Bayesian Unsupervised Topic Segmentation

Jacob Eisenstein and **Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge MA 02139
{jacobe,regina}@csail.mit.edu

Abstract

This paper describes a novel Bayesian approach to unsupervised topic segmentation. Unsupervised systems for this task are driven by *lexical cohesion*: the tendency of well-formed segments to induce a compact and consistent lexical distribution. We show that lexical cohesion can be placed in a Bayesian context by modeling the words in each topic segment as draws from a multinomial language model associated with the segment; maximizing the observation likelihood in such a model yields a lexically-cohesive segmentation. This contrasts with previous approaches, which relied on hand-crafted cohesion metrics. The Bayesian framework provides a principled way to incorporate additional features such as cue phrases, a powerful indicator of discourse structure that has not been previously used in unsupervised segmentation systems. Our model yields consistent improvements over an array of state-of-the-art systems on both text and speech datasets. We also show that both an entropy-based analysis and a well-known previous technique can be derived as special cases of the Bayesian framework.[1]

1 Introduction

Topic segmentation is one of the fundamental problems in discourse analysis, where the task is to divide a text into a linear sequence of topically-coherent segments. Hearst's TEXTTILING (1994) introduced the idea that unsupervised segmentation can be driven by *lexical cohesion*, as high-quality segmentations feature homogeneous lexical distributions within each topic segment. Lexical cohesion has provided the inspiration for several successful systems (e.g., Utiyama and Isahara, 2001; Galley et al.2003; Malioutov and Barzilay, 2006), and is currently the dominant approach to unsupervised topic segmentation.

But despite the effectiveness of lexical cohesion for unsupervised topic segmentation, it is clear that there are other important indicators that are ignored by the current generation of unsupervised systems. For example, consider *cue phrases*, which are explicit discourse markers such as "now" or "however" (Grosz and Sidner, 1986; Hirschberg and Litman, 1993; Knott, 1996). Cue phrases have been shown to be a useful feature for supervised topic segmentation (Passonneau and Litman, 1993; Galley et al., 2003), but cannot be incorporated by current unsupervised models. One reason for this is that existing unsupervised methods use arbitrary, hand-crafted metrics for quantifying lexical cohesion, such as weighted cosine similarity (Hearst, 1994; Malioutov and Barzilay, 2006). Without supervision, it is not possible to combine such metrics with additional sources of information. Moreover, such hand-crafted metrics may not generalize well across multiple datasets, and often include parameters which must be tuned on development sets (Malioutov and Barzilay, 2006; Galley et al., 2003).

In this paper, we situate lexical cohesion in a Bayesian framework, allowing other sources of information to be incorporated without the need for labeled data. We formalize lexical cohesion in a generative model in which the text for each seg-

[1] Code and materials for this work are available at http://groups.csail.mit.edu/rbg/code/bayesseg/.

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 334–343,
Honolulu, October 2008.

ment is produced by a distinct lexical distribution. Lexically-consistent segments are favored by this model because probability mass is conserved for a narrow subset of words. Thus, lexical cohesion arises naturally through the generative process, and other sources of information – such as cue words – can easily be incorporated as emissions from the segment boundaries.

More formally, we treat the words in each sentence as draws from a language model associated with the topic segment. This is related to topic-modeling methods such as latent Dirichlet allocation (LDA; Blei et al. 2003), but here the induced topics are tied to a linear discourse structure. This property enables a dynamic programming solution to find the exact maximum-likelihood segmentation. We consider two approaches to handling the language models: estimating them explicitly, and integrating them out, using the Dirichlet Compound Multinomial distribution (also known as the multivariate Polya distribution).

We model cue phrases as generated from a separate multinomial that is shared across all topics and documents in the dataset; a high-likelihood model will obtain a compact set of cue phrases. The addition of cue phrases renders our dynamic programming-based inference inapplicable, so we design a sampling-based inference technique. This algorithm can learn in a completely unsupervised fashion, but it also provides a principled mechanism to improve search through the addition of declarative linguistic knowledge. This is achieved by biasing the selection of samples towards boundaries with known cue phrases; this does not change the underlying probabilistic model, but guides search in the direction of linguistically-plausible segmentations.

We evaluate our algorithm on corpora of spoken and written language, including the benchmark ICSI meeting dataset (Janin et al., 2003) and a new textual corpus constructed from the contents of a medical textbook. In both cases our model achieves performance surpassing multiple state-of-the-art baselines. Moreover, we demonstrate that the addition of cue phrases can further improve segmentation performance over cohesion-based methods.

