $\psi$ to be a phrase if it occurs more frequently than our model would dictate—in this case, if the words are not independent. Here we outline our approach to defining a set of candidate bigrams; we are currently working on adding in words to build phrases of length $n$. The first step is to compute the following contingency table for all bigrams in the corpus:

| | # first word is $w_1$ | # first word is not $w_1$ | row total |
|---:|:---:|:---:|:---:|
| # second word is $w_2$ | $a$ | $b$ | $a+b$ |
| # second word is not $w_2$ | $c$ | $d$ | $c+d$ |
| column total | $a+c$ | $b+d$ | $n$ |

Previous methods have 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 five, a $\chi^2$ approximation can be used in Pearson's $\chi^2$ test; however, under an independence assumption the expected values are usually much lower, so we use a Yates' $\chi^2$ test:

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

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

Testing for independence can also be performed in a Bayesian setting—arguably a more appropriate one, given that most modern topic models are Bayesian. Here, we assume that word pairs can arise 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 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, including 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 $\chi^2$ tests, Bayes factors do not rely on the asymptotic approximations inherent in $\chi^2$ approximations. This makes Bayes factors especially favorable for this setting, where expected table entries can be close to zero.

We choose to use Bayes factors with a Dirichlet prior for the multinomial parameters, as this is a common prior for topic models like LDA [4]. Bayes factors testing contingency table row/column independence under a Dirichlet prior have been studied by Gunel and Dickey [12], who proposed the following model. To enhance model tractability, the counts in each cell of the contingency table (i.e., $a$, $b$, $c$, and $d$) are modeled as independent Poisson random variables conditioned on the total table count $n$ 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 parameters 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: $\boldsymbol{\pi} \sim \mathrm{Dir}_4(\alpha_a, \alpha_b, \alpha_c, \alpha_d)$ and $\bar{\lambda} \sim \Gamma(\bar{\alpha}, \beta)$, where $\bar{\alpha} = \alpha_a + \alpha_b + \alpha_c + \alpha_d$. Under the null model, row and column probabilities—$\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_c$, respectively—are modeled independently. Both are given independent 2-dimensional Dirichlet (or beta) priors, i.e., $\boldsymbol{\pi}_c \sim \mathrm{Dir}_2(\alpha_a + \alpha_c - 1, \alpha_b + \alpha_d - 1)$ and $\boldsymbol{\pi}_r \sim \mathrm{Dir}_2(\alpha_a + \alpha_b - 1, \alpha_c + \alpha_d - 1)$, while the count is given a gamma prior: $\bar{\lambda} \sim \Gamma(\bar{\alpha} - 1, \beta)$.

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 $n$, which removes the dependency on the Gamma scaling parameter, $\beta$. According to Gunel and Dickey [12], this factor is

$$B_{01}(a,b,c,d\,|\,n) = \frac{\Gamma(a+b+\alpha_a+\alpha_b-1)\Gamma(c+d+\alpha_c+\alpha_d-1)\Gamma(\bar{\alpha}-2)}{\Gamma(n+\bar{\alpha}-2)\Gamma(\alpha_a+\alpha_b-1)\Gamma(\alpha_c+\alpha_d-1)} \times \frac{\Gamma(a+c+\alpha_a+\alpha_c-1)\Gamma(b+d+\alpha_b+\alpha_d-1)\Gamma(\bar{\alpha}-2)}{\Gamma(n+\bar{\alpha}-2)\Gamma(\alpha_a+\alpha_c-1)\Gamma(\alpha_b+\alpha_d-1)} \times \frac{\Gamma(\alpha_a)\Gamma(\alpha_b)\Gamma(\alpha_c)\Gamma(\alpha_d)\Gamma(n+\bar{\alpha})}{\Gamma(\bar{\alpha})\Gamma(a+\alpha_a)\Gamma(b+\alpha_b)\Gamma(c+\alpha_c)\Gamma(d+\alpha_d)}.$$

We used a symmetric Dirichlet prior, with $\alpha_a = \alpha_b = \alpha_c = \alpha_d = 1$ and $\bar{\alpha} = 4$. We set the threshold at $1/10$, meaning that the odds ratio for all selected phrases is greater than or equal to ten.

## 4 Phrase Selection

Words or phrases that contain a lot of information about the topic should be: 1) precise, as the word or phrase should be identify the topic with little ambiguity, and 2) recognizable, as the word or phrase should be common enough that somebody with some subject expertise has a reasonable probability of recognizing it. Precision can be viewed as the ability of a word or phrase to indicate a given topic, but not other topics. Mathematically, we say that a word or phrase $\psi$ has high precision for topic $t$ if it greatly changes the KL divergence between the distribution over topics given $\phi$ from the unconditional distribution over topics. This definition should eliminate high probability words or phrases that are common over all topics. In contrast, recognizability, which is highly correlated with the commonness of a word or phrase, guards against high precision phrases that are topic specific but very rare. The more a word or phrase is used, the more likely it is that the word or phrase is recognizable to a relatively large group of people. Mathematically, we say that a word or phrase $\psi$ is recognizable if $p(\psi)$—the empirical probability of that word or phrase in the corpus—is high.

