Global Models of Document Structure Using Latent Permutations

Harr Chen, S.R.K. Branavan, Regina Barzilay, David R. Karger
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{harr, branavan, regina, karger}@csail.mit.edu

Abstract

We present a novel Bayesian topic model for learning discourse-level document structure. Our model leverages insights from discourse theory to constrain latent topic assignments in a way that reflects the underlying organization of document topics. We propose a *global* model in which both topic selection and ordering are biased to be similar across a collection of related documents. We show that this space of orderings can be elegantly represented using a distribution over permutations called the *generalized Mallows model*. Our structure-aware approach substantially outperforms alternative approaches for cross-document comparison and single-document segmentation.[1]

1 Introduction

In this paper, we introduce a novel latent topic model for the unsupervised learning of document structure. Traditional topic models assume that topics are randomly spread throughout a document, or that the succession of topics in a document is Markovian. In contrast, our approach takes advantage of two important discourse-level properties of text in determining topic assignments: first, that each document follows a progression of nonrecurring coherent topics (Halliday and Hasan, 1976); and second, that documents from the same domain tend to present similar topics, in similar orders (Wray, 2002). We show that a topic model incorporating these long-range dependencies outperforms alternative approaches for segmentation and cross-document comparison.

For example, consider a collection of encyclopedia articles about cities. The first constraint captures the notion that a single topic, such as Architecture, is expressed in a contiguous block within the document, rather than spread over disconnected sections. The second constraint reflects our intuition that all of these related articles will generally mention some major topics associated with cities, such as History and Culture, and will often exhibit similar topic orderings, such as placing History before Culture.

We present a Bayesian latent topic model over related documents that encodes these discourse constraints by positing a single distribution over a document's *entire* topic structure. This global view on ordering is able to elegantly encode discourse-level properties that would be difficult to represent using local dependencies, such as those induced by hidden Markov models. Our model enforces that the same topic does not appear in disconnected portions of the topic sequence. Furthermore, our approach biases toward selecting sequences with similar topic *ordering*, by modeling a distribution over the space of topic permutations.

Learning this ordering distribution is a key technical challenge in our proposed approach. For this purpose, we employ the *generalized Mallows model*, a permutation distribution that concentrates probability mass on a small set of similar permutations. It directly captures the intuition of the second constraint, and uses a small parameter set to control how likely individual topics are to be reordered.

We evaluate our model on two challenging

[1] Code, data, and annotations used in this work are available at http://groups.csail.mit.edu/rbg/code/mallows/

document-level tasks. In the *alignment* task, we aim to discover paragraphs across different documents that share the same topic. We also consider the *segmentation* task, where the goal is to partition each document into a sequence of topically coherent segments. We find that our structure modeling approach substantially outperforms state-of-the-art baselines for both tasks. Furthermore, we demonstrate the importance of explicitly modeling a distribution over topic permutations; our model yields significantly better results than variants that either use a fixed ordering, or are order-agnostic.

2 Related Work

Topic and Content Models Our work is grounded in topic modeling approaches, which posit that latent state variables control the generation of words. In earlier topic modeling work such as *latent Dirichlet allocation* (LDA) (Blei et al., 2003; Griffiths and Steyvers, 2004), documents are treated as bags of words, where each word receives a separate topic assignment; the topic assignments are auxiliary variables to the main task of language modeling.

More recent work has attempted to adapt the concepts of topic modeling to more sophisticated representations than a bag of words; they use these representations to impose stronger constraints on topic assignments (Griffiths et al., 2005; Wallach, 2006; Purver et al., 2006; Gruber et al., 2007). These approaches, however, generally model Markovian topic or state transitions, which only capture local dependencies between adjacent words or blocks within a document. For instance, content models (Barzilay and Lee, 2004; Elsner et al., 2007) are implemented as HMMs, where the states correspond to topics of domain-specific information, and transitions reflect pairwise ordering preferences. Even approaches that break text into contiguous chunks (Titov and McDonald, 2008) assign topics based on local context. While these locally constrained models can implicitly reflect some discourse-level constraints, they cannot capture long-range dependencies without an explosion of the parameter space. In contrast, our model captures the entire sequence of topics using a compact representation. As a result, we can explicitly and tractably model global discourse-level constraints.