In addition to the practical advantages demonstrated by these experimental results, our model reveals interesting theoretical properties. Other researchers have observed relationships between discourse structure and entropy (e.g., Genzel and Charniak, 2002). We show that in a special case of our model, the segmentation objective is equal to a weighted sum of the negative entropies for each topic segment. This finding demonstrates that a relationship between discourse segmentation and entropy is a natural consequence of modeling topic structure in a generative Bayesian framework. In addition, we show that the benchmark segmentation system of Utiyama and Isahara (2001) can be viewed as another special case of our Bayesian model.

2 Related Work

Existing unsupervised cohesion-based approaches can be characterized in terms of the metric used to quantify cohesion and the search technique. Galley et al. (2003) characterize cohesion in terms of lexical chains – repetitions of a given lexical item over some fixed-length window of sentences. In their unsupervised model, inference is performed by selecting segmentation points at the local maxima of the cohesion function. Malioutov and Barzilay (2006) optimize a normalized minimum-cut criteria based on a variation of the cosine similarity between sentences. Most similar to our work is the approach of Utiyama and Isahara (2001), who search for segmentations with compact language models; as shown in Section 3.1.1, this can be viewed as a special case of our model. Both of these last two systems use dynamic programming to search the space of segmentations.

An alternative Bayesian approach to segmentation was proposed by Purver et al. (2006). They assume a set of documents that is characterized by some number of hidden topics that are shared across multiple documents. They then build a linear segmentation by adding a switching variable to indicate whether the topic distribution for each sentence is identical to that of its predecessor. Unlike Purver et al., we do not assume a dataset in which topics are shared across multiple documents; indeed, our model can be applied to single documents individually. Additionally, the inference procedure of Purver et al. requires sampling multiple layers of hidden variables. In contrast, our inference procedure leverages the nature of linear segmentation to search only in the space of segmentation points.

The relationship between discourse structure and cue phrases has been studied extensively; for an early example of computational work on this topic, see (Grosz, 1977). Passonneau and Litman (1993) were the first to investigate the relationship between cue phrases and linear segmentation. More recently, cue phrases have been applied to topic segmentation in the supervised setting. In a supervised system that is distinct from the unsupervised model described above, Galley et al. (2003) automatically identify candidate cue phrases by mining labeled data for words that are especially likely to appear at segment boundaries; the presence of cue phrases is then used as a feature in a rule-based classifier for linear topic segmentation. Elsner and Charniak (2008) specify a list of cue phrases by hand; the cue phrases are used as a feature in a maximum-entropy classifier for conversation disentanglement. Unlike these approaches, we identify candidate cue phrases automatically from unlabeled data and incorporate them in the topic segmentation task without supervision.

3 Lexical Cohesion in a Bayesian Framework

The core idea of lexical cohesion is that topically-coherent segments demonstrate compact and consistent lexical distributions (Halliday and Hasan, 1976). Lexical cohesion can be placed in a probabilistic context by modeling the words in each topic segment as draws from a multinomial language model associated with the segment. Formally, if sentence t is in segment j, then the bag of words $\mathbf{x}_t$ is drawn from the multinomial language model θ_j. This is similar in spirit to hidden topic models such as latent Dirichlet allocation (Blei et al., 2003), but rather than assigning a hidden topic to each word, we constrain the topics to yield a linear segmentation of the document.

We will assume that topic breaks occur at sentence boundaries, and write z_t to indicate the topic assignment for sentence t. The observation likelihood is,

$$p(\mathbf{X}|\mathbf{z}, \Theta) = \prod_{t}^{T} p(\mathbf{x}_t|\theta_{z_t}), \tag{1}$$

where $\mathbf{X}$ is the set of all T sentences, $\mathbf{z}$ is the vector of segment assignments for each sentence, and Θ is the set of all K language models.[2] A linear segmentation is ensured by the additional constraint that z_t must be equal to either z_{t-1} (the previous sentence's segment) or $z_{t-1} + 1$ (the next segment).

To obtain a high likelihood, the language models associated with each segment should concentrate their probability mass on a compact subset of words. Language models that spread their probability mass over a broad set of words will induce a lower likelihood. This is consistent with the principle of lexical cohesion.

Thus far, we have described a segmentation in terms of two parameters: the segment indices $\mathbf{z}$, and the set of language models Θ. For the task of segmenting documents, we are interested only in the segment indices, and would prefer not to have to search in the space of language models as well. We consider two alternatives: taking point estimates of the language models (Section 3.1), and analytically marginalizing them out (Section 3.2).

