# Topic Segmentation with a Structured Topic Model

**Lan Du**
Department of Computing
Macquarie University
Sydney, Australia
lan.du@mq.edu.au

**Wray Buntine**
Canberra Research Lab
National ICT Australia
Canberra, Australia
wray.buntine@nicta.com.au

**Mark Johnson**
Department of Computing
Macquarie University
Sydney, Australia
mark.johnson@mq.edu.au

## Abstract

We present a new hierarchical Bayesian model for unsupervised topic segmentation. This new model integrates a point-wise boundary sampling algorithm used in Bayesian segmentation into a structured topic model that can capture a simple hierarchical topic structure latent in documents. We develop an MCMC inference algorithm to split/merge segment(s). Experimental results show that our model outperforms previous unsupervised segmentation methods using only lexical information on Choi's datasets and two meeting transcripts and has performance comparable to those previous methods on two written datasets.

## 1 Introduction

Documents are usually comprised of topically coherent text segments, each of which contains some number of text passages (*e.g.*, sentences or paragraphs) (Salton et al., 1996). Within each topically coherent segment, one would expect that the word usage demonstrates more consistent lexical distributions (known as lexical cohesion (Eisenstein and Barzilay, 2008)) than that across segments. A linear partition of texts into topic segments may reveal information about, for example, themes of segments and the overall thematic structure of the text, and can subsequently be useful for text analysis tasks, such as information retrieval (*e.g.*, passage retrieval (Salton et al., 1996)), document summarisation and discourse analysis (Galley et al., 2003).

In this paper we consider how to automatically find a topic segmentation. It involves identifying the most prominent topic changes in a sequence of text passages, and splits those passages into a sequence of topically coherent segments (Hearst, 1997; Beeferman et al., 1999). This task can be cast as an unsupervised machine learning problem: placing topic boundaries in unannotated text.

Although a variety of cues in text can be used for topic segmentation, such as cue phases (Beeferman et al., 1999; Reynar, 1999; Eisenstein and Barzilay, 2008)) and discourse information (Galley et al., 2003), in this paper, we focus on lexical cohesion and use it as the primary cue in developing an unsupervised segmentation model. The effectiveness of lexical cohesion has been demonstrated by TextTiling (Hearst, 1997), c99 (Choi, 2000), MinCut (Malioutov and Barzilay, 2006), PLDA (Purver et al., 2006), Bayesseg (Eisenstein and Barzilay, 2008), TopicTiling (Riedl and Biemann, 2012), *etc.*

Our work uses recent progress in hierarchical topic modelling with non-parametric Bayesian methods (Du et al., 2010; Chen et al., 2011; Du et al., 2012a), and is based on Bayesian segmentation methods (Goldwater et al., 2009; Purver et al., 2006; Eisenstein and Barzilay, 2008) using topic models. This can also be viewed as a multi-topic extension of hierarchical Bayesian segmentation (Eisenstein, 2009), although our use of hierarchies is used to improve the performance of linear segmentation, rather than develop hierarchical segmentation.

Recently, topic models are increasingly used in various text analysis tasks including topic segmentation. Previous work (Purver et al., 2006; Misra et al., 2008; Sun et al., 2008; Misra et al., 2009; Riedl and Biemann, 2012) has shown that using

*Proceedings of NAACL-HLT 2013*, pages 190–200,
Atlanta, Georgia, 9–14 June 2013. ©2013 Association for Computational Linguistics

topic assignments or topic distributions instead of word frequency can significantly improve segmentation performance. Here we consider more advanced topic models that model dependencies between (sub-)sections in a document, such as structured topic models (STMs) presented in (Du et al., 2010; Du et al., 2012b). STMs treat each text as a sequence of segments, each of which is a set of text passages (*e.g.*, a paragraph or sentence). Text passages in a segment share the same prior distribution on their topics. The topic distributions of segments in a single document are then encouraged to be similar via a hierarchical prior. This gives a substantial improvement in modelling accuracy. However, instead of explicitly learning the segmentation, STMs just leverage the existing structure of documents from the given segmentation.

Given a sequence of text passages, how can we automatically learn the segmentation? The word boundary sampling algorithm introduced in (Goldwater et al., 2009) uses point-wise sampling of word boundaries after phonemes in an utterance. Similarly, the segmentation method of PLDA (Purver et al., 2006) samples segment boundaries, but also jointly samples a topic model. This is different to other topic modelling approaches that run LDA as a precursor to a separate segmentation step (Misra et al., 2009; Riedl and Biemann, 2012). While conceptually similar to PLDA, our non-parametric approach built on STM required new methods to implement, but the resulting improvement by the standard segmentation scores is substantial.

This paper presents a new hierarchical Bayesian unsupervised topic segmentation model, integrating a point-wise boundary sampling algorithm with a structured topic model. This new model takes advantage of the high modelling accuracy of structured topic models (Du et al., 2010) to produce a topic segmentation based on the distribution of latent topics. We show that this model provides high quality segmentation performance on Choi's dataset, as well as two sets of meeting transcripts and written texts.

In the following sections we describe our topic segmentation model and an MCMC inference algorithm for the non-parametric split/merge process. The rest of the paper is organised as follows. In Section 2 we review recent related work in the topic segmentation literature. Section 3 presents the new topic segmentation model, followed by the derivation of a sampling algorithm in Section 4. We report the experimental results by comparing several related topic segmentation methods in Section 5. Section 6 concludes the paper.

## 2 Related Work

We are interested in unsupervised topic segmentation in either written or spoken language. There is a large body of work on unsupervised topic segmentation of text based on lexical cohesion. It can be characterised by how lexical cohesion is modelled.

One branch of this work represents the lexical cohesion in a vector space by exploring the word co-occurrence patterns, *e.g.*, TF or TF-IDF. Work following this line includes TextTiling (Hearst, 1997), which calculates the cosine similarity between two adjacent blocks of words purely based on the word frequency; C99 (Choi, 2000), an algorithm based on divisive clustering with a matrix-ranking scheme; LSeg (Galley et al., 2003), which uses a lexical chain to identify and weight word repetitions; U00 (Utiyama and Isahara, 2001), a probalistic approach using dynamic programming to find a segmentation with a minimum cost; MinCut (Malioutov and Barzilay, 2006), which casts segmentation as a graph cut problem, and APS (Kazantseva and Szpakowicz, 2011), which uses affinity propagation to learn clustering for segmentation.