Modeling Ordering Constraints Sentence ordering has been extensively studied in the context of probabilistic text modeling for summarization and generation (Barzilay et al., 2002; Lapata, 2003; Karamanis et al., 2004). The emphasis of that body of work is on learning ordering constraints from data, with the goal of reordering new text from the same domain. Our emphasis, however, is on applications where ordering is already observed, and how that ordering can improve text analysis. From the methodological side, that body of prior work is largely driven by local pairwise constraints, while we aim to encode global constraints.

3 Problem Formulation

Our document structure learning problem can be formalized as follows. We are given a corpus of D related documents. Each document expresses some subset of a common set of K topics. We assign a single topic to each paragraph,[2] incorporating the notion that paragraphs are internally topically consistent (Halliday and Hasan, 1976). To capture the discourse constraint on topic progression described in Section 1, we require that topic assignments be contiguous within each document.[3] Furthermore, we assume that the underlying topic sequences exhibit similarity across documents. Our goal is to recover a *topic assignment* for each paragraph in the corpus, subject to these constraints.

Our formulation shares some similarity with the standard LDA setup, in that a common set of topics is assigned across a collection of documents. However, in LDA each word's topic assignment is conditionally independent, following the bag of words view of documents. In contrast, our constraints on how topics are assigned let us connect word distributional patterns to document-level topic structure.

4 Model

We propose a generative Bayesian model that explains how a corpus of D documents, given as sequences of paragraphs, can be produced from a set of hidden topic variables. Topic assignments to each

[2] Note that our analysis applies equally to other levels of textual granularity, such as sentences.

[3] That is, if paragraphs i and j are assigned the same topic, every paragraph between them must have that topic.

paragraph, ranging from 1 to K, are the model's final output, implicitly grouping topically similar paragraphs. At a high level, the process first selects the bag of topics to be expressed in the document, and how they are ordered; these topics then determine the selection of words for each paragraph.

For each document d with N_d paragraphs, we separately generate a *bag of topics* $\mathbf{t}_d$ and a *topic ordering* π_d. The unordered bag of topics, which contains N_d elements, expresses how many paragraphs of the document are assigned to each of the K topics. Note that some topics may not appear at all. Variable $\mathbf{t}_d$ is constructed by taking N_d samples from a distribution over topics τ, a multinomial representing the probability of each topic being expressed. Sharing τ between documents captures the intuition that certain topics are more likely across the entire corpus.

The topic ordering variable π_d is a permutation over the numbers 1 through K that defines the order in which topics appear in the document. We draw π_d from the *generalized Mallows model*, a distribution over permutations that we explain in Section 4.1. As we will see, this particular distribution biases the permutation selection to be close to a single centroid, reflecting the discourse constraint of preferring similar topic structures across documents.

Together, a document's bag of topics $\mathbf{t}_d$ and ordering π_d determine the topic assignment $z_{d,p}$ for each of its paragraphs. For example, in a corpus with $K = 4$, a seven-paragraph document d with $\mathbf{t}_d = \{1, 1, 1, 1, 2, 4, 4\}$ and $\pi_d = (2\ 4\ 3\ 1)$ would induce the topic sequence $\mathbf{z}_d = (2\ 4\ 4\ 1\ 1\ 1\ 1)$. The induced topic sequence $\mathbf{z}_d$ can never assign the same topic to two unconnected portions of a document, thus satisfying the constraint of topic contiguity.

As with LDA, we assume that each topic k is associated with a language model θ_k. The words of a paragraph assigned to topic k are then drawn from that topic's language model θ_k.

Before turning to a more formal discussion of the generative process, we first provide background on the permutation model for topic ordering.

4.1 The Generalized Mallows Model

A central challenge of the approach we take is modeling the distribution over possible topic permutations. For this purpose we use the generalized Mallows model (GMM) (Fligner and Verducci, 1986; Lebanon and Lafferty, 2002; Meilă et al., 2007), which exhibits two appealing properties in the context of this task. First, the model concentrates probability mass on some "canonical" ordering and small perturbations of that ordering. This characteristic matches our constraint that documents from the same domain exhibit structural similarity. Second, its parameter set scales linearly with the permutation length, making it sufficiently constrained and tractable for inference. In general, this distribution could potentially be applied to other NLP applications where ordering is important.