3.1 Setting the language model to the posterior expectation

One way to handle the language models is to choose a single point estimate for each set of segmentation points $\mathbf{z}$. Suppose that each language model is drawn from a symmetric Dirichlet prior: $\theta_j \sim \text{Dir}(\theta_0)$. Let $\mathbf{n}_j$ be a vector in which each element is the sum of the lexical counts over all the sentences in segment j: $n_{j,i} = \sum_{\{t:z_t=j\}} m_{t,i}$, where $m_{t,i}$ is the count of word i in sentence t. Assuming that each $\mathbf{x}_t \sim \theta_j$, then the posterior distribution for θ_j is Dirichlet with vector parameter $\mathbf{n}_j + \theta_0$ (Bernardo and Smith, 2000). The expected value of this distribution is the multinomial distribution $\hat{\theta}_j$, where,

$$\hat{\theta}_{j,i} = \frac{n_{j,i} + \theta_0}{\sum_i^W n_{j,i} + W\theta_0}. \tag{2}$$

In this equation, W indicates the number of words in the vocabulary. Having obtained an estimate for the language model $\hat{\theta}_j$, the observed data likelihood for segment j is a product over each sentence in the segment,

[2] Our experiments will assume that the number of topics K is known. This is common practice for this task, as the desired number of segments may be determined by the user (Malioutov and Barzilay, 2006).

$$p(\{\mathbf{x}_t : z_t = j\}|\hat{\theta}_j) = \prod_{\{t:z_t=j\}} \prod_{i \in \mathbf{x}_t} \hat{\theta}_{j,i} \quad (3)$$

$$= \prod_{\{t:z_t=j\}} \prod_{i}^{W} \hat{\theta}_{j,i}^{m_{t,i}} \quad (4)$$

$$= \prod_{i}^{W} \hat{\theta}_{j,i}^{n_{j,i}}. \quad (5)$$

By viewing the likelihood as a product over all terms in the vocabulary, we observe interesting connections with prior work on segmentation and information theory.

3.1.1 Connection to previous work

In this section, we explain how our model generalizes the well-known method of Utiyama and Isahara (2001; hereafter U&I). As in our work, Utiyama and Isahara propose a probabilistic framework based on maximizing the compactness of the language models induced for each segment. Their likelihood equation is identical to our equations 3-5. They then define the language models for each segment as $\hat{\theta}_{j,i} = \frac{n_{j,i}+1}{W+\sum_i^W n_{j,i}}$, without rigorous justification. This form is equivalent to Laplacian smoothing (Manning and Schütze, 1999), and is a special case of our equation 2, with $\theta_0 = 1$. Thus, the language models in U&I can be viewed as the expectation of the posterior distribution $p(\theta_j|\{\mathbf{x}_t : z_t = j\}, \theta_0)$, in the special case that $\theta_0 = 1$. Our approach generalizes U&I and provides a Bayesian justification for the language models that they apply. The remainder of the paper further extends this work by marginalizing out the language model, and by adding cue phrases. We empirically demonstrate that these extensions substantially improve performance.

3.1.2 Connection to entropy

Our model also has a connection to entropy, and situates entropy-based segmentation within a Bayesian framework. Equation 1 defines the objective function as a product across sentences; using equations 3-5 we can decompose this across segments instead. Working in logarithms,

$$\log p(\mathbf{X}|\mathbf{z}, \hat{\Theta}) = \sum_{t}^{T} \log p(\mathbf{x}_t|\hat{\theta}_{z_t})$$

$$= \sum_{j}^{K} \sum_{\{t:z_t=j\}} \log p(\mathbf{x}_t|\hat{\theta}_j)$$

$$= \sum_{j}^{K} \sum_{i}^{W} n_{j,i} \log \hat{\theta}_{j,i} \quad (6)$$

The last line substitutes in the logarithm of equation 5. Setting $\theta_0 = 0$ and rearranging equation 2, we obtain $n_{j,i} = N_j \hat{\theta}_{j,i}$, with $N_j = \sum_i^W n_{j,i}$, the total number of words in segment j. Substituting this into equation 6, we obtain

$$\log p(\mathbf{X}|\mathbf{z}, \hat{\Theta}) = \sum_{j}^{K} N_j \sum_{i} \hat{\theta}_{j,i} \log \hat{\theta}_{j,i}$$

$$= \sum_{j}^{K} N_j H(\hat{\theta}_j),$$

where $H(\hat{\theta}_j)$ is the negative entropy of the multinomial $\hat{\theta}_j$. Thus, with $\theta_0 = 0$, the log conditional probability in equation 6 is optimized by a segmentation that minimizes the weighted sum of entropies per segment, where the weights are equal to the segment lengths. This result suggests intriguing connections with prior work on the relationship between entropy and discourse structure (e.g., Genzel and Charniak, 2002; Sporleder and Lapata, 2006).