The other branch of this work characterises the lexical cohesion using topic models, to which the model introduced in Section 3 belongs. Lexical cohesion in this line of research is modelled by a probabilistic generative process. PLDA presented by Purver *et al.* (2006) is an unsupervised topic modelling approach for segmentation. It chains a set of LDAs (Blei et al., 2003) by assuming a Markov structure on topic distributions. A binary topic shift variable is attached to each text passage (*i.e.*, an utterance in (Purver et al., 2006)). It is sampled to indicate whether the $j^{th}$ text passage shares the topic distribution with the $(j-1)^{th}$ passage.

Using a similar Markov structure, SITS (Nguyen et al., 2012) chains a set of HDP-LDAs (Teh et al., 2006). Unlike PLDA, SITS assumes each text passage is associated with a speaker identity that is attached to the topic shift variable as supervising information. SITS further assumes speakers have different topic change probabilities that work as priors on topic shift variables. Instead of assuming documents in a dataset share the same set of topics, Bayesseg (Eisenstein and Barzilay, 2008) treats words in a segment generated from a segment specific multinomial language model, *i.e.*, it assumes each segment is generated from one topic, and a later hierarchical extension (Eisenstein, 2009) assumes each segment is generated from one topic or its parents. Other methods using as input the output of topic models include (Sun et al., 2008), (Misra et al., 2009), and (Riedl and Biemann, 2012).

In this paper we take a generative approach lying between PLDA and SITS. In contrast to PLDA, which uses a flat topic model (*i.e.*, LDA), we assume each text has a latent topic structure that can reflect the topic coherence pattern, and the model adapts its parameters to the segments to further improve performance. Unlike SITS that targets analysing multiparty meeting transcripts, where speaker identities are available, we are interested in more general texts and assume each text has a specific topic change probability, since (1) the identity information is not always available for all kinds of texts (*e.g.*, continuous broadcast news transcripts (Allan et al., 1998)), (2) even for the same author, topic change probabilities for his/her different articles might be different.

## 3 Segmentation with Topic Models

In documents, topically coherent segments usually encapsulate a set of consecutive passages that are semantically related (Wang et al., 2011). However, the topic boundaries between segments are often unavailable *a priori*. Thus we treat all passage boundaries (*e.g.*, sentence boundaries, paragraph boundaries or pauses between utterances) as possible topic boundaries. To recover the topic boundaries we develop a structured topic segmentation model by integrating ideas from the segmented topic model (Du et al., 2010, STM) and Bayesian segmentation models.

The basic idea of our model is that each document consists of a set of segments where text passages in the same segment are generated from the same topic distribution, called segment level topic distribution. The segment level topic distribution is drawn from a topic distribution associated with the whole document, called document level topic distribution. The relationships between the levels is managed using Bayesian non-parametric methods and a significant change in segment level topic distribution indicates a segment change.

Our unsupervised topic segmentation model is based on the premise that using a hierarchical topic model like the STM with a point-wise segment sampling algorithm should allow better detection of topic boundaries. We believe that (1) segment change should be associated with significant change in the topic distribution, (2) topic cohesion can be reflected in document topic structure, (3) the log-likelihood of a topically coherent segment is typically higher than an incoherent segment (Misra et al., 2008).

Assume we have a corpus of $D$ documents, each document $d$ consists of a sequence of $U_d$ text passages, and each passage $u$ contains a set of $N_{d,u}$ words denoted by $\boldsymbol{w}_{d,u}$ that are from a vocabulary $W$. Our model consists of:

**Modelling topic boundary:** We assume each document has its own topic shift probability $\pi_d$, a Beta distributed random variable, *i.e.*, $\pi_d \sim Beta(\lambda_0, \lambda_1)$. Then, we associate a boundary indicator variable $\rho_{d,u}$ with $u$, like the topic shift variable in PLDA and SITS. $\rho_{d,u}$ is Bernoulli distributed with parameter $\pi_d$, *i.e.*, $\rho_{d,u} \sim Bernoulli(\pi_d)$. It indicates whether there is a topic boundary after text passage $u$ or not. To sample $\rho_{d,u}$, we use a point-wise sampling algorithm. Consequently, a sequence of $\rho$'s defines a set of segments, *i.e.*, a topic segmentation of $d$. For example, let a $\rho$ vector $\boldsymbol{\rho} = (0, 0, 1, 0, 1, 0, 0, 1)$[1], it gives us three segments, which are $\{1, 2, 3\}$, $\{4, 5\}$ and $\{6, 7, 8\}$.

**Modelling topic structure:** Following the idea of the STM, we assume each document $d$ is associated with a document level topic distribution $\boldsymbol{\mu}_d$, which is drawn from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$; and text passages in topic segment $s$ in $d$ are generated from $\boldsymbol{\nu}_{d,s}$, a segment level topic distribution. The number of segments $S_d$ can be computed as $S_d = 1 + \sum_{u=1}^{U_d - 1} \rho_{d,u}$. Then, a Pitman-Yor

[1] The last 1 in $\boldsymbol{\rho}$ is the document boundary that is know a priori. This means one does not need to sample it.

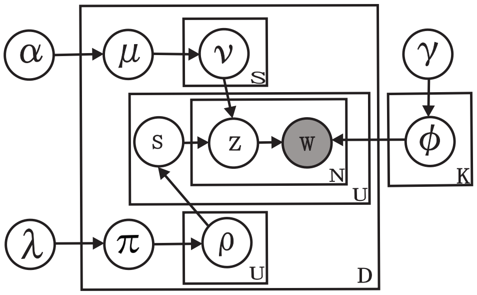

Figure 1: The topic segmentation model

process with a discount parameter $a$ and a concentration parameter $b$ is used to link $\boldsymbol{\mu}_d$ and $\boldsymbol{\nu}_{d,s}$ by $\boldsymbol{\nu}_{d,s}$~PYP($a$, $b$, $\boldsymbol{\mu}_d$), which forms a simple topic hierarchy. The idea here is that topics discussed in segments can be variants of topics of the whole document. Du *et al.* (2010) have shown that this topic structure can significantly improve the modelling accuracy, which should contribute to more accurate segmentation. This generative process is different from PLDA. PLDA does not assume the document level topic distribution and each time generates the segment level topic distribution directly from a Dirichlet distribution.