Permutation Representation Typically, permutations are represented directly as an ordered sequence of elements. The GMM utilizes an alternative representation defined as a vector $(v_1, \ldots, v_{K-1})$ of *inversion counts* with respect to the identity permutation $(1, \ldots, K)$. Term v_j counts the number of times a value greater than j appears before j in the permutation.[4] For instance, given the standard-form permutation $(3\ 1\ 5\ 2\ 4)$, $v_2 = 2$ because 3 and 5 appear before 2; the entire inversion count vector would be $(1\ 2\ 0\ 1)$. Every vector of inversion counts uniquely identifies a single permutation.

The Distribution The GMM assigns probability mass according to the distance of a given permutation from the identity permutation $\{1, \ldots, K\}$, based on $K - 1$ real-valued parameters $(\rho_1, \ldots \rho_{K-1})$.[5] Using the inversion count representation of a permutation, the GMM's probability mass function is expressed as an independent product of probabilities for each v_j:

$$\begin{aligned} \text{GMM}(\mathbf{v} \mid \rho) &= \frac{e^{-\sum_j \rho_j v_j}}{\psi(\rho)} \\ &= \prod_{j=1}^{n-1} \frac{e^{-\rho_j v_j}}{\psi_j(\rho_j)}, \end{aligned} \tag{1}$$

where $\psi_j(\rho_j)$ is a normalization factor with value:

$$\psi_j(\rho_j) = \frac{1 - e^{-(K-j+1)\rho_j}}{1 - e^{-\rho_j}}.$$

[4] The sum of a vector of inversion counts is simply that permutation's Kendall's τ distance to the identity permutation.

[5] In our work we take the identity permutation to be the fixed centroid, which is a parameter in the full GMM. As we explain later, our model is not hampered by this apparent restriction.

Due to the exponential form of the distribution, requiring that $\rho_j > 0$ constrains the GMM to assign highest probability mass to each v_j being zero, corresponding to the identity permutation. A higher value for ρ_j assigns more probability mass to v_j being close to zero, biasing j to have fewer inversions.

The GMM elegantly captures our earlier requirement for a probability distribution that concentrates mass around a global ordering, and uses few parameters to do so. Because the topic numbers in our task are completely symmetric and not linked to any extrinsic observations, fixing the identity permutation to be that global ordering does not sacrifice any representational power. Another major benefit of the GMM is its membership in the exponential family of distributions; this means that it is particularly amenable to a Bayesian representation, as it admits a natural conjugate prior:

$$\mathrm{GMM}_0(\rho_j \mid v_{j,0}, \nu_0) \propto e^{(-\rho_j v_{j,0} - \log \psi_j(\rho_j))\nu_0}. \quad (2)$$

Intuitively, this prior states that over ν_0 prior trials, the total number of inversions was $\nu_0 v_{j,0}$. This distribution can be easily updated with the observed v_j to derive a posterior distribution.[6]

4.2 Formal Generative Process

We now fully specify the details of our model. We observe a corpus of D documents, each an ordered sequence of paragraphs, and a specification of a number of topics K. Each paragraph is represented as a bag of words. The model induces a set of hidden variables that probabilistically explain how the words of the corpus were produced. Our final desired output is the distributions over the paragraphs' hidden topic assignment variables. In the following, variables subscripted with 0 are fixed prior hyperparameters.

1. For each topic k, draw a language model $\theta_k \sim$ Dirichlet(θ_0). As with LDA, these are topic-specific word distributions.
2. Draw a topic distribution $\tau \sim$ Dirichlet(τ_0), which expresses how likely each topic is to appear regardless of position.
3. Draw the topic ordering distribution parameters $\rho_j \sim \mathrm{GMM}_0(\rho_0, \nu_0)$ for $j = 1$ to $K - 1$. These parameters control how rapidly probability mass decays for having more inversions for each topic. A separate ρ_j for every topic allows us to learn that some topics are more likely to be reordered than others.
4. For each document d with N_d paragraphs:
   (a) Draw a bag of topics $\mathbf{t}_d$ by sampling N_d times from Multinomial(τ).
   (b) Draw a topic ordering π_d by sampling a vector of inversion counts $\mathbf{v}_d \sim \mathrm{GMM}(\rho)$.
   (c) Compute the vector of topic assignments $\mathbf{z}_d$ for document d's paragraphs, by sorting $\mathbf{t}_d$ according to π_d.[7]
   (d) For each paragraph p in document d:
      i. Sample each word $w_{d,p,j}$ according to the language model of p: $w_{d,p,j} \sim$ Multinomial$(\theta_{z_{d,p}})$.