3.2 Marginalizing the language model

The previous subsection uses point estimates of the language models to reveal connections to entropy and prior work on segmentation. However, point estimates are theoretically unsatisfying from a Bayesian perspective, and better performance may be obtained by marginalizing over all possible lan-

guage models:

$$p(\mathbf{X}|\mathbf{z},\theta_0) = \prod_j^K \prod_{\{t:z_t=j\}} p(\mathbf{x}_t|\theta_0)$$
$$= \prod_j^K \int d\theta_j \prod_{\{t:z_t=j\}} p(x_t|\theta_j)p(\theta_j|\theta_0)$$
$$= \prod_j^K p_{dcm}(\{\mathbf{x}_t : z_t = j\}|\theta_0), \quad (7)$$

where p_{dcm} refers to the Dirichlet compound multinomial distribution (DCM), also known as the multivariate Polya distribution (Johnson et al., 1997). The DCM distribution expresses the expectation over all multinomial language models, when conditioning on the Dirichlet prior θ_0. When θ_0 is a symmetric Dirichlet prior,

$$p_{dcm}(\{\mathbf{x}_t : z_t = j\}|\theta_0)$$
$$= \frac{\Gamma(W\theta_0)}{\Gamma(N_j + W\theta_0)} \prod_i^W \frac{\Gamma(n_{j,i} + W\theta_0)}{\Gamma(\theta_0)},$$

where $n_{j,i}$ is the count of word i in segment j, and $N_j = \sum_i^W n_{j,i}$, the total number of words in the segment. The symbol Γ refers to the Gamma function, an extension of the factorial function to real numbers. Using the DCM distribution, we can compute the data likelihood for each segment from the lexical counts over the entire segment. The overall observation likelihood is a product across the likelihoods for each segment.

3.3 Objective function and inference

The optimal segmentation maximizes the joint probability,

$$p(\mathbf{X}, \mathbf{z}|\theta_0) = p(\mathbf{X}|\mathbf{z}, \theta_0)p(\mathbf{z}).$$

We assume that $p(\mathbf{z})$ is a uniform distribution over valid segmentations, and assigns no probability mass to invalid segmentations. The data likelihood is defined for point estimate language models in equation 5 and for marginalized language models in equation 7. Note that equation 7 is written as a product over segments. The point estimates for the language models depend only on the counts within each segment, so the overall likelihood for the point-estimate version also decomposes across segments.

Any objective function that can be decomposed into a product across segments can be maximized using dynamic programming. We define $B(t)$ as the value of the objective function for the optimal segmentation up to sentence t. The contribution to the objective function from a single segment between sentences t' and t is written,

$$b(t', t) = p(\{\mathbf{x}_{t'} \dots \mathbf{x}_t\}|\mathbf{z}_{t'\dots t} = j)$$

The maximum value of the objective function is then given by the recurrence relation, $B(t) = \max_{t'<t} B(t')b(t'+1, t)$, with the base case $B(0) = 1$. These values can be stored in a table of size T (equal to the number of sentences); this admits a dynamic program that performs inference in polynomial time.[3] If the number of segments is specified in advance, the dynamic program is slightly more complex, with a table of size TK.

3.4 Priors

The Dirichlet compound multinomial integrates over language models, but we must still set the prior θ_0. We can re-estimate this prior based on the observed data by interleaving gradient-based search in a Viterbi expectation-maximization framework (Gauvain and Lee, 1994). In the E-step, we estimate a segmentation $\hat{\mathbf{z}}$ of the dataset, as described in Section 3.3. In the M-step, we maximize $p(\theta_0|\mathbf{X}, \hat{\mathbf{z}}) \propto p(\mathbf{X}|\theta_0, \hat{\mathbf{z}})p(\theta_0)$. Assuming a non-informative hyperprior $p(\theta_0)$, we maximize the likelihood in Equation 7 across all documents. The maximization is performed using a gradient-based search; the gradients are dervied by Minka (2003). This procedure is iterated until convergence or a maximum of twenty iterations.