The complete probabilistic generative process, shown as a graph in Figure 1 is as follows:

1. For each topic $k \in \{1, \ldots, K\}$, draw a word distribution $\boldsymbol{\phi}_k \sim$ Dirichlet$_W$ $(\boldsymbol{\gamma})$.
2. For each document $d \in \{1, \ldots, D\}$,
   (a) Draw topic shift probability $\pi_d \sim$ Beta$(\lambda_0, \lambda_1)$.
   (b) Draw $\boldsymbol{\mu}_d \sim$ Dirichlet$_K$ $(\boldsymbol{\alpha})$.
   (c) For each text passage (except last) $u \in \{1, \ldots, U_d - 1\}$, draw $\rho_{d,u} \sim$ Bernoulli$(\pi_d)$.
   (d) Compute $S_d$ the number of segments as $1 + \sum_{u=1}^{U_d-1} \rho_{d,u}$.
   (e) For each segment $s \in \{1, \ldots, S_d\}$, draw $\boldsymbol{\nu}_{d,s} \sim$ PYP$(a,\, b,\, \boldsymbol{\mu}_d)$.
   (f) For each text passage $u \in \{1, \ldots, U_d\}$,
      i. Set segment $s_{d,u} = 1 + \sum_{v=1}^{u-1} \rho_{d,v}$.
      ii. For each word index $n \in \{1, \ldots, N_{d,u}\}$,
         A. Draw topic $z_{d,u,n} \sim$ Discrete$_K$ $\left(\boldsymbol{\nu}_{d,s_{d,u}}\right)$.
         B. Draw word $w_{d,u,n} \sim$ Discrete$_K(\boldsymbol{\phi}_{z_{d,u,n}})$.

where $s_{d,u}$ indicates which segment text passage u belongs to. We assume the dimensionality of the Dirichlet distribution (*i.e.*, the number of topics) is known and fixed, and word probabilities are parameterized with a $K \times W$matrix $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K)$. In future work we plan to investigate replace the Dirichlet prior $\boldsymbol{\alpha}$ on $\boldsymbol{\mu}$ with a Pitman-Yor prior (Pitman and Yor, 1997) to make the model fully nonparametric, like SITS.

Table 1: List of statistics

| | |
|---|---|
| $M_{k,w}$ | total number of words with topic $k$. |
| $\boldsymbol{M}_k$ | a vector of $M_{k,w}$. |
| $n_{d,s,k}$ | total number of words with topic $k$ in segment $s$ in document $d$. |
| $N_{d,s}$ | total number of words in segment $s$. |
| $t_{d,s,k}$ | table count of topic $k$ in the CRP for segment $s$ in document $d$. |
| $\boldsymbol{t}_{d,s}$ | a vector of $t_{d,s,k}$ for segment $s$ in $d$. |
| $T_{d,s}$ | total table count in segment $s$. |
| $c_{d,1}$ | total number of topic boundaries in $d$. |
| $c_{d,0}$ | total number of non-topic boundaries in $d$. |

## 4 Posterior Inference

In this section we develop a collapsed Gibbs sampling algorithm to do an approximate inference by integrating out some latent variables (*i.e.*, $\boldsymbol{\mu}$'s, $\boldsymbol{\nu}$'s and $\pi_d$'s). The hierarchy in our model can be well explained with the Chinese restaurant franchise metaphor introduced in (Teh et al., 2006). For easier understanding, terminologies of the Chinese Restaurant Process (CRP) will be used throughout this section, *i.e.*, customers, dishes and restaurants, correspond to words, topics, and segments respectively. Statistics used are listed in Table 1.

To integrate out the $\boldsymbol{\nu}_{d,s}$'s generated from the PYP, we use the technique presented in (Chen et al., 2011), which computes the joint posterior for the PYP by summing out all the possible seating arrangements for a sequence of customers (Teh, 2006). In this technique an auxiliary binary variable, called *table indicator* ($\delta_{d,u,n}$), is introduced to facilitate computing table count $t_{d,s,k}$ for topic $k$. This method has two effects: (1) faster mixing of the sampler, and (2) elimination of the need for dynamic memory to store the populations/counts of each table in the CRP. In the CRP each word $w_{d,u,n}$ in topic $k$ (*i.e.*, where $z_{d,u,n}$=$k$) contributes a count to $n_{d,s,k}$ for $u \in s$; and, if $w_{d,u,n}$, as a customer, also opens a new table to the CRP, it leads to increasing $t_{d,s,k}$ by one. In this case, $\delta_{d,u,n}$=1 indicates $w_{d,u,n}$ is the first customer on the table, called *table head*. Thus,

$$t_{d,s,k} = \sum_{u \in s} \sum_{n=1}^{N_{d,u}} \delta_{d,u,n} 1_{z_{d,u,n}=k} \; . \qquad (1)$$

Note the two constraints on these two counts, *i.e.*,

$$n_{d,s,k} \geq t_{d,s,k} \geq 0 \text{ and } t_{d,s,k}=0 \textit{ iff } n_{d,s,k}=0 \qquad (2)$$

can be replaced be a simpler constraint in the table indicator representation.