5 Inference

The variables that we aim to infer are the topic assignments z of each paragraph, which are determined by the bag of topics $\mathbf{t}$ and ordering π for each document. Thus, our goal is to estimate the marginal distributions of $\mathbf{t}$ and π given the document text.

We accomplish this inference task through Gibbs sampling (Bishop, 2006). A Gibbs sampler builds a Markov chain over the hidden variable state space whose stationary distribution is the actual posterior of the joint distribution. Each new sample is drawn from the distribution of a single variable conditioned on previous samples of the other variables. We can "collapse" the sampler by integrating over some of the hidden variables in the model, in effect reducing the state space of the Markov chain. Collapsed sampling has been previously demonstrated to be effective for LDA and its variants (Griffiths and Steyvers, 2004; Porteous et al., 2008; Titov and McDonald, 2008). Our sampler integrates over all but three sets

[6] Because each v_j has a different range, it is inconvenient to set the prior hyperparameters $v_{j,0}$ directly. In our work, we instead fix the mode of the prior distribution to a value ρ_0, which works out to setting $v_{j,0} = \frac{1}{\exp(\rho_0)-1} - \frac{K-j+1}{\exp((K-j+1)\rho_0)-1}$.

[7] Multiple permutations can contribute to the probability of a single document's topic assignments $\mathbf{z}_d$, if there are topics that do not appear in $\mathbf{t}_d$. As a result, our current formulation is biased toward assignments with fewer topics per document. In practice, we do not find this to negatively impact model performance.

of hidden variables: bags of topics $\mathbf{t}$, orderings π, and permutation inversion parameters ρ. After a burn-in period, we treat the last samples of $\mathbf{t}$ and π as a draw from the true posterior.

Document Probability As a preliminary step, consider how to calculate the probability of a single document's words $\mathbf{w}_d$ given the document's paragraph topic assignments $\mathbf{z}_d$, and other documents and their topic assignments. Note that this probability is decomposable into a product of probabilities over individual paragraphs, where paragraphs with different topics have conditionally independent word probabilities. Let $\mathbf{w}_{-d}$ and $\mathbf{z}_{-d}$ indicate the words and topic assignments to documents other than d, and W be the vocabulary size. The probability of the words in d is then:

$$\begin{aligned} &P(\mathbf{w}_d \mid \mathbf{z}, \mathbf{w}_{-d}, \theta_0) \\ &= \prod_{k=1}^{K} \int_{\theta_k} P(\mathbf{w}_d \mid \mathbf{z}_d, \theta_k) P(\theta_k \mid \mathbf{z}, \mathbf{w}_{-d}, \theta_0) d\theta_k \\ &= \prod_{k=1}^{K} \text{DCM}(\{\mathbf{w}_{d,i} : z_{d,i} = k\} \\ &\qquad \mid \{\mathbf{w}_{-d,i} : z_{-d,i} = k\}, \theta_0), \end{aligned} \tag{3}$$

where DCM$(\cdot)$ refers to the *Dirichlet compound multinomial* distribution, the result of integrating over multinomial parameters with a Dirichlet prior (Bernardo and Smith, 2000). For a Dirichlet prior with parameters $\alpha = (\alpha_1, \ldots, \alpha_W)$, the DCM assigns the following probability to a series of observations $\mathbf{x} = \{x_1, \ldots, x_n\}$:

$$\text{DCM}(\mathbf{x} \mid \alpha) = \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{i=1}^{W} \frac{\Gamma(N(\mathbf{x}, i) + \alpha_i)}{\Gamma(|\mathbf{x}| + \sum_j \alpha_j)},$$

where $N(\mathbf{x}, i)$ refers to the number of times word i appears in $\mathbf{x}$. Here, $\Gamma(\cdot)$ is the Gamma function, a generalization of the factorial for real numbers. Some algebra shows that the DCM's posterior probability density function conditioned on a series of observations $\mathbf{y} = \{y_1, \ldots, y_n\}$ can be computed by updating each α_i with counts of how often word i appears in $\mathbf{y}$:

$$\begin{aligned} &\text{DCM}(\mathbf{x} \mid \mathbf{y}, \alpha) \\ &= \text{DCM}(\mathbf{x} \mid \alpha_1 + N(\mathbf{y}, 1), \ldots, \alpha_W + N(\mathbf{y}, W)). \end{aligned} \tag{4}$$