4 Cue Phrases

One of the key advantages of a Bayesian framework for topic segmentation is that it permits the principled combination of multiple data sources, even

[3] This assumes that the objective function for individual segments can also be computed efficiently. In our case, we need only keep vectors of counts for each segment, and evaluate probability density functions over the counts.

without labeled data. We are especially interested in cue phrases, which are explicit markers for discourse structure, such as "now" or "first" (Grosz and Sidner, 1986; Hirschberg and Litman, 1993; Knott, 1996). Cue phrases have previously been used in supervised topic segmentation (e.g., Galley et al. 2003); we show how they can be used in an unsupervised setting.

The previous section modeled lexical cohesion by treating the bag of words in each sentence as a series of draws from a multinomial language model indexed by the topic segment. To incorporate cue phrases, this generative model is modified to reflect the idea that some of the text will be topic-specific, but other terms will be topic-neutral cue phrases that express discourse structure. This idea is implemented by drawing the text at each topic boundary from a special language model ϕ, which is shared across all topics and all documents in the dataset.

For sentences that are not at segment boundaries, the likelihood is as before: $p(\mathbf{x}_t|\mathbf{z}, \Theta, \phi) = \prod_{i \in \mathbf{x}_t} \theta_{z_t,i}$. For sentences that immediately follow segment boundaries, we draw the first ℓ words from ϕ instead. Writing $\mathbf{x}_t^{(\ell)}$ for the ℓ cue words in $\mathbf{x}_t$, and $\tilde{\mathbf{x}}_t$ for the remaining words, the likelihood for a segment-initial sentence is,

$$p(\mathbf{x}_t|z_t \neq z_{t-1}, \Theta, \phi) = \prod_{i \in \mathbf{x}_t^{(\ell)}} \phi_i \prod_{i \in \tilde{\mathbf{x}}_t} \theta_{z_t,i}.$$

We draw ϕ from a symmetric Dirichlet prior ϕ_0. Following prior work (Galley et al., 2003; Litman and Passonneau, 1995), we consider only the first word of each sentence as a potential cue phrase; thus, we set $\ell = 1$ in all experiments.

4.1 Inference

To estimate or marginalize the language models Θ and ϕ, it is necessary to maintain lexical counts for each segment and for the segment boundaries. The counts for ϕ are summed across every segment in the entire dataset, so shifting a boundary will affect the probability of every segment, not only the adjacent segments as before. Thus, the factorization that enabled dynamic programming inference in Section 3.3 is no longer applicable. Instead, we must resort to approximate inference.

Sampling-based inference is frequently used in related Bayesian models. Such approaches build a stationary Markov chain by repeatedly sampling among the hidden variables in the model. The most commonly-used sampling-based technique is Gibbs sampling, which iteratively samples from the conditional distribution of each hidden variable (Bishop, 2006). However, Gibbs sampling is slow to converge to a stationary distribution when the hidden variables are tightly coupled. This is the case in linear topic segmentation, due to the constraint that $z_t \in \{z_{t-1}, z_{t-1} + 1\}$ (see Section 3).

For this reason, we apply the more general Metropolis-Hastings algorithm, which permits sampling arbitrary transformations of the latent variables. In our framework, such transformations correspond to moves through the space of possible segmentations. A new segmentation $\mathbf{z}'$ is drawn from the previous hypothesized segmentation $\mathbf{z}$ based on a *proposal distribution* $q(\mathbf{z}'|\mathbf{z})$.[4] The probability of accepting a proposed transformation depends on the ratio of the joint probabilities and a correction term for asymmetries in the proposal distribution:

$$p_{\text{accept}}(\mathbf{z} \rightarrow \mathbf{z}') = \min \left\{ 1, \frac{p(\mathbf{X}, \mathbf{z}'|\theta_0, \phi_0)}{p(\mathbf{X}, \mathbf{z}|\theta_0, \phi_0)} \frac{q(\mathbf{z}|\mathbf{z}')}{q(\mathbf{z}'|\mathbf{z})} \right\}.$$

The Metropolis-Hastings algorithm guarantees that by accepting samples at this ratio, our sampling procedure will converge to the stationary distribution for the hidden variables $\mathbf{z}$. When cue phrases are included, the observation likelihood is written:

$$p(\mathbf{X}|\mathbf{z}, \Theta, \phi) = \prod_{\{t: z_t \neq z_{t-1}\}} \prod_{i \in \mathbf{x}_t^{(\ell)}} \phi_i \prod_{i \in \tilde{\mathbf{x}}_t} \theta_{z_t,i} \times \prod_{\{t: z_t = z_{t-1}\}} \prod_{i \in \mathbf{x}_t} \theta_{z_t,i}.$$

As in Section 3.2, we can marginalize over the language models. We obtain a product of DCM distributions: one for each segment, and one for all cue phrases in the dataset.