The sampler we develop is an MCMC sampler on the space $\boldsymbol{\theta} = \{\boldsymbol{z}, \boldsymbol{\delta}, \boldsymbol{\rho}\}$ where $\boldsymbol{z}$ defines the topic assignments of words, $\boldsymbol{\delta}$ maintains the needed CRP configuration (from which $\boldsymbol{t}$ is derived) and $\boldsymbol{\rho}$ defines the segmentation. Moreover, it is not a traditional Gibbs sampler changing one variable at a time, but is a block Gibbs sampler where two different kinds of blocks are used. The first block is $(z_{d,u,n}, \delta_{d,u,n})$ (for each word $w_{d,u,n}$), which can be sampled with a table indicator variant of a hierarchical topic sampler (Du et al., 2010), described in Section 4.1. This corresponds to Equation (6) in (Purver et al., 2006). The second kind of block is a boundary indicator $\rho_{d,u}$ together with a particular constrained set of table counts designed to handle splitting and merging, which corresponds to Equation (7) in (Purver et al., 2006). Sampling this second kind of block is harder in our non-parametric model requiring a potentially exponential summation, a problem we overcome using symmetric polynomials, shown in Section 4.2.

### 4.1 Sampling Topics

One step in our model is to sample the assignments of topics to words conditioned on all $\rho$'s. As discussed in Section 3, given the sequence of $\rho_{d,u}$'s, $\boldsymbol{\rho}_d$, one can figure out which segment $s$ text passage $u$ belongs to. Thus, conditioned on a set of segments $\boldsymbol{s}$ given by $\boldsymbol{\rho}$, the joint posterior distribution of $\boldsymbol{w}$, $\boldsymbol{z}$ and $\boldsymbol{\delta}$ is computed as $p(\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\delta} \mid \boldsymbol{\rho}, \boldsymbol{\Phi}, a, b, \boldsymbol{\gamma})$

$$= \prod_d \frac{\text{Beta}_K\left(\boldsymbol{\alpha} + \sum_s \boldsymbol{t}_{d,s}\right)}{\text{Beta}_K\left(\boldsymbol{\alpha}\right)} \prod_k \frac{\text{Beta}_W\left(\boldsymbol{\gamma} + \boldsymbol{M}_k\right)}{\text{Beta}_W\left(\boldsymbol{\gamma}\right)} \prod_d \prod_{s \in \boldsymbol{s}} \frac{(b|a)_{T_{d,s}}}{(b)_{N_{d,s}}} \prod_k \mathcal{S}^{n_{d,s,k}}_{t_{d,s,k},a} \binom{n_{d,s,k}}{t_{d,s,k}}^{-1}, \quad (3)$$

where $\text{Beta}_K(\cdot)$ is a $K$-dimension Beta function, $(x|y)_n$ the Pochhammer symbol[2], and $\mathcal{S}^n_{t,a}$ the generalised Stirling number of the second kind (Hsu and Shiue, 1998)[3] precomputed in a table so costing $O(1)$ to use (Buntine and Hutter, 2012).Eq (3) is an indicator variant of Eq (1) in (Du et al., 2010) with applying Theorem 1 in (Chen et al., 2011).

Given the current segmentation and topic assignments for all other words, using Bayes rule, we can derive the following two conditionals from Eq (3):

1. The joint probability of assigning topic $k$ to word $w_{d,u,n}$ and $w_{d,u,n}$ being a *table head*, $p(z_{d,u,n} = k,\, \delta_{d,u,n} = 1 \mid \boldsymbol{\theta}')$

$$= \frac{\gamma_{w_{i,j,n}} + M_{k,w_{i,j,n}}}{\sum_w (\gamma_w + M_{k,w})} \frac{\alpha_k + \sum_s t_{d,s,k}}{\sum_k \alpha_k + \sum_{s,k} t_{d,s,k}} \frac{b + aT_{d,s}}{b + N_{d,s}} \frac{S^{n_{d,s,k}+1}_{t_{d,s,k}+1,a}}{S^{n_{d,s,k}}_{t_{d,s,k},a}} \frac{t_{d,s,k}+1}{n_{d,s,k}+1} \quad (4)$$

2. The joint probability of assigning $k$ to $w_{d,u,n}$ and $w_{d,u,n}$ not being a *table head*, $p(z_{d,u,n} = k,\, \delta_{d,u,n} = 0 \mid \boldsymbol{\theta}')$

$$= \frac{\gamma_{w_{i,j,l}} + M_{k,w_{i,j,l}}}{\sum_w \gamma_w + M_{k,w}} \frac{1}{b + N_{d,s}} \frac{S^{n_{d,s,k}+1}_{t_{d,s,k},a}}{S^{n_{d,s,k}}_{t_{d,s,k},a}} \frac{n_{d,s,k} + 1 - t_{d,s,k}}{n_{d,s,k}+1} \quad (5)$$

where $\boldsymbol{\theta}' = \{\boldsymbol{z}^{-z_{d,u,n}}, \boldsymbol{w}, \boldsymbol{\delta}^{-\delta_{d,u,n}}, \boldsymbol{\rho}, \boldsymbol{\alpha}, a, b, \boldsymbol{\gamma}\}$. From the two conditionals, we develop a blocked Gibbs sampling algorithm for $(z_{d,u,n}, \delta_{d,u,n})$.

### 4.2 Sampling Segmentation Boundaries

In our model, each segment corresponds to a Chinese restaurant in the CRP. Sampling topic boundaries corresponds to splitting/merging restaurant(s). This is different from the split-merge process proposed by Jian and Neal (2004), where one actually splits/merges table(s). To our knowledge, there has been no method developed to split/merge restaurant(s). We tried different approximations, such as the minimum-path-assumption (Wallach, 2008), which in our case assumes one table for each topic $k$, and all words in $k$ are placed in the same table. Although this simplifies the split-merge process, it yielded poor results. We instead developed a novel approximate block Gibbs sampling algorithm using symmetric polynomials. Its segmentation performance worked well in our development dataset.

For simplicity, we consider a passage $u$ in document $d$, and assume: (1) If $\rho_{d,u}$=1, there are two segments, $s_l$ and $s_r$; $s_l$ ends at text passage $u$, and $s_r$ starts at text passage $u$+1. (2) If $\rho_{d,u}$=0, there is one

[2] The Pochhammer symbol $(x|y)_n$ denotes the rising factorial with a specified increment, *i.e.*, $y$. It is defined as $(x|y)_n = x(x+y)...(x+(n-1)y)$.