A metric that balances precision and recognizability is the expected KL divergence between the distribution over topics given the word or phrase $\psi$, i.e., $p(t\,|\,\psi)$ and the unconditional distribution over topics $p(t)$ implied by the topic model, perhaps via a set of topic assignments:

$$Q(\psi,t) = p(\psi)\left(\sum_{s=t,\neg t} p(s\,|\,\psi)\log\frac{p(s\,|\,\psi)}{p(s)}\right) + p(\neg\psi)\left(\sum_{s=t,\neg t} p(s\,|\,\neg\psi)\log\frac{p(s\,|\,\neg\psi)}{p(s)}\right), \quad (1)$$

where $p(\psi) = \frac{\#\ \psi \text{ s.t. all terms in same topic}}{\#\ n\text{-grams s.t. all terms in same topic}}$, $p(t\,|\,\psi) = \frac{\#\psi \text{ s.t. all terms in topic } t}{\#\psi \text{ s.t. all terms in same topic}}$, $p(t\,|\,\neg\psi) = \frac{\#\ n\text{-grams excluding } \psi \text{ s.t. all terms in topic } t}{\#\ n\text{-grams excluding } \psi \text{ s.t. all terms in same topic}}$, and "n-gram" refers to either a word or phrase as determined by $\psi$. Note that the occurrence of any phrase can change the distribution over topics, regardless of identity of that phrase. The first part of (1) is similar to the saliency metric of Chuang et al. [5], although the latter is over the entire distribution over topics rather than a single topic. This weights the KL divergence of the topics given that $\psi$ has been seen from the unconditional distribution with the probability of $\psi$. The second part of (1) weights the KL divergence between the distribution over topics given that $\phi$ is absent and the unconditional distribution by the probability that $\psi$ is absent. The second term should always be close to 0 for bigrams and unigrams. Since $Q(\psi, t)$ is not dependent on the length of a phrase, it can be used to compare phrases of differing lengths.

## 5 Results

We applied our phrase generation and selection methods to the output of LDA[1] on two corpora: transcripts from Federal Open Market Committee Meetings [2] and previously restricted documents made available by the Clinton Library[3]. Both the $\chi^2$ and Bayes factor hypothesis tests were used to generate candidate phrases; these phrase lists were then used to generate descriptive phrases. We show

[1] Run with MALLET, which uses Gibbs sampling.

[2] Data source: `http://poliinformatics.org/data/`

[3] Data source: `http://www.clintonlibrary.gov/previouslyrestricteddocs.html`

the top five candidate phrases in table 1. The Bayes factor test tends to give higher scores to phrases which occur often, while the $\chi^2$ test often gives high scores to phrases that occur only a handful of times; this difference is due to the influence of the prior in the Bayes factor test. Otherwise, the phrase lists are very similar. Finally, we show descriptive phrases for several topics in tables 2 and 3. Selected phrases can include ligature errors, such as "certi cation" and "signi cantly"; common phrases, like "bully pulpit"; and uncommon phrases not included in the top 10 single words, like "tri-party repo." In the latter situations, these phrases may direct users to new lines of inquiry.

| $\chi^2$ | | | 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 1: Candidate phrase generation for FOMC meetings.

| Top words | Descriptive phrases | KL values |
|---|---|---|
| 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 |
| 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 |
| 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 |
| 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 |

Table 2: Descriptive phrases for topics inferred from FOMC meetings.

| Top words | Descriptive phrases | KL values |
|---|---|---|
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |

Table 3: Descriptive phrases for topics inferred from Clinton documents.

## References


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

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

[3] Bischof, J. and Airoldi, E. M. [2012], Summarizing topical content with word frequency and exclusivity, *in* Proceedings of the 29th International Conference on Machine Learning, pp. 201–208.

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

[5] Chuang, J., Manning, C. D. and Heer, J. [2012], Termite: Visualization techniques for assessing textual topic models, *in* Proceedings of the International Working Conference on Advanced Visual Interfaces, ACM, pp. 74–77.

[6] Church, K., Gale, W., Hanks, P. and Hindle, D. [1991], Using statistics in lexical analysis, *in* U. Zernik, ed., Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, Lawrence Erlbaum, Hillsdale, NJ, pp. 115–164.

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

[8] Church, K. W. and Hanks, P. [1990], Word association norms, mutual information, and lexicography, *Computational linguistics* **16**(1), 22–29.

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

[10] Geraci, F., Pellegrini, M., Maggini, M. and Sebastiani, F. [2006], Cluster generation and cluster labelling for web snippets: A fast and accurate hierarchical solution, *in* String Processing and Information Retrieval, Springer, pp. 25–36.

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

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

[13] Lau, J. H., Grieser, K., Newman, D. and Baldwin, T. [2011], Automatic labelling of topic models, *in* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 1536–1545.

[14] Mei, Q., Shen, X. and Zhai, C. [2007], Automatic labeling of multinomial topic models, *in* Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 490–499.

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

[16] Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. [2009], 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, pp. 248–256.

[17] Rizoiu, M.-A. and Velcin, J. [2011], Topic extraction for ontology learning, *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances* pp. 38–61.

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

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

[20] Treeratpituk, P. and Callan, J. [2006], Automatically labeling hierarchical clusters, *in* Proceedings of the 2006 international conference on Digital government research, Digital Government Society of North America, pp. 167–176.

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