Equation 3 and 4 will be used again to compute the conditional distributions of the hidden variables.

We now turn to a discussion of how each individual random variable is resampled.

Bag of Topics First we consider how to resample $t_{d,i}$, the ith topic draw for document d conditioned on all other parameters being fixed (note this is *not* the topic of the ith paragraph, as we reorder topics using π_d):

$$\begin{aligned} &P(t_{d,i} = t \mid \ldots) \\ &\propto P(t_{d,i} = t \mid \mathbf{t}_{-(d,i)}, \tau_0) P(\mathbf{w}_d \mid \mathbf{t}_d, \pi_d, \mathbf{w}_{-d}, \mathbf{z}_{-d}, \theta_0) \\ &\propto \frac{N(\mathbf{t}_{-(d,i)}, t) + \tau_0}{|\mathbf{t}_{-(d,i)}| + K\tau_0} P(\mathbf{w}_d \mid \mathbf{z}, \mathbf{w}_{-d}, \theta_0), \end{aligned}$$

where $\mathbf{t}_d$ is updated to reflect $t_{d,i} = t$, and $\mathbf{z}_d$ is deterministically computed by mapping $\mathbf{t}_d$ and π_d to actual paragraph topic assignments. The first step reflects an application of Bayes rule to factor out the term for $\mathbf{w}_d$. In the second step, the first term arises out of the DCM, by updating the parameters τ_0 with observations $\mathbf{t}_{-(d,i)}$ as in equation 4 and dropping constants. The document probability term is computed using equation 3. The new $t_{d,i}$ is selected by sampling from this probability computed over all possible topic assignments.

Ordering The parameterization of a permutation π as a series of inversion values v_j reveals a natural way to decompose the search space for Gibbs sampling. For a single ordering, each v_j can be sampled independently, according to:

$$\begin{aligned} &P(v_j = v \mid \ldots) \\ &\propto P(v_j = v \mid \rho_j) P(\mathbf{w}_d \mid \mathbf{t}_d, \pi_d, \mathbf{w}_{-d}, \mathbf{z}_{-d}, \theta_0) \\ &= \text{GMM}_j(v \mid \rho_j) P(\mathbf{w}_d \mid \mathbf{z}_d, \mathbf{w}_{-d}, \mathbf{z}_{-d}, \theta_0), \end{aligned}$$

where π_d is updated to reflect $v_j = v$, and $\mathbf{z}_d$ is computed according to $\mathbf{t}_d$ and π_d. The first term refers to the jth multiplicand of equation 1; the second is computed using equation 3. Term v_j is sampled according to the resulting probabilities.

GMM Parameters For each $j = 1$ to $K - 1$, we resample ρ_j from its posterior distribution:

$$\begin{aligned} &P(\rho_j \mid \ldots) \\ &= \text{GMM}_0\left(\rho_j \,\middle|\, \frac{\sum_i v_{j,i} + v_{j,0}\nu_0}{N + \nu_0}, N + \nu_0\right), \end{aligned}$$

where GMM_0 is evaluated according to equation 2. The normalization constant of this distribution is unknown, meaning that we cannot directly compute and invert the cumulative distribution function to sample from this distribution. However, the distribution itself is univariate and unimodal, so we can expect that an MCMC technique such as *slice sampling* (Neal, 2003) should perform well. In practice, the MATLAB black-box slice sampler provides a robust draw from this distribution.