[3] A Stirling number of the second kind is used to study the number of ways of partitioning a set of $n$ objects into $k$ nonempty subsets. The generalised version given by Hsu and Shiue (1998) has a linear recursion which in our case is $\mathcal{S}^{n+1}_{m,a} = \mathcal{S}^n_{m-1,a} + (n - ma)\mathcal{S}^n_{m,a}$.

segment, $s_m$, where $u$ is is somewhere in the middle of $s_m$. The split-merge choice we sample is one to many, for a given split pair $(s_l, s_r)$ we consider a set of merged states $s_m$ (represented by different possible table counts). Then, to compute the Gibbs probability for splitting/merging restaurant(s), we consider the probability of the single split, the probability of the corresponding set of merges, and then if a merge is selected, we have to sample from the set of merges. These are as follows:

**Splitting:** split $s_m$ into $s_r$ and $s_l$ by placing a boundary after $u$. Since passages have a fixed order in each document, all the words are put into $s_r$ and $s_l$ based on which passages they belong to. Then, given all the topic assignments, we first sample all table indicators $\delta_{d,u',n}$, for $n \in \{1, ..., N_{d,u'}\}$ and $u' \in s_m$ using Bernoulli sampling without replacement. It runs as follows: 1) sample $\delta_{d,u',n}$ according to probability $t_{d,s_m,k}/n_{d,s_m,k}$; 2) decrease $t_{d,s_m,k}$ if $\delta_{d,u',n} = 1$, otherwise, just decrease $n_{d,s_m,k}$. Using the sampled $\delta_{d,u',n}$'s we compute the inferred table counts $t_{d,s,k}$ (from Eq (1)) and customer counts $n_{d,s,k}$ respectively for segments $s{=}s_l$ and $s_r$ and topics $k$. The computation may result in the following cases: for a given topic $k$,

(I) Both $s_l$ and $s_r$ have $n_{d,s,k}{>}0$ and $t_{d,s,k}{\geq}1$, which means both segments have words assigned to $k$ and words being labelled with *table head*. According to constraints (2), after splitting, restaurants corresponding to $s_l$ and $s_r$ are valid. We do not make any change on table counts.

(II) Either $s_l$ or $s_r$ has $n_{d,s,k}{=}0$ and $t_{d,s,k}{=}0$. In this case, for example, all the words assigned to $k$ in $s_m$ are in $s_l$ after splitting, and all those labelled with *table head* should also be in $s_l$. $s_r$ has no words assigned to $k$. Thus, there is no need to change table counts.

(III) Either $s_l$ or $s_r$ has $n_{d,s,k}{>}0$ and $t_{d,s,k}{=}0$. Both segments have words assigned to $k$, but those labelled with *table head* only exist in one segment. For instance, if they only exist in $s_l$ then $s_r$ has no table head, which means the restaurant of $s_r$ has customers eating a dish, but no tables serving that dish. Thus, we set $t_{d,s_r,k}{=}1$ to make the constraints (2) satisfied.

The Gibbs probability for splitting a segment is

$$p(\rho_{d,u} = 1 \,|\, \boldsymbol{\theta}'') \propto \frac{\lambda_1 + c_{d,1}}{\lambda_0 + \lambda_1 + c_{d,0} + c_{d,1}} \tag{6}$$

$$\mathrm{Beta}_K\big(\boldsymbol{\alpha} + \sum_{s=1}^{S_d} \boldsymbol{t}_{d,s}\big) \prod_{s\in\{s_l,s_r\}} \frac{(b|a)_{T_{d,s}}}{(b)_{N_{d,s}}} \prod_k \mathcal{S}^{n_{d,s,k}}_{t_{d,s,k},a} ,$$

where $\boldsymbol{\theta}'' = \{\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\delta}, \boldsymbol{\rho}^{-\rho_{d,u}}, \boldsymbol{\alpha}, a, b, \lambda_0, \lambda_1\}$.

**Merging:** remove the boundary after $u$, and merge $s_r$ and $s_l$ to one segment $s_m$. For this case, both $s_r$ and $s_l$ satisfy constraints (2) for all $k$'s, and set $n_{d,s_m,k}{=}n_{d,s_r,k} + n_{d,s_l,k}$. The following cases are considered: for a topic $k$

(I) Both $s_l$ and $s_r$ have $n_{d,s,k}{>}0$ and $t_{d,s,k}{>}1$. We compute $t_{d,s_m,k}$ using Eq (7). Thus table counts before and after merging are equal.

(II) Either $s_l$ or $s_r$ has $n_{d,s,k}{=}0$ and $t_{d,s,k}{=}0$. Similar to the above case, we use Eq (7).

(III) Both $s_l$ and $s_r$ have $n_{d,s,k}{>}0$, and either of them has $t_{d,s,k}{=}1$ or both. We have to choose between Eq (7) and Eq (8), *i.e.*, to decide whether a table should be removed or not.

$$t_{d,s_m,k} = t_{d,s_l,k} + t_{d,s_r,k} \tag{7}$$

$$t_{d,s_m,k} = t_{d,s_l,k} + t_{d,s_r,k} - 1 \tag{8}$$

Note that choosing Eq (8) means we need to decrease the table count $t_{d,s_m,k}$ by one. The idea here is that we sample to decide whether the remove table was added due to splitting case (III) or not. Clearly, we have a one-to-many split-merge choice. To compute the probability of a set of possible merges, we use elementary symmetric polynomials as follows: let $\mathcal{KS}$ be a set of topic-segment combinations that satisfy the condition in merging case (III), for $(k, s) \in \mathcal{KS}$, we sample either Eq (7) or Eq (8). Let $\mathcal{T} = \{t_{d,s,k} : (k, s) \in \mathcal{KS}\}$ be the set of table counts affected by the changes of Eq (7) or Eq (8). The Gibbs probability for merging two segments is

$$p(\rho_{d,u} = 0 \,|\, \boldsymbol{\theta}''') = \sum_{\mathcal{T}} p(\rho_{d,u} = 0, \mathcal{T} \,|\, \boldsymbol{\theta}''') \tag{9}$$

$$\propto \sum_{\mathcal{T}} \Bigg( \frac{\lambda_0 + c_{d,0}}{\lambda_0 + \lambda_1 + c_{d,0} + c_{d,1}} \mathrm{Beta}_K\big(\boldsymbol{\alpha} + \sum_{s=1}^{S_d} \boldsymbol{t}_{d,s}\big) \frac{(b|a)_{T_{d,s_m}}}{(b)_{N_{d,s_m}}} \prod_k \mathcal{S}^{n_{d,s_m,k}}_{t_{d,s_m,k},a} \Bigg) ,$$

where $\boldsymbol{\theta}''' = \{\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{t} - \mathcal{T}, \boldsymbol{\rho}^{-\rho_{d,u}}, \boldsymbol{\alpha}, a, b, \lambda_0, \lambda_1\}$. This is converted to a sum on $|\mathcal{T}|$ booleans with independent terms and evaluated recursively in $O(|\mathcal{T}|^2)$ by symmetric polynomials. If a merge is chosen, one then samples according to the terms in the sum using a similar recursion.