4.2 Proposal distribution

Metropolis-Hastings requires a proposal distribution to sample new configurations. The proposal distri-

[4] Because the cue phrase language model ϕ is used across the entire dataset, transformations affect the likelihood of all documents in the corpus. For clarity, our exposition will focus on the single-document case.

bution does not affect the underlying probabilistic model – Metropolis-Hastings will converge to the same underlying distribution for any non-degenerate proposal. However, a well-chosen proposal distribution can substantially speed convergence.

Our basic proposal distribution selects an existing segmentation point with uniform probability, and considers a set of local moves. The proposal is constructed so that no probability mass is allocated to moves that change the order of segment boundaries, or merge two segments; one consequence of this restriction is that moves cannot add or remove segments.[5] We set the proposal distribution to decrease exponentially with the move distance, thus favoring incremental transformations to the segmentation.

More formally, let $d(\mathbf{z} \rightarrow \mathbf{z}') > 0$ equal the distance that the selected segmentation point is moved when we transform the segmentation from $\mathbf{z}$ to $\mathbf{z}'$. We can write the proposal distribution $q(\mathbf{z}' \mid \mathbf{z}) \propto c(\mathbf{z} \rightarrow \mathbf{z}')d(\mathbf{z} \rightarrow \mathbf{z}')^{\lambda}$, where $\lambda < 0$ sets the rate of exponential decay and c is an indicator function enforcing the constraint that the moves do not reach or cross existing segmentation points.[6]

We can also incorporate declarative linguistic knowledge by biasing the proposal distribution in favor of moves that place boundaries near known cue phrase markers. We multiply the unnormalized chance of proposing a move to location $\mathbf{z} \rightarrow \mathbf{z}'$ by a term equal to one plus the number of candidate cue phrases in the segment-initial sentences in the new configuration $\mathbf{z}'$, written $\textsf{num-cue}(\mathbf{z}')$. Formally, $q_{\text{ling}}(\mathbf{z}' \mid \mathbf{z}') \propto (1 + \textsf{num-cue}(\mathbf{z}'))q(\mathbf{z}' \mid \mathbf{z})$. We use a list of cue phrases identified by Hirschberg and Litman (1993). We evaluate our model with both the basic and linguistically-enhanced proposal distributions.

4.3 Priors

As in section 3.4, we set the priors θ_0 and ϕ_0 using gradient-based search. In this case, we perform gradient-based optimization after epochs of 1000 Metropolis-Hasting steps. Interleaving sampling-based inference with direct optimization of parameters can be considered a form of Monte Carlo Expectation-Maximization (MCEM; Wei and Tanner, 1990).

5 Experimental Setup

Corpora We evaluate our approach on corpora from two different domains: transcribed meetings and written text.

For multi-speaker meetings, we use the ICSI corpus of meeting transcripts (Janin et al., 2003), which is becoming a standard for speech segmentation (e.g., Galley et al. 2003; Purver et al. 2006). This dataset includes transcripts of 75 multi-party meetings, of which 25 are annotated for segment boundaries.

For text, we introduce a dataset in which each document is a chapter selected from a medical textbook (Walker et al., 1990).[7] The task is to divide each chapter into the sections indicated by the author. This dataset contains 227 chapters, with 1136 sections (an average of 5.00 per chapter). Each chapter contains an average of 140 sentences, giving an average of 28 sentences per segment.

Metrics All experiments are evaluated in terms of the commonly-used P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002) scores. Both metrics pass a window through the document, and assess whether the sentences on the edges of the window are properly segmented with respect to each other. WindowDiff is stricter in that it requires that the number of intervening segments between the two sentences be identical in the hypothesized and the reference segmentations, while P_k only asks whether the two sentences are in the same segment or not. P_k and WindowDiff are penalties, so lower values indicate better segmentations. We use the evaluation source code provided by Malioutov and Barzilay (2006).

System configuration We evaluate our Bayesian approach both with and without cue phrases. Without cue phrases, we use the dynamic programming inference described in section 3.3. This system is referred to as BAYESSEG in Table 1. When adding

[5] Permitting moves to change the number of segments would substantially complicate inference.