6 Experimental Setup

Data Sets We evaluate our model on two data sets drawn from the English Wikipedia. The first set is 100 articles about large cities, with topics such as History, Culture, and Demographics. The second is 118 articles about chemical elements in the periodic table, including topics such as Biological Role, Occurrence, and Isotopes. Within each corpus, articles often exhibit similar section orderings, but many have idiosyncratic inversions. This structural variability arises out of the collaborative nature of Wikipedia, which allows articles to evolve independently. Corpus statistics are summarized below.

Corpus	Docs	Paragraphs	Vocab	Words
Cities	100	6,670	41,978	492,402
Elements	118	2,810	18,008	191,762

In each data set, the articles' *noisy section headings* induce a reference structure to compare against. This reference structure assumes that two paragraphs are aligned if and only if their section headings are identical, and that section boundaries provide the correct segmentation of each document. These headings are only used for evaluation, and are not provided to any of the systems.

Using the section headings to build the reference structure can be problematic, as the same topic may be referred to using different titles across different documents, and sections may be divided at differing levels of granularity. Thus, for the Cities data set, we manually annotated each article's paragraphs with a consistent set of section headings, providing us an additional reference structure to evaluate against. In this *clean section headings* set, we found approximately 18 topics that were expressed in more than one document.

Tasks and Metrics We study performance on the tasks of alignment and segmentation. In the former task, we measure whether paragraphs identified to be the same topic by our model have the same section headings, and vice versa. First, we identify the "closest" topic to each section heading, by finding the topic that is most commonly assigned to paragraphs under that section heading. We compute the proportion of paragraphs where the model's topic assignment matches the section heading's topic, giving us a *recall* score. High recall indicates that paragraphs of the same section headings are always being assigned to the same topic. Conversely, we can find the closest section heading to each topic, by finding the section heading that is most common for the paragraphs assigned to a single topic. We then compute the proportion of paragraphs from that topic whose section heading is the same as the reference heading for that topic, yielding a *precision* score. High precision means that paragraphs assigned to a single topic usually correspond to the same section heading. The harmonic mean of recall and precision is the summary *F-score*.

Statistical significance in this setup is measured with *approximate randomization* (Noreen, 1989), a nonparametric test that can be directly applied to nonlinear metrics such as F-score. This test has been used in prior evaluations for information extraction and machine translation (Chinchor, 1995; Riezler and Maxwell, 2005).

For the second task, we take the boundaries at which topics change within a document to be a segmentation of that document. We evaluate using the standard penalty metrics P_k and WindowDiff (Beeferman et al., 1999; Pevzner and Hearst, 2002). Both pass a sliding window over the documents and compute the probability of the words at the ends of the windows being improperly segmented with respect to each other. WindowDiff requires that the number of segmentation boundaries between the endpoints be correct as well.[8]

Our model takes a parameter K which controls the upper bound on the number of latent topics. Note that our algorithm can select fewer than K topics for each document, so K does not determine the number

[8] Statistical significance testing is not standardized and usually not reported for the segmentation task, so we omit these tests in our results.

of segments in each document. We report results using both $K = 10$ and 20 (recall that the cleanly annotated Cities data set had 18 topics).

Baselines and Model Variants We consider baselines from the literature that perform either alignment or segmentation. For the first task, we compare against the *hidden topic Markov model* (HTMM) (Gruber et al., 2007), which represents topic transitions between adjacent paragraphs in a Markovian fashion, similar to the approach taken in content modeling work. Note that HTMM can only capture local constraints, so it would allow topics to recur noncontiguously throughout a document.

We also compare against the structure-agnostic approach of clustering the paragraphs using the CLUTO toolkit,[9] which uses repeated bisection to maximize a cosine similarity-based objective.

For the segmentation task, we compare to BayesSeg (Eisenstein and Barzilay, 2008),[10] a Bayesian topic-based segmentation model that outperforms previous segmentation approaches (Utiyama and Isahara, 2001; Galley et al., 2003; Purver et al., 2006; Malioutov and Barzilay, 2006). BayesSeg enforces the topic contiguity constraint that motivated our model. We provide this baseline with the benefit of knowing the correct number of segments for each document, which is not provided to our system. Note that BayesSeg processes each document individually, so it cannot capture structural relatedness across documents.