## 5 Experiments

To demonstrate the effectiveness of our model (denoted by TSM) in topic segmentation tasks, we evaluate it on three different kinds of corpora[4]: a set of synthetic documents, two meeting transcripts and two sets of text books (see Tables 2 and 3); and compare TSM with the following methods: two baselines (the Random algorithm that places topic boundaries uniformly at random, and the Even algorithm that places a boundary after every $m^{th}$ text passage, where $m$ is the average gold-standard segment length (Beeferman et al., 1999)), C99, MinCut, Bayesseg, APS (Kazantseva and Szpakowicz, 2011), and PLDA.

**Metrics:** We evaluated the segmentation performance with PK (Beeferman et al., 1999) and WindowDiff (WD$^r$) (Pevzner and Hearst, 2002), which are two common metrics used in topic segmentation. Both move a sliding window of fixed size $k$ over the document, and compare the inferred segmentation with the gold-standard segmentation for each window. The window size is usually set to the half of the average gold-standard segment size (Pevzner and Hearst, 2002). In addition, we also used an extended WindowDiff proposed by Lamprier *et al.* (2007), denoted by WD$^e$. One problem of WD$^r$ is that errors near the two ends of a text are penalised less than those in the middle. To solve the problem WD$^e$ adds $k$ fictive text passages at the beginning and the end of the text when computing the score. We evaluated all the methods with the same Java code for the three metrics.

**Parameter Settings:** In order to make all the methods comparable, we chose for each method the parameter settings that give the gold-standard number of segments[5]. Specifically, we used a $11 \times 11$ rank mask for C99, as suggested by Choi (2000), the configurations included in the code (`http://groups.csail.mit.edu/rbg/code`) for Bayesseg and manually tuned parameters for MinCut. For APS, a greedy approach was used to search parameter settings that can approximately give the gold-standard number of segments. For PLDA, two randomly initialised Gibbs chains were used. Each chain ran for 75,000 burn-in iterations, then 1000 samples were drawn at a lag of 25 from each chain. For TSM, 10 randomly initialised Gibbs chains were used. Each chain ran for 30,000 iterations with 25,000 for burn-in, then 200 samples were drawn. The concentration parameter $b$ in TSM was sampled using the Adaptive-Reject sampling scheme introduced in (Du et al., 2012b), the discount parameter $a = 0.2$, and $\lambda_0 = \lambda_1 = 0.1$. To derive the final segmentation for PLDA and TSM, we first estimated the marginal probabilities of placing boundaries after text passages from the total of 2000 samples. These probabilities were then thresholded to give the gold-standard number of segments. Precisely, we apply a small amount of Gaussian smoothing to the marginal probabilities (except for Choi's dataset), like Puerver *et al.* (2006) does. Finally, we used a symmetric Dirichlet prior in PLDA and STM, the one on topic distributions is $\alpha = 0.1$, the other on word distributions $\gamma = 0.01$.

Table 2: The Choi's dataset

| Range of n | | 3-11 | 3-5 | 6-8 | 9-11 |
|---|---|---|---|---|---|
| #docs | | 400 | 100 | 100 | 100 |
| DocLen | mean | 69.7 | 39.3 | 69.6 | 98.6 |
| | std | 8.2 | 2.6 | 2.9 | 3.5 |
| SegLen | mean | 7 | 4 | 7 | 10 |
| | std | 2.57 | 0.84 | 0.87 | 1.03 |

Table 3: Real dataset statistics

| | | ICSI | Election | Fiction | Clinical |
|---|---|---|---|---|---|
| # doc | | 25 | 4 | 84 | 227 |
| DocLen | mean | 994.5 | 144.3 | 325.0 | 139.5 |
| | std | 354.5 | 16.4 | 230.1 | 110.4 |
| SegLen | mean | 188 | 7 | 22 | 35 |
| | std | 219.1 | 8.9 | 23.8 | 41.7 |

### 5.1 Evaluation on Choi's Dataset

Choi's dataset (Choi, 2000) is commonly used in evaluating topic segmentation methods. It consists of 700 documents, each being a concatenation of 10 segments. Each segment is the first $n$ sentences of a randomly selected document from the Brown corpus, *s.t.* $3 \leq n \leq 11$. Those documents are divided into 4 subsets with different range of $n$, as shown in Table 2. We ran PLDA and STM with 50 topics. Results in Table 4 show that our model significantly outperforms all the other methods on the four subsets over all the metrics. Furthermore, comparing to other published results, this also outperforms (Misra et al., 2009) (see their table 2), and (Riedl and Biemann, 2012) (they report an average of 1.04 and 1.06 in Tables 1 and 2, whereas TSM averages 0.93). This gives TSM the best reported results to date.

[4] For preprocessing, we only removed stop words.

[5] The segments learnt by those methods will differ, but just the segment count will be the same as the gold-standard count.