[6] We set $\lambda = -\frac{1}{\textsf{max-move}}$, where $\textsf{max-move}$ is the maximum move-length, set to 5 in our experiments. These parameters affect the rate of convergence but are unrelated to the underlying probability model. In the limit of enough samples, all non-pathological settings will yield the same segmentation results.

[7] The full text of this book is available for free download at http://onlinebooks.library.upenn.edu.

cue phrases, we use the Metropolis-Hastings model described in 4.1. Both basic and linguistically-motivated proposal distributions are evaluated (see Section 4.2); these are referred to as BAYESSEG-CUE and BAYESSEG-CUE-PROP in the table.

For the sampling-based systems, results are averaged over five runs. The initial configuration is obtained from the dynamic programming inference, and then 100,000 sampling iterations are performed. The final segmentation is obtained by annealing the last 25,000 iterations to a temperature of zero. The use of annealing to obtain a maximum *a posteriori* (MAP) configuration from sampling-based inference is common (e.g., Finkel 2005; Goldwater 2007). The total running time of our system is on the order of three minutes per document. Due to memory constraints, we divide the textbook dataset into ten parts, and perform inference in each part separately. We may achieve better results by performing inference over the entire dataset simultaneously, due to pooling counts for cue phrases across all documents.

Baselines We compare against three competitive alternative systems from the literature: U&I (Utiyama and Isahara, 2001); LCSEG (Galley et al., 2003); MCS (Malioutov and Barzilay, 2006). All three systems are described in the related work (Section 2). In all cases, we use the publicly available executables provided by the authors.

Parameter settings For LCSEG, we use the parameter values specified in the paper (Galley et al., 2003). MCS requires parameter settings to be tuned on a development set. Our corpora do not include development sets, so tuning was performed using the lecture transcript corpus described by Malioutov and Barzilay (2006). Our system does not require parameter tuning; priors are re-estimated as described in Sections 3.4 and 4.3. U&I requires no parameter tuning, and is used "out of the box." In all experiments, we assume that the number of desired segments is provided.

Preprocessing Standard preprocessing techniques are applied to the text for all comparisons. The Porter (1980) stemming algorithm is applied to group equivalent lexical items. A set of stop-words is also removed, using the same list originally employed by several competitive systems (Choi, 2000; Utiyama and Isahara, 2001; Malioutov and Barzilay, 2006).

Textbook	P_k	WD
U&I	.370	.376
MCS	.368	.382
LCSEG	.370	.385
BAYESSEG	**.339**	**.353**
BAYESSEG-CUE	**.339**	**.353**
BAYESSEG-CUE-PROP	.343	.355
Meetings	P_k	WD
U&I	.297	.347
MCS	.370	.411
LCSEG	.309	.322
BAYESSEG	.264	.319
BAYESSEG-CUE	.261	.316
BAYESSEG-CUE-PROP	**.258**	**.312**

Table 1: Comparison of segmentation algorithms. Both metrics are penalties, so lower scores indicate better performance. BAYESSEG is the cohesion-only Bayesian system with marginalized language models. BAYESSEG-CUE is the Bayesian system with cue phrases. BAYESSEG-CUE-PROP adds the linguistically-motivated proposal distribution.

6 Results

Table 1 presents the performance results for three instantiations of our Bayesian framework and three competitive alternative systems. As shown in the table, the Bayesian models achieve the best results on both metrics for both corpora. On the medical textbook corpus, the Bayesian systems achieve a raw performance gain of 2-3% with respect to all baselines on both metrics. On the ICSI meeting corpus, the Bayesian systems perform 4-5% better than the best baseline on the P_k metric, and achieve smaller improvement on the WindowDiff metric. The results on the meeting corpus also compare favorably with the topic-modeling method of Purver et al. (2006), who report a P_k of .289 and a WindowDiff of .329.

Another observation from Table 1 is that the contribution of cue phrases depends on the dataset. Cue phrases improve performance on the meeting corpus, but not on the textbook corpus. The effectiveness of cue phrases as a feature depends on whether the writer or speaker uses them consistently. At the

Meetings		Textbook	
okay*	234.4	**the**	1345.9
I	212.6	**this**	14.3
so*	113.4	it	4.1
um	91.7	these	4.1
and*	67.3	a	2.9
yeah	10.5	on	2.1
but*	9.4	most	2.0
uh	4.8	heart	1.8
right	2.4	creating	1.8
agenda	1.3	hundred	1.8

Table 2: Cue phrases selected by our unsupervised model, sorted by chi-squared. Boldface indicates that the chi-squared value is significant at the level of $p < .01$. Asterisks indicate cue phrases that were extracted by the supervised procedure of Galley et al. (2003).

same time, the addition of cue phrases prevents the use of exact inference techniques, which may explain the decline in results for the meetings dataset.