To investigate the importance of our ordering model, we consider two variants of our model that alternately relax and tighten ordering constraints. In the *constrained* model, we require all documents to follow the same canonical ordering of topics. This is equivalent to forcing the topic permutation distribution to give all its probability to one ordering, and can be implemented by fixing all inversion counts $\mathbf{v}$ to zero during inference. At the other extreme, we consider the *uniform* model, which assumes a uniform distribution over all topic permutations instead of biasing toward a small related set. In our implementation, this can be simulated by forcing the GMM parameters ρ to always be zero. Both variants still enforce topic contiguity, and allow segments across documents to be aligned by topic assignment.

Evaluation Procedures For each evaluation of our model and its variants, we run the Gibbs sampler from five random seed states, and take the 10,000th iteration of each chain as a sample. Results shown are the average over these five samples. All Dirichlet prior hyperparameters are set to 0.1, encouraging sparse distributions. For the GMM, we set the prior decay parameter ρ_0 to 1, and the sample size prior ν_0 to be 0.1 times the number of documents.

For the baselines, we use implementations publicly released by their authors. We set HTMM's priors according to values recommended in the authors' original work. For BayesSeg, we use its built-in hyperparameter re-estimation mechanism.

7 Results

Alignment Table 1 presents the results of the alignment evaluation. In every case, the best performance is achieved using our full model, by a statistically significant and usually substantial margin.

In both domains, the baseline clustering method performs competitively, indicating that word cues alone are a good indicator of topic. While the simpler variations of our model achieve reasonable performance, adding the richer GMM distribution consistently yields superior results.

Across each of our evaluations, HTMM greatly underperforms the other approaches. Manual examination of the actual topic assignments reveals that HTMM often selects the same topic for disconnected paragraphs of the same document, violating the topic contiguity constraint, and demonstrating the importance of modeling global constraints for document structure tasks.

We also compare performance measured on the manually annotated section headings against the actual noisy headings. The ranking of methods by performance remains mostly unchanged between these two evaluations, indicating that the noisy headings are sufficient for gaining insight into the comparative performance of the different approaches.

Segmentation Table 2 presents the segmentation experiment results. On both data sets, our model

[9] http://glaros.dtc.umn.edu/gkhome/views/cluto/

[10] We do not evaluate on the corpora used in their work, since our model relies on content similarity across documents in the corpus.

		Cities: clean headings			Cities: noisy headings			Elements: noisy headings		
		Recall	Prec	F-score	Recall	Prec	F-score	Recall	Prec	F-score
$K = 10$	Clustering	0.578	0.439	$\ast$ 0.499	0.611	0.331	$\ast$ 0.429	0.524	0.361	$\ast$ 0.428
	HTMM	0.446	0.232	$\ast$ 0.305	0.480	0.183	$\ast$ 0.265	0.430	0.190	$\ast$ 0.264
	Constrained	0.579	0.471	$\ast$ 0.520	0.667	0.382	$\ast$ 0.485	0.603	0.408	$\ast$ 0.487
	Uniform	0.520	0.440	$\ast$ 0.477	0.599	0.343	$\ast$ 0.436	0.591	0.403	$\ast$ 0.479
	Our model	**0.639**	**0.509**	**0.566**	**0.705**	**0.399**	**0.510**	**0.685**	**0.460**	**0.551**
$K = 20$	Clustering	0.486	0.541	$\ast$ 0.512	0.527	0.414	$\ast$ 0.464	0.477	0.402	$\ast$ 0.436
	HTMM	0.260	0.217	$\ast$ 0.237	0.304	0.187	$\ast$ 0.232	0.248	0.243	$\ast$ 0.246
	Constrained	0.458	0.519	$\ast$ 0.486	0.553	0.415	$\ast$ 0.474	0.510	0.421	$\ast$ 0.461
	Uniform	0.499	0.551	$\ast$ 0.524	0.571	0.423	$\ast$ 0.486	0.550	0.479	$\diamond$ 0.512
	Our model	**0.578**	**0.636**	**0.606**	**0.648**	**0.489**	**0.557**	**0.569**	**0.498**	**0.531**

Table 1: Comparison of the alignments produced by our model and a series of baselines and model variations, for both 10 and 20 topics, evaluated against clean and noisy sets of section headings. Higher scores are better. Within the same K, the methods which our model significantly outperforms are indicated with $\ast$ for $p < 0.001$ and $\diamond$ for $p < 0.01$.