Table 4: Comparison on Choi's datasets with WD and PK (%)

| | 3-11 | | | 3-5 | | | 6-8 | | | 9-11 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK |
| Random | 51.7 | 49.1 | 48.7 | 51.4 | 50.0 | 48.4 | 52.5 | 49.9 | 49.2 | 52.4 | 48.9 | 49.2 |
| Even | 49.1 | 46.7 | 49.0 | 46.3 | 45.8 | 46.3 | 38.8 | 37.3 | 38.8 | 30.0 | 28.6 | 30.0 |
| MinCut | 30.4 | 29.8 | 26.7 | 41.6 | 41.5 | 37.3 | 28.2 | 27.4 | 25.5 | 23.6 | 22.7 | 21.6 |
| APS | 40.7 | 38.8 | 38.4 | 32.0 | 30.6 | 31.8 | 34.4 | 32.6 | 32.7 | 34.5 | 32.2 | 33.2 |
| C99 | 13.5 | 12.3 | 12.3 | 11.3 | 10.2 | 10.8 | 10.2 | 9.3 | 9.8 | 8.9 | 8.1 | 8.6 |
| Bayesseg | 11.6 | 10.9 | 10.9 | 11.8 | 11.5 | 11.1 | 7.7 | 7.2 | 7.3 | 6.1 | 5.7 | 5.7 |
| PLDA | 2.4 | 2.2 | 1.8 | 4.0 | 3.9 | 3.3 | 3.6 | 3.5 | 2.7 | 3.0 | 2.8 | 2.0 |
| TSM | **0.8** | **0.8** | **0.6** | **1.3** | **1.3** | **1.0** | **1.4** | **1.4** | **0.9** | **1.9** | **1.8** | **1.2** |

Table 5: Comparison on the meeting transcripts and written texts with WD and PK (%)

| | ICSI | | | Election | | | Fiction | | | Clinical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK | WD$^r$ | WD$^e$ | PK |
| Random | 46.3 | 41.7 | 44.1 | 51.0 | 49.7 | 45.1 | 51.0 | 48.7 | 47.5 | 45.9 | 38.5 | 44.1 |
| Even | 48.3 | 43.0 | 46.4 | 56.0 | 55.1 | 51.2 | 48.1 | 45.9 | 46.3 | 49.2 | 42.0 | 48.8 |
| C99 | 42.9 | 37.4 | 39.9 | 43.1 | 41.5 | 37.0 | 48.1 | 45.1 | 42.1 | 39.7 | 31.9 | 38.7 |
| MinCut | 40.6 | 36.9 | 36.9 | 43.6 | 43.3 | 39.0 | 40.5 | 39.7 | 37.1 | 38.2 | 36.2 | 36.8 |
| APS | 58.2 | 49.7 | 54.6 | 47.7 | 36.8 | 40.6 | 48.0 | 45.8 | 45.1 | 39.9 | 32.8 | 39.6 |
| Bayesseg | 32.4 | 29.7 | 26.7 | 41.1 | 41.3 | 34.1 | **33.7** | **32.8** | **27.8** | 35.0 | **28.8** | 34.0 |
| PLDA | 32.6 | 28.8 | 29.4 | 40.6 | 41.1 | 32.0 | 43.0 | 41.3 | 36.1 | 37.3 | 32.1 | 32.4 |
| TSM | **30.2** | **26.8** | **25.8** | **38.1** | **38.9** | **31.3** | 40.8 | 38.7 | 32.5 | **34.5** | 29.1 | **30.6** |

Note the lexical transitions in these concatenated documents are very sharp (Malioutov and Barzilay, 2006). The sharp transitions lead to significant change in segment level topic distributions, which further implies the variance of these distributions is large. In TSM, a large variance causes a small concentration parameter $b$. We observed that the sampled $b$'s (about 0.1) are indeed small for the four subsets, which shows there is no topic sharing among segments. Therefore, TSM is able to recognise the segments are unrelated text.

### 5.2 Evaluation on Meeting Transcripts

We applied our model to segmenting the two meeting transcripts, which are the ICSI meeting transcripts (Janin et al., 2003) and the 2008 presidential election debates (Boydstun et al., 2011). The ICSI meeting has 75 transcripts, we used the 25 annotated transcripts provided by Galley *et al.* (2003) for evaluation. For the election debates, we used the four annotated debates used in (Nguyen et al., 2012). The statistics are shown in Table 3. PLDA and TSM were trained with 10 topics on the ICSI and 50 on the Election. In this set of experiments, we show that our model is robust to meeting transcripts.

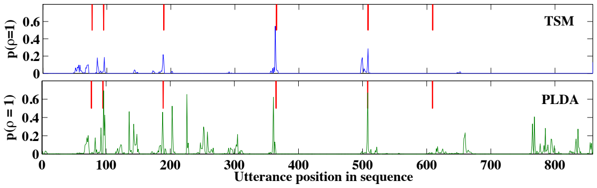

Figure 2: Probability of a topic boundary, compared with gold-standard segmentation (shown in red and at the top of each diagram) on one ICSI transcript.

As shown in Table 5, topic modelling based methods (*i.e.*, Bayesseg, PLDA and TSM) outperform those using either TF or TF-IDF, which is consistent with previously reported results (Misra et al., 2009; Riedl and Biemann, 2012). Among the topic model based methods, TSM achieves the best results on all the three metrics. On the ICSI transcripts, TSM performs 6.8%, 9.7% and 3.4% better than Bayesseg on the WD$^r$, WD$^e$ and PK metrics respectively. Figure 2 shows an example of how the inferred topic boundary probabilities at utterances compare with the gold-standard boundaries on one ICSI meeting transcript. The gold-standard segmentation is {77, 95, 189, 365, 508, 609, 860}, TSM and PLDA infer {85, 96, 188, 363, 499, 508, 860} and {96, 136,

Table 6: Sampled concentration parameters

| | Choi | ICSI | Election | Fiction | Clinical |
|---|---|---|---|---|---|
| b | 0.1 | 5.2 | 5.4 | 18.4 | 4.8 |

203, 226, 361, 508, 860} respectively. Both models miss the boundary after the $609^{th}$ utterance, but put a boundary after the $508^{th}$ utterance. Note the boundaries placed by TSM are always within 10 utterances with respect to the gold standard.

Although TSM still performs the best on the debates, all the methods have relatively worse performance than on the ICSI meeting transcripts. Nguyen *et al.* (2012) pointed out that the ICSI meetings are characterised by pragmatic topic changes, in contrast, the debates are characterised by strategic topic changes with strong rewards for setting the agenda, dodging a question, *etc*. Thus, considering the properties of debates might further improve the segmentation performance.