To investigate the quality of the cue phrases that our model extracts, we list its top ten cue phrases for each dataset in Table 2. Cue phrases are ranked by their chi-squared value, which is computed based on the number of occurrences for each word at the beginning of a hypothesized segment, as compared to the expectation. For cue phrases listed in bold, the chi-squared value is statistically significant at the level of $p < .01$, indicating that the frequency with which the cue phrase appears at the beginning of segments is unlikely to be a chance phenomenon.

As shown in the left column of the table, our model has identified several strong cue phrases from the meeting dataset which appear to be linguistically plausible. Galley et al. (2003) performed a similar chi-squared analysis, but used the true segment boundaries in the labeled data; this can be thought of as a sort of ground truth. Four of the ten cue phrases identified by our system overlap with their analysis; these are indicated with asterisks. In contrast to our model's success at extracting cue phrases from the meeting dataset, only very common words are selected for the textbook dataset. This may help to explain why cue phrases improve performance for meeting transcripts, but not for the textbook.

7 Conclusions

This paper presents a novel Bayesian approach to unsupervised topic segmentation. Our algorithm is capable of incorporating both lexical cohesion and cue phrase features in a principled manner, and outperforms state-of-the-art baselines on text and transcribed speech corpora. We have developed exact and sampling-based inference techniques, both of which search only over the space of segmentations and marginalize out the associated language models. Finally, we have shown that our model provides a theoretical framework with connections to information theory, while also generalizing and justifying prior work. In the future, we hope to explore the use of similar Bayesian techniques for hierarchical segmentation, and to incorporate additional features such as prosody and speaker change information.

Acknowledgments

The authors acknowledge the support of the National Science Foundation (CAREER grant IIS-0448168) and the Microsoft Research Faculty Fellowship. Thanks to Aaron Adler, S. R. K. Branavan, Harr Chen, Michael Collins, Randall Davis, Dan Roy, David Sontag and the anonymous reviewers for helpful comments and suggestions. We also thank Michel Galley, Igor Malioutov, and Masao Utiyama for making their topic segmentation code publically available. Any opinions, findings, and conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the NSF.

References

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

José M. Bernardo and Adrian F. M. Smith. 2000. *Bayesian Theory*. Wiley.

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

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

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL*, pages 26–33.

Micha Elsner and Eugene Charniak. 2008. You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In *Proceedings of ACL*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. *Proceedings of ACL*, pages 562–569.

Jean-Luc Gauvain and Chin-Hui Lee. 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of ACL*, pages 199–206.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of ACL*, pages 744–751.

Barbara Grosz and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Barbara Grosz. 1977. The representation and use of focus in dialogue understanding. Technical Report 151, Artificial Intelligence Center, SRI International.

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

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of ACL*, pages 9–16.

Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, et al. 2003. The ICSI Meeting Corpus. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 1.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. 1997. *Discrete Multivariate Distributions*. Wiley.

Alistair Knott. 1996. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. Ph.D. thesis, University of Edinburgh.

Diane J. Litman and Rebecca J. Passonneau. 1995. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the ACL*, pages 108–115.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL*, pages 25–32.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Thomas P. Minka. 2003. Estimating a dirichlet distribution. Technical report, Massachusetts Institute of Technology.

Rebecca Passonneau and Diane Litman. 1993. Intention-based segmentation: Human reliability and correlation with linguistic cues. In *Proceedings of ACL*, pages 148–155.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

M. F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14:130–137.

M. Purver, T.L. Griffiths, K.P. Körding, and J.B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of ACL*, pages 17–24.

Caroline Sporleder and Mirella Lapata. 2006. Broad coverage paragraph segmentation across languages and domains. *ACM Transactions on Speech and Language Processing*, 3(2):1–35.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of ACL*, pages 491–498.

H. Kenneth Walker, W. Dallas Hall, and J. Willis Hurst, editors. 1990. *Clinical Methods : The History, Physical, and Laboratory Examinations*. Butterworths.

Greg C. G. Wei and Martin A. Tanner. 1990. A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411), September.