		Cities: clean headings			Cities: noisy headings			Elements: noisy headings		
		P_k	WD	# Segs	P_k	WD	# Segs	P_k	WD	# Segs
BayesSeg		0.321	0.376	† 12.3	0.317	0.376	† 13.2	0.279	0.316	† 7.7
$K = 10$	Constrained	0.260	**0.281**	7.7	0.267	0.288	7.7	0.227	0.244	5.4
	Uniform	0.268	0.300	8.8	0.273	0.304	8.8	0.226	0.250	6.6
	Our model	**0.253**	0.283	9.0	**0.257**	**0.286**	9.0	**0.201**	**0.226**	6.7
$K = 20$	Constrained	0.274	0.314	10.9	0.274	0.313	10.9	0.231	0.257	6.6
	Uniform	0.234	0.294	14.0	0.234	0.290	14.0	0.209	0.248	8.7
	Our model	**0.221**	**0.278**	14.2	**0.222**	**0.278**	14.2	**0.203**	**0.243**	8.6

Table 2: Comparison of the segmentations produced by our model and a series of baselines and model variations, for both 10 and 20 topics, evaluated against clean and noisy sets of section headings. Lower scores are better. †BayesSeg is given the true number of segments, so its segments count reflects the reference structure's segmentation.

outperforms the BayesSeg baseline by a substantial margin regardless of K. This result provides strong evidence that learning connected topic models over related documents leads to improved segmentation performance. In effect, our model can take advantage of shared structure across related documents.

In all but one case, the best performance is obtained by the full version of our model. This result indicates that enforcing discourse-motivated structural constraints allows for better segmentation induction. Encoding global discourse-level constraints leads to better language models, resulting in more accurate predictions of segment boundaries.

8 Conclusions

In this paper, we have shown how an unsupervised topic-based approach can capture document structure. Our resulting model constrains topic assignments in a way that requires global modeling of entire topic sequences. We showed that the generalized Mallows model is a theoretically and empirically appealing way of capturing the ordering component of this topic sequence. Our results demonstrate the importance of augmenting statistical models of text analysis with structural constraints motivated by discourse theory.

Acknowledgments

The authors acknowledge the funding support of NSF CAREER grant IIS-0448168, the NSF Graduate Fellowship, the Office of Naval Research, Quanta, Nokia, and the Microsoft Faculty Fellowship. We thank the members of the NLP group at MIT and numerous others who offered suggestions and comments on this work. We are especially grateful to Marina Meilă for introducing us to the Mallows model. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.

References

Regina Barzilay and Lillian Lee. 2004. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *Proceedings of NAACL/HLT*.

Regina Barzilay, Noemie Elhadad, and Kathleen McKeown. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34:177–210.

José M. Bernardo and Adrian F.M. Smith. 2000. *Bayesian Theory*. Wiley Series in Probability and Statistics.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

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

Nancy Chinchor. 1995. Statistical significance of MUC-6 results. In *Proceedings of the 6th Conference on Message Understanding*.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of EMNLP*.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Proceedings of NAACL/HLT*.

M.A. Fligner and J.S. Verducci. 1986. Distance based ranking models. *Journal of the Royal Statistical Society, Series B*, 48(3):359–369.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of ACL*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235.

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *Advances in NIPS*.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden topic markov models. In *Proceedings of AISTATS*.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman.

Nikiforos Karamanis, Massimo Poesio, Chris Mellish, and Jon Oberlander. 2004. Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. In *Proceedings of ACL*.

Mirella Lapata. 2003. Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of ACL*.

Guy Lebanon and John Lafferty. 2002. Cranking: combining rankings using conditional probability models on permutations. In *Proceedings of ICML*.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL*.

Marina Meilă, Kapil Phadnis, Arthur Patterson, and Jeff Bilmes. 2007. Consensus ranking under the exponential model. In *Proceedings of UAI*.

Radford M. Neal. 2003. Slice sampling. *Annals of Statistics*, 31:705–767.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28:19–36.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of SIGKDD*.

Matthew Purver, Konrad Körding, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of ACL/COLING*.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of ACL*.

Hanna M. Wallach. 2006. Topic modeling: beyond bag of words. In *Proceedings of ICML*.

Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press, Cambridge.