### 5.3 Evaluation on Written Texts

We further tested TSM on two written text datasets, Clinical (Eisenstein and Barzilay, 2008) and Fiction (Kazantseva and Szpakowicz, 2011). The statistics are shown in Table 3. Each document in the Clinical dataset is a chapter of a medical textbook. Section breaks are selected to be the true topic boundaries. For the Fiction dataset, each document is a fiction downloaded from Project Gutenberg, the true topic boundaries are chapter breaks. We trained PLDA and TSM with 25 topics on the Fiction and 50 on the Clinical. Results are shown in Table 5. TSM compares favourably with Bayesseg and outperforms the other methods on the Clinical dataset, but it does not perform as well as Bayesseg on the Fiction dataset.

In fiction books, the topic boundaries between sections are usually blurred by the authors for reasons of continuity (Reynar, 1999). We observed that the sampled concentration (or inverse variance) parameter $b$ in TSM is about 18.4 on Fiction, but 4.8 on Clinical, as shown in Table 6. This means the variance of segment level topic distributions $\boldsymbol{\nu}$ learnt by TSM is not large for the fiction, so chapter breaks may not necessarily indicate topic changes. For example, there is a document in the Fiction dataset where gold-standard topic boundaries are placed after each block of text. In contrast, Bayesseg assumes each segment has its own distribution over words, *i.e.*, one topic per segment, which means topics are not shared among segments. We hypothesize that for certain kinds of documents where the change in topic distribution is subtle, such as fiction, assuming one topic per segment can capture subtle changes in word usage. This is an area for future investigation.

## 6 Conclusion

In this paper, we have presented a hierarchical Bayesian model for unsupervised topic segmentation. This new model takes advances of both Bayesian segmentation and structured topic modelling. It uses a point-wise boundary sampling algorithm to sample a topic segmentation, while concurrently building a structured topic model. We have developed a novel approximation to compute the Gibbs probabilities of spliting/merging segment(s). Our model shows prominent segmentation performance on both written or spoken texts.

In future work, we would like to make the model fully nonparametric and investigate the effects of adding different cues in texts, such as cue phrases, pronoun usage, prosody, *etc*. Currently, our model uses marginal boundary probabilities to generate the final segmentation. Instead, we could develop a Metropolis-Hasting sampling algorithm to move one boundary at a time, given the gold-standard number of segments. To further study the effectiveness of our model, we would like to compare it with other methods, like SITS (Nguyen et al., 2012) and to run on more datasets, like email (Joty et al., 2010). For example, in order to compare with SITS, one can make an assumption that each document just has one speaker.

## Acknowledgments

The authors would like to thank all the anonymous reviewers for their valuable comments. This research was supported under Australian Research Council's Discovery Projects funding scheme (project numbers DP110102506 and DP110102593). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## References

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

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

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

A.E. Boydstun, C. Phillips, and R.A. Glazier. 2011. Its the economy again, stupid: Agenda control in the 2008 presidential debates.

W. Buntine and M. Hutter. 2012. A Bayesian review of the Poisson-Dirichlet process. Technical Report arXiv:1007.0296v2, *ArXiv*, Cornell.

Changyou Chen, Lan Du, and Wray Buntine. 2011. Sampling for the Poisson-Dirichlet process. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database*, pages 296–311.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 26–33.

Lan Du, Wray Buntine, and Huidong Jin. 2010. A segmented topic model based on the two-parameter Poisson-Dirichlet process. *Mach. Learn.*, 81(1):5–19.

Lan Du, Wray Buntine, and Huidong Jin. 2012a. Modelling sequential text with an adaptive topic model. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 535–545.

Lan Du, Wray Buntine, Huidong Jin, and Changyou Chen. 2012b. Sequential latent Dirichlet allocation. *Knowledge and Information Systems*, 31(3):475–503.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 334–343.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 353–361. The Association for Computational Linguistics.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–53.

Marti A. Hearst. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

Leetsch C. Hsu and Peter Jau-Shyong Shiue. 1998. A unified approach to generalized Stirling numbers. *Adv. Appl. Math.*, 20:366–384, April.

Sonia Jain and Radford Neal. 2004. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13(1):158–182.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *Proceedings of 2003 IEEE International Conference on Acoustics, Speech, and Signal (ICASSP '03)*, pages 364–367.

Shafiq Joty, Giuseppe Carenini, Gabriel Murray, and Raymond T. Ng. 2010. Exploiting conversation structure in unsupervised topic segmentation for emails. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 388–398.

Anna Kazantseva and Stan Szpakowicz. 2011. Linear text segmentation using affinity propagation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 284–293.

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, and Frederic Saubion. 2007. On evaluation methodologies for text segmentation algorithms. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02*, ICTAI '07, pages 19–26.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 25–32.

Hemant Misra, Olivier Cappe, and Francois Yvon. 2008. Using LDA to detect semantically incoherent documents. In *Proceedings of CoNLL-08*, pages 41–48.

Hemant Misra, François Yvon, Joemon M. Jose, and Olivier Cappe. 2009. Text segmentation via topic modeling: an analytical study. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1553–1556.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2012. SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 78–87.

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

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

Matthew Purver, Thomas L. Griffiths, Konrad P. Körding, and Joshua B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 17–24.

Jeffrey C. Reynar. 1999. Statistical models for topic segmentation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 357–364.

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Gerard Salton, Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Automatic text decomposition using text segments and text themes. In *Proceedings of the the seventh ACM conference on Hypertext*, pages 53–65.

Qi Sun, Runxin Li, Dingsheng Luo, and Xihong Wu. 2008. Text segmentation with LDA-based Fisher kernel. In *Proceedings of ACL-08: HLT, Short Papers*, pages 269–272.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Y. W. Teh. 2006. A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore.

Masao Utiyama and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 499–506.

H.M. Wallach. 2008. Structured topic models for language. *doctoral dissertation, Univ. of Cambridge*.

Hongning Wang, Duo Zhang, and ChengXiang Zhai. 2011. Structural topic model for latent topical structure analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1526–1535.